A RAND NOTE

THREE METHODS FOR PROCESSING
LIFE-HISTORY DATA

Terry Fain

July 1980

N-1544-AID

Prepared For

The Agency for International Development

Rand
SANTA MONICA, CA. 90406

This research was supported by the Agency for International Development under Contract No. AID/OTR-G-1744.

Published by The Rand Corporation

A RAND NOTE

THREE METHODS FOR PROCESSING LIFE-HISTORY DATA

Terry Fain

July 1980

N-1544-AID

Prepared For

The Agency for International Development

Rand
SANTA MONICA, CA. 90406

PREFACE

This Note is one of several publications growing out of a survey and research project to investigate the influence of certain economic and institutional factors on couples' fertility behavior in Peninsular Malaysia. The project was funded by the U.S. Agency for International Development and was conducted by The Rand Corporation in collaboration with, initially, the Department of Statistics of the Government of Malaysia, and subsequently, Survey Research Malaysia, Sdn. Bhd. The purpose of the project was to identify factors within the range of direct public policy influence that affect birthspacing and family size, and to estimate the magnitude of statistical relations between these factors and fertility outcomes.

The present Note discusses three approaches to computer processing of the data gathered and analyzed during the project.

SUMMARY

It is often the case that the richness of life-history data is matched by its complexity. Thus, handling and processing such data presents a challenge to researchers and programmers. This Note looks at three methods for dealing with the complexity of life-history data, based on experience with a particular dataset, the Malaysian Family Life Survey.

The approaches considered are:

(1) SIR, a commercially available package, designed for editing and processing hierarchical datasets.

(2) RETRO, a custom-written program, for use with a particular dataset but designed to be utilized by persons at the research assistant level.

(3) Custom programming by a resident computer professional.

No single method can be considered superior, in the sense that it is clearly advantageous and cost-effective in every situation. Instead, each of the methods has both advantages and shortcomings, and the Note looks at each method in the context of scope, data preparation and storage, program preparation, execution and output, and relative costs. This approach allows prospective users of life-history data to evaluate the three methods based on their own data and the designs of their research.

ACKNOWLEDGMENTS

I wish to thank Bill Butz and Julie DaVanzo for help and encouragement on this Note, and for always being the kind of people who are a pleasure to work with.

I would also like to thank Sue Polich for insightful and constructive comments on an earlier draft, and Rosemary Rhoades and Marilyn LaPrell for help in preparing the manuscript.

Finally, I am grateful to Rosemarie Rothwell-Konningsburger for the flow charts presented in Appendix C.

A slightly different version of this Note was read at the Eleventh Summer Seminar in Population, held at the East-West Population Institute, East-West Center, Honolulu, in June 1980. Research was partially supported by the U.S. Agency for International Development under Contract No. AID/DS/PE/C/0002 with the Institute. Comments and suggestions from participants, resource persons, and coordinators of the Seminar were helpful in preparing the final version.

CONTENTS

I. INTRODUCTION

In recent years, researchers in the social sciences have become increasingly aware of the limitations of data which reflect the characteristics of an individual or family at a single point in time. Consequently, we now see an increasing number of datasets that give information about an individual or family over part or all of its life cycle. There are two types of such datasets:

(1) longitudinal panel data
(2) retrospective life-history data.

Longitudinal panel data come from multiple interviews with the subject over a period of time, while retrospective life-history data are collected at one time but relate to a historical reference period. Longitudinal data are generally more reliable, but require considerable time and expense to gather; retrospective data are relatively quick and inexpensive to collect. Both methods produce life histories in data. Although this Note focuses on retrospective life history data files, much of the discussion is applicable to longitudinal panel data, as well.

Unfortunately, the complexity of life-history datasets means that data processing is correspondingly more difficult than with ordinary datasets. Many of the conventional data processing approaches, geared to fixed-length records, cannot be used in the analysis of life-history data without considerable prior data manipulation to produce the required input formats. Thus, much of the programming support for analysis of life-history data necessarily concerns data handling and manipulation.

In response to the challenges of life-history data, new techniques have been developed. These techniques fall into two main types: those that perform analysis on the raw data directly, and those that manipulate the raw data in such a way as to make analysis possible by conventional program packages such as SPSS, SAS, BMDP, etc. Certain methods address only one of these requirements; others attempt, at least to some extent, to do both.

This Note will discuss three such methods used in the analysis of a particularly complex life-history dataset, the Malaysian Family Life Survey.

II. THE MALAYSIAN FAMILY LIFE SURVEY (MFLS)

The Malaysian Family Life Survey (MFLS) is an analysis dataset for a research project that focuses on economic and sociological factors influencing the fertility and child-rearing behaviors of couples in Peninsular Malaysia. It was developed by economists William P. Butz and Julie DaVanzo of The Rand Corporation, under a grant from the Agency for International Development (AID). It was administered by Survey Research Malaysia (SRM), in co-operation with the Department of Statistics of the Malaysian government, to 1262 households in Peninsular Malaysia. Three rounds of data were gathered, beginning in Fall of 1976; average interview time for the first round was three hours and twenty minutes per household. The entire analysis dataset for all three rounds consists of approximately 300,000 card-image records; there are sixty-eight distinct types of record.

Over the three rounds, ten questionnaires were administered to each household.* Some were given in only one round; others were updated or administered in full in all three rounds. The central focus (and almost always the primary respondent) of each household is an ever-married woman (EMW) aged 15 to 50. A brief description of each questionnaire may be found in Appendix A.

This discussion will focus on data gathered by three question-

---

* In addition, there is a community-level questionnaire (MF11), administered to village chiefs, midwives, and other knowledgeable persons.

naires in Round One: the Household Roster (MF1), the Female Retrospective History (MF2), and the Male Retrospective History (MF3). The household roster contains demographic information for each member of the household, while the two retrospective questionnaires contain life-history data for the EMW, and for her husband if she is currently married. Information gathered includes a pregnancy history; a record of marital changes and separations; a history of migrations, house characteristics and changes, and living arrangements; data related to child care; and separate education and employment histories for the EMW and husband.

In the analysis stage, the data have been organized by round, by household, and by instrument (questionnaire). Within instrument, records are organized as follows: a summary record gives information about the interview itself and tells how many records of each of several types will follow; the specified numbers of records of each type then follow in sequence. The resulting dataset thus consists of a variable number of fixed-length records; such an arrangement has come to be known as a "hierarchical" dataset.*

---

* The analysis dataset for the INCAP-Rand Guatemala Survey is a hierarchical dataset similar to that of the MFLS, but without summary records. Much of the discussion of this Note is as applicable to the Guatemala Survey as to the MFLS.

For a description of the content of the INCAP-Rand Guatemala Survey, see Appendix B.

III. PROCESSING THE DATA: THREE APPROACHES

Three primary methods have been used to extract meaningful information from the MFLS. These are:

- SIR (Scientific Information Retrieval), a package which is designed specifically for use with hierarchical datasets. SIR uses an SPSS-like pseudo-language, and is capable of both data manipulation and computation of simple statistics. The MFLS was chosen as an ideal dataset for testing a pre-release version of this package for IBM computers.

- RETRO, a program written by Iva Maclennan of The Rand Corporation specifically for use with the retrospective questionnaires of the MFLS and the earlier INCAP-Rand Guatemala Survey (see Appendix B). RETRO produces fixed-length records for input to other packages; it does no data analysis.

- Custom programming, that is to say, in-house programming support performed by an individual or group familiar with the structure of the data, and geared to specific requests by researchers for a given product. This approach can, of course, produce either statistical analysis or fixed-length output for use with other packages.

Each of these specific methods represents a general overall approach to working with life-history data. SIR is the comprehensive package designed to be used by data processors rather than programmer/analysts. As such, it may be utilized by the researcher himself or by a research assistant. Custom programming, on the other hand, calls for the skills of a professional expert, assuming that such a person is ultimately more cost-effective in producing a volume of analysis within a reasonable amount of time. RETRO represents the compromise between the other two methods; that is, it uses a highly skilled programmer to create a product designed for a specific body of data, which can then be used by less specialized persons for the

actual analyses.*

Each approach will now be discussed with respect to scope, data preparation, data storage, program preparation, execution and output, and relative costs.

SCOPE

SIR purports to be a complete data-handler for hierarchical datasets, enabling data manipulation and retrieval, as well as the production of simple analyses without recourse to other packages. It can link data from several different types of data records into a single output record; it can use selectivity criteria to produce, for example, one output record per pregnancy, or one record per Chinese household; it can compute means and frequency distributions and generate data listings. SIR also interfaces directly with SPSS and BMDP, thus enabling the user to access the full potential of these popular packages.

RETRO is the most limited of the three approaches in scope. It was designed to work only with the household roster and the two retrospective instruments of the MFLS and the INCAP-Rand Guatemala Survey. Within these limits, however, it is capable of quite sophisticated retrievals. The ability to link the male and female retrospective questionnaires is especially useful. For example, between a respondent's first and second pregnancy outcomes, RETRO can tell you how many times she changed residences, what her husband's

---

* Flow charts showing the steps from raw data to completed analysis for each of the three methods are given in Appendix C.

salary was, what kind of birth control she used, and any demographic information about either herself or her husband.

Like SIR, RETRO is capable of generating output records according to user-supplied specifications. RETRO also allows for the re-coding of variables, arithmetic combinations of variables, and the specification of selectivity criteria either before or after record creation. For example, one may produce one record per family, one per pregnancy, or one for each migration among only those women who have been divorced.

Custom programming, of course, offers the widest scope of all, limited only by the ability of the programmers and the time/cost constraints of the research.

DATA PREPARATION

All three approaches require a high level of data reliability; in my experience, considerable time has always been needed to edit the data before any actual processing can begin. Custom programming is especially susceptible to problems of record counting; if an error exists on a summary record, the wrong number of subsequent records will be read--a condition which inevitably produces garbage. If custom programming uses an approach other than counting records, problems can still arise unless all variables used for sequencing and identifying record type have been carefully edited.

SIR distinguishes one record type from another by a nested set of sequencing variables, which implies that all records must reserve certain positions for these sequencing variables. In practice, this may mean that some data manipulation is required before SIR can be

used. Moreover, it may be desirable to re-sort the data before input to SIR in order to reduce the cost of the initial run.

RETRO works in conjunction with another program called CONVERT. CONVERT reads the raw data, extracts only those variables that will ever be of concern to RETRO, and stores them in two compactly formatted files which are accessed jointly or separately during an actual RETRO run. CONVERT need be run only once (so long as there are no corrections to the raw data), and the resulting "CONVERTed" files stored. As with custom programming, it is critical that there be no errors in record identification and sequencing in order to produce meaningful results.

It is worth noting that a researcher may save considerable time in the data preparation stage by deciding before coding his data what approach he will use for subsequent analyses. This enables him to tailor the coding to the requirements of the method he has chosen.

DATA STORAGE

This is a consideration that might be easily overlooked, but it is in fact of no small importance. Modern computer installations primarily use disk and magnetic tape for data storage; the relative cost advantage of one type of storage over the other is installation-specific. In addition, the installation imposes certain fixed limitations on the availability of storage, in that only so much disk space or so many tape drives are available to the user. This is important because both SIR and RETRO require disk storage of

specially prepared files.* Thus, the user must either leave these files resident on disk storage, or else copy them from tape to disk before each use. Almost certainly this will produce overhead costs for using either method.

SIR has tackled this problem by providing for an "unloaded" dataset, designed as a tape backup. Utilities which load the SIR dataset onto disk and unload it to tape are built into the package and are easy to use. It is also likely that future versions of SIR (now in the development stage) will be able to access tape files directly.

Clearly, custom programming has the greatest amount of flexibility with regard to data storage. With the MFLS, we have divided the data into subfiles (one per questionnaire). Since most retrievals require data from a single questionnaire, this approach allows data to be stored on tape and reduces both the complexity of the programming and the volume of data read in a given retrieval.

PROGRAM PREPARATION

Although SIR and RETRO are accessible to users of lesser skill than that required for custom programming, their use is not a simple matter. SIR requires the mastery of its own pseudo-language and a familiarity with the conventions of the package, which are not always intuitively obvious. RETRO demands that the user prepare a number of

---

* There is one exception to this: If no cross-referencing of the male and female retrospectives is required, RETRO can read directly from tape. But joint processing of the two questionnaires--one of RETRO's most significant features--requires both CONVERTed files to reside on disk.

control cards to guide the retrieval; preparation of these can be complex and tedious, partly because of the program's ability to perform complicated retrievals. Moreover, documentation for both SIR and RETRO could stand improvement. Custom programming, of course, requires the greatest amount of skill (and time); indeed, its success depends almost entirely on the ability of the programmer and the capacity of the machine.

Another consideration in program preparation is this: No matter how flexible methods such as SIR and RETRO are, and no matter how carefully the research has been designed and executed, there inevitably comes a time when the researcher wants to do something beyond the capacity of these methods. As an example, we wanted to know, for each pregnancy outcome in the MFLS, the salary of the husband at his next-to-last job within one year of the date of outcome. This was a little too subtle for either RETRO or SIR, and we had to get this information by custom programming. One suspects that there will always be some retrievals that are possible only through custom programming.

EXECUTION AND OUTPUT

Execution speed and output readability are always in part a function of user ability. But even if one assumes an equal ability to manipulate each of the three methods, differences will still exist due to the techniques themselves. RETRO, for example, is designed to read a fixed amount of data, make certain calculations and selections, and write out the resulting variables. This design makes for a retrieval whose execution time and input/output (I/O) costs are

relatively independent of the number of variables retrieved. SIR, on the other hand, provides for a hierarchy of variables, so that certain commonly needed data items, such as demographic variables, are readily accessible without having to read the entire dataset.

With custom programming, execution time and output characteristics are more variable than with the other methods. The type of data structure, the kind of software used, and the creativity of the programmer have a profound influence on output quantity and quality. Yet there is no doubt that custom programming is potentially the most flexible of the three methods.

OVERALL COSTS

In figuring the overall cost of a given approach, we must take into account not only the cost of all the stages discussed above, but also the time-cost of the users themselves and the inevitable cost of false starts, mistakes, infinite loops, and other data processing nightmares. In other words, we must consider the cost of failures as well as the cost of successes. We must also bear in mind that a given problem in retrieval may dictate the use of a given technique, or at least not be equally amenable to all three.

To complicate matters even further, there is considerable variation among computer installations in cost accounting.* Another consideration is the fixed cost of leasing the SIR package.

---

* A good example of the importance of this can be found in the historical development of SIR. Originally optimized for a CDC computer, SIR was at first prohibitively expensive on Rand's IBM 3032, because it required so many I/O's. Enlarging the blocksize reduced the number of I/O's and made the cost of SIR comparable with other methods.

No systematic cost-effectiveness studies have been done to compare SIR, RETRO, and custom programming. In the absence of such studies, the best one can hope for here is an impressionistic consensus among researchers at a specific installation (Rand), working with a single dataset (the MFLS). The fact is that we use a combination of the three methods, and some specific tasks are accomplished much more easily by one method than by another. But our overall impression is that custom programming is the most cost-effective way to go, given our particular skills and our particular requirements.

IV. CONCLUSIONS

At the risk of lapsing into platitude, the "best" technique for processing life-history data necessarily depends on one's resources. A researcher who wishes to do his own data processing may find SIR a powerful tool in analyzing hierarchical datasets. A program such as RETRO, designed for a specific dataset, can offer considerable sophistication at a relatively modest cost. And the possibilities of custom programming are limited only by the capacity of the machine and the imagination and skill of the programmer.

All three methods discussed here have been useful in work with the MFLS. It is much easier to say which method is better for a specific task than to say which one is "best" overall. The optimal solution would seem to be the selective use of all three. But it does appear that whatever combination of methods one uses, at least some custom programming will be required.

BIBLIOGRAPHY

Butz, William P., and Julie Davanzo, The Malaysian Family Life Survey: Summary Report, The Rand Corporation, R-2351-AID, March 1978.

Corona, Henry L, INCAP-Rand Guatemala Survey: Code Book and User's Manual, The Rand Corporation, P-6181, August 1978.

Fain, Terry, and Tan Poh Kheong, The Malaysian Family Life Survey: Appendix B, Round One Codebook, The Rand Corporation, R-2351/2-AID, March 1978.

Fain, Terry, and Tan Poh Kheong, The Malaysian Family Life Survey: Appendix E, Rounds Two and Three Codebook, The Rand Corporation, R-2351/5-AID (forthcoming).

Maclennan, Iva, RETRO: A Computer Program for Processing Life History Data, The Rand Corporation, R-2363-AID/RF, March 1978.

Robinson, Barry N., Gary D. Anderson, Eli Cohen, and Wally F. Gazdzik, SIR Scientific Information Retrieval User's Manual, SIR, Inc., Evanston, Ill., 1979.

APPENDIX A: SURVEY QUESTIONNAIRES FOR THE MALAYSIAN FAMILY LIFE SURVEY (MFLS)

QUESTIONNAIRE	ELIGIBLE RESPONDENTS	ROUND(S) IN WHICH ADMINISTERED	AVERAGE INTERVIEW LENGTH IN ROUND 1
MF1: Household Roster	Selected ever-married women (EMW) less than 50 yr old, or other eligible adult female	Administered completely in 1; updated in 2 and 3	20 minutes
MF2: Female Retrospective	EMW	Administered completely in 1; updated in 2 and 3	60 minutes*
MF3: Male Retrospective	Present husbands of EMW	Administered completely in 1; updated in 2 and 3	40 minutes*
MF4 and MF5: Female and Male Time Budgets	EMW and their present husbands	Administered completely in 1, 2, and 3	25 minutes - MF4 13 minutes - MF5
MF6: Income and Wealth	Male heads of household or other members of household that contains an EMW less than 50 yr old	Administered completely in 1, 2, and 3	43 minutes
MF7 and MF8: Female and Male Attitudes and Expectations	EMW and their present husbands	Administered in 2 only	
MF9: Networks of Economic Support	EMW	Administered in 3 only	
MF10: Migration and Urban Assimilation	Present husbands of EMW	Administered in 3 only	
MF11: Community Information	Village chiefs, midwives, and other knowledgeable persons (several questionnaires per Primary Sampling Unit)	Administered throughout the survey	

*Round 2 and 3 updates take considerably less time.

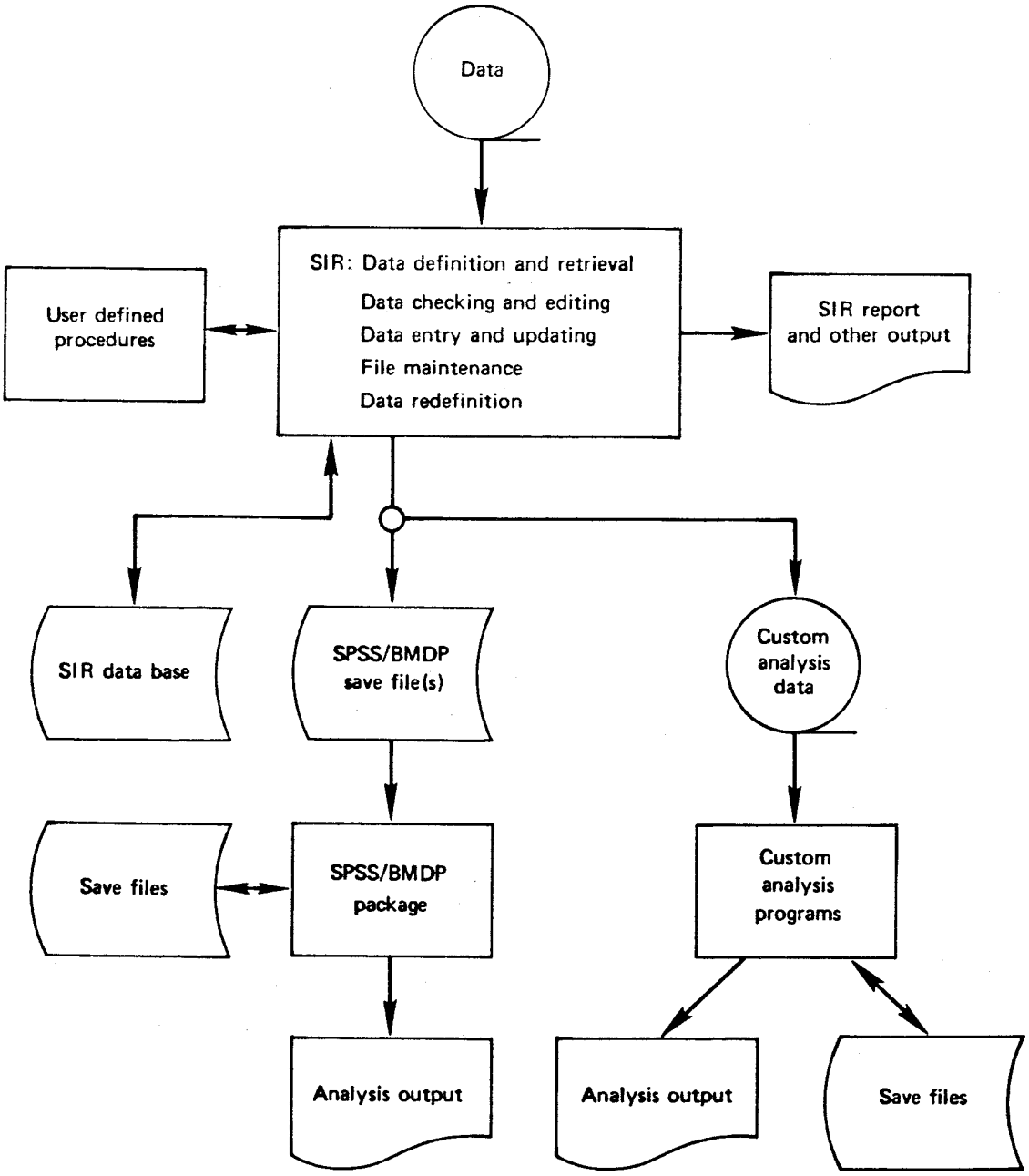
APPENDIX B: SURVEY QUESTIONNAIRES FOR THE INCAP-RAND GUATEMALA SURVEY

QUESTIONNAIRE	TOTAL SAMPLE SIZE	SAMPLE IN 4 RURAL VILLAGES	SAMPLE IN 2 URBAN VILLAGES
R03: Female Retrospective Life History	1097	All ever-united or ever-pregnant women aged 14-49	Same
R05: Home Stimulation of Children	393	50% sample of all pregnant women, and all women with a child less than 7	50% sample of all pregnant women, and all women with child less than 3
R06: Modernity of Family	393	Same as R05	Same as R05
R07: Vocabulary Test for Mothers	393	Same as R05	Same as R05
R08: Schooling	393	Same as R05	Same as R05
R09: Time Budget (4 survey rounds)	Round 1: 795 Round 2: 793 Round 3: 764 Round 4: 876	All pregnant women and all women with a child less than 7	All pregnant women and all women with a child less than 3
R09 Fifties: Female Opportunity Structure (4 survey rounds)	Same as R09	Same as R09	Same as R09
R10: Income, Wealth, and Agricultural Production (1st Round)	1356	All heads of nuclear families	All heads of nuclear families surveyed in R09 plus heads of all remaining agricultural families plus random 25% of remaining family heads
R10B: Community Prices and Variables	5	Three key informants in each village	Same
R11: Female Attitudes and Expectations	847	All women surveyed for R09 plus random 50% of remaining female heads of families	Random 50% of women surveyed for R09 plus random 25% of remaining female heads of families
R12: Male Attitudes and Expectations	497	All male spouses of women surveyed for R11	Same

APPENDIX B: SURVEY QUESTIONNAIRES FOR THE INCAP-RAND GUATEMALA SURVEY

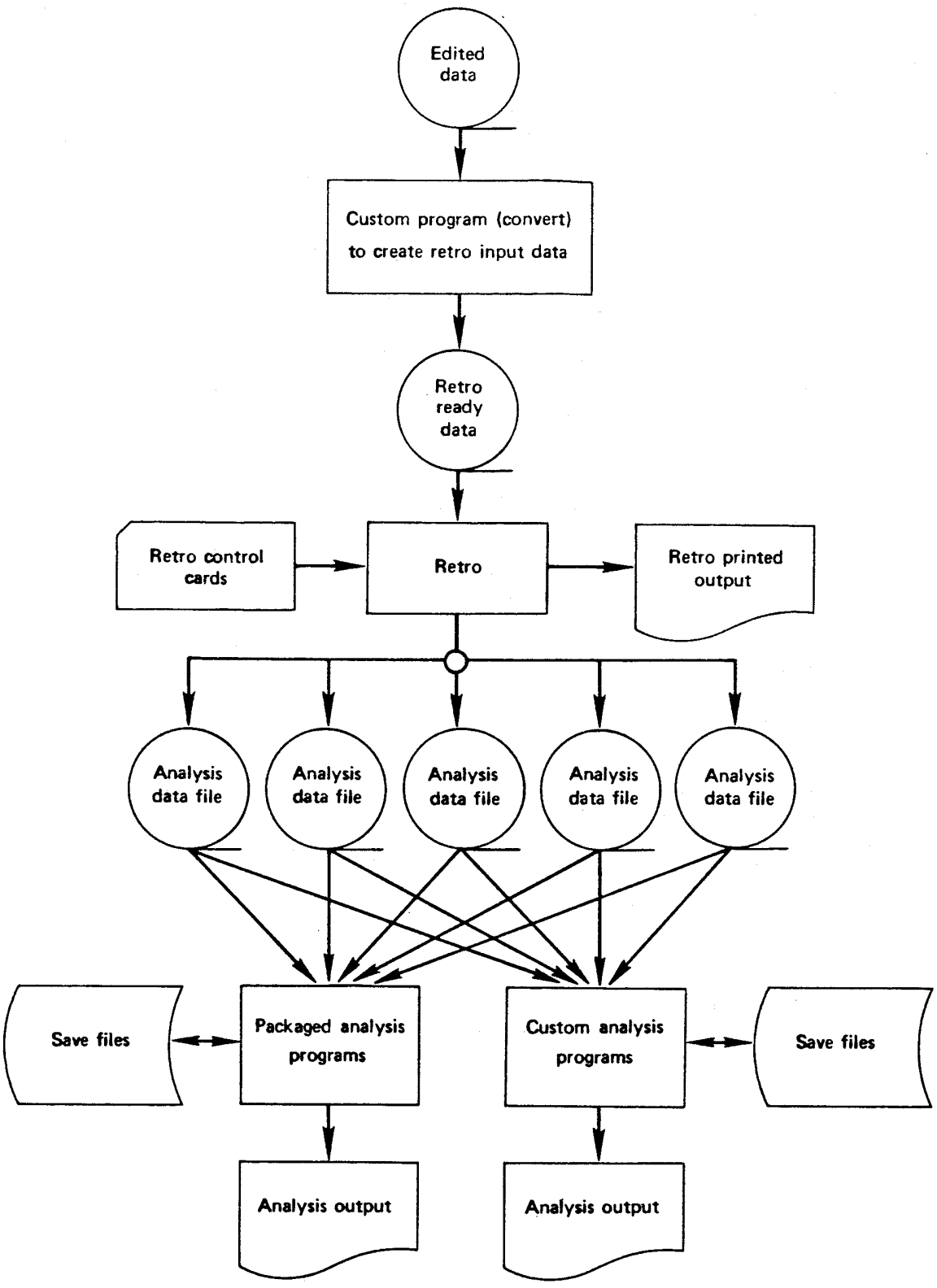
QUESTIONNAIRE	TOTAL SAMPLE SIZE	SAMPLE IN 4 RURAL VILLAGES	SAMPLE IN 2 URBAN VILLAGES
R13: Male Retrospective Life History	467	All male spouses less than 60 years old of women surveyed for R11	Random 50% of male spouses less than 60 years old of women surveyed for R11
R14: Income, Wealth, and Agricultural Production (2nd Round)		Random 50% sample of respondents for R10 in village 03. Random 20% of respondents for R10 in 3 other villages	Random 25% of respondents for R10
425: Census	1830	All families living in the villages	Same

Source: Corona, Rand Paper P-6181, pp. 1-2.

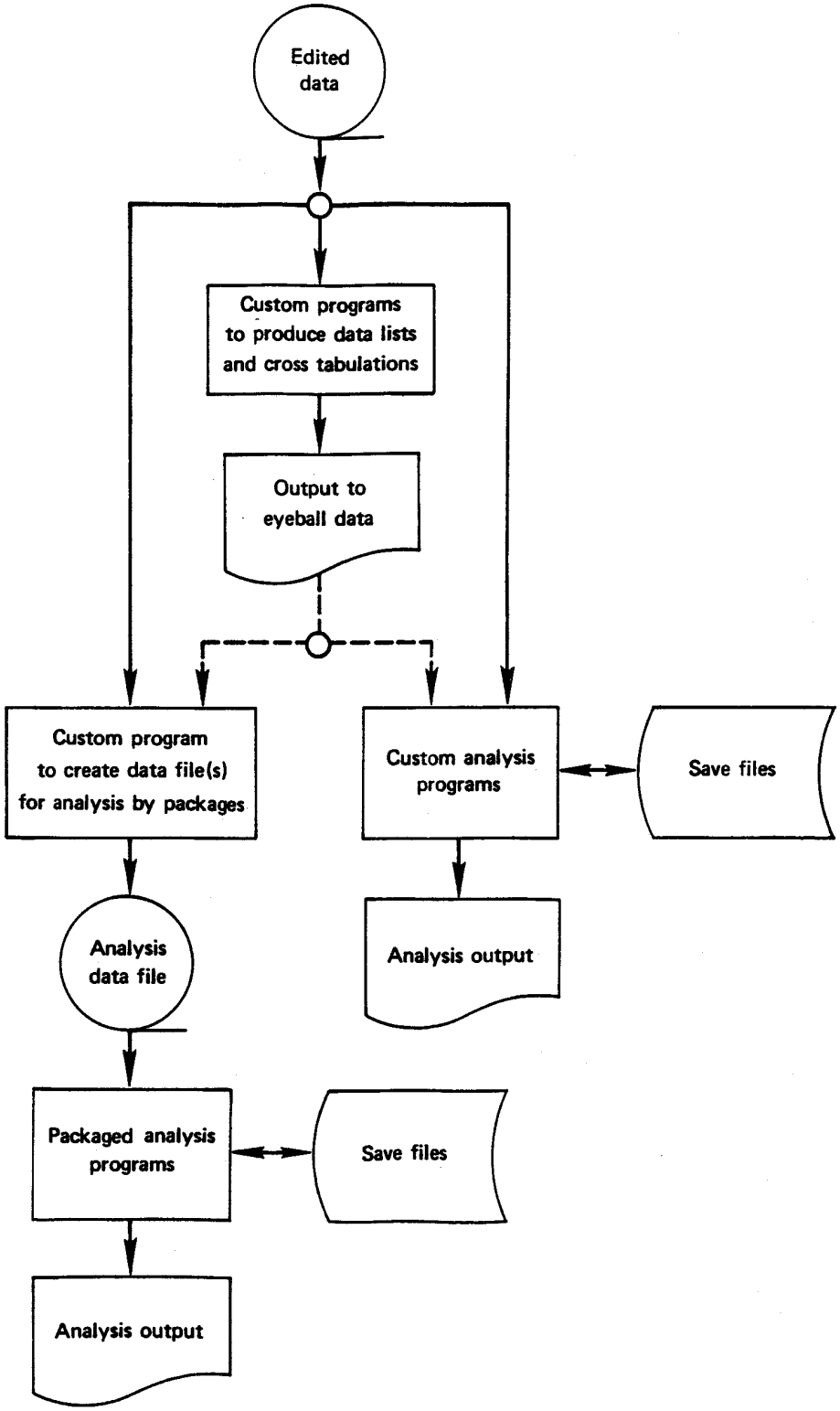


Appendix C: Flow charts of the three methods

I. S.I.R.


Edited data
Custom program (convert) to create retro input data
Retro ready data
Retro control cards
Retro
Retro printed output
Analysis data file
Analysis data file
Analysis data file
Analysis data file
Analysis data file
Save files
Packaged analysis programs
Custom analysis programs
Save files
Analysis output
Analysis output

II. Retro



III. Custom programming