UNITED NATIONS

ECONOMIC AND SOCIAL COUNCIL

Distr.
LIMITED

E/CN.14/SM/39
8 May 1981

Original: ENGLISH

ECONOMIC COMMISSION FOR AFRICA

Working Group on Organization, Content and Methodology of Household Surveys

Addis Ababa, 29 June - 3 July 1981

# SURVEY DATA ANALYSIS

C O N T E N T S



M81-1510

## I. INTRODUCTION

1. A short paper on survey data evaluation and analysis (E/CN.14/SM/27) was produced for the second meeting of the working group on organization, content and methodology of household surveys in October 1979. Its purpose was to draw attention to some of the basic requirements and the possible means of fulfilling them. Most of the main considerations will be repeated in the present paper with additional material arising from further thought and consultations.

2. However it must be said at the outset that the paper will not provide a set of technical guidelines for analysing survey data. Its limited objective is simply to identify the main analytical options in the light of probable applications and to put them as far as possible into a practical working context. In doing this it will be necessary to take into account other relevant aspects of survey operations and perhaps look at statistical and planning activities more broadly because the analysis itself cannot be considered in isolation.

3. During earlier discussions it was concluded that not enough is known about the subject to produce a satisfactory paper at the present time and that it would be better to concentrate on the initial steps of constructing survey data bases and the preparation of tabulations needed for analytical development.

4. There are two main reasons for the limited nature of the remarks in the present paper. First, data collection from households and other statistical units is becoming increasingly regarded as an integrated and continuing operation; conventional analysis is not fully satisfactory in this situation and new methods have to be developed. Second, data analysis in general has been badly neglected in most developing countries with the result that there is very little practical experience on which to base conclusions.

5. Nevertheless survey data collection is making considerable progress throughout the region and it is necessary to develop adequate analysis without delay so that the new data can achieve their full potential in planning and other applications. The paper will try to go a little beyond the data base and tabulation aspects but there is no guarantee that fully valid conclusions will be reached.

## II. ANALYTICAL REQUIREMENTS AND RESPONSIBILITIES

6. After participating in the 1970 round of population censuses a large number of African countries turned their attention to household surveys and the incomplete records at ECA show that more than half the countries of the region now have significant activities in this field. The African Household Survey Capability Programme (AHSCP) was a response to the renewed interest in surveys and provides assistance to individual countries in establishing or improving permanent field survey organizations for the continuous production of integrated economic and social data.

7. The need for information on employment, other productive activity, income, consumption and expenditure and the related demographic and social characteristics of households has therefore already been clearly demonstrated. We have now reached the stage where it is necessary to consider the uses of the data more carefully in order to determine the ways in which it should be analyzed.

8. Most of the previous documentation on AHSCP and its global counterpart NHSCP has pointed out that households are the suppliers of all labour and they are the recipients of the end-product of development efforts and economic problems. It is also notable that in recent years there has been growing concern with questions such as manpower availability and poverty.

9. It follows that the uses of household data relate broadly to the economic activities of people and to their levels of living. A little thought will show that the data applications in these two areas are an essential basis for most aspects of planning, administration and entrepreneurial activity.

10. It should also be mentioned that since the national survey programmes will generate a lot of integrated demographic, social and economic data, there is a whole range of themes or topics which can form the subject of analysis. For example it is possible from the subject-fields so far identified in some of the project proposals formulated to obtain enough data for such research topics as basic needs, rural development, poverty, economic and socio-demographic determinants of development, and agrarian reform. The type of analysis undertaken in connexion with the survey programme will thus be related to the specific objectives of the research.

11. In this situation it is clear that the analytical requirements in respect of household data are very wide indeed. For convenience two kinds of analysis can be distinguished. Firstly there is that which is concerned with the general account of economic and social structure and trends, dealing mainly in aggregated information. Secondly there is the deeper analysis of particular aspects of the economic and social situation which is largely based on the inter-relationships between variables. It covers the complex array of factors which determine the quality of life and the interaction of these factors with economic activity.

12. For aggregative data analytical frameworks already exist. On the economic side there is the UN System of National Accounts (SNA), the Material Product System (MPS) and various other arrangements, plus more detailed or specialized configurations such as the social accounting matrix and the more recent food accounting matrix. In the case of social data there are proposals for a framework in the System of Social and Demographic Statistics (SSDS) but there is currently a move towards a simpler approach in the absence of any common unit in which most of the data can be expressed.

13. For in-depth analysis based on inter-related data the position is much less satisfactory and guidelines are available only in a few subject areas. The World Fertility Survey has produced a great deal of good material in its own field. There are also similar techniques for the analysis of other demographic and social variables and the International Labour Office has provided guidelines for the analysis of income, consumption, expenditure and labour force data. However the overall position appears to be that, while recommendations on basic data requirements and tabulations are available for most subject fields, analytical guidelines are still limited and are strongest in the population field. The work undertaken by WFS, ILO and other agencies clearly has to be extended to additional subject areas.

14. However the main gap in available technical guidelines relates to the inter-relationship of data between different subject fields. This is something new arising from the effort to establish continuing and integrated programmes of surveys and it calls for new analytical methodology. The requirements in this respect will be kept in view throughout the remarks in the present paper.

15. A provisional list of the data to be collected for the analytical purposes outlined above has already been published in the paper on household survey data requirements, E/CN.14/SM/22, which is being circulated to the meeting as a reference document.

16. The second question to be considered in this section of the paper is who should be responsible for survey data analysis. At its first session in 1980 the Joint Conference of African Planners, Statisticians and Demographers considered means of strengthening the relationship between statisticians and planners and came to the conclusion that data analysis, neglected in many countries, is at least to some extent the missing link. The Conference felt that analysis should involve both the producers and users of data. There are clearly advantages for the quality of the analysis and the effectiveness of planning and other applications if this can be done.

17. The extent to which statisticians and data users can collaborate in survey data analysis will have to be determined in the light of practical experience and the technical expertise of the persons available. All the basic data processing and much of the computing work leading to the analysis will of course be the task of the statisticians but, as a minimum initial step, the relevant users must be consulted when a tabulation programme or analysis is planned and they should participate in the interpretation of the results.

## III. ACTIVITIES PRIOR TO ANALYSIS

18. There are a number of survey activities before the analysis stage which should be mentioned because they have a bearing on the analysis itself. They relate mainly to the quality of data and it is also useful to have an idea of where analysis begins after these preliminary steps.

19. The normal procedures for quality control are of course most important at all stages of a survey but it should be noted that there is some difference between short single-visit surveys and those which are spread over a longer period and may involve single or multiple visits. The accuracy of results from a short survey is dependent mainly on the care put into its preparation and the adequacy of supervision during the fieldwork. There is no time for intervention by the central office running the survey.

20. In surveys which go on for a longer time there is the additional possibility of checking results while the fieldwork is in progress and amending recording arrangements, etc. before any detected error has done significant damage. The requirement in this respect is to make summaries of the records, probably at monthly intervals, which enable the performance of the survey and the consistency of the data to be kept under continuous review. There is an added advantage if the summaries can be compared with data from independent sources such as population census records, marketing board purchases and exports. A process of continuous checking during a long survey can greatly improve the quality of data. It is surprising that so many African statisticians take the unnecessary risk of not watching survey progress in a systematic manner.

21. Survey records are subject to the same kind of manual and automatic editing as is used for population censuses. However there is a very much smaller bulk of data from a survey and the editing process ought not to introduce significant corrections if field supervision has been adequate. In particular it is desirable to keep the imputation of missing data and correction of biases to a minimum because there can never be any certainty that such adjustments are improving the results. Whenever any changes are made in the original data it is important that they should be recorded in detail.

22. Data evaluation is probably the most essential of the preliminary steps towards analysis. It comes mainly towards the end of the data processing stage but should be regarded as an integral part of the work throughout the survey. One important aspect has already been mentioned under quality control and the editing procedure also serves as a guide to the accuracy of records. The calculation of sampling errors is another part of the evaluation exercise. Other means of evaluation include the use of inter-penetrating sub-samples and re-interviewing a sub-sample of households but they have to be incorporated in the survey design,

particularly when there is a continuing programme of surveys. All of the evaluation leads to decisions on the extent to which the survey data are good enough to support detailed analysis.

23. There is another topic worth mentioning which concerns estimation procedures rather than the quality of survey data. For various practical reasons the sample units in a number of African surveys have not been selected in accordance with the survey designs. This departure from planned selection probabilities leads to biased estimates if the error is not detected, or to a great deal of additional work in re-weighting survey results. Such a situation is a serious handicap for analysis and is avoidable.

24. As already indicated the tabulation programme for data processing has to be prepared in collaboration with the users of the survey results, otherwise it may not be relevant to the analysis. The tabulations will usually be reproduced in the survey report. The text of the report describes the survey and its results and, depending on the extent to which it attempts an interpretation of the results, parts of it may be regarded as analytical. It is therefore difficult to say precisely where data processing ends and data analysis begins. For the purposes of this paper data analysis is considered to be everything that is done to the survey results after the reporting stage.

## IV. ANALYTICAL OPTIONS

25. As indicated in the section on requirements it is convenient to distinguish two kinds of analysis: that concerned with aggregative data and that dealing with the inter-relationships between variables. The second of these can be sub-divided according to whether the data come from the same or different surveys.

26. Not much needs to be said here about analysis in terms of aggregative data. Survey results are simply used as a data source for appropriate components of a standard framework. To the extent that several surveys may contribute to the same framework, the framework itself serves as a means of integrating the data and may help in showing relationships between topics.

27. The inter-relationship of variables investigated in the same survey is also a straightforward matter. Tabulations prepared at the data processing stage will suggest possible relationships that need to be examined in more detail and the analysis will be of a multi-variate nature. In addition there is the specialized analysis applicable to various subject fields which has already been mentioned.

28. The problem area is the inter-relationship of variables from different surveys and it is here that practical experience is lacking. Some methods that have so far been identified are mentioned below.

29. It is necessary for all surveys to use the standard concepts, definitions and classifications appropriate to the subjects investigated. This is a pre-requisite for any kind of data comparison rather than a method of analysis.

30. A provisional set of core questions has been published in document E/CN.14/SM/22. If these questions are asked in all surveys irrespective of the subject investigated they provide common variables which enable survey results to be classified in the same ways, thus enabling the examination of inter-relationships.

31. A special case is that of "core surveys" which investigate agriculture and other important topics continuously and deal with miscellaneous inquiries on the basis of modules. These still need core questions of the kind indicated above if the sample at the household level is changed fairly frequently.

32. In cases where the samples of different surveys have common units at one or more stages it is possible to relate the survey results aggregated for those units. It is not expected that this will be feasible at the household level except where partial replacement arrangements are used.

33. In all methods of the kinds indicated above it is necessary to bear in mind the different timing of the various surveys. In the case of "core surveys" the main point of interest may be to explain changes in the same variables over time and there is no problem. When different variables are examined in relation to one another the times at which data collection took place need rather careful consideration in interpreting the results of the analysis. Similar considerations of course apply when estimates from different surveys are pooled or when a time series analysis is undertaken.

## V. THE DATA BASE APPROACH

34. As the data processing of most surveys is now computerized most countries should consider the construction and use of a permanent data base of survey records. It will presumably contain one micro-data file for each of the surveys undertaken.

35. In the light of the remarks in earlier paragraphs the main functions of each data file will be as follows:

(1) Preparation of the survey tabulations at the data processing stage.
(2) Calculation of estimates, sampling errors, etc as envisaged in the sample design.
(3) Additional specialized analysis recommended for the subject field concerned.
(4) Multi-variate analysis of the variables recorded in the survey.
(5) Classification of the survey variables by selected core variables.

(6) Aggregation of survey results for individual sample units at all levels.
(7) Provision of data for ad hoc purposes and research.

36. The interesting point about the seven file functions is that all relate to operations performed on individual files. The only situation where two or more files might have to be used simultaneously or merged is where different surveys have used the same sample of households which, as indicated above, is likely only when partial replacement is used. Even for items (5) and (6), which are directly concerned with the inter-relationship of variables recorded in different surveys, the data needed can be extracted from individual files before the analysis is actually carried out.

37. The implication is that there is no rigid constraint in terms of a need to impose a fully standardized file structure for all surveys. However for ease in handling the data and to reduce the problem of preparing software it is of course desirable that all files should be as similar as possible. The important point that emerges here is that the possibility of some flexibility in the construction of individual survey files is a considerable asset when it is borne in mind that different sample designs and different kinds and quantities of data have to be accommodated and that people's ideas on how to do the job necessarily develop over time.

38. It is envisaged that the individual survey files would normally be of a simple sequential nature, with each record carrying all the data for one ultimate-stage unit, and the basic storage medium would be magnetic tape. This remark does not apply to data from the preliminary enumeration of penultimate-stage units which obviously requires separate files; they are important but cannot be considered in the present paper.

39. The status of the data base itself as a permanent source of information also calls for some comment. At present most survey data are disseminated in formal survey reports which are of only limited use for further analysis. It can be expected that the computerized data bases will progressively take over and add to the functions of the reports. The methods of operating such data bases will be a matter of concern for the future.

40. The basic survey files will contain a very large quantity of micro-data which are needed only for the file functions (1) -- (6) listed above and perhaps a few others of a similar nature. These data cannot be made available outside the statistical service because of the confidentiality clause in the statistical legislation of most countries. Therefore for any detailed research undertaken by external agencies as envisaged in file function (7) it is necessary to prepare edited micro-data files which meet the confidentiality regulations.

41. For the dissemination of more general information files at micro-data level would be too cumbersome. The easiest solution is a summary file based on some of the data already discussed and suitable for on-line transmission. Some thought and experimentation is needed in determining the file content but topics of interest seem to be the general results of individual surveys, conclusions regarding the inter-relationship of economic, social and demographic variables derived from the analysis of all surveys, plus data on the smallest geographical areas for which the surveys can provide reasonable estimates. The details of item (6) are not of general interest because they relate only to a sample. It is assumed that there would be separate general-purpose files for national accounts, other economic and social data but they are not discussed here.

## VI. FURTHER NOTES ON ANALYSIS

42. According to the definition of survey data analysis given earlier in the paper this section should deal with the data file functions (3) - (6). It is not proposed to consider item (3) because recommendations for special analysis in some subject fields are available elsewhere and revised proposals for labour force data are being presented in a separate paper to the meeting.

43. Item (4) relating to multi-variate analysis of data from individual surveys is the next consideration. Ideally the work should start by calculating the correlations of all the pairs of variables, including core questions, recorded in the survey and then proceed through a selection of the strongest relevant relationshipsto the establishment of multi-variate regression equations. Fortunately the amount of arithmetic can be reduced because the initial tabulations should give some indication of data relationships and some groups of variables, e.g. those on income and expenditure, can be dealt with on an aggregated basis. It is doubtful whether the analysis needs to go as far as model-building, at least in the earlier stages, because the main purpose of investigating the inter-relationships is to find an explanation of specific economic and social conditions for policy and planning purposes.

44. One example of the kind of work that has to be done comes from the World Fertility Survey and is reported in its paper "Strategies for the Analysis of WFS Data" (WFS Technical Paper No. 9, January 1977). WFS classified all the variables included in its surveys as dependent, intermediate (proximate) and explanatory with respect to fertility. On this basis it was able to establish a framework for fertility analysis. Because the field of study was narrow only a few of the explanatory variables appear to be useful in establishing linkages with other data, but it is now recognized that there was a problem in identifying and using these variables. However the main point of interest is the extent to which variables can be classified as explanatory or dependent in the multi-subject operation envisaged by AHSCP. For individual subject fields the position is presumably

much the same as that encountered by WFS and a satisfactory distinction should be possible but perhaps the main requirement with respect to the broader question of linkages between subjects is to find the strongest statistical relationships. It may also be mentioned that similar studies on demographic, social and economic interrelationships are implied in the research work of ILO's World Employment Programme involving the application of the Bachue model.

45. Item (5) is concerned with the classification of survey variables by selected core variables recorded in the same surveys, which is the first step in examining the inter-relationships of data from different surveys. The basic questions are: which core variables should be used and is the existing list of core questions satisfactory?

46. The list of core questions given in document E/CN.14/SM/22 is of a provisional nature and contains two kinds of questions. There are those which tentatively can be expected to serve as a means of linking data from different surveys and others which have been inserted because they might be useful in this connexion but in any case provide information which is valuable enough to be recorded at all opportunities. There is therefore still a great deal of work to be done in identifying the true core questions which can be used for analytical purposes.

47. Internal analysis of individual surveys will be one of the principal means of selecting these questions. When a particular variable is classified as explanatory in several subject fields it obviously has possibilities for analysis provided its correlations with variables in those subject fields are strong enough. It can be envisaged that the end-result will be several lists of core questions applicable to overlapping groups of subjects and that the lists will be rather short. The questions themselves will probably relate to main indicators of the socio-economic conditions and activities of households, plus basic factors such as size and composition of households needed for the standardisation of records to the extent possible. A few other kinds of core variable may emerge during the analysis outlined above and there is also the more general question of the extent to which socio-economic indicators will be useful. Nothing more can be said until countries have experimented with their survey results.

48. The method of linking results from different surveys by using core questions is of course indirect because the relationship between variables has to be examined through their individual relationships with one or more other variables, i.e. the core questions. It is for this reason that the analysis has to concentrate on the strongest meaningful statistical relationships that can be found.

49. There is just one comment in respect of file function (6) which deals with the aggregation of survey results for individual sample units. The aggregations will probably be most useful at the one or two area stages used in most African surveys and they do have to take the form of estimates of population values for

the individual units. As already indicated this method of linking the results of different surveys is applicable only when the surveys have the same sampling units at the stages concerned. However the unit estimates also serve as a short cut in calculating sampling errors.

## VII. CONCLUSION

50. The present paper is only a limited improvement on E/CN.14/SM/27 presented to the 1979 meeting of the African working group on household surveys but it does manage to look at survey data analysis in terms of the practical steps which have to be taken and more emphasis has been placed on the importance of continuous collaboration with data users.

51. Only one new concept has been introduced and it relates to the establishment of national survey data bases. Two fairly important conclusions in this connexion are: (a) it is not too difficult to construct the data base because most of its uses can be confined initially to individual survey files and (b) the discipline imposed by a data base and the related identification of file functions make it much easier to organize analytical work in a systematic manner.

52. There is no claim that the paper is a comprehensive review of the current situation and some deliberate omissions have been mentioned in the text. However the paper is probably one of the first to attempt an overall look at the analytical treatment of data produced by continuing multi-subject programmes of surveys.

53. Comments on the prospects for linking survey variables have probably been unduly conservative. There are obvious relationships between the economic, social and demographic situations affecting household productive activities and levels of living and the kind of analysis proposed will help to identify and examine them. The main misgiving is that some other kinds of analysis have not been taken into account. One is the Living Standards Measurement Study of the World Bank and another is the Food Accounting Matrix of the Food Supply Analysis Group at Oxford University. There is clearly a need for integration of thinking by interested agencies and it seems that the results should be applied in terms of advice on the construction and operation of survey data bases at national level.

54. In this situation the main practical requirement is an attempt by individual countries to analyze their own survey data. Until some numerical results are available international agencies able to provide technical advice will be groping in the dark.