14952

Distr.
LIMITED

UNIDO/IS.558
9 September 1985

UNITED NATIONS
INDUSTRIAL DEVELOPMENT ORGANIZATION

ENGLISH

INDUSTRIAL STATISTICS FOR RESEARCH PURPOSES* ;

Methcdology Applied in Compiling UNIDO's International Data
on the Number of Employees, Wages and Salaries,
Gross Output and Value Added.

Prepared by the
Statistics and Survey Unit
Division tor Industrial Studies

* This document has been reproduced without formal editing. The designations employed and the presentation of material do not imply the expression of any opinion whatsoever on the part of the Secretariat of the United Nations concerning the legal status of any country or its authorities, or concerning the delimitation of its frontiers.

CONTENTS

INTRODUCTION

The availability of reliable and comparable data is an essential requirement for an accurate assessment of economic progress or technical assistance needs. Given UNIDO's mandate in these two fields, the Organization undertook the establishment of the UNIDO Data Base (UDB) in 1977. The project was originally motivated by the fact that existing industrial statistics suffered from certain drawbacks that limited their usefulness. This problem, of course, is not unique to industry; other types of data concerning trade, agriculture, labour, national accounts, etc. suffer from many of the same difficulties. However, it is arguable that the magnitude and extent of such problems is somewhat more severe for industrial statistics than is encountered in other fields of statistics. Consequently, the development of the UDB was undertaken by the Statistics and Survey Unit (SSU) of the Division for Industrial Studies. The purpose of the project was to centralize the Organization's statistical activities and to develop a set of detailed and comparable industrial data for users both inside UNIDO and outside.

Since its inception, the statistical programme of SSU has emphasized the need to provide an internationally comparable set of industrial data to be used in the organization's programmes for technical assistance and applied research. Two important aspects should be immediately noted. First, UNIDO is primarily a user (rather than a producer) of industrial statistics. The organization makes extensive use of the data collection efforts of other statistical organizations whether international, supra-regional, regional or national. Second, the UDB is primarily intended to meet the statistical needs of researchers engaged in international, or cross-country, studies rather than country-specific or time series investigations. Accordingly, UNIDO statisticians give priority to the development of a set of international statistics which meets acceptable standards of consistency and comparability in terms of the statistical definitions and concepts used in compiling the country data.

The UDB now constitutes the major source of data for several recurrent publications produced by SSU. These include: A Statistical Review of the

World Industrial Situation, the *Handbook of Industrial Statistics* and the *Industrial Development Survey*, as well as many *ad hoc* studies. But the UDB is also available, on-line, to users throughout UNIDO who have need for industrial data.

Finally, copies in machine-readable form of the UDB can be purchased by users outside the organization on an annual basis. For information on purchase procedures, readers should refer to the Statistics and Survey Unit, UNIDO, P.O. Box 300, A-1400 Vienna, Austria.

I. THE UNIDO DATA BASE

A. Contents of the UNIDO Data Base

The UDB includes annual figures measuring five variables: the average number of employees, wages and salaries, gross output, value added and index numbers of industrial production. All these statistics are compiled by country and are reported on an annual basis spanning the period 1963 to latest year. The data are presented according to the ISIC[1/] which provides for 28 industrial branches within the manufacturing sector.

While relying largely on data compiled by the Statistical Office of the United Nations Secretariat (UNSO), the UDB is original in the sense that it includes UNIDO estimates accounting for over one quarter of the total number of entries. UNIDO estimates may be based on national sources, they may include figures taken from, or based on, supplementary statistical sources of national or international origin or they may have been collected in the course of field work by UNIDO staff and consultants.

Although the UDB contains over 200,000 entries, coverage is considerably less than the theoretical maximum which would assume that data were available for all cells. Frequently, observations related to one or several variables are not available for all years and for all branches. And for a few countries, many of the desired observations are missing.[2/] Many of the missing observations occur in countries or branches which are of minor importance with respect to their contribution to MVA. More important, however, is the fact that in a number of cases, no activity in a particular industry is actually carried out by the country in question. The frequency of such occurrences is not accurately known because the countries rarely indicate

[1/] For a complete description of the ISIC, see International Standard Industrial Classification of All Economic Activities, Statistical Papers, Series M, No.4, Rev.2 (Sales No. E.68.XVII.8), United Nations, New York.

[2/] For a detailed listing of the contents of the UDB see "An Inventory of Industrial Statistics: UNIDO Data Base 1985" (UNIDO/IS.528).

zerc values to signal the absence of a branch, but it can be safely assumed that this fact would explain a significant number of the missing observations.

B. Structure of SSU's Statistical Programme

The statistical programme is composed of three basic functions: development of the data base, improvements in the international comparability and consistency of the data included. In practice these functions are often performed simultaneously, however, for the sake of clarity, they will be described separately.

Development of the data base consists in enlarging the number of entries in the data base from several sources of information. One source is the annual update incorporating official data supplied to UNSO by national statistical offices and forwarded to the SSU on magnetic tape. These entries include additional data corresponding to the latest year as well as revised data corresponding to previous years. Another source is the entry of data from supplementary suppliers or publications. Finally, estimations of missing observations relating to certain variables, branches or years are carried out. This activity also includes the systematic production of provisional estimates covering the most recent years for which information is expected at a later date but is not yet available. These activities are carried out in five procedural stages, based on an ordering of the source of information which is described in detail in chapter II.

The second function, improvements in the international comparability of the data, constitutes perhaps the most intricate aspect of SSU's programme. Data incomparability arises from differences in the national reporting practices of countries. These differences involve mainly three factors. First, the classification used in the UDB is the 1968 version of ISIC at the major group (three-digit) level. And where the national classification either differs from ISIC or is less disaggregated than the three-digit level, SSU attempts to convert the data to the desired system and level. Second, coverage is frequently incomplete, that is data items are gathered only for statistical units defined by certain characteristics (size, type of ownership, types of legal organization, and so on). To the extent that these characteristics vary from one country to another international comparisons are jeopardized. However, SSU is endeavouring to estimate figures relating to the desired coverage of all establishments with five or more employees. Third, definition of data items may differ across countries. Output, for instance, is sometimes expressed in producers' prices, sometimes in factor values. Because some branches are highly sensitive to the definition attached to the data, this source of variation may impair comparability. Here again, efforts have been made to improve the situation.

The third function involves data consistency and results not only from the diversity of data sources used but also the lack of internal consistency affecting many of these sources. The classification, coverage and definition of data published by a given source may differ according to years, branches and variables; furthermore, errors may have slipped into the data publishing and dissemination processes. To enable users of the UDB to construct time series and to calculate indicators combining several variables, consistency must be ensured. To this purpose SSU implements a systematic screening of the data and attempts to redress identified inconsistencies.

II. DETAILED PROCEDURES

A. Maintenance and Development of the Data Base

Performance of this function is closely related to the design and internal structure of the UDB which incorporates five distinct and separately identifiable stages.[1/] At each stage, the data are subject to various forms of examination and/or adjustment. The purpose of the structure is twofold. First, the layout facilitates the organization of the work of UNIDO statisticians - both with regard to statistical methods and data sources. A 'stages approach' serves to impose priorities on the work of statisticians with regard to the sources which they utilize and, indirectly, the statistical methods they employ. As explained below, information from national sources (published and unpub'ished) receives a higher priority than similar data supplied by international sources. Accordingly, different sources for data modification are associated with each stage in the data base. Second, the stages approach lends itself to more extensive use of the computer for data screening, analysis, editing, etc. This is a natural consequence of the fact that development of the data base is broken into a series of specially defined tasks which are carried out in a pre-determined sequence. The following discussion summarizes the maintenance and development of the UDB in terms of the stages in the data base.

Stage I - Responses to national questionnaires compiled by UNSO

At this stage the data is a duplicate, in machine-readable form, of that provided to UNIDO by UNSO. Information has been compiled from replies to questionnaires sent by UNSO to national offices. Prior to its receipt by UNIDO, the data has already been subject to a certain amount of screening by UNSO to determine consistency and accuracy. Upon completion of this exercise,

[1/] Additional stages may be added at a later date. The incorporation of later stages is contingent on the successful development of standardized methods of estimation, using both time series and cross section approaches, to deal with outstanding types of inconsistencies which remain in the data base after completion of stage IV.

UNSO publishes the results annually in the Yearbook of Industrial Statistics - Volume I, General Industrial Statistics. Consequently, the condition of the data already reflects the results of considerable work to identify and to document any known departures from the International Recommendations for Industrial Statistics.

Stage II - The incorporation of national data

Work begins by drawing upon UNIDO's collection of national statistical publications. These consist of industrial censuses, surveys and a smaller number of input-output tables. Statisticians compare the information in the national publication with that communicated to UNSO for the corresponding year(s). In doing so, they rely heavily on the earlier work of UNSO to identify and document departures from the ISIC. In some instances the two sets of data are identical. However, the approach and orientation of the two organizations differs.[1] But the mandate of UNIDO statisticians is to take account of - rather than to document - significant departures from the international classification. Examples of the adjustments carried out by UNIDO at this stage include the following: (a) the reconciliation of discrepancies previously noted by UNSO; (b) the inclusion of final, rather than preliminary results; (c) the incorporation of data from additional national sources; and (d) the inclusion of statistics compiled through UNIDO field work.[2]

Stage III - The inclusion of data from additional international sources

In extending the coverage of available data, UNIDO gives highest priority to the UNSO data (stage I), as supplemented by national data (stage II).

[1] UNSO follows the International Recommendations for Industrial Statistics and the concepts and definitions developed in connection with the International Standard Industrial Classification (ISIC). The approach, as stated by UNSO, is as follows: "Deviations from these standards, where known, are mentioned in the introductory note to each country chapter or in the footnotes to each table." In other words, UNSO indicates discrepancies but does not adjust or correct them.

[2] For further discussion, see UNIDO, "The UNIDO Data Base: Primary Sources and Data Base Design" (UNIDO/IS.463, pp. 14-15).

However, other international institutions also compile industrial data concerning a variety of subjects and, sometimes, in considerable detail andthese sources have been found useful in further development of the UNIDO database. Initially, international data sources are carefully screened by UNIDO statisticians to determine their quality, coverage, scope and definitions. Where UNIDO standards are met, the information is considered by statisticians in their work on the data base. Uses for such data may include adjustments in the data emerging from stage II. The purpose of these adjustments may be to ensure compatibility in terms of coverage, account for departures from the ISIC, include previously missing observations, etc.[1] In addition, unpublished data, mainly in machine-readable form, are obtained from various institutions and are incorporated in stage III when this is practical.

Stage IV - The estimation of data from results obtained in previous stages

Upon completion of stage III, the various sources of industrial statistics, in descending order of priority, are UNSO, national publications and international sources. At this point all adjustments have been made with the help of supplementary information from exogenous sources.

The most obvious way of extending the work from stages II and III is to make use of these results to estimate data for which exogenous information was not available. Thus, information gathered in stages II and III is utilized to make further adjustments for years other than those considered in earlier stages.

[1] Published sources which have been utilized in connection with the work at stage III include Eurostat, ACP, Statistical Yearbook; ECLA, Statistical Yearbook for Latin America; ECWA, Statistical Abstract of the Region of the Economic Commission for Western Asia; ESCAP, Statistical Yearbook for Asia and the Pacific; ILO, Yearbook of Labour Statistics; UNESCO, Statistical Yearbook; United Nations, Yearbook of National Accounts Statistics, Statistical Yearbook and The 1973 World Programme of Industrial Statistics; United Nations Office for Development Research and Policy Analysis, Handbook of World Development Statistics.

Stage V - Provisional estimates for latest years

Official statistics are often reported with a time lag of several years. The actual duration of the lag will vary from country to country and even a particular set of national statistics may differ depending on variable and industry. Under the best circumstances, the lag will be at least two years (i.e. in 1985 the latest available observation would refer to 1983). However, the number of countries reporting with this minimum lag are few. Roughly one half of the developed countries and only 15 per cent of the developing countries report with a two-year delay. And even among these countries the statistical coverage for the latest year is often incomplete. In fact, the minimum time lag required to obtain coverage for 90 per cent of the countries in the UDB is the latest five years.

Such delays obviously jeopardize the usefulness of statistical data. In order to redress this problem, operations in stage V are concerned with the development of estimates for more recent years for the following variables: number of employees, wages and salaries, gross output and value added. The estimates are all provisional and are eliminated as official statistics become available.

Five steps are involved in compiling the provisional updates included in stage V. The first step concerns only the adjustment of time series for each country to a common terminal year. Often, the latest reported year in a time series will differ and, in such cases, statisticians extend the abbreviated series. The approach is based on the assumption that ratios between variables (all of which are observable from national data available for previous years with adequate coverage) will be applicable to the more recent years. After completion of this step, the data for each country has been extended to a common terminal year. However, the terminal year may differ from country to country and subsequent operations are intended to align the time series for all countries to one common terminal year.

In step 2, value added in manufacturing in every country is extended to a common terminal year - a lag of only two years in relation with the current year. In doing so, use is made of the national accounts compiled by the United Nations Office for Development Research ar Policy Analysis (DRPA). These series have a time lag of only two years and provide estimates of MVA (at current prices) expressed in terms of national accounting definitions for most of the countries in the UDB.[1] Under the assumption that total value added, calculated according to industrial census concepts, will grow at the same rate as the corresponding figure when expressed in a national accounting framework, the growth rates of DRPA series are used to estimate total MVA updates for countries with a time-lag of more than two years.

In a third step the estimates obtained for value added in manufacturing are used to derive the corresponding totals for other variables on the assumption that the ratio between variables (e.g. value added as a share of gross output or wages as a share of value added) remain stable over short periods.

At this point in the exercise, the outcome is a set of estimates which lag two years behind the current year and refer to the total for the entire manufacturing sector and to each of the four variables. In a fourth step, updated series of manufacturing totals are extended for one additional year by an extrapolation which uses a constant growth function fitted over the last seven years. The results are then a set of sector-wide estimates for each of the four variables which are lagged by only one year. In the final step these totals are disaggregated in 28 branches on the basis of the inter-branch distribution for the respective countries in previous years.

In conclusion, the five stages which compose the data base not only provide an operational framework for SSU statisticians but also present advantages to the data user. Each data item on the UDB tape is accompanied by a numerical "source" code, from 1 to 5, indicating the stage at which the figure was ultimately derived. Codes "1" and "2" refer to national data sources (belonging to stages I and II, respectively); code "3" means either

[1] No information is provided for centrally planned economies.

that the data as supplied by national sources did not meet SSU standards, or that no national information was available, but that a satisfactory estimate could be generated from international sources; code "4" indicates that the gaps left after stage III are completed with the only help of data derived at stages I, II or III; code "5" means that the data are provisional estimates.

B. Improving International Comparability

Efforts to adjust the data for greater international comparability constitute the bulk of SSU's data development programme. Although work on this aspect is continuous, data for all but ten of the 150 countries and areas included in the UDB have been adjusted or supplemented in some way by SSU, using a wide range of additional sources.

Inter-country differences in the reporting of industrial statistics derive mainly from three factors: (i) the use of national classifications which do not conform to the ISIC; (ii) incomplete coverage or total absence of national data relating to certain variables, branches or years; and (iii) variations in concepts or definitions used. Such differences may also emerge within time series for individual countries and, because the UDB begins with data for 1963, these too can affect the continuity of a country time series. The following sections review the potential sources of incomparability, discuss their numerical effects where applicable, and describe the methods used by SSU for data adjustment.

1. The industrial classification

Most countries either use the ISIC or a compatible classification. Even so, 106 of the 150 countries or areas included in the data base report at least some data which are combined for two or more (3-digit) branches of the ISIC. The reasons for this vary. In some cases the practice reflects vertical integration of industrial activities within reporting units lacking records for their statistical separation; in others, national disclosure rules may require that activity in one or more branches of industry not be shown separately, especially if the number of reporting units in the branch or branches is very small; in a few cases, the national industrial classification may not be convertible to the ISIC.[1] While the reasons are valid, their effects, i.e. "combined" data, present serious limitations for cross-country comparisons. especially if data users are attempting to analyse international data on a specific industry or a few selected branches of industry.

Roughly half of the adjustments made by SSU have involved disaggregation of data referring to two or more industrial branches. The choice of a series for estimation purposes depends not only upon data availability but also upon the goodness of fit between the disaggregated data and the combined data. In the use of supplementary information, the preferred approach was proportionate distribution of combined data according to shares of data for the same or a proxy variable from another source which reports for each branch separately for the time period required. Alternatives included the similar application of data for other years, or the use of "related" variables[2] (e.g. the use of data on gross output to disaggregate value added, or of data on the number of employees to disaggregate wages and salaries). Methods are discussed below.

[1] Two less frequent but related problems are the inclusion of certain mining industries with the data for manufacturing, and the assignment of data for certain manufacturing activities to branches of the mining sector. The former condition usually applies to the data for ISIC 35, where the extraction activities that supply raw materials for further processing (crude petroleum and natural gas) may not have been reported separately, and occasionally to ISIC 36, where the extraction of other non-metallic mineral products may not have been separated from associated downstream activities. The latter condition refers to the assignment of petroleum refining and the petrochemical industries to the mining sector in a few major crude oil-producing countries.

[2] The terms "proxy" and "related" variables, as used in this report, do not have the same meaning. Proxy variables are those which measure roughly the same dimension of industrial activity. Related variables are pairs of variables which, although they do not measure the same dimension of industrial activity, are so closely tied that one might be predicted from the other. The distinction is necessary because the two types require a different method of treatment for estimation purposes. For further discussion, see below.

(i) Same variable, same years: For any country, two sets of data from different sources or vintage purporting to show information on the same variable, branch and year may report somewhat different figures because of possible differences in coverage, in sources of the data, in concept/definition, or in the stage of revision or correction of the data. Even so, for many applications, and particularly for those to which UNIDO is oriented, it may be assumed that such data sets are reasonably compatible. Consequently, the use of branch shares from one data set to disaggregate ISIC combinations in another is an obvious way of adjusting the original data for international comparability.

(ii) Same variable, other years: This approach has been used for the disaggregation of combined observations carried out in stage IV of the UDB. If the combined data were surrounded in time by branch-level data, shares from the surrounding years were interpolated. In all other cases, shares in the nearest available year were applied. The use of structural patterns in neighbouring years, while less desirable than resorting to exogenous information for the precise year in question, is reasonable on the grounds that structural changes occur only gradually.

(iii) Proxy variables: Proxies which measure roughly the same dimension of industrial activity as the problem variable were also utilized. For example, the main difference between gross sales and gross output is the change in stocks of finished goods and work-in-progress. This, of course, would vary among industrial branches and from year to year, but in the absence of precise information on gross output, gross sales have been accepted as a proxy for use in estimation.

(iv) Related variables: Related variables are pairs of variables (such as employment and wages and salaries, or value added and gross output) which are so closely tied that estimates for one of them might reasonably be predicted from figures for the other. However, unlike proxy variables, related variables do not purport to measure the same dimension of industrial activity. Therefore, they should not be used in the same way for splitting branch combinations.

The above point needs further explanation. At the branch level of industry, although wage rates (wages and salaries per employee) or value added/gross output ratios may be expected to change gradually through time, it is widely recognized that considerable variation can exist across branches within a given year, due to inter-branch differences in capital intensity, technology, the level of manpower skills required, etc. This is true even for branches within the same ISIC 2-digit division,[1/] where ISIC branch combinations are commonly found. The use of derived indicators (i.e. ratios[2/]) of related variables - rather than their shares - to disaggregate combinations would allow these inter-branch differences to be taken into account. In addition, the approach requires only the reasonable assumption that year-to-year changes in a derived indicator will be the same for all branches appearing in the original combination.

In order to use this approach, it was necessary to have (i) complete branch-specific data on one of the related variables for all years (ii) branch-specific information (either within the data base or from a supplementary source) for the chosen derived indicator (e.g. value added/gross output ratios or wage rates) in at least one year. Then, by linking the derived indicators in the overlapping year or years using ratios to the same indicator for combined branches, branch-specific estimates for missing years could be derived. These in turn could be applied to the related variable for which branch data were complete for all years, to yield disaggregated

[1/] For example, in ISIC division 32, which includes textiles (321), wearing apparel (322), leather and products (323), and footwear (324), wage rates are known to vary widely.

[2/] A "derived indicator" is simply a ratio, but is not so labelled because the word "ratio" must be reserved for another step in this approach.

estimates for the variable.[1/]

2. Data coverage

Often the original national data are known to exclude a significant portion of industrial activity for one or more years, either because the coverage of small-scale establishments may be incomplete in one or more years or the data may refer only to a geographic sub-region of the country or to a part of the manufacturing sector (e.g. publicly-owned enterprises, selected branches of industry, etc.). This characteristic is certainly the most challenging of all sources of data incomparability because adjusting for coverage involves the attempt to quantify what is not there. The problem of data coverage may be broken down into three parts: (i) incomplete or varying degrees of coverage of establishments; (ii) non-reporting of data, and (iii) the failure to adjust for non-response.

[1/] The steps are summarized by the following:

$$\frac{di_o}{DI_o} \cdot DI_1 = di_1 \qquad \text{(the linking step)}$$

$$di_1 \cdot x_1 = y_1 \qquad \text{(the branch estimate step)}$$

The elements are defined as follows:

$di_o = y_o/x_o$; $di_1 = y_1/x_1$; $DI_o = Y_o/X_o$ and $DI_1 = Y_1/X_1$,

where x = a figure for the related variable with branch-level data available in all years; and

X = the corresponding figure for the ISIC combination.

y = a figure for the related variable where branch-level data must be estimated for some years; and

Y = the corresponding figure for the ISIC combination.

The subscript "o" refers to the overlapping year or years; the subscript "1" refers to the year or years for which disaggregated data are needed.

(i) Cut-off points. A cut-off point is a theoretical limit below which no attempt is made to measure industrial activity. It is usually defined in terms of the employment size of the establishment or enterprise but a variety of other criteria may sometimes be used, ranging from the amount of annual turnover, to the use of motor power or modern accounting systems, to type of ownership. Even among those countries that define the cut-off point on the basis of employment size, there is wide variation. Moreover, as a measure of data coverage, any single employment criterion may have a different significance from country to country, due to the varying size characteristics of manufacturing establishments in each country. In general, the goal of standardizing international data to a common cut-off point, of establishments with five or more persons engaged, has been adopted by SSU. However, national data based on a lower cut-off point (e.g. establishments with three or more persons engaged or data covering all establishments) have usually been retained in the data base.

Adjustment for high or varying cut-off points has been the subject of a special project. The goal was to bring the coverage of international data either to the ideal cut-off point or to the closest possible cut-off point, depending upon data availability. The first step in the adjustment process for each country was a complete search of available supplementary sources - mainly national census publications - to identify the cut-off point closest to the ideal. For those years where this supplementary branch-level information was readily available, the data were entered directly in the data base.

Adjusted data for available years were then used to estimate comparable figures for the remaining years. The basic approach depended upon whether independent totals for manufacturing at the desired cut-off point were provided in supplementary sources. If no such totals were available (the more common case), the adjusted data were compared with original data base figures at the branch level in the overlapping years, by calculating ratios of the "uncovered" portion (i.e. the difference between adjusted and original data base figures) to the "covered" portion (the original figures). Ratios, if available for more than one year, were examined for patterns, and interpolations were made if the required estimates referred to intervening years. To derive estimates for remaining years, either ratios for a benchmark year or a mean of ratios for available years were used. These estimated branch-level ratios were then applied to the original data base figures, and the total for ISIC 3000 was calculated as the sum of the adjusted branches.

The following example illustrates the procedure described above. For one country the reported data on employment, wages and salaries, gross output and value added covered only establishments with 10 or more persons engaged, but national publications also contained data for 1973 and 1978 for establishments with 5 to 9 persons engaged. It was therefore possible to calculate ratios between the uncovered and covered portions for these two years, to interpolate, and to generate branch-level estimates for the uncovered portion (i.e. those referring to establishments with 5 to 9 persons engaged) from 1974 through 1977. For the years before 1973, the 1973 ratios were used as a benchmark; after 1978, the 1978 ratios were applied. Totals for manufacturing were then derived as the sum of the branches.

Where independent totals for manufacturing were available - or could be reasonably estimated - in years where the adjustment was required, a different approach was used. Branch shares of the uncovered portion in the overlapping year or years were calculated and applied to the residual between the independent totals for manufacturing and the sum of the branches for the covered portion in each year. In another instance, for example, the reported data on all variables refer to establishments with 10 or more employees. However, for one year, 1977, branch level data for "small industrial establishments" were available. Corresponding totals for manufacturing were also published, covering the years from 1973 to 1976, and totals for earlier years were estimated by SSU using the ratio between uncovered and covered portions of all manufacturing establishments in 1973. Branch shares of small establishments in 1977 were then applied to all totals for small establishments - whether published by the country in question (1973-76) or estimated by SSU (before 1973).

In cases where supplementary information at the desired cut-off point was available for some but not all variables in the years where an adjustment was required, estirates for remaining variables were based on ratios between data for unadjusted and adjusted variables, using the smallest category of establishments for which supplementary informatior on both types of variable was available.

(ii) Non-reporting of data. Missing data may be due to difficulty in enumeration (perhaps because of a large number of small establishments or the lack of an up-to-date register of establishments), to conceptual differences in accounting systems which preclude measurement of certain indicators or to confidentiality (i.e. disclosure rules). Alternatively, for the most recent years, missing data may be only a transitory problem resulting from time lags in data preparation for a particular branch.

The treatment of non-reporting depended upon whether all national data for a particular variable or only a part of the data set were missing. These two conditions are discussed in turn. Some countries do not report to UNSO - or even collect - data on certain variables. In the latter case, of course, nothing can be done but SSU is attempting to identify and redress cases that belong to the former category.

The general approach used was to enter directly all annual data available from supplementary sources - adjusted, where necessary, to the 3-digit (branch) level of the ISIC - and to extend the series for other years using proxy variables, where possible. Efforts to fill such data gaps are continuing.

A more common - and generally more tractable - form of non-reporting relates to cases where country data for only some industrial branches and/or some years are missing. The choice of an est mation approach for this type of problem involved the appraisal and reconciliation of five factors:

a) The number of years, and number and importance (i.e. relative weight) of the branch(es) for which data were missing;

b) the availability of independent totals in country/variable/years where branch data were missing;

c) the internal consistency of existing data;

d) the configuration of missing items within the data base matrix for each variable; and

e) the availability and goodness of fit of data on the same or proxy or related variables - within the UDB or from supplementary sources - for the missing country/branch/years.

Each of these is discussed below.

a) Extent of missing data. This factor was the primary determinant of whether efforts to develop estimates for missing data were worthwhile. At the two extremes, the choice was fairly clear. In cases where almost all data were missing, no effort was warranted unless new supplementary sources of rather complete information were available; where almost all data were present, every possibility was to be explored to fill the small number of data gaps remaining. For cases between these two extremes, the relative importance of missing branches was examined but no fixed criteria were set. The decision was left to statisticians' judgement, based largely upon the other factors listed.

b) Availability of independent sub-totals or totals. Among the supplementary statistical sources that were consulted for additional information, it was sometimes found that data for missing branches were included with those for other branches in larger aggregates. Thus, the estimation exercise could be reduced to one of splitting branch combinations, i.e. with a known but undistributed residual to be allocated among the missing branches. Readily available sources of this type of information were, for employment, data on the number of employees as compiled by the International Labour Organization (ILO) and for value added, ISIC 2-digit data on GDP originating in marufacturing at current prices as shown in the Yearbook of National Accounts Statistics. National publications were also an important source of data at the 2-digit level of ISIC.

c) Internal consistency of existing data. Internal consistency was measured on the basis of the regularity of year-to-year changes in branch shares, or in ratios between branch data for related variables over time, etc. These patterns were used as an indicator of how far existing data could be depended upon to yield reasonable estimates for missing data. This was important because even a few isolated branch estimates for missing data would result in changes in the totals for manufacturing, thus affecting in turn the relative shares of data for all other branches as well. It was critically important when data for entire variable/years were missing and estimates were being made on the basis of related variables.

d) Configuration of missing items. Missing data for a variable may arise: (i) in isolated branch/years; (ii) in most or all branches for one or more years; or (iii) in one or a few branches for many or all years. Solutions for isolated missing items were usually quite easy to find, using derived indicators (where data on related variables were available) or branch shares of the same variable in surrounding years or a neighbouring year. Pattern (ii) was also relatively straightforward, using branch shares in a benchmark year - or interpolated shares in surrounding years - if totals for manufacturing during the missing years were available. If totals were not available but branch-level data for a related variable were in place during the years being estimated, derived indicators were usually applied. However, pattern (iii) could only be addressed - again using derived indicators or branch shares - if some data for the missing variable/branches were available. If the branch accounted for a small proportion of manufacturing value added, even a share based on a single year was sometimes regarded as acceptable. Otherwise, no solution was possible.

e) Availability of branch-level data. Supplementary sources of information (usually national publications) were heavily exploited and, if the coverage of supplementary data matched that described in the country notes prepared by UNSO, the data could be entered directly into the data base. If the coverage did not match, supplementary data were still sometimes used, in the form of shares or derived indicators. If supplementary data for only one or some variables relating to missing years were available, efforts were made to fill data gaps among the remaining variables using derived indicators. In the absence of supplementary sources, data already in the data base were sometimes used, but with due prior consideration given to other factors, especially the internal consistency of existing data.

The multiplicity of factors to be taken into account may suggest an exceedingly complex situation. In practice, however, the number of alternatives, along with the evaluation process, were restricted by the limitations of available data. Indeed, statisticians have sometimes had to exercise their final option, i.e. that of not producing an estimate at all.

(iii) Non-response. Non-response may be due to any of several factors: incomplete registers of establishments, failures in the mechanisms for ensuring compliance among reporting units, weaknesses in follow-up procedures for missing or incomplete establishment returns, etc. The chief problem is that non-response is not systematic, and is therefore best addressed before the final results of an inquiry are processed.

While the question of the treatment of non-response is basic for the data user, it has not received the attention that it deserves among many national data producers. Some countries adjust their data for non-response, and others do not. The latter usually provide some measure of the extent of non-response, and SSU has attempted to make use of this information, where possible. However, some countries fail to address the question altogether. The International Recommendations[1/] specifically request such information, and perhaps this is an area where improvements in national reporting practices may be anticipated.

3. Concepts and definitions

Differences in concept or definition are variable-specific although their numerical effects may vary across branches. In reporting their industrial data, most countries conform to the United Nations' recommendations. Even

[1/] International Recommendations for Industrial Statistics, Statistical Papers, Series M, No.48, Rev.1, United Nations, 1983, paragraph 74.

among those countries that do not, the international standards provide a convenient reference point for comparing all variations in national reporting practices.

(i) Employment. The United Nations recommendations contain two basic definitions for data on employment: the "number of persons engaged" and the "number of employees."[1] National data are often reported according to both definitions but SSU gives preference to the average number of employees, where available, for entry into the UDB's single employment data field.

Of the 139 countries and areas for which employment data are available in the UDB, data for 88 countries are defined as the number of employees. For 30 countries data refer to the number of persons engaged and for 21 countries the definition changes for one or more years. However, because in some branches - and especially among developing countries - the number of working proprietors and unpaid family workers can be significant, SSU is exploring the possibility of adjusting the data to a single concept (number of employees), segregating data on the number of persons engaged, and eventually making both data sets available. In the meantime, numerical differences between the two types of data for any branch would depend upon the number of small establishments operating and the cut-off point set for the enumeration. Such differences would be relatively insignificant in most aggregations of data for groups of countries. However, users should exercise caution when making comparisons at the country level, especially those involving any of the 21 countries for which the definition of the employment series changes during the period.

(ii) Wages and salaries. In the reporting of wages and salaries, the most common differences between national practices and the international recommendations relate to the inclusion of payments to homeworkers and of employers' contributions to social security schemes or the exclusion of

[1] There is a third variable, the "number of operatives," which refers to production workers only. However, screening and adjustment of the data on the number of operatives has not yet begun.

payments-in-kind. The numerical effects of these differences, although not known, are probably of small consequence both within and between countries, compared to the effects of differences in cut-off point.

(iii) Gross output and value added. Among the variations in concept that may apply to the data on gross output and value added, the most important are: (i) whether data are based on the national accounting or the industrial census concept, and (ii) valuation of the data. The main difference between the national accounting concept and the industrial census concept is in the treatment of non-industrial services. This difference can be significant, and should be taken into account especially if comparisons between individual countries are being made.[1/] Valuation of the reported data on gross output or value added may be at producers' prices or factor values.[2/] Although the United Nations' International Recommendations for Industrial Statistics give priority to the collection of data at producers' prices, the choice of valuation is a matter of country discretion (as of course are the national policies that determine which branches of industry should receive subsidies and how indirect taxes should be levied).

The results of UNIDO's work on this aspect suggest that the amalgamation of values on different definitions produces inconsistent aggregate statements of regional shares in total world output, and even more significant distortions in the case of commodities like alcoholic beverages, tobacco, and petroleum products which are generally the ones most heavily taxed. These

[1/] Countries known to report their data according to the national accounting concept are: Belgium, Costa Rica (1976-80), France, Gambia (1976-8), Germany, Federal Republic of, Haiti, Ivory Coast, Jamaica, Malaysia (1963-7[illegible]), Mexico, Netherlands (1963-79), Nicaragua (1973-8), Peru (1972-80), Portugal (1963-70) and Togo (1968-9).

[2/] There are several other types of valuation in use by some countries. However, the two types mentioned here are the most common in industrial statistics, and at present are the only ones for which a distinction is available on the computer tapes. A third category, labelled "not specified," is used for all data that cannot be assigned to either category, or for which there is insufficient information on the valuation.

differences will also affect growth rates. Since many of the countries which account for a significant share of world manufacturing value added report their data at factor cost, separate data sets at each valuation would be desirable. However, because of the paucity of published statistical detail and the lack of systematic year-to-year patterns (even within countries) in the data that do exist, such a goal is not realistic at present.[1/]

C. Ensuring data consistency

As evidenced in the two preceding sections, SSU's statisticians take considerable care in ensuring data consistency in the process of enlarging the UDB and in improving the international comparability of its content. However, due to inconsistency inherent to many series reported by primary sources as well as to the wide variety of sources used, it is felt that a final screening of the data is needed. The purpose of this final screening is to diagnose and display 'abnormal' entries in the UDB, to allow for possible corrections. The final screening takes place in two phases. First, possible abnormalities are identified through a computerized procedure. Second, SSU's statisticians redress, to the extent possible, the identified abnormalities.

Identifying possible abnormalities

Each of the four variables in the Data Base - gross output (GO), value added (VA), wages and salaries (W&S) and employment (E) - per country, branch (ISIC at the 3-digit level, combinations of branches, and ISIC 3000) and year is treated as one observation. An observation (one variable) or a combination of two observations in the form of a ratio of two variables pertaining to the

[1/] Of the 130 countries for which data on gross output are available in the UDB, data for 61 countries are reported at producers' prices and for 32 countries at factor values. Among 31 countries the valuation is classed as "not specified" and in 6 countries it changes from one category to another at some time during the period. Data on value added are available for 120 countries, of which 59 report at producers' prices and 34 at factor values. For 17 countries the valuation is classed as "not specified" and for 10 countries it changes at some time during the period.

same country, branch (or combination of branches) and year is considered to be abnormal if it appears to be implausible on logical, statistical or economic grounds.

The criteria used in the tests apply to:

(i) Individual observations;

(ii) combinations (ratios) of observations pertaining to the same country, year and branch;

(iii) (i) and (ii) in relation to other branches, and

(iv) (i), (ii) and (iii) in relation to other years.

The criteria consists of acceptable ranges for (i) to (iv) above (see appendix). Other than purely logical ones, these ranges were set by screening a sample of countries having dissimilar economies, data collecting and reporting procedures. Some of the acceptable ranges are allowed to take different values depending on the degree of specialization and volatility of the manufacturing sector in a country.

The abnormality may stem from one or more of the following:

(a) An outright mistake, e.g. a typo;

(b) a problem related to definitions and/or methods used in collecting and processing data, or changes in those definitions or methods over time;

(c) actual extraordinary economic circumstances.

Only a fraction of the tests unambiguously point to a mistake. In all other cases the diagnosed abnormality may stem from any combination of the three problems stated above.

APPENDIX

Details of the tests applied in the course of the final screening process

I. Wages and salaries/value added (W&S/VA)

Flag 3 = 1 if $W\&S/VA > 1$

II. Value added/gross output (VA/GO)

Flag 3 = 1 if $VA/GO > 1$
Flag 3 = 2 if $VA/GO < 0.03$

Flag 4 = 1 if $VA/GO > 0$ and $\triangle\ e^{(VA/GO\ .\ 10)} > 30$.

III. Wages and salaries/employment (W&S/E)

For each year (and country) calculate unweighted mean W&S/E for all industries (excluding branches with negative W&S/E) and define an index WE for each industry, as the ratio of its W&S/E to this mean value.

Calculate coefficient of variation (CV), i.e. standard deviation/mean of this index for each year.

Flag 1 = 1 if $CV \leq 0.25$
and
$0.50 > WE > 2.00$
and
$\triangle\ WE > 0.15$

(where $\triangle$ WE is the rate of change in WE from one year to the next, when the first year of a series is involved the test drops, in case of missing years, annual rate of change is calculated from compound rate of change).

Flag 1 = 2 if $CV > 0.25$

and

$0.25 > WE > 4.00$

and

$\Delta WE > 0.30$

Flag 3 = 1 if $CV \leq 0.25$

and

$\Delta WE > 0.25$

Flag 3 = 2 if $CV > 0.25$

and

$\Delta WE > 0.50$

For total manufacturing T (ISIC 300)

Flag 2 = 1 if $CV \leq 0.25$

and

$0.85 > TWE > 1.15$

Flag 2 = 2 if $CV > 0.25$

and

$0.70 > TWE > 1.30$

IV. Value added/employment (VA/E)

Exclude ISIC 353 and 354 and calculate the mean VA/E, and an index VE as defined in III above (also Δ VE same as Δ WE).

For 26 branches

Flag 1 = 1 if $0.25 > VE > 5.00$

and

$\Delta VE > 0.50$

Flag 3 = 1 if $\Delta VE \geq 1.00$

For ISIC 353 and 354

Calculate VE also for ISIC 353 and 354 (the mean value excluding these two industries).

Flag 4 = 1 if $0.50 > VE > 15.00$
and
$\triangle VE \geq 1.00$

Flag 5 = 1 if $\triangle VE > 2.00$

For total manufacturing T (ISIC 300)

Flag 2 = 1 if $0.50 > TVE > 2.00$
(where the mean excludes ISIC 353 and 354)

V. Employment (E) (subscripts refer to two adjacent years and E is rate of change from year to year - no compound rate for missing years.)

Flag 0 = 1 if $E_o = 0$ and $E_1 > 1{,}000$
Flag 0 = 2 if $E_o > 1{,}000$ and $E_1 = 0$

Flag 1 = 1 if $E_o \leq 1{,}000$ and $E_1 > 10{,}000$
Flag 1 = 2 if $E_o > 10{,}000$ and $E_1 \leq 1{,}000$

Flag 2 = 1 if $1{,}000 < E_o \leq 10{,}000$ and $E_1 > 20{,}000$
Flag 2 = 2 if $E_o > 20{,}000$ and $1{,}000 < E_1 \leq 10{,}000$

Flag 3 = 1 if $(10{,}000 < E_o < 100{,}000$ and $E_1 > 10{,}000)$ or
$(10{,}000 < E_1 < 100{,}000$ and $E_o > 10{,}000)$
and
$\triangle E > 1.00$

Flag 4 = 1 if $(E_o$ and $E_1) > 100{,}000$ and $\triangle E > 0.25$

Flag 5 = 1 if $E < 0$

(If either E_0 or E_1 has a negative value, none of the tests above - except the last one - are undertaken.)

VI. Consistency check

For any country, year, ISIC, if one of the four variables (GO, VA, W&S, E) has a zero value, the other variables should also have zero values (or missing values):

Flag = 1 if not.

VII. Check for total manufacturing versus sum of branches

Flag GO = 1 if GO for ISIC 3000/sum of GO for all branches (Σ) $\geq$ 1.005

or

if ISIC 3000 GO/Σ GO$\leq$ 0.955

Flag VA = 1 if ISIC 3000 VA/Σ VA$\geq$1.005

or

if$\leq$ 0.995

Flag WS = 1 if ISIC 3000 W+S/Σ W&S$\geq$1.005

or

if $\leq$ 0.995

Flag E = 1 if ISIC 3000 E/Σ E$\geq$1.005

if$\leq$0.995

VIII. Check for consistency with the "base weights"

Flag = 1 if VA75 $>$0 (and not missing) and if any of the four variables = 0 in 1975 or in any following year

or

if VA75 = 0 and if any of the four variables = 0 or not missing in 1975 or in any preceding year.

IX. Check for negative W&S, GO and VA

Flag 1 = 1 if W&S is negative.

Flag 2 = 1 if VA is negative.

Flag 3 = 1 if GO is negative.