OCCASION

This publication has been made available to the public on the occasion of the 50th anniversary of the United Nations Industrial Development Organisation.

DISCLAIMER

FAIR USE POLICY

CONTACT

Please contact publications@unido.org for further information concerning UNIDO publications.

For more information about UNIDO, please visit us at www.unido.org

UNITED NATIONS INDUSTRIAL DEVELOPMENT ORGANIZATION
Vienna International Centre, P.O. Box 300, 1400 Vienna, Austria
Tel: (+43-1) 26026-0 · www.unido.org · unido@unido.org

19134

Distr.
RESTRICTED

PPD.192(SPEC.)
8 April 1991

UNITED NATIONS
INDUSTRIAL DEVELOPMENT ORGANIZATION

ORIGINAL: ENGLISH

INDUSTRIAL STATISTICS FOR RESEARCH PURPOSES

Methodology Applied in the Development and Maintenance of the UNIDO Industrial Statistics Data Base*

Prepared by the
Industrial Statistics and Sectoral Surveys Branch
Industrial Policy and Perspectives Division

V.91 23197

1/2

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 Industrial Statistics 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 Industrial Statistics and Sectoral Surveys Branch (STAT) of the Industrial Policy and Perspectives Division, aiming at the dissemination of useful industrial data among users both inside UNIDO and outside.

Since its inception, the statistical programme of STAT 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. The UDB is primarily intended to meet the statistical needs of researchers engaged in international, or cross-country, studies rather than country-specific investigations. Accordingly, UNIDO statisticians give priority to the compilation and the development of statistics which meet standards of consistency over time and cross-country 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 STAT. These include: the Handbook of Industrial Statistics, the African Industry in Figures, the Sectoral Survey series, as well as many ad hoc studies.

The dissemination of industrial statistics among users within UNIDO is made by providing extracts from the UDB according to standardized formats and by maintaining a system of on-line access and data processing. Outside STAT, the Global Report is probably the main single user of UDB data within UNIDO but the Country Briefs that are prepared by utilizing UDB data are more widely disseminated.

STAT also supplies selected statistical indicators for use in recurrent publications of other international organizations including the World Development Report and the World Tables of the World Bank and the Handbook of International Trade and Development Statistics of the UNCTAD.

Finally, copies (either on tape or diskettes for the whole UDB or on diskettes for subsets) are purchased by users outside the Organization. For information on purchase procedures, readers should refer to the Industrial Statistics and Surveys Branch, UNIDO, P.O. Box 300, A-1400 Vienna, Austria.

I. THE UNIDO INDUSTRIAL STATISTICS DATA BASE

A. Contents of the 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, covering about 150 countries and areas. The data are presented according to the three-digit level of the ISIC[1] which provides for the 28 industrial branches of the manufacturing sector.

Four major sources of information are used in compiling, cleaning and developing the data contained in the UDB: The primary source consists of country questionnaires completed by national statistical offices. The secondary source is threefold. First in importance is national statistical publications of industrial censuses, annual industrial surveys and other statistical surveys. Second, international sources, both published and unpublished, are used. Third is national data compiled by statisticians engaged by UNIDO to work in specific countries. Any data found able to fill a gap is checked, standardized and incorporated in the UDB but unavoidably the exploitation of supplementary sources brings about only a marginal improvement in data availability. In addition to the primary and secondary source data, the UDB includes a number of estimates that have been made by STAT.

Although the UDB contains over 340,000 entries (as in February 1991), coverage is considerably less than the corresponding theoretical maximum (some 570,000) which would assume that data were available for all cells. Frequently, observations 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 manufacturing value added (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 zero 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 STAT'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

---

[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 1991" (PPD.186).

described separately.

Development of the data base consists in enlarging the number of entries in the data base from several sources of information and from imputation as mentioned earlier. The activities of incorporating available data are carried out in five stages, based on an ordering of the source of information as is described in detail in chapter II.

The second function, improvement in the international comparability of the data, constitutes perhaps the most intricate aspect of STAT's programme. Although the international community sustains continuous efforts to promote international standards, divergences in national practices persist in the key areas of industrial classification, establishment coverage and statistical concept and definition. STAT makes every effort to improve the comparability of reported data. The industrial 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, STAT attempts to convert the data to the desired system and level. Establishment coverage is frequently incomplete, that is data items are gathered only for statistical units defined by certain characteristics (e.g., size of establishment, 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. STAT is endeavouring to estimate figures relating to the desired coverage of all establishments with five or more employees.

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 STAT implements a systematic screening of the data and attempts to redress identified inconsistencies.

II. DETAILED PROCEDURES

A. Maintenance and Development of the Data Base

The design and structure of the UDB are closely related to the performance of this function. The UDB is organized in five stages. At each stage, the data are subject to various forms of examination and adjustment. The purpose of the stage-organization is twofold. First, the layout allows to retrieve the data according to the degree of confidence they deserve. The first layer contains only data officially communicated by the national statistical offices and screened by STAT; the fifth layer cumulates the data contained in the four previous layers thus estimates made by STAT. The intermediate layers add data of decreasingly authoritative sources. Second, the stage organization serves to select the methods employed with regard to data screening, analysis, editing, etc..

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 the Statistical Office of the United Nations (UNSO) once a year. 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, 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 departures from the International Recommendations for Industrial Statistics.

Stage II - The incorporation of national data

Prior to 1987 the UDB was updated with information obtained from national questionnaires only once a year. Since 1987 when UNIDO started receiving copies of completed national questionnaires via UNSO, updating has been continuous. Supplementary to data obtained through questionnaires, UNIDO draws on national statistical publications such as reports of industrial censuses and surveys, and statistical yearbooks. UNIDO statisticians use these sources to carry out various adjustments to eliminate departures from the international standards. Data compiled through UNIDO field work is also incorporated at this stage.

Stage III - The inclusion of data from additional international sources

In extending the coverage of available data, UNIDO gives highest priority to data compiled through country questionnaires supplemented by other national sources. However, other international institutions also compile industrial data and these sources have been found useful in further development of the UDB. Initially, these data are carefully screened by UNIDO statisticians to determine their quality, coverage, scope and definitions. Where UNIDO standards are met, the information is used for 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.[3] 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 UDB cumulates data from questionnaires, national publications and international sources. At this point all adjustments have been made with the help of supplementary information from exogenous sources. These data are then used to make imputations for missing observations.

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 varies from country to country and even, sometimes, from variable to variable. Under the best circumstances, the lag will be two years. However, this minimum lag is achieved only in roughly one half of the developed countries and 15 per cent of the developing countries.

---

[3] Data provided by ILO, OECD, the World Bank, and various UN regional economic commissions have been utilized.

Furthermore, the data coverage for the latest year is often incomplete. In fact, the minimum lag required to obtain coverage for 90 per cent of the countries in the UDB is five years.

Delays obviously jeopardize the usefulness of 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.

Stage V begins with estimation of the missing observations for gross output based on production indexes.[4] Then the other three variables are estimated on the basis of past trends in the relationships between gross output and the three variables.

Four steps are involved in this approach. The first step is to extrapolate the reported gross output for each branch by applying corresponding production index and MVA deflator or consumer price index (No branch-specific deflator is generally available.). In symbols, the estimated gross output (EGO) for a given branch can be expressed as:

$$EGO_t = GO_0 * IN_t * MVADEF_t, \qquad t=1,2,.....k$$

where GO_0 is the reported gross output for the latest year (i.e., t=0), IN is a production index with the base year 0, MVADEF is a MVA deflator with the base year 0, and t is a year (year k is the latest year for which production index is available).

The second step is to estimate missing data on value added (VA) by estimating a country- and branch-specific time-series regression equation;

$$VA = a + b*GO + u$$

where GO is gross output, a is the intercept, b is the regression coefficient on GO and is assumed to be in the range between 0 and 1, and u is a residual. Estimates on missing VA for up to the year k are derived from this equation by using the estimated GO.

In the third step, similar to the case of VA, missing data on wages and salaries (WS) are estimated by the following regression equation;

$$WS = c + d*VA + v$$

where c is the intercept, d is the regression coefficient on VA and v is a residual.

Finally in step four, missing data on employment (E) are estimated from the following time-series regression equation;

---

[4] As the result of the direct incorporation of the production indexes which are, in general, more timelily reported in national questionnaires then other statistics and STAT,s estimation work (See Section C of this chapter), there are usually many cases where period coverage of production index is larger than that of gross output.

$$\ln E = e + f*\ln IN + w$$

assuming a constant elasticity of employment with respect to production. In this equation, e is the intercept, f is the regression coefficient which is expected to be less than unity.

After completion of the above estimation procedures, still many missing observations remain unestimated. Provisional estimation of these missing data for latest years consists of four steps. The first step concerns only the extension of time series for each country to a common terminal year. The approach is based on the assumption that ratios between variables for which data were available for previous years 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 data compiled by the UNSO and the World Bank. These series have a time lag of only two years and provide estimates of MVA.[5] These series are used to update MVA in countries with a time-lag of more than two years.

In a third step the estimates obtained for MVA are used for deriving 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. Employment in total manufacturing is estimated based on employment data that are extracted from various statistical publications.

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 the fourth step these totals are disaggregated in 28 branches on the basis of the inter-branch distribution for the respective countries in previous years. The assumption is that all 28 branches grew at a same rate during the estimation period.

In conclusion, the five stages which compose the data base not only provide an operational framework for STAT 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 that the data as supplied by national sources did not meet STAT 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.

---

[5] No information is provided for Eastern European countries and the USSR.

B. Improving International Comparability

Together with the processing of incoming questionnaires completed by national statistical offices, efforts to adjust the data for greater international comparability constitute the bulk of STAT's data development programme. Although work on this aspect is continuous, data for almost all of the 150 countries and areas included in the UDB have been adjusted or supplemented in some way by STAT, 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, thus, 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 STAT for data adjustment.

1. The industrial classification

Most countries either use the ISIC or a compatible classification. Even so, more than two thirds 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 particularly for earlier years. The reasons for this vary. In some cases the practice reflects multiple 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. Their effects, i.e. "combined" data, present serious limitations for cross-country comparisons of a specific industry.

Roughly half of the adjustments made by STAT have involved disaggregation of data referring to two or more industrial branches. The employed methods are discussed below:

(i) Same variable, same years: For some countries, it may happen that disaggregated data are available from extra-official sources. For many applications, and particularly for those to which UNIDO is oriented, it may be assumed that such data set is reasonably compatible with aggregated data from official sources. 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 in the majority of the cases. If the combined data were surrounded in time by branch-level data, shares from the surrounding years were interpolated.

(iii) Proxy variables[6]: 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 are used with caution for splitting branch combinations.

2. Data coverage

Often the original national data are known to exclude a significant portion of industrial activity, either because the coverage of small-scale establishments may be incomplete in one or more years or the data may refer only to a certain area of the country(e.g., urban area, metropolitan area) 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.

(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 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 STAT. 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.

---

[6] 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. It should be noted, however, that disaggregated data on proxy or related variables are not frequently available. For further discussion, see below.

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 target 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 target. 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.

Where independent totals for manufacturing were available 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.

(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 in questionnaires - or even collect - data on certain variables. In the latter case, of course, nothing can be done but STAT 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 estimation 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. 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.

b) Availability of independent sub-totals or totals. Among the supplementary statistical sources, 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.

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 MVA, 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 was judged to be acceptable, the data could be entered directly into the data base. Otherwise, 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.

(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, which is used by STAT to assess the quality of the data. However, some countries fail to address the question altogether. The International Recommendations[7] 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 among those countries that do not, the international standards provide a convenient reference point for comparing all variations in national reporting practices.

(i) Employment. For the majority of the countries, employment data refer to number of employees. However, in some cases data refer to number of persons engaged. For a few countries, the definition changes over time. In general, no supplementary information for standardization of reported employment data is available. Any use of employment data, therefore, requires caution, particularly in those cases where definition changes over time.

(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 family workers and of employers' contributions to social security schemes or the exclusion of payments-in-kind. The numerical effects of these differences, although not

---

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

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. Valuation of the reported data on gross output or value added may be at producers' prices or factor values.[8] 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 differences will also affect growth rates. Since many of the countries which account for a significant share of world MVA 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.

There are some instances where reported value added (VA) is smaller than corresponding reported wages and salaries (WS) or VA is reported to be even negative. By definition, gross value added consists of the three major components - WS, operational surplus, depreciation cost - of which only operational surplus can be negative. Therefore, particularly when VA is valuated in terms of producers' prices, it is always possible that VA turns out to be smaller than WS due to operational deficit.

The data on VA which are smaller than corresponding data on WS or even negative have important indication concerning the branch's business performance. However, if VA (as well as gross output) is to be an indicator of production, data on VA excluding operational loss (or the other extreme, monopoly profit) would be more useful for multi-country analysis. In practice, adjustment of reported VA in this line is not feasible. Instead, if the reported WS was judged to be acceptable and the reported VA was smaller than the WS, then the reported VA was replaced with the reported WS and, consequently, the difference between the adjusted VA and the reported VA was

---

[8] 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.

added to the reported gross output (GO).*

C. Estimation of Production Indexes

All above discussions refer mainly to the four variables - employment, wages and salaries, gross output and value added. As in the case of these variables, the primary source of data on production indexes is country questionnaires completed by national statistical offices. The reported indexes usually needed to be re-based, however. In many aspects, the way STAT treats reported production indexes is somewhat different from the other four variables. The main reason of this is that there is, in general, little information concerning the comparability and consistency of the reported production indexes except only for national deviations from the ISIC.

One of the hazards of working with production indexes is that it is possible to create an index from any time series. The challenge is to find a reasonable indicator of real change in net output. The highest priority is given to the data reported by national statistical offices as in the case of other variables. In the cases where production indexes were not reported either by the primary source or by any supplementary sources, the following estimation procedures were employed.

A widely accepted indicator of change in industrial output over time is the one which is based on commodity production series expressed in physical units at a highly detailed level. Theoretically, a set of commodity production series which represent the major output of the corresponding branch could be weighted by base-year (e.g., 1980) prices to form a highly reliable index of industrial production. Similarly, data on value added in current prices could be used to form a production index with a simple application of a deflator to adjust for price changes over time. However, in practice, neither base-year price weights nor appropriate deflator are generally available.

The paucity of data on prices and price deflator presents a very serious limitation on the utility of these indicators for estimation purposes. Nevertheless, certain methodological concessions have been made to allow their use. These methods have been accepted only with the condition that the resultant indexes be subjected to careful scrutiny and evaluation.

Use of commodity production data. The most common source of commodity production data is United Nations, Yearbook of Industrial Statistics, Vol.II. In the absence of price weights for combining the series, unweighted geometric means were calculated. Each quantity series was converted to an index (1980=100) for all appropriate 6-digit ISIC groups which represent part of the output of the 3-digit ISIC group being estimated. Unweighted geometric means of the commodity series were then calculated to form the estimated production indexes, and linked to existing production index data.

The choice and treatment of commodity data are somewhat subjective, in that some series were rejected if the absolute figures were small or if

---

* However, if the country- and branch-specific ratio of WS to VA for other years was stable throughout the data period, an average ratio of WS to VA across these years was employed to adjust the VA.

interruptions in many of the primary series would require too many links in the combined series. (Price data, if available, would have eliminated some of these problems.) Use of the commodity approach has generally been contingent upon a certain degree of consonance (i.e., parallel movement through time) among the individual series themselves, thereby reducing the dangers of combining quantity data without weighing.

D. Ensuring Data Consistency

As evidenced in the two preceding sections, UNIDO 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, UNIDO statisticians redress, to the extent possible, the identified abnormalities.

Each of the four variables in the data base - gross output (GO), value added (VA), wages and salaries (WS) 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 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 (WS/VA)

Flag 3 = 1 if WS/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 $\Delta\, e^{(VA/GO \cdot 10)} > 30$.

III. Wages and salaries/employment (WS/E)

For each year (and country) calculate unweighted mean WS/E for all industries and define an index WE for each industry, as the ratio of its WS/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 <= 0.25
and
(WE < 0.50) or (WE > 2.00)
and
ΔWE > 0.15

(where ΔWE 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
(WE < 0.25) or (WE > 4.00)
and
ΔWE > 0.30

Flag 3 = 1 if CV <= 0.25
and
ΔWE > 0.25

Flag 3 = 2 if CV > 0.25
and
ΔWE > 0.50

For total manufacturing T (ISIC 300)

Flag 2 - 1 if CV <= 0.25
and
(TWE < 0.85) or (TWE > 1.15)

Flag 2 - 2 if CV > 0.25
and
(TWE < 0.70) or (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 ΔVE same as ΔWE).

For 26 branches

Flag 1 - 1 if (VE < 0.25) or (VE > 5.00)
and
ΔVE > 0.50

Flag 3 - 1 if ΔVE >= 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 (VE < 0.50) or (VE > 15.00)
and
ΔVE >= 1.00

Flag 5 - 1 if ΔVE > 2.00

For total manufacturing T (ISIC 300)

Flag 2 - 1 if (TVE < 0.50) or (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 $E_0 = 0$ and $E_1 > 1{,}000$
Flag 0 - 2 if $E_0 > 1{,}000$ and $E_1 = 0$

Flag 1 - 1 if $E_0 <= 1{,}000$ and $E_1 > 10{,}000$
Flag 1 - 2 if $E_0 > 10{,}000$ and $E_1 <= 1{,}000$

Flag 2 - 1 if $1{,}000 < E_0 <= 10{,}000$ and $E_1 > 20{,}000$
Flag 2 - 2 if $E_0 > 20{,}000$ and $1{,}000 < E_1 <= 10{,}000$

Flag 3 - 1 if ($10{,}000 < E_0 < 100{,}000$ and $E_1 > 10{,}000$) or
($10{,}000 < E_1 < 100{,}000$ and $E_0 > 10{,}000$)
and
$\Delta E > 1.00$

Flag 4 = 1 if (E_o and E_1) > 100.000 and $\Delta E > 0.25$

Flag 5 = 1 if $E < 0$

(If either E_o 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, WS, 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 ($\sum$) >= 1.005
or
if ISIC 3000 GO/ $\sum$GO <= 0.955

Flag VA = 1 if ISIC 3000 VA/ $\sum$VA >= 1.005
or
if <= 0.995

Flag WS = 1 if ISIC 3000 WS/ $\sum$WS >= 1.005
or
if <= 0.995

Flag E = 1 if ISIC 3000 E/ $\sum$E >= 1.005
or
if <= 0.995

VIII. Check for consistency with the "base weights"

Flag = 1 if VA80 > 0 (and not missing) and if any of the four variables = 0 in 1980 or in any following year
or
if VA80 = 0 and if any of the four variables = 0 or not missing in 1980 or in any preceding year.

IX. Check for negative WS, GO and VA

Flag 1 = 1 if WS is negative.
Flag 2 = 1 if VA is negative.
Flag 3 = 1 if GO is negative.