

**UNIDO INDUSTRIAL STATISTICS  
DATABASE in  
Revision 2 of ISIC**

**Methodological Notes**

Prepared in October 2003 by

Statistics Section  
Strategic Research and Economics Branch

---

The designations employed and the presentation of the material in this document do not imply the expression of any opinion whatsoever on the part of the Secretariat of the United Nations Industrial Development Organization (UNIDO) concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers and boundaries. Mention of company names and commercial products does not imply the endorsement of UNIDO. This document has not been edited.

## CONTENTS

	<u>Page</u>
INTRODUCTION .....	3
I. THE UNIDO INDUSTRIAL STATISTICS DATABASE .....	4
A. Contents of the database .....	4
B. Structure of the Industrial Statistics Branch's statistical programme for the database .....	5
II. DETAILED PROCEDURES .....	6
A. Maintenance and development of the database .....	6
Stage I - Responses to national questionnaires compiled by UNIDO and OECD .....	6
Stage II - Correction and adjustment of the data reported via questionnaires; incorporation of published national data .....	6
Stage III - Disaggregation of data; improvement of data based on supplementary information .....	7
Stage IV - Automatic disaggregation and interpolation of data .....	7
Stage V - Estimation of provisional data for latest years .....	7
B. Existing data problems and improving international comparability .....	11
1. The industrial classification .....	11
2. Data coverage .....	12
3. Concepts and definitions .....	15
C. Estimation of production indexes at the 3-digit level of ISIC .....	16
D. Ensuring data consistency .....	17

## INTRODUCTION

The availability of reliable and comparable data is an essential requirement for an accurate assessment of economic progress or for the formulation and implementation of viable development policies, programmes and projects. Given UNIDO's mandate in these two fields, the Organization undertook the establishment of the UNIDO Industrial Statistics Database (INDSTAT) 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 INDSTAT was undertaken by the Statistics Section (STAT) of the Strategic Research and Economics Branch, Programme Coordination and Field Operations Division, aiming at the dissemination of useful industrial data among users inside UNIDO and outside.

Until the beginning of the 1990s, the maintenance of the INDSTAT depended heavily on the data provided by the Statistical Division of the United Nations Secretariat (UNSD) which collected those data from national sources through its General Industrial Statistics Questionnaire. However, in accordance with the recommendations of the United Nations Statistical Commission at its twenty-seventh session, the responsibility for collection and dissemination of worldwide general industrial statistics was transferred, in 1994, from the United Nations Secretariat to UNIDO and the Organisation for Economic Co-operation and Development (OECD): UNIDO assumes responsibility for the collection of data from non-OECD member countries while OECD collects data from its member countries and provides those data to UNIDO to complete the worldwide coverage of the INDSTAT. Along with the new arrangement, UNIDO started the annual collection of eight basic industrial statistics from national statistical offices through the newly designed UNIDO General Industrial Statistics Questionnaire.<sup>1</sup>

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 co-operation and applied research. The INDSTAT 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 across countries.

The INDSTAT constitutes the major source of data for several recurrent publications produced by STAT. These include: the *International Yearbook of Industrial Statistics*, the *Statistical Country Brief* series, as well as many *ad-hoc* research publications. With

---

<sup>1</sup> UNSD continues to compile national data on the index number of industrial production and provides these data to UNIDO for inclusion of them in the INDSTAT.

respect to UNIDO recurrent publications produced outside STAT, the *Industrial Development Report* is also the major user of UISDB.

The dissemination of industrial statistics among users within UNIDO is made by providing extracts from the INDSTAT according to standardized formats and by maintaining a system of on-line access and data processing. In addition, the INDSTAT's dissemination products, including its CD-ROM versions, a hardcopy commercial publication – *International Yearbook of Industrial Statistics* and the Internet-based *Statistical Country Briefs* are widely disseminated within and outside UNIDO.

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

For information on purchase procedures of INDSTAT sales products, readers should refer to the UNIDO website [www.unido.org](http://www.unido.org).

## I. THE UNIDO INDUSTRIAL STATISTICS DATABASE

### A. Contents of the Database

The INDSTAT in accordance with Revision 2 of the International Standard Classification of All Economic Activities (ISIC)<sup>2</sup> consists of two data sets; the data set arranged in accordance with the 3-digit level of the code of the International Standard Industrial Classification of All Economic Activities (Revision 2) (ISIC), which provides for the 28 industrial branches of the manufacturing sector, and that at the 4-digit level of the ISIC code, which provides for 81 manufacturing industries.<sup>3</sup>

The ISIC 3-digit level data set of the INDSTAT includes annual figures measuring the following eight variables:

- (1) Number of establishments
- (2) Number of employees
- (3) Number of female employees
- (4) Wages and salaries
- (5) Output

---

<sup>2</sup> INDSTAT database system includes also INDSTAT at the 3- and 4-digit levels of ISIC(Revision 3). However, this ISIC(Rev.3)-based database is beyond the scope of this document.

<sup>3</sup> 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.

- (6) Value added
- (7) Gross fixed capital formation
- (8) Index numbers of industrial production

All these statistics are compiled by country and are reported on an annual basis spanning the period 1963 (in the cases of the number of establishments and female employment, 1981) to latest year, covering more than 160 countries and areas.

On the other hand, the ISIC 4-digit level data set, which is consistent with the 3-digit data set, includes annual time series on only seven out of the above eight statistics (the exception being the index numbers of industrial production). These statistics are also compiled by country but covering a shorter period, namely, 1981 - latest year, and less countries (114 countries and areas as of December 2002).

The primary source of information consists of country questionnaires completed by national statistical offices.<sup>4</sup> The data collected from this primary source is supplemented with a secondary source which is threefold: First in importance is national statistical publications concerning industrial censuses, annual industrial surveys and other statistical surveys. Second, international sources, both published and unpublished, are used. Third is national data collected by statisticians engaged by UNIDO to work in specific countries. Any data found appropriate to fill a gap is checked, standardized and incorporated in the INDSTAT but unavoidably the exploitation of supplementary sources brings about only a marginal improvement in data availability. In addition to the data from the primary and secondary sources, the database includes a number of estimates made by STAT.

## **B. Structure of STAT's Statistical Programme for the Database**

The statistical programme for the INDSTAT is composed of three basic functions: (i) expansion of the database, (ii) improvements in the international comparability and (iii) consistency of the data included. In practice, these functions are often performed simultaneously. However for the sake of clarity, they will be described separately.

(i) Expansion of the database consists in enlarging the number of entries in the database from the national questionnaires, several other sources of information and from imputation made by STAT. 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.

---

<sup>4</sup> A large portion of the data for the years up to 1992 maintained in the INDSTAT are those compiled from national statistical offices through the UNSD/UNIDO joint questionnaire for general industrial statistics. With regard to data for recent years, most of the data referring to the countries and areas other than the OECD member countries were compiled through the UNIDO General Industrial Statistics Questionnaire. Information referring to the OECD member countries is based on data compiled by OECD via questionnaire and provided to UNIDO. With respect to production indexes, primary data source is UNSD.

(ii) 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. With respect to the improvement in the international comparability of the data, STAT, in principle, focuses on the improvement in the comparability of reported data in terms of industrial classification. The industrial classification used in the INDSTAT is the 1968 version (i.e., Revision 2) of ISIC at the three-digit and the four-digit levels<sup>5</sup>. And where the national classification either differs from ISIC or is less disaggregated than the three-digit (or four-digit) level, STAT attempts to convert the data to the desired system and level.

(iii) 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 (or industries) and variables; furthermore, errors may have slipped into the data publishing and dissemination processes. To enable users of the INDSTAT 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 Database**

The design and structure of the INDSTAT are closely related to the performance of this function. The database is organized in several 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; the last layer cumulates the data contained in the previous layers and 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 UNIDO and OECD**

Data for non-OECD member countries are collected directly from national statistical offices in these countries through the UNIDO General Industrial Statistics Questionnaire

---

<sup>5</sup> A number of countries (as in December 2002, over ninety countries) have made a switch-over in their data reporting system from Revision 2 to Revision 3 of ISIC. After screening, these data have been stored in the INDSTAT database in accordance with ISIC(Rev.3), at the same time, converted, whenever possible, to those arranged according to ISIC (Revision 2) and entered in the INDSTAT database in accordance with the 3-digit level of ISIC(Rev.2).

while those for OECD member countries are collected from these countries' national statistical offices by OECD through its Information System on Industrial Statistics Questionnaire and provided to UNIDO. These data are in their original form, not adjusted by STAT. After correction of obvious errors in reporting, the data reported by national statistical offices are stored in the database. The data at Stage-I are to be pre-filled in the following edition of the questionnaire.

Stage II - Correction and adjustment of the data reported via questionnaires; incorporation of published national data

The reported data through questionnaires are, whenever necessary, corrected or adjusted in order to maintain data consistency. Whenever possible, required statistics are generated from reported statistics (i.e., Stage-I data). For instance, if per-employee output is reported (instead of total output) together with the number of employees, then data on output will be generated from the reported data on per-employee output and those on employment. Supplementary to data obtained through questionnaires, UNIDO draws on national statistical publications such as reports of industrial censuses and surveys and statistical yearbooks. Data compiled from national statistical offices through UNIDO field work are also incorporated at this stage.

The data at Stage I as well as those at Stage II are considered to be official. UNIDO publishes the cumulated data up to this stage in its annual publication, the *International Yearbook of Industrial Statistics*.

Stage III - Disaggregation of data; improvement of data based on supplementary information

Available supplementary sources, both national and international, are used to carry out various adjustments to eliminate departures from the international standards for industrial classification. Based on relevant supplementary information that has been carefully screened by UNIDO statisticians, reported data for combined ISIC categories are disaggregated for individual ISIC categories, inconsistent data are adjusted and missing data are estimated.

Stage IV - Automatic disaggregation and interpolation of data  
(for ISIC 3-digit data only)

After Stage III, there will be still a number of cases where reported data for combined ISIC categories have not been disaggregated due to lack of relevant supplementary information. At this stage, those aggregates are split to data for individual ISIC categories by applying the proportion between the corresponding individual ISIC categories for other years or the proportion between those ISIC categories derived from other related variables. In addition, whenever possible, automatic interpolation of time series is also made.

### Stage V - Estimation of provisional data for latest years

Official statistics are often reported with a time lag of several years. The actual duration of the time 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 industrialized countries and less than twenty per cent of the developing countries. Furthermore, the data coverage for the latest year is often incomplete.

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 (**E3**)
- wages and salaries (**WS3**)
- output (**GO3**)
- value added (**VA3**)

The estimates derived at Stage V are all provisional and are eliminated as corresponding data become available at an earlier stage.

#### **(a) ISIC 3-digit data:**

Stage V begins with estimation of the missing observations for output based on available production indexes.<sup>6</sup> Then the other three variables are estimated on the basis of past trends in the relationships between output and the three variables.

Four steps are involved in this approach. The first step is to extrapolate the reported GO3 for each branch by applying corresponding IN3 and MVA deflator or, if MVA deflator is not available, GDP deflator or consumer price index. (The reason for the use of such overall deflator is that no branch-specific deflator is generally available.) In symbols, the estimated GO3 (**EGO3**) for a given branch can be expressed as;

$$EGO3_t = GO3_0 * IN3_t * MVADEF_t, \quad t=1,2,\dots,k$$

(*Note: If MVADEF is not available, then GDP deflator or consumer price index is used instead of MVADEF.*)

where  $GO3_0$  is the reported GO3 for the latest year ( $t=0$ ), IN3 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 IN3 is available).

---

<sup>6</sup> As the result of the direct incorporation of the production indexes which are, in general, more timely published by national statistical offices than 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 output.

The second step is to estimate missing data on VA3 by estimating a country- and branch-specific time-series regression equation (throughout this paper, the residual terms of regression equations are omitted for simplicity);

$$VA3 = a + b*GO3$$

where a is the intercept and b is the regression coefficient on GO3 and is assumed to be in the range between 0 and 1. Estimates on missing VA3 for the years up to the year k are derived from this equation by using the estimated GO3.

In the third step, similar to the case of VA3, missing data on WS3 are estimated by the following regression equation;

$$WS3 = c + d*VA3$$

where c is the intercept and d is the regression coefficient on VA3.

Finally in step four, missing data on E3 are estimated from the following regression equation over time;

$$\ln(E3) = e + f*\ln(IN3)$$

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

After completion of the above estimation procedures, still many missing observations will remain unestimated. The provisional estimation of these missing data for latest years consists of again four steps. Step 1 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 have 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 UNSD, the World Bank and OECD. These series have a time lag of only two years and provide estimates of MVA. These series are used to update MVA in countries with a time-lag of more than two years.

In Step 3, 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 output, or wages and salaries 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.

Finally in Step 4, 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 period.

**(b) ISIC 4-digit data:**

Due to different data situations, somewhat different estimation procedures are applied to ISIC 4-digit data. With respect to this data set, individual reported time series (i.e., data at Stage I or II) is extended for 's' years by the following data-projection procedures:

Similar to the case of the ISIC 3-digit data, the data projection is carried out with four steps.

While the estimation work for ISIC 3-digit data begins with utilizing available production indexes, that for ISIC 4-digit data is based on the past trends in structural changes since production indexes at the 4-digit level of ISIC are generally not available.

Suppose that time-series data at the ISIC 4-digit level are available for the period up to year T. Then, given the estimates for VA3, GO3, WS3 and E3 for years T+1, T+2, ..., T+s being available, the following method for projecting ISIC 4-digit data for years T+1, T+2, is employed:

**Step 1 - Projection of value added (VA4):**

For each industry (i.e., ISIC 4-digit category) initial estimates of VA4<sub>t</sub> (t=T+1, T+2, ..., T+s) are obtained in the following estimated time-series regression equation which is industry- as well as country-specific:

$$\ln(\text{VA4}/\text{VA3})_t = a + b \cdot \ln(t), \quad t=1, 2, \dots, T.$$

Initial estimates of VA4<sub>T+1</sub>, VA4<sub>T+2</sub>, ... and VA4<sub>T+s</sub> for all k industries within a given branch are obtained through this method. Since these individual estimates are obtained independently from industry to industry, it is unlikely that they would sum up to the VA3 for the year in question. Based on the distribution of these initial estimates in the corresponding branch (i.e., ISIC 3-digit category), the available VA3 for the year T+i (i=1, 2, ..., s) is disaggregated proportionally.

**Step 2 - Projection of output (GO4):**

Initial estimates of  $GO4_{T+1}$ ,  $GO4_{T+2}$ , ... and  $GO4_{T+s}$  are obtained from a time-series regression equation:

$$GO4_t = c + d*VA4_t, \quad t=1, \dots, T$$

The linear relationship between the two variable assumes that regardless of the values of coefficient  $c$  and  $d$ , the elasticity of  $GO4$  with respect to  $VA4$  approaches unity as  $VA4$  increases. In the current exercise, it is desired that ' $c$ ' be positive and ' $d$ ' be greater than unity since the value added ratio generally increase in the long run but never exceed unity.

As in the case of  $VA4$ , each of the available  $GO3$  for the year  $T+i$  ( $i=1, \dots, s$ ) is disaggregated proportionally based on the distribution of these  $k$  initial estimates for the year  $T+i$ .

Step 3 - Projection of wages and salaries (WS):

Initial estimates of  $WS4_{T+i}$  ( $i=1, 2, \dots, s$ ) are obtained from a time-series regression equation:

$$WS4_t = e + f*VA4_t, \quad t=1, \dots, T$$

As in the cases of  $VA4$  and  $GO4$ , each of the available  $WS3$  for the year  $T+i$  ( $i=1, \dots, s$ ) is disaggregated proportionally based on the distribution of these  $k$  initial estimates for the year  $T+i$ .

Step 4 - Projection of the number of employees (E4):

Desirably the number of employees be related to the level of real output. However, due to the lack of data on real output or production indexes, the level of employment will be related to total wage bill in real term. Thus, initial estimates of the number of employees for the individual  $k$  industries for the year  $T+i$  ( $i=1, \dots, s$ ) are obtained from the following time-series regression equations assuming constant elasticities with respect to the explanatory variables:

$$\ln(E4_t) = g + h*\ln(WS4_t/CPI_t) + i*\ln(t), \quad t=1, 2, \dots, T$$

The second term on the right-hand side refers to the 'general trend over time'. The inclusion of this term is due to the fact that the level of employment is generally not determined strictly by the level of output which explains the level of wage bill.

Each of the available  $E3$  for the year  $T+i$  ( $i=1, \dots, s$ ) is disaggregated proportionally based on the distribution of these  $k$  initial estimates for the year  $T+i$ .

In conclusion, the five stages which compose the database not only provide an operational framework for STAT statisticians but also present advantages to the data user. Each data item on the INDSTAT CD-ROM products is accompanied by a numerical "source" code, from 1 to 5, indicating the stage at which the figure was ultimately derived. For instance, codes "1" and "2" refer to national data sources (belonging to Stage I and II, respectively); code 5 means that the data are provisional estimates.

## **B. Existing Data Problems and 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 many countries and areas included in the INDSTAT 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, for instance, in the case of data at the 3-digit level of ISIC, more than two thirds of the countries and areas included in the database report at least some data which are combined for two or more branches 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 ISIC categories, not be shown separately, especially if the number of reporting units in the category or categories 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 branch or industry.

Large part of the adjustments made by STAT has involved disaggregation of data referring to two or more ISIC categories. 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 reasonable compatible with aggregated data from official sources. Consequently, the use of the shares of ISIC categories 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 data for individual categories, shares from the surrounding years were interpolated.

(iii) Proxy variables<sup>7</sup>: Proxies which measure roughly the same dimension of industrial activity as the problem variable were also utilized. For example, the main difference between sales and output is the change in stocks of finished goods and work-in-progress. This, of course, would vary among ISIC categories and from year to year, but in the absence of precise information on output, 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 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 ISIC 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 county (e.g., urban area, metropolitan area) or to a part of the manufacturing sector (e.g., publicly-owned enterprises, selected branches or industries, 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.

---

<sup>7</sup> 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.

(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 these 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 view of attaching the first priority to officially reported data via national questionnaires and because of general lack of official supplementary information, STAT does not adjust reported data in terms of cut-off point.

(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 system 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 ISIC category (branch or industry).

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) or 4-digit (industry) 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 ISIC categories 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 ISIC category/categories for which data were missing;
- b) The availability of independent totals in country/variable/years where disaggregated data were missing;
- c) The internal consistency of existing data;
- d) The configuration of missing items within the database matrix for each variable; and

- e) The availability and goodness of fit of data on the same or proxy or related variables - within the INDSTAT or from supplementary sources - for the missing country, ISIC category or 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 ISIC categories were included with those for other ISIC categories in larger aggregates. Thus, the estimation exercise could be reduced to one of splitting ISIC combinations, i.e., with a known but undistributed residual to be allocated among the missing ISIC categories.

c) Internal consistency of existing data. Internal consistency was measured on the basis of the regularity of year-to-year changes in the shares of ISIC categories, or in ratios between data for related variables for each ISIC category 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 isolated estimates for missing data for even a few ISIC categories would result in changes in the totals for manufacturing, thus affecting in turn the relative shares of data for all other ISIC categories 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 ISIC categories or years; (ii) in most or all ISIC categories for one or more years; or (iii) in one or a few ISIC categories 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 (or industry) shares of the same variable in surrounding years or a neighbouring year. Pattern (ii) was also relatively straightforward, using branch (or industry) shares in a bench-mark year - or interpolated shares in surrounding years - if totals for manufacturing during the missing years were available. If totals were not available but disaggregated 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 shares of individual ISIC categories - if some data for the missing variable or ISIC categories were available. If the branch (or industry) 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 data for individual categories of ISIC. 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 database. 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 database 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/obsolete 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<sup>8</sup> 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 ISIC categories. 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

---

<sup>8</sup> *International Recommendations for Industrial Statistics*, Statistical Papers, Series M, No.48, Rev.1, United Nations, 1983, paragraph 7.

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) Output and value added. Among the variations in concept that may apply to the data on 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 output or value added may be at producers' prices or factor values<sup>9</sup>. 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 and/or industries 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 valued 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 (or industry's) business performance. However, if VA (as well as output) is to be an indicator of production, data on VA excluding operational loss (or the other extreme, monopoly profit) would be more useful

---

<sup>9</sup> 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 tape/diskettes. 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.

for multi-country analysis. However, in practice, adjustment of reported VA in this line is not feasible.<sup>10</sup>

### C. Estimation of Production Indexes at the 3-digit level of ISIC

All above discussions refer mainly to the four variables - employment, wages and salaries, output and value added. While the primary compiler of international data on these variables as well as those on number of establishments and gross fixed capital formation is UNIDO, the primary source of data on production indexes is UNSD which compiles these data from national sources through questionnaire. The reported indexes usually need to be re-based, however. In many aspects, the way STAT treats reported production indexes is somewhat different from the other 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., 1990) 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, *Industrial Commodity Statistics Yearbook*. In the absence of price weights for combining the series, unweighted geometric means were calculated. Each

---

<sup>10</sup> However, if the country- and branch- (or industry-) 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.

quantity series was converted to an index (with a selected common base year) 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 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 INDSTAT 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 database, 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 database - output (GO), value added (VA), wages and salaries (WS) and employment (E) - per country, ISIC category, combinations of ISIC categories, and total manufacturing (ISIC 300) 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, ISIC category (or combination of ISIC categories) 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 ISIC category;
- (iii) (i) and (ii) in relation to other ISIC categories; and
- (iv) (i), (ii) and (iii) in relation to other years.

The criteria consist of acceptable ranges for (i) to (iv) above. 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.

/\*