CHAPTER VI. DATA SOURCES AND DATA COMPILATION METHODS

6.1. This chapter contains general recommendations on data sources and data compilation methods for use in distributive trade statistics. More detailed guidance on the relevant good practices will be provided in *Distributive Trade Statistics: Compilers Manual*, which is to be issued as a follow-up publication to the current recommendations.

A. Data sources

6.2. *Data sources for compilation of distributive trades statistics*. The generation of distributive trade statistics is based on data collected from numerous sources describing production activities and selected balance sheet items of units engaged in distributive trade as classified in Section G of ISIC, Rev.4. Two basic categories of data sources can be distinguished according to their purpose or the provider of the statistical information. In either category, however, the original sources of the data are the same, namely the records kept by the trade units. These two data sources are:

(a) Statistical data sources that provide data collected specifically for statistical purposes, such as census and survey data;

(b) Administrative data sources that provide data created originally for purposes other than the production of statistical data.

1. Statistical data sources

6.3. *Statistical surveys*. Statistical surveys of units concerned are traditionally the main source of information for compiling distributive trade statistics. The surveys are done either by enumerating all the units in the population (census) or by eliciting response only from few representative units scientifically selected from the population (sample survey).

6.4. The main advantages of statistical surveys as compared to the administrative data sources are that the planning, execution of the surveys, data collection and the processing procedures are under the control of the statistical office itself. In principle, respondents have less reason to deliberately misreport the data as the statistical office guarantees that the data it collects are strictly confidential and that they will not be used for other than statistical purposes. The survey approach, however, has certain disadvantages such as the resource intensity (both financial and manpower), additional respondent burden, higher non-response rates and sampling errors.

6.5. *Economic census*. In general, an economic census[1] is a statistical survey that is conducted at infrequent intervals of time (usually every five or ten years) aiming at collecting comprehensive and detailed statistics about the operating characteristics and structure of units engaged in all (or some) of the economic activities. Some of the main objectives of an economic census are to establish and maintain the business register and provide a sampling frame for more frequent statistical surveys.

6.6. *Census of trade units.* The census of trade units can be conducted either as a part of an economy-wide census, including all economic activities, or as an independent census for distributive trade sector/activities only. It should be noted that the census planning and organization and the subsequent transformation of census's basic data into distributive trade statistics data items is a time consuming and resource intensive exercise. This approach is costly; imposes a high burden on respondent units; may reduce the response rates and thus affects the quality of collected information. Conduct of a complete census of trade units may be useful in cases when a particular country does not maintain an up-to-date business register or there is a significant users' interest for detailed statistical data by geographical area. Censuses of trade units should not be conducted if there are other ways of collecting and producing distributive trade statistics of a high enough quality.

6.7. Censuses of trade units tend to provide a complete enumeration of units engaged with trade activity (including the small units of informal sector) at a particular point of time and are an appropriate approach for generation of trade statistics required at longer intervals of time. Censuses, however, are limited in terms of data content. For countries implementing censuses of distributive trade units as part of their data collection strategy it is recommended that the censuses are followed as closely as possible by periodic (annual, quarterly or monthly) sample surveys, providing a continuous measure of trade activity and collecting more detailed sector specific data.

6.8. In some countries, enterprise survey frames are derived from lists created during economic censuses or from a specially maintained area frame. This is not a recommended practice. At the very least it is recommended that countries establish a permanent business register (see para. 6.30) containing all trade enterprises.

6.9. *Sample surveys*. Statisticians often use a sample survey technique to obtain data about a large population of statistical units by selecting and measuring a sample from that population. Due to the variability of characteristics among units in the population, scientific sample designs in the sample selection process are applied in order to reduce the risk of a distorted view of the population. Conclusions about the total population of units are made on the basis of the estimates obtained from the sample survey data. The sample survey technique is less costly way of data collection as comparing to the economic census. It may be used in conjunction with a cut-off point or not.

[1] There is no an internationally agreed definition of economic census. Countries may have different names and understandings for one and same statistical survey type. Some of the known variations of this term are "census of economic units", "establishment census" and "establishment and enterprise census"

6.10. *Sample surveys for distributive trade*. In most of the national statistical offices, the wholesale and retail trade sample surveys are rarely restricted to one standard form, but tend to be a combination of forms, differentiated by periodicity and major characteristics, namely:

(a) the activity, size, legal form, type of operation and the type of variables asked (turnover, expenditures, employment, other specialized variables);

(b) occasionally an extra characteristic, such as the geographical location of the unit, may influence the contents of a survey.

6.11. *Size threshold to determine the target population*. When considering the trade surveys size thresholds play an important role in determining the target population and, where relevant, the sample population of units. Most of the sample surveys are conducted for units above a certain size threshold. The reasons for this are diverse and include the desire to limit the size of the survey, to limit the response burden on businesses and also to take account of the problems of maintaining registers for smaller units. There is no international recommendation for an appropriate size threshold. The decision is left to the judgment of each national statistical office and may vary between surveys for different trade activities and periodicity. However, countries are encouraged to make periodic assessments of the under coverage of the surveys due to the thresholds and to include a description of such thresholds in country's metadata which will be made available to users.

6.12. *Types of surveys for collecting data about trade units*. In general, three types of sample surveys are appropriate for collecting data about trade units depending on the units sampled and/or contacted, namely *enterprise survey, household surveys,* and *mixed household-enterprise surveys*. Choice of the type of survey to be conducted depends upon the statistical system of a country and the resources available to its statistical office.

6.13. *Enterprise surveys* are those in which the sampling units comprise enterprises (or statistical units belonging to these enterprises such as establishments or kind-of-activity units) in their capacity as the reporting and observation units from/about which data are obtained. In the *household surveys* on the other hand the households are the sampled, reporting and observation units. In *mixed household-enterprise surveys*, a sample of households is selected and each household is asked whether any of its members own and operate an unincorporated enterprise (also called informal sector enterprise in developing countries). The list of enterprises thus compiled is used as the basis for selecting the enterprises from which desired data are finally collected. Mixed household-enterprise surveys are useful to cover only unincorporated (or household) enterprises which are numerous and cannot be easily registered.

6.14. *Sampling frame*. Availability of a sampling frame of the statistical units is a prerequisite for conducting given survey as it provides a basis for selection of sample units. Depending upon the source of the sampling frame surveys may also be classified as

either *list based* or *area based.* In a list based survey, the initial sample is selected from a pre-existing list of enterprises or households. In an area based survey, the initial sampling units are a set of geographical areas. After one or more stages of selection, a sample of areas is identified within which enterprises or households are listed. From this list, the sample is selected and data collected.

6.15. *Enterprise surveys.* Enterprise surveys assume the availability of a sampling frame of trade enterprises. The sampling frame is made available from the business register, if such a register is maintained by the statistical office to support a range of surveys (see para. 6.30). For countries not maintaining a current up-to-date business register, the list of enterprises drawn from the latest economic census is recommended to be used as a sampling frame. In an area based enterprise survey, a sample of areas is selected first, and then selected areas are enumerated for compiling the list of enterprises operating in the area which serves as the sampling frame for selection of the enterprise in the sample and to collect the required information. It is recommended that for surveys of distributive trade enterprises the list based enterprise surveys be generally preferred to area based surveys for the following reasons:

(a) A list-based survey is more efficient from a sampling perspective in terms of sample size. The area based approach involves cluster sampling which require a larger sample than in the case of list based survey sample in order to achieve a given level of accuracy.

(b) It may be difficult to enumerate the enterprises within an area. While many enterprises are likely to be readily identifiable, household-based enterprises that carry out their work within the household or do not have a fixed location are usually difficult to identify.

(c) Maintenance of a list of enterprises via a general purpose statistical business register is cheaper than maintenance of an area based list.

(d) Area based sampling is inappropriate for large or medium sized enterprises that operate in several areas because of the difficulty of collecting data from just those parts of the enterprises that lie within the areas actually selected. Furthermore, in order to avoid inadvertently missing parts of the enterprise, it is usually considered preferable to collect data from the whole of an enterprise not just a part of it

6.16. It is recommended that countries use area based enterprise survey approach for collection of data from small trade enterprises generally operating in informal or unorganized segment of the economy. For such enterprises satisfactory register or list is normally not available.

6.17. *Household surveys.* Household enterprises which are unincorporated producer units are not recognized as a legal entity separate from their owners (see para. 2.44). Fixed and other assets used in the production by these enterprises do not belong to the enterprises but to their owners. Compiling a satisfactory list of such enterprises is either

not feasible or is a very resource intensive exercise. Household surveys are recommended in providing coverage of production of such enterprises.

6.18. As household surveys exist for the purposes of collecting labour force and household expenditure data, additional questions related to production activities can be added at relatively little extra cost. This makes the use of a household survey generally cheaper than conducting an area based enterprise survey for the same purpose. It should be noted however, that the responding unit is a person in a household, not an enterprise and the data that can be collected about trade activities of the enterprise may be correspondingly more limited. Some statistical offices maintain, or can access, population or household registers, at least for urban areas, and thus can conduct list-based household surveys. However, there are few such registers, so most household surveys are area-based.

6.19. *Disadvantages of household surveys*. The main disadvantage of the use of household surveys for collecting data from the unincorporated trade enterprise is that the sample of such surveys is not designed to provide a representative coverage of trade activities, but on the distribution of households. Although, it is possible that the retail trade which by definition sells goods and provides services to final consumers (households) may be spread across areas in a similar way as the population, in many cases the two distributions are different, as trade activities tend to be concentrated in commercial and market zones.

6.20. *Mixed household-enterprise surveys*. In the mixed household-enterprise surveys, a sample of households is selected and each household is asked whether any of its members is an *entrepreneur, i.e.,* the sole proprietor of, or a partner in, an unincorporated enterprise engaged with economic (including trade) activity. Data for all the enterprises thereby identified (or for a sub-sample of them) are then collected – either immediately from the respondent reporting on behalf of the enterprise or in a subsequent stage of data collection. Thus the feature of a mixed household-enterprise survey that distinguishes it from a household survey is that it collects information about enterprises *per se*, whereas a household survey collects information about the persons in a household, including possibly their personal contributions to enterprises.

6.21. Mixed household-enterprise surveys are also recommended for providing data on small enterprises that are not included in list-based enterprise surveys. Countries should be aware that they suffer from similar to area based enterprise surveys disadvantages, namely the inefficiency of the sample design and the difficulty of handling enterprises with production units in more than one location. In general, the mixed household-enterprise surveys are recommended as the preferred to household surveys or area based enterprise surveys approach for collecting the data and estimating the output of small trade units that are excluded from list-based enterprise surveys.

6.22. *Modified mixed household-enterprise surveys approach.* To avoid the limitations of the mixed household-enterprise survey approach (see para. 6.21), some countries[2]

[2] For example India and Philippines

adopt a modified version of the approach, which involves a dual, mutually exclusive, listing of (i) households and household-based business operators; and (ii) establishments in the sample areas. At the listing stage, each structure of the selected area units is visited to identify and prepare a complete list of all establishments falling in the domain of the survey. Modified mixed household-enterprise surveys approach is recommended as preferred to an area-based enterprise survey as it improves the quality of data of micro and small units specially the mobile units as compared with those with fixed location.

6.23. *Respondent burden*. Minimizing the respondent burden should be an important objective for the national statistical offices when distributive trade surveys are designed and conducted.

6.24. Special attention should be made to the issue of the respondent burden. As a way of reducing the respondent burden it is recommended that countries co-ordinate data collection both internally at the statistical office, by central supervision of the delimitation of sampling frames and selection of the samples drawn, and externally by using existing sources of information, such as administrative registers, to the largest possible extent.

2. Administrative data sources

6.25. *Administrative data sources* are set up in response to legislation and/or regulation. Each regulation (or related group of regulations) results in a register of the units – enterprises, persons, etc. – bound by that regulation and in data resulting from application of the regulation. The register and data are referred to collectively by the statistical offices as an *administrative source*. The administrative authorities keep records of the units in response to legislated administrative requirements or simply for internal purposes to assist the units in managing their operations. The data emanating from the administrative source can be used by the statistical offices. It is recommended when countries use administrative data sources for statistical purposes they pay special attention of their limitations and describe them in their metadata.

6.26. *Privately controlled administrative data sources*. Besides from the administrative data sources set up in response to legislation and/or regulation, statistical offices may obtain certain data from a private sector data supplier. Private sector data suppliers[3] operate on a commercial basis so the transfer of data from them to the statistical offices takes the form of a contract with a payment of a fee.

6.27. *Main advantages of the administrative data sources.* The following is the list of the most important advantages of administrative data sources:

(a) Complete coverage of the population to which the administrative process applies and perceived as low non-response;

[3] An example of a private sector data supplier is Dun and Bradstreet in the United Kingdom

(b) Avoidance of response burden. The responding units make available the information as part of the administrative procedure;

(c) Cheaper for the statistical office to acquire data from an administrative source than to conduct a survey;

(d) Suitable for covering the smallest segment of the units' population which contributes relatively little to the estimates but makes up a substantial percentage of the number of units in the population;

(e) Smaller than a survey sampling errors;

(f) Some data may be more accurate because of intense data checks by administrative authorities

6.28. *Main disadvantages of the administrative source* include the following:

(a) Discrepancy between administrative concepts and statistical concepts. As the administrative processes are not under statistical office control concepts regarding variables and units in respect of data coverage, content, quality and consistency comply with the administrative objectives. This limits the use of administrative data for statistical estimation and analysis purposes.

(b) Poor integration with other data of the statistical systems. This is in particular a problem when administrative units do not correspond to statistical units either because of difference in the concept or because of deviating identification numbers. Even if the variables existing in the administrative register perfectly fit to the needs of the statistical office, matching problem can prevent from using them.

(c) Risks with respect to stability. Administrative processes are subject to change in response to new legislation without much (or any) regard for the impact on the statistical series. This may cause systematic bias.

(d) Data may become available with unacceptable delay.

(e) Legal constraints with respect to access and confidentiality.

6.29. It is recommended that compilers of distributive trade statistics identify and review the available administrative data source in their countries and use the most appropriate of them for compilation of distributive trade statistics. This can be of a great help in reducing significantly the response burden and the surveying costs. The relative advantages and disadvantages mentioned above have no absolute value. It depends on the specific situation whether they apply and to what extent. Therefore, the review has to be seen as a checklist which can be used in the process of decision making. Examples of the most appropriate administrative sources are the tax authorities (any fiscal or VAT information on units), customs authorities, social security registers etc.

3. Business register as a frame for statistical surveys

6.30. *Need for a business register.* The organization and conduct of any enterprise survey of distributive trade units assumes availability of an adequate sampling frame, i.e., the set of units subject to sampling together with the details about them that will be used for stratification, sampling and contact purposes. In principle, the sampling frame should contain all the units that are in the survey target population, without duplication or omissions. A business register, maintained by the countries for statistical purposes is recommended as the most appropriate source for deriving the sampling frame for distributive trade surveys.

6.31. *Statistical business registers*. In general, statistical business register means a comprehensive list of all enterprises and other units together with their characteristics that are active in a national economy. It is a tool for the conduct of statistical surveys as well as a source for statistics in its own right. The establishment and maintenance of a statistical business register in most of the cases is based on legal provisions as its scope and coverage is determined by country specific factors. As the best option it is recommended that the frame for every list-based enterprise survey for distributive trade is derived from a single general purpose, statistical business register maintained by the statistical office, rather than the option of using stand-alone registers for each individual survey. There are two basic reasons for using a single statistical business register. First, and most importantly, the statistical business register operationalises the selected model of statistical units and facilitates classification of units according to the agreed conceptual standards for all surveys. Second, it is more efficient for a single organizational unit within the national statistical office to be responsible for frame maintenance than to create units responsible for the frames of each survey separately.

6.32. *Establishment of a statistical business register*. The starting point for the establishment of a statistical business register should be the available administrative registers that are registers of enterprises created and maintained to support the administration of certain legislation or regulations. If only one administrative register is used, the resulting statistical business register would likely to be deficient in terms of coverage and content and would not provide an adequate sampling frame for subsequent statistical surveys. Countries are encouraged to work towards improvement of the coverage and content of their statistical business registers by incorporating data from several administrative sources. Each administrative register should be examined carefully by statistical offices in terms of units' coverage and quality of data before being used. It should be mentioned that combining the data would be possible only if a single business number for all enterprises is introduced.

6.33. *Maintenance of the register*. To be a central sampling and weighting frame for all statistical surveys, including distributive trade surveys, the statistical business register should be up-to-date and with satisfactory quality. In practice, however, the enterprises in it do not remain the same over time - the legal units that own them may merge or split up or go out of business; they may change production activities or move the location; or new

enterprises may be created (births) and existing enterprises may cease to exist (deaths). For these reasons it is recommended that the statistical business registers be regularly maintained and updated to take note of the changes in the enterprise dynamics. Unless the business register is regularly maintained, it will quickly loose its value as it becomes dated and ceases to adequately reflect the real world.

6.34. *Sources for the establishment and maintenance of a statistical business register*. In principle, the sources used for establishing a statistical business register usually are used also for its maintenance. They include the following:

(a) *Economic census.* Economic censuses (see para. 6.5) provide in practice the most comprehensive list of units and links between them in a given country on which basis a statistical business register can be established and maintained. Economic census is recommended for use in the cases explained in para. 6.6.

(b) *Administrative data sources.* Administrative data sources are one of the most important sources for establishing a statistical business register (see para. 6.32), however, there are also a number of problems associated with their use for its maintenance. Common examples of administrative data sources that may be used to establish and maintain business registers include business registration systems, VAT tax systems, payroll tax systems, and records maintained by the governments for the administration of unemployment insurance, social security or other programmes. Such records however, need careful review to determine their completeness, suitability and accuracy; they are not designed primarily to serve economic survey needs. These sources are known to contain inactive units; they may also be deficient in terms of activity classification of units, of contact information, and of the ability to track an unincorporated enterprise through a change of owners.

(c) *Feedback from enterprise surveys.* Feedback from enterprise surveys is a vital source for establishing and updating the statistical business register as it provides new information on contact address changes, closure of business, change in the economic activity of the unit, etc.

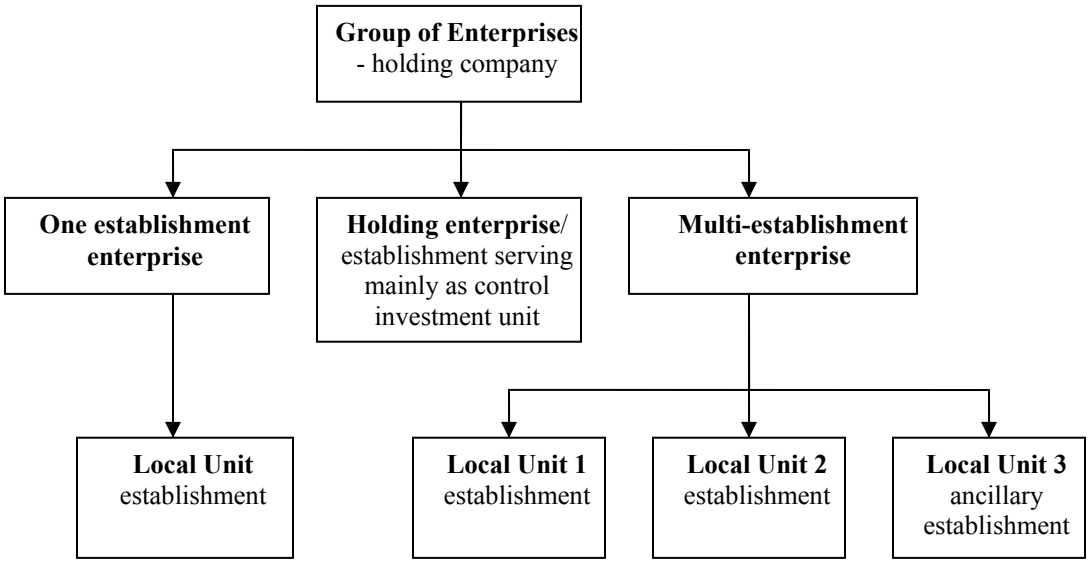
(d) *Business register surveys*. Register updating information that cannot be obtained from the administrative source on which the register is based, or from survey feedback, has to be obtained by *business register surveys* (sometimes termed *nature of business surveys*) and profiling operations conducted by business register staff.

(e) *Other potential sources.* These include information maintained in trade associations about their members, telephone directories or special listings prepared by telephone companies etc.. Each type has its own special characteristics which must be studied carefully before a decision is made on how to use it.

6.35. In general, the statistical business register is set up using one record for each establishment and one record for each enterprise with the link identifiable between each establishment and its parent enterprise. For multi-establishment enterprises, this means

that there will also be a record for the central office, and each establishment should be cross-referenced to the central office. It is recommended that countries assign proper coding to the enterprises and establishments as to establish hierarchical link between them as shown in the graph below. The coding of relationship would allow for the allocation of the operating surplus of the main establishment to its supporting ancillary units and the imputation of the outputs of ancillary units as intermediate consumption to consuming establishments. Holding companies are not ancillary units because the functions they perform to control and direct subsidiary companies are not ancillary activities. The 2008 SNA treats holding companies as financial corporations.

Figure 3. Hierarchical links between a group of enterprises, an enterprise and its constituting establishments



6.36. As a minimum, the statistical business register should include the following information about trade units:

- name and physical location of each enterprise;
- mailing address, which may be different from its physical location;
- name and address of the central office or the headquarter of the enterprise and establishments that are part of multi-establishment enterprise;
- kind of economic activity, description or code;

- legal organisation - incorporated and unincorporated;

- type of ownership: public (by central, state and local governments); national private and foreign controlled;

- number of persons employed;

- volume of sales or value of output;

- source and date of information.

B. Data compilation methods

6.37. Data as they have been received from the respondents to the statistical surveys is the starting point for the compilation of distributive trade statistics. The process of data compilation comprises more than just aggregating the questionnaire items. Statistical offices perform a number of checks, validation and statistical procedures on collected data with the aim to bring them to the level of the intended statistical output. The most important of these procedures are explained in the paragraphs below.

6.38. *Data validation and editing.* Like any other survey respondent, a trade statistics respondent is prone to commit errors while completing a statistical questionnaire. Thus, data collected in best of statistical surveys are affected by response and non-response errors of different kinds. To resolve these problems of missing, invalid or inconsistent responses, editing and imputation have become an integral part of all types of statistical surveys data processing operations. Editing is the systematic examination of data collected from respondents for the purpose of identifying and eventually modifying the inadmissible, inconsistent and highly questionable or improbable values, according to predetermined rules. It is an essential process for assuring quality of the collected information. Micro editing (also called input editing) focuses on the individual record or questionnaire, as opposed to macro editing where checks are performed on aggregated data.

6.39. *Selective (significance) editing*[4]. Selective editing is an approach for prioritizing and further reducing costs of editing, which is one of the most resource-consuming processes in the production of official statistics. It is a procedure which targets only those of the micro data items or records that would have a significant impact on the distributive trade surveys results. It is recommended that while deciding on allocation of resources on various stages of statistical process countries give priority to the use of selective editing as a more efficient method of editing of distributive trade data.

[4] For more details see OECD STES Timeliness Framework: Selective (or Significance) Editing at: http://www.oecd.org/document/21/0,2340,en_2649_34257_30214485_1_1_1_1,00.html

6.40. The data editing may take place during (input editing) or after the data entry phase (output editing). The following edit checks are recommended as useful for detecting errors in distributive trade data:

(a) *Routine checks* - used to test whether all questions which should have been answered in fact do have been answered.

(b) *Validation checks* - used to test whether answers are permissible. Response to a particular data item in the questionnaire is checked against a valid value range specified for the purpose. Any observation lying outside the valid value range should be reviewed by the compilers of data and corrective actions taken.

(c) *Rational checks* - set of checks based on the statistical analysis of respondent data. Many checks take form of a ratio between two variables, which should be within specified limits. Another type of rational check is the arithmetic check, for instance specifying that a sum of variables should equal a total.

6.41. Large random errors by respondents can usually be picked up through plausibility checks on the data, for example by comparing the data reported with previous values, or the ratios of data reported with reasonable bounds for the types of enterprise. Not all errors committed by respondents can be traced by the statistical office and therefore even exhaustive data editing will never result in error-free data file. For example, sustained systematic errors, such as under reporting of turnover and over reporting of expenditures by trade units can hardly be detected.

6.42. *Influential observations.* Some particular data item responses have most significant impact upon the main estimates. These are often termed as *influential observations.* Editing efforts should generally be more focused on such data item responses. In particular, very large enterprises are usually a source of influential observations and their data should be individually checked.

6.43. *Imputations.* Missing data is often encountered in most of the trade surveys which creates problems for the data editing. The data may either be missing for a particular data item of the questionnaire (item non-response) or the selected unit may not return the filled-in questionnaire at all (unit non-response). The technique of imputation is used for estimating the missing data in case of item non-response. The problem of unit non-response is dealt with either by re-weighting or imputing the data from previous available periods of that unit (substitution), or on the basis of the available administrative information for it.

6.44. *Item non-response*. Item non-response or partial non-response occurs when the sampled unit has not answered all relevant questions, but did respond to only part of them. Cases may arise wherein a respondent has reported on all questions but either some of the answers may not be logically correct or there may be inconsistencies between some of the answers provided by a respondent. Presence of such item non-response and invalid

data in the data set ultimately affect the quality of the survey results. Much of these are removed by following appropriate editing rules.

6.45. *Unit non-response*. Though the units selected in the sample are legally required to provide response to the survey conducted by the statistical offices and are liable to be penalized in case of a non-response this does not efface the problem of non-response. The non-response remains a problem which may occur for one reason or the other, namely, non existence of the unit included in the survey, lack of appreciation of the importance of the data on part of the respondents, refusal, not knowing how to respond, lack of resources and non-availability of the desired information.

6.46. There are ways to minimize the non-response including the awareness of the importance of the data to be collected and appeal to the respondents to cooperate with the statistical authorities through the print and electronic media at the launch of the survey, reminders to the non-respondents and resorting to the enforcement measures laid down in the national legislation.

6.47. *Approaches for dealing with item non-response*. Presence of non-response requires that steps should be taken to reduce its effect on the estimates. There are two general strategies to deal with missing data item (non-response):

(a) All forms with missing values are ignored and confined to analysis of the fully completed forms; or

(b) Missing data are imputed so that the data matrix is complete. Statistical analysis techniques are applied on the full data set completed with the help of imputation.

6.48. Adopting the first strategy leads to discarding even the valid data contained in the partially complete forms. Thus, it is desirable to adopt the second strategy to deal with the item non-response. The values of individual data items that are missing from the original response or believed to be in error should not be automatically interpreted as zeroes; rather appropriate methods for imputation should be applied. When all of the data have been edited using the predetermined rules and the file is found to have missing data, then imputation is usually done as a separate step. It takes care of inconsistencies that remain unresolved in the earlier stages of manual and computer-aided scrutiny.

6.49. Imputation consists in replacing one or more erroneous responses or non-responses in a record or more than one record with plausible and internally consistent values. It is the process of filling the gaps and eliminating inconsistencies and the means of producing a complete and consistent file containing imputed data. There is variety of methods for imputation, ranging from simple and intuitive to rather complicated statistical procedures. The choice of the appropriate method depends on the objective of the analysis and on the type of missing data. Some of the commonly used imputation methods include:

(a) *Subjective treatment* - impute on the basis of values which appear reasonable. For example, one might deduce the labour costs if the number of employees are known;

(b) *Mean/modal value imputation* - impute the mean value of a variable for missing data. For categorical data impute the modal value. An improvement may be to impute the median in order to eliminate the effect of the extreme values;

(c) *Post stratification* - more precision will be achieved in keeping the imputed value closer to the true value if the mean/mode/median are imputed using the observations from those units which are homogeneous with the one with missing data. For this purpose, post stratification is used, i.e. the sample is divided into strata and then stratum mean/mode/median is imputed.

(d) *Substitution* - relies on the availability of comparable data. Imputed data can be the value for the enterprise from the same survey occasion in the previous year, adjusted to reflect the average increase (decrease) of the data item in the stratum;

(e) *Cold deck* - makes use of a fixed set of values, which covers all of the data items. Values can be constructed with the use of historical data, subject-matter expertise, etc. A 'perfect' questionnaire is created in order to answer complete or partial imputation requirements;

(f) *Hot deck* - replaces each missing value by the available value from a 'donor', i.e. a similar participant in the same survey. The donor can be randomly selected from a pool of donors with the same set of predetermined characteristics. A list of possible donors matching these criteria is created and one of them is randomly selected. Once a donor is found, the donor response (for example, the yearly income) replaces the corresponding missing or invalid response;

(g) *Nearest-neighbour imputation or distance function matching* - assigns an item value for a failed edit record from a "nearest" passed edit record. In this case, the "nearest" is defined using a distance function in terms of other known variables. The closest unit to the missing value is then used as the donor;

(h) *Sequential hot deck imputation* - this method also uses classes and requires single pass. The values from passed edit records are stored and the missing value is replaced by a function of the stored values. It begins with a cold deck value. The main disadvantage of this method is that it often leads to multiple uses of donors, thus affecting the distribution.

(i) *Regression (model based) imputation* - a set of predictor variables of the passed records are used to regress the variable. The regression equation is then used to impute the values for the missing or inconsistent item values regression technique is used to impute the missing data for defining the predictor variable suitably.

6.50. In most imputation systems, a mix of imputation methods is used. The following are the desirable properties of all imputation methods:

(a) The imputed records should closely resemble the failed edit record, retaining as much respondent data as possible. Thus, a minimum number of variable (or fields) should be imputed.

(b) The imputed records should satisfy all edit checks.

(c) It is desirable to flag the imputed values and identify the methods and sources of imputation.

6.51. *Approaches for dealing with unit non-response.* Re-weighting is the most commonly applied approach when no response to a statistical questionnaire is received from the respondent unit. This case is referred to as unit non-response. The sample is re-weighted as to include only the responding sample units. It is common practice for the statistical offices to attach weights to the elements in the sample. These weights are used, amongst other attributes, to expand the sample information to the level of target population. Alternatively, the problem of unit non-response can also be dealt with approaches similar to those used for item non-response, namely various forms of imputing either from previous available periods of that unit (substitution), or on the basis of the available administrative information for it.

6.52. *Grossing up procedures, aggregation.* The data after it has been treated through editing for the non-response etc. is used to estimate the level of the variable. The grossing up comprises raising the sample value with a factor based on the sampling fraction (or the factor using returned data) for each cell in the stratified sample for obtaining the levels of data for the frame population. The grossing up will use edited data to calculate a value representative of all units. In case information on auxiliary variable related to the variable under study are available for units in the sample as well as in the sampling frame, more sophisticated statistical techniques can be used for using this information for grossing up.

6.53. *Outlier values*. Outliers are a particular category of influential observations which are correct but are unusual in the sense that they do not represent the sampled population and hence will tend to distort the estimates. Therefore, it is recommended that outlier values be identified and handled carefully as it may affect the estimates significantly. If the grossing up factor is large and outlier value is included in the sample, the final estimate will be substantially large and unrepresentative as it is driven by one extreme value. The simplest way to deal with the outlier is to reduce its weight in the sample so that it represents itself only. Alternatively, statistical techniques can be used to calculate more appropriate weight for the outlier unit.

C. Data collection strategy

6.54. All units in the economy engaged in economic activities within the scope of the distributive trade sector (Section G of ISIC, Rev.4) should be covered by the statistical surveys and/or administrative data sources for the purpose of collecting and compiling distributive trade statistics. This embraces units of all sizes and types including corporations and unincorporated (household) units. The household units include small trade enterprises that are household-based, operate outside the household at a separate location (i.e. fixed stall at a market place), or have no fixed location (i.e. a movable stall along a public road or a street vendor). An unincorporated household unit (called also an informal sector unit) is a term utilized in developing countries. In most of the developed countries, a household unit generally takes a more formal form of small enterprise and is incorporated. Some small household units, however, may still remain unincorporated.

6.55. In order to ensure a complete coverage of distributive trade activity, countries should develop their own data collection strategy based on an integrated approach covering in principle all trade units across all class sizes enterprises, commensurate with their specific statistical and organizational circumstances. The legal organisation (incorporated or unincorporated), size (from large multi-establishment enterprises with more than 250 employees to small single establishment enterprises with less than 10 employees) and the ownership pattern (public sector, privately owned and foreign controlled) of units within the scope of distributive trade statistics differ significantly. An illustration of a general data collection strategy for different segments of the economy is presented in the diagram below.

6.56. At one end of the spectrum are the corporate trade units which are incorporated under the statute of a country (public enterprises) and are comparatively large, while at the other end are the unincorporated trade enterprises characterised by a low level of organisation.

(a) *Public incorporated enterprises.* These enterprises are quite organised and are required to keep account of their transactions and to present their annual statements to the authorities with whom they are registered. A directory of such units is always available. Their number is not expected to be large and such enterprises should be covered on a complete enumeration basis.

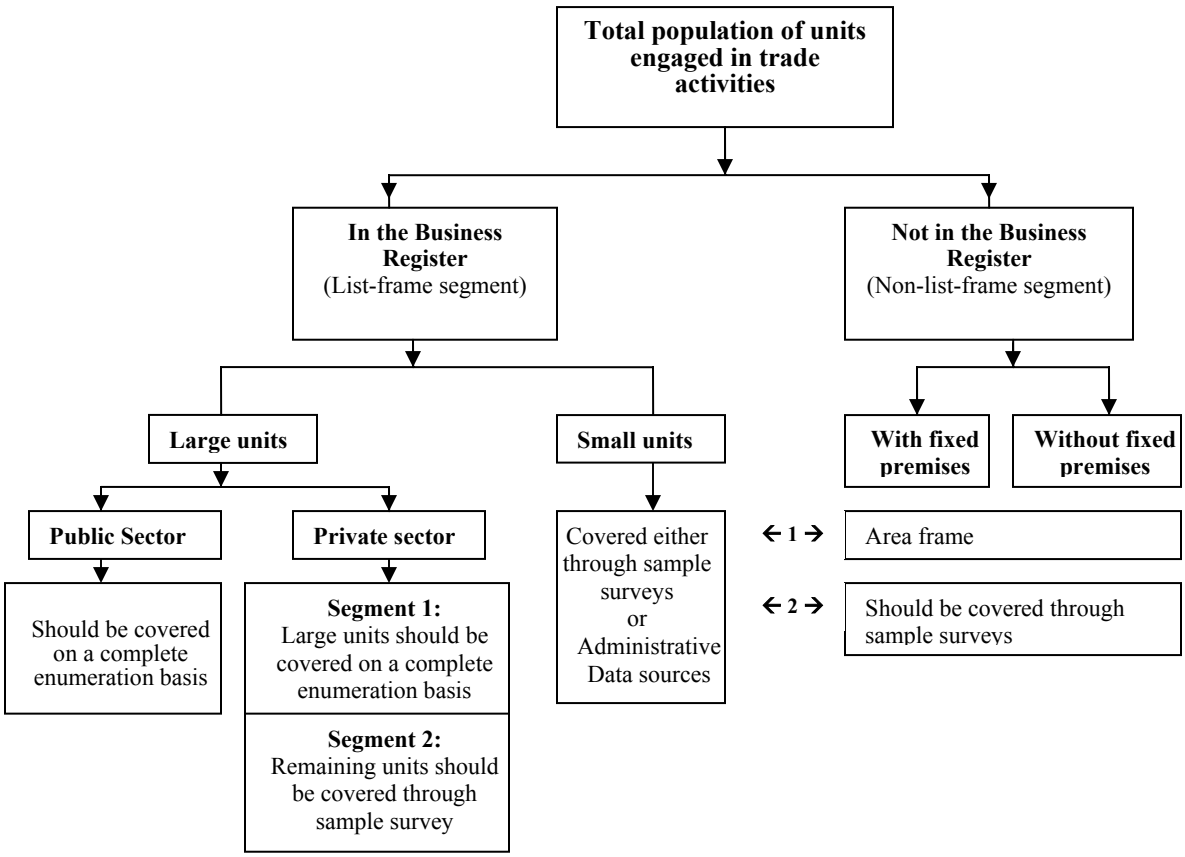
(b) *Private and foreign controlled incorporated enterprises.* The coverage of these enterprises should be achieved by dividing them into two segments – one containing the large-scale units and the second containing the rest. It might be considered that the large-scale segment of the economy is not suited for sample surveys because the differentiation in size and activity is great compared with number of units involved. Enterprises in large-scale segment, therefore, should be covered on a complete enumeration basis if possible. The smaller enterprises, whose number tends to be much larger, are relatively homogenous as compared to the large-scale segment counterparts. Sample survey can gainfully be used to cover this segment of enterprises

(c) *Small enterprises*. The segment of small incorporated enterprises or unincorporated household enterprises can be covered by two approaches, depending on whether the enterprises are registered or not:

(i) Through sample surveys if these are on the statistical business register or through the use of administrative data (tax returns of small enterprises).

(ii) Through the Fully Integrated Rational Survey Technique (FIRST) if the register of unincorporated enterprises is not available (see section D below).

Figure 4. Data collection strategy for different segments of the economy[5]



[5] All units on the business register are excluded from the area frame (i.e. non-list frame segment); All units in the sample that are part of a list frame segment and included therein are excluded from the sample of non-list frame segment.

D. Survey method

6.57. Countries are encouraged to review the Fully Integrated Rational Survey Technique (FIRST)[6] as an option for a survey programme that efficiently capture comprehensive statistical information from all enterprises, including distributive trade enterprises, of all sizes operating in an economy. Application of this survey technique requires two basic statistical sets of information, namely: (i) some census enumeration, preferably an economic census, to establish the complete statistical population of units for construction of sampling frame and sample selection. In the absence of an economic census, a population census will generally be also sufficient; and (ii) good supporting documentation on sample areas/enumeration blocks for the benchmark enumeration. Once these two basic requirements are met, the field conditions should determine the selection of the most appropriate design for any particular distributive trade survey.

6.58. The FIRST methodology requires the statistical universe to be divided into two parts:

(a) A list-frame of a relatively small number of large units (hereinafter called the *'list-frame segment'*) that are clearly distinguished by their legal status from the rest of the units. For this segment either a complete enumeration or a uni-stage (most often stratified) sampling scheme is adopted; and

(b) The rest of the units (hereinafter called the '*non-list-frame segment'*) for which drawing an exhaustive list is not feasible and thus can be covered only by an (geographical) area frame approach. A two-stage (in specific cases may be multi-stage) sample design is adopted for this segment.

6.59. The FIRST methodology allows for covering all economic activities of the economy in an integrated manner and has a distinct advantage over conducting a set of separate activity surveys (each carried out independently on a single group of economic activities) to cover the same domain. Reducing survey costs is one of the main advantages of FIRST. Besides it, an integrated survey ensures a non-overlapping coverage of groups of establishments by kind of economic activity. Each establishment is classified in one and only one sector. The FIRST methodology provides comprehensive information collected in a short time-span with relatively modest means. If properly implemented, FIRST obviates the need for trade-offs between survey contents and the timeliness of release of results that often plays an important role in survey designing.

6.60. *List-frame based survey of the 'list-frame segment'*. In the surveys conducted using FIRST, the list frame is usually drawn from a business register or a directory of units that consists of all the units of the '*list-frame segment'* using the criterion of the legal and/or administrative status that distinguishes the 'large' units from the rest. This list is used for carrying out a FIRST survey preferably by mailed questionnaire with

[6] *Strategies for Measuring Industrial Structure and Growth*, United Nations, 1994, Studies in Methods, Series F, No.65 (United Nations Publication, Sales No. E.94.XVII.11).

follow-up visits where required. The definition of large-scale used here is based on practical considerations and differs from country to country. The ease of maintaining the list frame forms the single most important criterion for the definition of the large-scale sub-sector. The list frame is usually made up of the following groups which are easily identifiable:

(a) Publicly traded companies (i.e. companies listed on a stock exchange);

(b) Non-traded companies (i.e. companies registered with a government agency such as the Justice Department, Ministry of Commerce or the like);

(c) Government-owned enterprises (public enterprises which may also have been included under (a) or (b) above.

6.61. Besides a single unduplicated frame, it is essential to use an integrated sample design to ensure complete and unduplicated coverage of the large-scale units. Availability of a list-frame permits a single-stage sampling for this sub-sector. However, further stratification of the list-frame is necessary if additional details by economic activities or geographical location are required.

6.62. The population of units in the large-scale segment tends to be very heterogeneous in its size and characteristics. A relatively small number of units often account for a major share of the production or value added of the economy. Inclusion of all such units in the sample is expected to provide estimates of higher efficiency. Therefore, for most establishment surveys, all units above a certain cut-off point are included in the survey, while only a sample is drawn from the rest of the units. The stratum constituted of all such units is referred to as the 'certainty' or 'self-representing' stratum.

6.63. The units falling outside the self-representing stratum within the list-frame segment can gainfully be covered on a sample basis for both the annual and infra-annual surveys. Adopting an integrated sample design for the both kinds of surveys often help resolve the problems of inconsistency between the two sets of estimates obtained from them. Estimates of both annual and infra-annual change parameters as well as level parameters can be obtained using a suitably framed *rotating panel sample design* for the integrated survey. A rotating panel design has a number of advantages over *repeated cross sectional design* (independent samples on different occasions) and *fixed panel sample design*, namely:

(a) It is cost effective and strikes a balance between the conflicting objectives of obtaining reliable estimates of annual and infra-annual estimates.

(b) Level of co-operation of the respondents tends to decline progressively with increasing number of revisits, thereby affecting the quality of response. Sample rotation eases the burden on respondents participating in the survey.

(c) The series of estimates obtained from repeated surveys employing a rotation panel-sampling scheme is usually free from large and unrealistic temporal variations. Moreover, use of rotation sampling permits use of composite estimates that further restricts such temporal variations resulting from sampling error.

(d) This provides the scope of including the new units in survey coverage.

6.64. All units *not* covered in the *'list-frame segment'* fall within the part of the universe described as the *'non-list-frame segment'*. Data collection for this sub-sector requires sampling of area units from an area frame formed from the data collected in the latest economic or population census.

6.65. *Area-frame based survey of the 'non-list frame segment'*. The FIRST methodology of integrated surveys for the *'list-frame segment'* and *'non-list-frame segment'* captures complete data of all economic activities for an economy as a whole in a consistent manner. This requires devising an operational rule to ensure that the units on the business register are excluded from the area frame for *'non-list-frame segment'*. Those establishments whose activities are consolidated in a parent company's accounts have to be deleted from the area sample. This refers, for example, to warehouses or depots operated by trade companies in different parts of the country.

6.66. The FIRST is an establishment-type survey in principle, but, for the *'non-list-frame segment'* uses area sampling techniques. In an area sampling technique of surveying households and establishments, a sample of area units is selected at the first stage. Next, in each of the selected first stage unit, it is required to identify and list all establishments operating in the selected area that are neither included nor linked to any enterprise appearing in the list frame used for the survey of the *'list frame segment'*. The establishments thus identified and falling in the coverage of the survey are then classified by kind-of-activity and a sample of units is drawn from the listed establishments for each kind of activity.

6.67. The group of activities that are given special treatment in this approach is that of the mobile units such as those in trade and some other services activities, which form an important group in most developing countries. This approach permits covering of the enterprises/establishments that are run by the households, even those without fixed premises.

6.68. In this approach, all identifiable establishments outside the owners' home located in the selected area unit as well as household-based enterprises located within home are listed by a house-to-house (structure-to-structure) visit. In addition, the units without any fixed premises of operation like hawkers, street vendors and service providing free-lancers (mobile units) are identified through additional questions put to the households at the listing stage and are listed against the household where the proprietor (or a partner of a partnership concern) resides. This way it is ensured that all establishments in the selected areas that are within the scope of the survey are included in the list which is then used for selection of sample of establishments.

E. Scope and coverage of distributive trade surveys

6.69. *Annual surveys*. All countries, regardless of the development of their statistical system, conduct annual distributive trade surveys. It is recommended that through annual surveys countries endeavour to provide estimates that cover all wholesale and retail trade establishments. This recommendation does not imply that a comprehensive survey is always necessary. Countries may apply one of the following options: (i) the survey may completely enumerate all establishments above a given cut-off point (for example, based on size criteria) and sample the others; (ii) all units may receive a survey form, but an abbreviated version may be used for the small establishments; and (iii) estimates for the small establishments may be made from administrative data or from other statistical inquiries such as mixed household-enterprise surveys.

6.70. *Infra-annual surveys*. The coverage of the infra-annual distributive trade surveys (quarterly or monthly) is necessarily more restricted than that of the annual surveys. Even in countries with a highly developed statistical system, the coverage of small establishments with monthly or quarterly surveys for the production of short-term distributive trade statistics is not feasible. If small establishments are significant in a particularly important distributive trade activity class and there is no reliable administrative data source to cover them, then these units should be included in the coverage of infra-annual surveys, by using the appropriate sampling techniques.

6.71. *Infrequent surveys*. In addition to annual and infra-annual surveys some countries may conduct infrequent surveys of distributive trade units. These surveys are used for collection of data items on specialised topics or in greater details. In general, the use of infrequent benchmark surveys, usually conducted at 5 to 10 year interval, is not appropriate for the purpose of collecting and compiling structural type distributive trade statistics.

6.72. For the countries with significant contribution of unincorporated units to distributive trade activity (the non-list frame segment in figure 4.) it is essential to collect data for these units as well. As explained in the previous sections, the coverage of these units requires conducting surveys based on area-frame sampling, which are resource intensive and time consuming. A benchmark survey normally carried out every 5 years can be used for comprehensive structural type data collection, while similar or fewer data are collected through annual or more frequent inquiries. The benchmark estimates may be projected forward using the estimates of change and growth obtained from annual and infra-annual surveys, either on the non-list frame segment or any other surveys of relevance.

F. Reference period

6.73. *Reference period for annual surveys*. In general, the data compiled in annual surveys should relate to a 12-month period. This 12-month period should preferably be

the calendar year beginning with 1 of January and ending with 31 December. However, where data are more readily available for particular establishments on a different fiscal-year basis, it may be necessary to accept data on that basis. In such instances, it would be desirable that some items of data, such as wages and salaries are collected on both a fiscal-year and calendar-year basis to facilitate building up calendar year aggregates. If a fiscal year different from the calendar year is the normal accounting period for most establishments, the data may be compiled uniformly on a fiscal year rather than a calendar year basis. There are advantages if all establishments can submit returns covering an identical 12-month period, particularly in integrating the annual data with monthly or quarterly data. In many countries, the closing dates of the financial years of companies are spread widely over the year, and statistical offices find it difficult to obtain returns from establishments for a consistent 12-month period. If reporting periods differ in this way, a supplementary table may be prepared in the published report showing the distribution of end-year dates by months, which will help users of the figures to estimate the period over which they are centred.

6.74. *Reference period for infra-annual surveys* Corresponding calendar month/quarter is recommended as the reference period for infra-annual surveys. However, some establishments work in quarterly periods of four, four and five weeks, and in such cases it will be necessary for the statistical offices to make every efforts to standardize the information provided in the monthly returns by some estimation procedures.