STATISTICS IN TRANSITION-new series, Summer 2013
Vol. 14, No. 2, pp. 287–318

COHERENCE AND COMPARABILITY AS CRITERIA OF QUALITY ASSESSMENT IN BUSINESS STATISTICS[1][2]

Andrzej Młodak[3]

ABSTRACT

The problems of coherence and comparability exceed the classical notion of analysis of survey errors, because they do not concern single surveys or variables but the question of how results of two or more surveys can be used together and how relevant data can effectively be compared to obtain a better picture of social and economic phenomena over various aspects, e.g. space or time. This paper discusses characteristics of the main concepts of coherence and comparability as well as a description of differences and similarities between these two notions. Types of coherence and various aspects of perception of these notions in business statistics are analysed. Main sources of lack of coherence and comparability, factors affecting them (e.g. methodology, time, region, etc.) and methods of their measurement in context of information obtained from businesses will be also presented.

Key words: coherence, comparability, mirror statistics, benchmarking, data precision, complex indices.

1. Introduction

Coherence and comparability of data are one of the most important aspects of each survey, especially, if it is regularly repeated over time. It is obvious that properly coherent and consistent methodology guarantees efficient long–time

[1] The paper was prepared for presentation at the Conference *Statistics–Knowledge–Development* organized to celebrate the International Year of Statistics in Łódź, Poland, 17th – 18th October 2013.

2 The paper is based on the output of the ESSnet project Methodology for Modern Business Statistics – MeMoBuSt (Specific Grant Agreements No. 61001.2010.006-2010.702 and 61001.2010.006-2012.273). The author expresses his gratitude to Mrs. Judit Vigh (Hungarian Central Statistical Office) as well as to two anonymous reviewers for valuable comments and suggestions.

3 Address: Statistical Office in Poznań, Small Area Statistics Centre, Branch in Kalisz, pl. J. Kilińskiego 13, 62–800 Kalisz, Poland. E-mail: a.mlodak@stat.gov.pl.

analysis combining many variables and taking into account various cross–sections, divisions or clusters of investigated objects (economic entities, spatial areas, branch groups of enterprises, etc.). In modern reality, where many globalization processes based on highly developed strict interregional and international economic connections are observed, a collection and use of multi–aspect coherent and comparable statistical data is one of key priorities. They enable efficient monitoring of occurring changes and reacting when some problems or threats are recognized.

The role of coherence and comparability of data in European statistics was appreciated by the Statistical Office for the European Union (Eurostat), which in its Code of Practice formulated these requirements as the Principle 14, expressed as follows (Eurostat (2005)):

European statistics should be consistent internally, over time and comparable between regions and countries; it should be possible to combine and make joint use of related data from different sources.

Indicators

- *Statistics are internally coherent and consistent (e.g. arithmetic and accounting identities observed).*
- *Statistics are coherent or reconcilable over a reasonable period of time.*
- *Statistics are compiled on the basis of common standards with respect to scope, definitions, units and classifications in the different surveys and sources.*
- *Statistics from the different surveys and sources are compared and reconciled.*
- *Cross–national comparability of the data is ensured through periodical exchanges between the European statistical system and other statistical systems; methodological studies are carried out in close cooperation between* the Member States and Eurostat.

This paper discusses and analyses the basic characteristics of the main concepts of coherence and comparability as well as a description of differences and similarities between these two notions. It is worth noting that "coherence" has much more wider sense than 'comparability'. Particular types of coherence and various aspects of perception of these notions in business statistics will be analysed. Main sources of lack of coherence and methods of its assessment in the context of information obtained from businesses will be analyzed. Similarly, problems concerning comparability will be characterized indicating possible reasons of its lack and factors affecting it, such as time, region, national accounts and economic benchmark as well as complex methods of measurement of comparability level.

The paper consists of two main sections devoted to coherence and comparability. The first of them (Section 2) discusses methodological fundaments of concepts of coherence in the context of accuracy and recognition of incoherence. Types of coherence taking into account the character of statistics and sources of incoherence as well as modelling of assessment of coherence using

various factors having an impact on it are here also presented. Section 3 concerns the comparability and shows its relation to the coherence, methodological and empirical sources of incomparability, basic directions of appraisal of the level of this property. At the end (Section 4), some indices which can measure coherence and comparability in a complex way, characteristics of accuracy and related problems are proposed. In Section 5 some conclusions are formulated.

2. Coherence

To obtain business statistics of satisfactory quality it is necessary to guarantee coherence between relevant data. *Coherence* is a general term referring to the consistency between a set of statistical variables describing finite population parameters in terms of various but – from the methodological point of view – mutually connected social or economic phenomena observed in business reality. More precisely, the level of coherence informs us whether and to what degree some statistics can be analysed jointly and how to indicate their 'optimum' levels. That is, if two variables are strictly methodologically connected and changes of one variable affect values of the other (and vice versa) then they can and should be analysed jointly. For example, such a joint analysis may concern production output and employment, production output and foreign trade turnover, sold production and wages and salaries, etc. These variables may also be combined to obtain one synthetic result.

2.1. Definition of coherence

According to Eurostat (2009), **the coherence of two or more statistical outputs refers to the degree to which the statistical processes by which they were generated relied on the same concepts – classifications, definitions, and target populations – and harmonised methods**. If statistical variables are coherent, it means that they have the potential to be validly combined and used jointly. As it was noted earlier, examples of such joint use reflect the situation where statistical outputs refer to the same level of aggregation forms (population, reference period, territorial level etc.) but concern different sets of data items (e.g. data on wages and salaries and production). Coherence can also occur in the opposite situation, i.e. where the investigated variables comprise the same data items (e.g. employment data) but are collected for different reference periods, regions or other domains.

Various data are usually collected using different processes (e.g. different reporting forms are used for the survey of employment and production output for medium and large economic entities – in addition, the former is conducted monthly and the latter quarterly, hence, in this situation one can look for coherence only for quarters). According to Eurostat (2009), the term coherence is usually used when assessing the extent to which outputs from different statistical processes have the potential to be reliably used in combination. More precisely,

coherence concerns the possibility of combining the two aforementioned variables for the same population and time period. Coherence depends on statistical processes leading to a given output and its level is assessed on the basis of final results in terms of quality of these processes.

Another aspect of coherence is **statistical accuracy**. In order to use several sets of variables or one large set of various variables it is very important to indicate which ingredients of the definition of these variables are constant and which vary and to what extent. For example, it is desirable to know what methodological rules were used to collect statistics about production output, the number of employees or turnover in various surveys and how they can be compared across various geographical areas or types of enterprises. The precision and accuracy of collected data can be a significant factor. For example, if one of several successive surveys contains missing values about turnover for many units, it can be dropped from the study (provided, of course, this action does not affect the resulting analysis).

It is worth noting that **there is a significant difference between coherence and accuracy**. Namely, if results of two different processes or of the same process at different aggregation levels are compared, differences between them are an effect of inconsistency of relevant estimates due to their various precision. This is an example of inaccuracy, which is different from incoherence, since it does not involve analyzing the possibility of combining these processes. In general (cf. Eurostat (2009)), coherence/comparability refers to descriptive (design) metadata (i.e., concepts and methods) about the processes, whereas accuracy is measured and assessed in terms of operational metadata (sampling rates, data capture error rates, etc.) associated with the actual operations that produced the data. Accuracy also concerns the problem of differences between various estimates of target variables and their errors (especially if error profiles are incomplete or unknown). The authors of the document by Eurostat (2009) note that where error profiles of the statistical processes are known and included within the description of accuracy there is no need for further reference to them under coherence (and also comparability). They give an example where it is supposed that sampling error bounds are published for two values of the same data item for adjacent time periods indicating the range within which a movement from one period to the next may be due to chance alone and does not reflect any actual change in the phenomenon being measured. If and only if the measured movement is larger than these bounds, one can discuss whether the movement is real or due to the lack of coherence (and comparability). On the other hand, the assessment of coherence should include elements not analysed in terms of accuracy, e.g. on non-response or non-sampling errors if they were not taken into account when assessing accuracy.

Körner and Puch (2011) discuss Eurostat's attempts to define coherence and note that two sources providing non-deviating results are therefore not necessarily coherent (as various effects in the underlying processes might mutually compensate each other). Therefore, it is easier to recognize incoherence than

coherence. Their operational definition starts from such deviations of results of two different processes and takes the expectations of users regarding consistency of results into account. It assumes that statistical outputs referring to identical concepts have to be numerically consistent. That is, two related surveys are coherent if the value of some basic variable observed in both surveys does not deviate (given that identical concepts are being used) from that expected by the user. For instance, it can be the number of enterprises in monthly employment surveys and quarterly reports on financial results).

The authors of the document by Eurostat (2009) view the relationship between coherence/comparability and accuracy by noting that the numeric consistency of estimates depends on two factors:

- logical consistency (called here coherence/comparability) of the processes that generated those estimates; and
- errors that actually occurred in those processes in generating the estimates,

and conclude that coherence (and comparability) is a prerequisite for numerical consistency. The degree of coherence/comparability determines the potential for numerical consistency (but does not guarantee it, since it also depends on errors).

2.2. Types of coherence

In terms of business statistics, two types of coherence are usually distinguished:

- ***internal coherence*** – coherence within a uniform set of annual and short–term business statistics or between data derived from different sources (surveys, administrative INTRASTAT, etc.);
- ***external coherence*** – consistency between business statistics and main macroeconomic indicators, e.g. with national accounts, statistics on prices and wages, external trade, etc.

In the first case, there are many possible causes of divergent trends between short-term indices and annual business statistics. Davies (2000) and Bergdahl *et al.* (2001) analyze several aspects in terms of which coherence should be perceived.

First, it is important how a variable is defined. A typical definition includes the main features of the variable to be collected such as statistical measure (total, mean, median, etc.), unit of measurement (e.g. different units are used to measure production output, number of hours worked, number of employees, wages and salaries, etc.), unit of observation (enterprise, KAU, LKAU, local units, etc.), domain (definition of subpopulation – e.g. using the NACE classification – or classes, e.g. by number of employees) and reference times (a time point or period which units and variable values relate to).

The problem of coherence with respect to these aspects is equivalent to the comparability of data over such factors. That is, one expects comparability over time, between countries, between non–geographical groups or functional areas

and between statistics from several surveys. These issues will be described broadly in Section 3. Now it is enough to note that in practice reference times are mainly time intervals, such as calendar year, quarter, or month, sometimes they are also points in time, e.g. 1st January of the reference year. The reference time should be the same for all units and variables. Usually reference times agree for all variables and units in a FPP[1]. For monthly statistics, for instance, this means that the delineation of units should refer to the current month.

The second type of coherence in business statistics (i.e. external coherence) can be perceived in terms of logical relationships between target statistics. For example, when both monthly and annual statistics on expenditures are used, the sum of twelve monthly values should sum up to the annual total.

The authors of the document by Eurostat (2009) also indicate several other types of coherence related strictly to business statistics:

- ***coherence between short-term (i.e. sub-annual) and annual statistics***: a good example are monthly and annual production data for the same industries in the same region
- ***coherence with the National Accounts*** – underlines the key importance of coherence for economic surveys using the national accounts; the National Accounts compilation process will detect the possible lack of coherence.
- ***coherence with other statistics***: for example, coherence between employment produced by a labour force survey of household members and the number of employees produced by an economic survey of enterprises.

Bergdahl *et al.* (2001) noted that difficulties encountered by the user often depend strongly on the 'distance' between statistics used jointly (since definitions of variables used even in the same survey can vary in terms of reference times – a period in one case, a point in time in another). Moreover, reference times may correspond to that of the sample frame or the variable and definitions of enterprises can vary in different countries across the EU.

2.3. Sources of incoherence

Körner and Puch (2011) note that deviating results have two main sources:

- ***differences due to concepts*** (like the target population, the reference period, or the definition of the items for analysis) and
- ***differences due to methods used*** (e.g. data collection methods and procedures, the data processing approach, or the sampling design).

They investigate the German Labour Force Survey and National Accounts by identifying and quantifying definitional differences and then recognizing methodological differences. Methodological differences can cover various

1 FPP – Finite Population Parameters, a general notion covering basic descriptive statistics for finite population which distinguish it from characteristics of the infinite population in statistical modelling, e.g. population total and average of a given variable, the ratio of the population averages of two variables, the population variance or the population median of a given variable, etc. (cf. Bergdahl *et al.* (2001)).

aspects, such as all elements of survey design and errors, but also the accounting rules and estimation methodology. If methodological differences are significant, then methods-related differences are usually not easily identified and some additional algorithms are necessary.

Sources of incoherence can also be combined. That is, errors of one type are usually to some extent inherent in the second. For example (Körner and Puch (2011)), the sampling error is a combination of effects due to the sampling design and deviations from it in practical implementation. Similarly, the survey mode, albeit part of the methodology, will directly influence the results which can be obtained.

Differences between short–term and annual statistics result from frames and various units used in the sources. Also, a lot depends on the choice of the domain for unit aggregation (e.g. in surveying the same statistics (variables), different data can be obtained depending on whether units are grouped by the type of activity or by location). It is worth noting that if short–term changes over time – i.e. the realization and trend of time series – are analysed, coherence can only refer to intra–annual results. Körner and Puch (2011) note that coherence of time series will therefore be covered in dedicated sections, focusing on aspects specific to the shorter (e.g. monthly and quarterly) time series.

Because coherence and comparability have usually similar reasons, their various aspects will be described in detail in the part of this paper devoted to comparability (i.e. Section 3). This subsection was only an overview of the general nature of these problems.

2.4. Measurement of coherence

The measurement of coherence (and more precisely, incoherence) is sophisticated. Of course, the simplest way involves computing differences between results of analyzed surveys or data sources. One should remember, however, that not all aspects of coherence are measurable and numerical assessment may be the result of a number of factors, such as properties of concepts and methods used. Cook (2007) notes that *coherence in economic statistics is increasingly dependent on integration of information obtained from individual large enterprises from different sources. Making balancing adjustments at industry or industry group level is of diminishing effectiveness as economic activity is concentrated in fewer enterprises, where coverage rates differ across surveys. In some industries, such as pension funds, the scale of inter-company transfers is so large that it makes it difficult to estimate industry change in an unbiased manner. Coherence is improved by survey design, and most countries have a long way to go, in doing as well as they could. Much improvement is still possible in the way that statistical frames are adopted in individual surveys*.

It is, therefore, very important to be aware of how these aspects affect inconsistencies. A good example of such quality assessment is the study

conducted by Casciano *et al.* (2012), who analyse coherence by studying the difference in the final estimate for turnover and value added. They compare the estimation of turnover obtained solely on the basis of the survey with that based on the survey combined with data from administrative sources. The new estimate is very close to the initial one with a percentage difference of +0.03%, although there is a high variability in results when a breakdown by economic activities and size classes is used. Other factors that have an impact on the assessment are weighting and non-response (the difference due to unit non-response is higher in construction activities and lower in service sector), which has produced a higher estimate of turnover for micro and small firms and a strong underestimate for medium enterprises especially in service sectors.

The measure of overall inaccuracy should concern all data sources, not only the sample. One should also expect significant (i.e. related to methodology or non-response) and non-significant (generated by, e.g. econometrically modelled 'white noise', i.e. random disturbances) inaccuracies. One can expect that relationships and connections between parameters will also be reflected in their estimates. For example, the sum of monthly estimates of the number of employees should be equal to the relevant annual estimate. Of course, in some situations (e.g. in the case of indices), consistency can only be approximate, but such approximations should be as precise as possible. Significant differences in estimates for the same parameter imply incoherence. One should, however, bear in mind that such problems can result not only from definitions but also from systematic errors. Therefore, sources of these problems should always be carefully analyzed and recognized.

P. Davies (2000) provides a comparison between short-term statistics, annual statistics, and national accounts in Sweden. In his example short-term statistics measures own production output, whereas the other two sets of statistics estimate the value added. He observed that the growth rate between different years was unusually large, but he also noticed that if growth rates for output and value added were assumed to be equal/similar, some short-term statistics for output could be regarded as a first estimate of the value added.

3. Comparability

In many professional papers the terms "coherence" and "comparability" are used interchangeably. Such usage can be justified by the fact that comparability is a special case of coherence and involves situations where statistical outputs refer to the same data items and the aim of combining them is to make comparisons over time or across domains. However, the idea of comparability has its special features, and therefore deserves a broader description and analysis. Therefore, its definition will be presented along with highlighting how it differs from the main concepts of coherence and discussing some methods that enable us to assess and improve the comparability of statistical outputs.

3.1. Definition of comparability and its relation to the problem of coherence

The authors of the document by Eurostat (2009) note that unlike coherence, comparability is used when assessing the extent to which outputs from (nominally) the same statistical process but for different time periods and/or for different regions have the potential to be reliably used in combination (recall that coherence concerns the possibility of combining different variables for the same region or time period). Therefore, coherence is a stronger notion. More precisely, if several different time periods and the same population and region are analysed then the validity of the combined use of, e.g. employment data for this structure will be perceived in the context of comparability, whereas combining this variable with, e.g. data on wages and salaries for the same reference points will be considered in terms of coherence.

Hence, the comparability refers to the possibility of making efficient comparisons between units and variables according to various aspects and criteria. One most frequently used basis of such comparisons is time. To make such comparisons possible, the stability of definitions in successive survey administrations should be ensured and the 'state of the art' in all time periods should be well described. To do so, one should remember that when a change is made, special measures are recommended to improve comparability, for example, producing statistics in both ways on one occasion or even re–estimating part of the old series in terms of new definitions (according to Bergdahl *et al.* (2001)).

3.2. Sources of incomparability

The document by Eurostat (2003) provides the following list of sources of the lack of comparability, which can be the basis for designing each quality analysis and quality report:

Concepts
- Statistical characteristics
- Statistical measure (indicator)
- Statistical unit
- Target population
- Frame population
- Reference period and frequency
- Study domains
- Geographical coverage (for comparability over time)
- Standards
- Structure effects
- Conceptual aspects specific for a domain under study

Measurement Quality Dimensions
- Sample design
- Data collection

- Data processing
- Estimation

Measurement aspects specific for a domain under study (these include characteristics pertinent to any specific domain, e.g. thresholds for Foreign Trade Statistics).

As one can see, the possible reasons are numerous, but the authors of the next version of this document (Eurostat (2009)) describe these problems in a much more consolidated form. They argue that possible reasons for the lack of comparability between outputs of statistical processes may be summarised under two broad headings used already to present causes of incoherence, i.e. – differences in concepts and differences in methods. This is the starting point for introducing more detailed classification. First, circumstances connected with concepts which affect comparability will be described.

Target population – units and coverage. Problems with comparability may result from the fact that target populations could differ for two statistical processes, or for the same process over time, in a variety of different ways. These differences can also concern the spatial aspect of definitions of a given notion or the lack of its harmonization. For example, many EU countries use a different definition of economically active population in Labour Force Survey (LFS). This can result from various legislations concerning the retirement age (for example, in Poland it is 60 years for women and 65 years for men, but in France the retirement age was established at 60 for both genders). In the USA persons waiting to start a new job are classified as employed, but in EU LFS – as unemployed. On the other hand, even in one country this definition can differ over time. For example, many countries are planning to extend the age for retirement – e.g. in Poland to 67 and in France to 62 for both genders. This fact will result in changes of the definition of the target population and the comparability of future statistics over time. Eurostat (2009) also provides another example – monthly statistics of industry might only include manufacturing enterprises, whereas another statistical output with the same name might include manufacturing and electricity, gas and water production. Two related but different surveys may use various units to be investigated – one can use an enterprise whereas another – local unit (LAU) or kind of activity unit (LKAU).

Geographic coverage – some spatial areas may be included in the survey of one country and excluded in advance from another. For example, wages and salaries in manufacturing can be investigated only in urban areas (where this kind of economic activity is mostly concentrated) or also in rural areas. Moreover, the internal structure of sampled territorial areas in one country may be different from those in another one. As a result, the coverage of functional areas will also be different.

Reference period – time points of measurement of a given feature within a given survey may differ. For example (cf. Eurostat (2009) in a survey of employees an enterprise might be asked for the number of full-time employees on

the third Monday in the month or on the first day of the month; an annual survey may refer to a fiscal year beginning in March, another to a calendar year.

Data item definitions, classifications – it is a very important cause of the lack of comparability. A variety of non-harmonised methodological definitions are in use. A good example are differences between countries, which seem to be unavoidable (e.g. in terms of accounting systems, definitions of units and variables, etc.). In the common EU methodology (e.g. expressed in URBAN AUDIT project) the concepts of "employees" and "employed persons" are not really distinguished. However, in Polish official statistics each of these terms has a different meaning. Namely, employees are defined as persons performing work and receiving earnings or income, i.e.:

- employees hired on the basis of employment contracts (labour contract, postings, appointment or election),
- employers and own account workers,
- outworkers,
- agents (including contributing family workers and persons employed by agents),
- members of agricultural co–operatives, (agricultural producers' co-operatives, other co-operatives engaged in agricultural production and agricultural farmers' co–operatives),
- clergy fulfilling their obligations.

Data about employees are presented without converting part-time employees into full-time employees and applying the principle that each person is listed once according to their main job. Instead, the number of employed persons is defined as the sum of the number of full-time employees and the number of part–time employees. Full–time employees are persons employed on a full–time basis, as defined by a given company or for a given position, as well as persons who, in accordance with regulations, work a shortened work-time period, e.g. due to hazardous conditions or a longer work-time period, e.g. property caretakers. Part-time employees are persons who, in accordance with labour contracts, regularly work on a part-time basis. In the metodology adopted in URBAN AUDIT, on the other hand, the category of 'employed persons' includes only persons performing work on the basis of a relevant contract – that is, it does not include employers, for example.

Eurostat (2009) also provides a definition of unemployed person in LFS, which includes any economically active person who does not work, is actively looking for a job and is available for employment during the survey or any economically active person who does not work, is actively looking for a job and is or will be available for employment in the period of up to two weeks after the survey's reference week. A well–known reason for incomparability are changes in classification schemes or in particular revisions in accordance with new versions of international standards. For instance, the current version of the NACE classification can have significant modifications in relation to its older own

outputs or variants used in particular countries (e.g. the Polish Classification of Activity is based on NACE but is more detailed at lowest levels and takes into account types of economic activity specific for this country). On the other hand (cf. Eurostat (2009)), even without a change in classification, the procedures for assigning classification codes may be different or change over time, for example with improved training of staff or the introduction of an automated or computer assisted schemes.

The next part of our classification will concern the impact of methods of data collection, processing and analysis on comparability.

Frame population – within the same survey based on a generally established target population, the actual coverage of a survey may depend on the frame used. That is, e.g. employment and wages and salaries in enterprises with fewer than 5 employees are surveyed only once a year and it is only a sample survey, whereas for larger entities such a survey is conducted monthly and in general (maybe with some small exceptions) it is exhaustive. Problems may also occur when frames are based on various administrative sources, e.g. one is constructed using the tax register and another one – using the business register. Apart from the difference resulting from various definitions of units to be registered in these databases, one can also consider possible changes in registration (which may occur in relation to one or both sources) and survey designs (one can be, e.g. cross sectional and the second – longitudinal, which produces significant difference in estimates). One should also account for possible changes of panels or rotation patterns or training or application of new methods (e.g. for creation, amalgamation, clustering, allocation, etc.) occurring over time or between various countries.

Source(s) of data and sample design – this problem concerns a situation when in one statistical structural survey data are obtained from an administrative data source and in another – from a direct survey, or when actual sample designs are different. A good example may be financial data for large and medium enterprises, which can be obtained either from the tax register or from the current reporting.

Data collection, capture and editing – the same survey can differ in terms of the incidence of non–response problems and possibilities of their reduction. These issues are described in details, e.g. by Yancheva and Iskrova (2011). At this point, let us just give one example: in one case the non-response rate may amount to 10% and in another (e.g. in another round of the same survey) – to 40%. This fact may result either from some changes in the frame, sample or questionnaire or from the application of more efficient methods or reduction of burdens.

Imputation and estimation – methods of imputation and estimation may differ depending on the place or time when a given survey is conducted. For example (cf. Eurostat (2009)), in one survey zeroes may be imputed for missing financial items, whereas in another survey non-zero values may be imputed based on a 'nearest neighbouring' record. Likewise, in dealing with missing records in an enterprise survey there are various options, such as assuming that the

corresponding enterprises are non-operational or assuming that they are similar to enterprises that have responded. However, if the number of missing data is relatively large, the hot–deck (such as nearest neighbour, for example) methods of imputation are recommended (cf. de Waal *et al.* (2011)).

To illustrate the problem of incomparability, an example based on empirical experience of Polish business statistics is given. During processing data on economic activity of entities, some important data are generalized on the level of group for some NACE Rev. 2 sections, others on the level of division. Data are generalized also on the ownership sectors for NACE Rev. 2 sections. In Poland the obligation of transmission of reports on economic activity pertains to all economic entities belonging to established NACE Rev 2. sections (they cover mainly manufacturing units) having more than 49 employees and sampled entities employing up to 49 persons. To generalize collected data, the number of employees is used. That is, for entities which have transmitted the report this number if taken from these reports (and is called the variable number of employees). On the other hand, the sampling frame contains data on number of employees collected from business register or earlier reports (called the variable number of employees). Thus, to obtain the estimates of various values for a population the following generalization coefficient is used

$$u_d \stackrel{\text{def}}{=} \frac{\sum_{j\in N_{dr}} \vartheta_j + \sum_{j\in N_d \setminus N_{dr}} \theta_j}{\sum_{j\in N_{dr}} \vartheta_j}$$

for a level of aggregation d, where N_d is the set of all units at the level d and N_{dr} is the set of units which have transmitted their reports belonging to the level d, ϑ_j is the variable number of employees of the j–th unit and θ_j – fixed numer of employees for j–th unit.

On the basis of this coefficient various quantities, such as revenues firm sale of products and services turnover in current process, gross wages and salaries, retail sale, etc., are estimated using the following general formula (with small variants in some particular cases):

$$\widehat{\varphi}_{(d)} \stackrel{\text{def}}{=} u_d \sum_{i\in N_{dr}} \varphi_i$$

for every level of aggregation d, where φ_i is the value of interest for i-th unit. Thus, two important problems concerning coherence and comparability occur:

a) the number of employees in register can be significantly different than the actual number of employees of a given entity,
b) the generalization coefficient and, in consequence, estimate depend on the level of aggregation. Thus, the sums of relevant estimates for, e.g. NUTS 4 units belonging to the NUTS 2 unit obtained on the basis of generalization for NACE groups can be significantly different from relevant estimates obtained when generalization is made at the level of sectors of ownership.

The first of these inconveniences can have an important impact on the quality of estimation due to an error occurring when unknown actual number of employees for those units which have not transmitted the report is replaced with the number of employees based on the business register. However, some average impact of such bias can be estimated as

$$v_d \stackrel{\text{def}}{=} \left| \frac{1}{\sum_{j \in N_{dr}} \vartheta_j} - \frac{1}{\sum_{j \in N_{dr}} \theta_j} \right| \frac{\sum_{j \in N_d \setminus N_{dr}} \theta_j}{u_d}$$

and expressed in percents. Summing (or averaging) these indices over domains d within a given larger domain one can estimate a level of incoherence/ incomparability in this case.

To reduce the second problem an imputation at the level of units is recommended. The aforementioned index can be useful to identify areas where the imputation is most necessary. Now, in Polish statistics some works on this were undertaken.

3.3. Types of comparability

It should be noted that the Handbook issued by Eurostat (2009) mentions the following types of coherence/comparability that are to be included in a reliable quality report:

- ***comparability over time*** – data should show whether collected information for a given region in several different time point was gathered under the same circumstances in terms of definitions, population, etc.
- ***comparability over region*** – for example, data for the same month from the structural business survey conducted in two Member States, indicate which regions are covered in both surveys and which not and why;
- ***comparability over other domains*** – domains over which comparisons are often made include economic activity group, occupational group, and sex. An example would be annual structural data for agriculture with annual structural data for manufacturing collected by a different survey.
- ***internal comparability*** – referring to data produced by a (single) statistical process (but possibly comprising several different segments) for a single time period and region.
- ***comparability between short-term (sub-annual) and annual statistics*** – for example monthly and annual production data for the same industries in the same region.
- ***comparability with the National Accounts*** – for economic surveys that feed into the national accounts, coherence is vital and, in so far as it is lacking, the National Accounts compilation process will detect it.
- ***comparability with other statistics*** – for example, coherence between employment produced by a labour force survey of members of households and numbers of employees produced by an economic survey of enterprises.

3.4. Assessment of comparability

An efficient assessment of comparability requires the knowledge of most useful methods including characteristic features of definitions and design. Therefore, the current type of incomparability according to the typology presented in Section 3.3 has to be recognized. This presentation of methods for assessing and reporting comparability will start from the general approach and next some problems will be addressed by referring to the aforementioned typology in the context of constructing quality reports.

The **General Approach** is intended to explain causes of the lack of coherence/comparability. Namely, in quality reports effects of the main sources of incomparability should be recognized and described. Their impact on estimates has to be determined. An attempt to reconcile these estimates on the basis of such knowledge is also required. The authors of the document by Eurostat (2009) emphasize that any general changes that have occurred, which may have an impact on comparability, should be reported, for example changes in legislation affecting data sources or definitions, reengineering or continual improvement of statistical processes, changes in operations resulting from reductions or increases in processing budget, deviations from relevant ESS legislation and other international standards, etc. Although in many cases comparability is equivalent to coherence, there are some situations when such equivalence does not hold. That is, even if coherent statistical data describes the same phenomenon, this information can be incomparable due to the occurrence of some errors.

A proper comparability analysis should contain a presentation of concepts and methods, explanation of what causes a given problem and each possible source of problems concerning comparability should be separately described. According to Eurostat (2009), the first step is to conduct a systematic assessment of possible reasons for the lack of comparability, based primarily on the examination of key metadata elements, and identification and analysis of differences. This action can enable us to make a picture concerning the magnitude of any lack of comparability. The second step consists in predicting the likely effect of such a difference on statistical outputs. The final step is to aggregate and summarise in some way the total possible effect, in other words, to form an impression of the degree of (lack of) comparability.

A different question is how to detect possible problems in methodology or response when only ‘raw’ data with many reports and possibilities to aggregate them in given categories (e.g. time, space, by unit incomes, by the number of employees, etc.) are at disposal. The **quality control of variables** should provide the first indication of such problems. One can use general rules proposed in the document by Eurostat (2008).

Multi–variate type of control – checking if calculated sums of different variables and their values coincide with the total provided. For example, one should verify whether the sum of the number of entities by the number of

employee groups corresponds to the total number of units. If not, the cause of this fact has to be looked for.

Hierarchy type of control – it consists in comparing sums of the same variable against the variable value provided at an upper spatial level. For example, it is checked whether in any case the sum of the number of units in NUTS 5 areas included in a given NUTS 4 area corresponds to the number provided at the relevant NUTS 5 level. To allow rounded numbers and estimates, sometimes a tolerance or deviation of 3% of the checked value from the control value is set. This type of control refers to the same variable but to different spatial levels.

In practice, many variables have the form of indices. In other words, they are obtained as relevant ratios of given data expressed in absolute values (e.g. average wage and salary per employee is the ratio of the total sum of paid wages and salaries and the total number of employees). The 'sum control' is inappropriate in this case and the extreme values have to be looked for instead. The authors of the document by Eurostat (2008) propose three ways of **detection of outliers** which are usually 'suspected' to be incredible:

Classical interval of variation – its construction is based on the average and the standard deviation of the indicator values over a specified population. The control range is determined by the interval $\bar{x} \pm z \cdot \sigma_x$, where $\bar{x}$ is the arithmetic mean and σ_x – the standard deviation of the index X. The control consists in checking whether the individual indicator value is out of the control range. The value of z is set by the user (more often it is set as $z = 2$). In extreme cases, when the indicator cannot have a negative value, z is set to the ratio of average and standard deviation.

Median interval of variation – a construction of an interval of variation based on the median. That is, the control range is determined by the interval $\mathrm{med}(X) \pm z \cdot \mathrm{mad}(X)$ where $\mathrm{med}(X)$ is the median of X and $\mathrm{mad}(X) = \mathrm{med}_{i=1,2,\dots,n}|x_i - \mathrm{med}(X)|$ (x_i is the i-th value of X, $i = 1,2,\dots,n$, n – number of observations) – its median absolute deviation. The parameter z is established as previously.

Growth rates – this method makes use of growth rates of the dataset over various periods. It is an extension of the method based on the classical interval of variation and takes into account gaps between the analysed years. The threshold limits are usually specified as the maximum allowed growth rates per year. If there are missing time points between the observed ones, this has to be taken into account. For example, if monthly data on employment are analyzed, months where the growth rate exceeds the arbitrarily accepted tolerance interval (e.g. +/- 20%) are indicated. For missing time point the limits should be adjusted respectively.

Some aspects of comparability quality will be recalled now by referring to its typology introduced in subsection 3.3.

Comparability over time is very important if statistical outputs are published at a number of consecutive time points. Hence, changes over time of economic or social phenomena can be observed and analysed and ensuring comparability over

time seems to be crucial for such an analysis to be efficient. The user (when using time series provided by the respondent or the statistical office or constructed by him-/herself on the basis of available 'raw' data) should obtain information about possible limitations and problems in data use concerning comparability with respect of time. This information also has to be included in the quality report. According to Eurostat (2009), in assessing comparability over time the first step is to determine (from the metadata) the extent of changes in the underlying statistical process that have occurred from one period to the next. There are three broad possibilities:

1) There have been no changes, in which case this should be reported;
2) There have been some changes but not enough to warrant the designation of a break in series;
3) There have been sufficient changes to warrant the designation of a break in series.

If changes result in negligible effects on statistical outputs, then it is sufficient to make a relevant note in the metadata. However, if the effect is significant, two cases should be considered. If an effect is too small to warrant a series break, then (cf. Eurostat (2009)) the NSI may wedge in the changes to the outputs over a period of time so that, between any two periods, the adjustment being made to move from old to new values is less than the sampling error and thus cannot by itself be detected and interpreted as a real change. In the second case, i.e. if changes are significant and sufficiently large to cause the break, the user should be precisely informed about this circumstance, their location and consequences. Moreover, the authors of the document by Eurostat (2009) recommend three possible ways to handle these inconveniences:

- The most comprehensive treatment is ***to carry forward both series for a period of time and/or to backcast the series***, i.e., to convert the old series to what it would have been with the new approach by duplicating the measurement in one time period using the original and the revised definitions/methods.
- A less expensive treatment is ***to provide the users with transition adjustment factors giving them the means of dealing with the break***, for example by doing their own backcasting.
- The least expensive treatment is t***o simply describe the changes that have occurred and provide only qualitative assessments of their probable impact upon the estimates***. Obviously, this is the least satisfactory from the user's perspective.

Comparability over time is especially important for short-term business statistics, which can have different priorities and are usually more sensitive to any changes in methodology and the current situation than, e.g. annual data. Given comparable data, users have more possibilities of adjusting relevant data to their individual preferences and priorities. For example, revenue or sold production of enterprises need to be analysed both in the current period and over a longer time.

Comparability in time enables the user to make a better assessment of the situation, the effect of current business strategies and future prospects. Therefore, users can also be interested in much more advanced tools such as trend separation, regular seasonal variations or random and non-random disturbances. Thus, time series modelling should also strongly rely on comparability over time.

Comparability over region is an aspect which could be assessed in two different ways: pairwise comparisons of metadata across regions or comparison of metadata for the region with a standard (such standard can be perceived in terms of norms valid in ESS or best practices of statistical institutes). According to Eurostat (2009), two broad categories of situation can be identified:

- where essentially the same statistical processes are used, e.g. a labour force survey designed in accordance with ESS standard, and differences across regions are expected to be quite small; and
- where a different sort of statistical process is used, for example a direct survey in one case and a register based survey in another. In such cases, differences are likely to be more profound.

A complex measure of differences could be constructed as a sum of partial absolute differences quantified by a scoring system. Of course, the simplest solution in this context is to recognize the key metadata elements for which a difference occurs and use the binary code: no difference or difference. If the aim is to categorize the levels of discrepancies, several categories, e.g. from 5 to 0 by 1, i.e. from the most essential difference (5) to no difference (0), have to be assumed. The intensity of such difference could be quantified using arbitrarily assumed classification based, e.g. on the range of values of a given variable or experts' opinions. Apart from this, one should also consider the possible effect of such differences. This differs from analyzing 'pure' differences because, for example, in some circumstances (e.g. for ratios) even significant 'pure' differences could generate negligible effects in terms of comparability. Therefore, assigning a weight to each metadata element where the difference occurs according to its potential effect on comparability and computing a weighted score across all metadata elements is desirable. Such assessment can be summarised not only within a given country but over all countries within ESS.

One can say that owing to the intensive integration of economies of countries across the world, precise interaction analyses are particularly desirable. They help to recognize the position of a particular company in a given country in the context of other countries, especially those affiliated in such prestigious international economic organizations as OECD, EU, EEA, EFTA, CEFTA, G–8, NAFTA, ASEAN, OPEC, etc. The variety of producers of statistical data, economic practices, legal regulations (e.g. in terms of conducting economic activity or taxes), methodologies used to collect and process statistical data result in many serious difficulties in obtaining internationally comparable data. In recent years many projects designed to improve international data comparability have been initiated. One should mention the so-called 2007 Operation, which has resulted in the harmonization of major economic classifications and nomenclatures used in

the EU with those used in other parts of the world. Some attempts are also undertaken to harmonize the programme of statistical monitoring of urban areas URBAN AUDIT (e.g. in terms of the scope of the functional impact of cities). M. Bergdahl *et al.* (2001) discuss other harmonization actions, e.g. in terms of NACE and PRODCOM applied in Germany and UK.

Comparability over other domains – areas or time periods are, of course, not the only domains, over which comparisons can be performed. Usually, official statistical and various statistical studies present breakdowns by, e.g. size groups of enterprises, economic activity group, occupational group, sex, education of employees, etc. These aggregates can be treated similarly to areas and thus methods for assessing comparability are usually analogous to those mentioned earlier. One should, however, consider the possible differences between various statistical tools.

Internal coherence/comparability – users of statistical data expect, which is obvious and logical, that each set of published outputs should be internally coherent, i.e. all the appropriate arithmetic and logical dependencies should be observed. However, this requirement is sometimes difficult to satisfy. For example, some efficient estimation methods have a drawback or the implemented process comprises more than one segment with data from different sources or for different units in each segment. The authors of the document by Eurostat (2009) suggest giving a brief explanation to users and mentioning these problems in a quality report, with the reasons for publishing non-coherent results explained.

Coherence/comparability between short-term (sub–annual) and annual statistics – another natural expectation on the part of users is that sub-annual and annual statistical outputs are consistent. On the one hand, it means that similar arithmetic dependencies should occur (sum of gross monthly revenue for economic entities should be equal to its annual revenue or the number of newly employed persons in a given month should not be greater than its total in the relevant quarter) and, on the other hand, statistical processes producing these data are often quite different. Thus, all causes of the lack of coherence need to be assessed and accounted for. Eurostat (2009) suggests that a comparison of sub-annual and annual estimates should be the starting point for assessing the magnitude of differences due to the lack of coherence:

- if both annual and sub-annual series measure the same phenomenon in absolute values or indices (other than growth rates), annual aggregates can be constructed from sub-annual estimates and compared to totals from the annual series;
- if one or other of the series produces only growth rates, then comparison can be made of year over year growth rates.

Possible discrepancies can often be explained in terms of sampling error or some other measures of accuracy. Sometimes, however, they have their source in other problems (e.g. methodological) and then their explanation requires an assessment of the possible causes by metadata comparison, as for all forms of coherence assessment.

Coherence/comparability with the National Accounts – the authors of the document by Eurostat (2009) note that the National Accounts compilation process is a method for detecting the lack of coherence in data received from its various source statistical processes, whether they be direct surveys, register based surveys or indices. Thus, the level of comparability can be efficiently assessed and studied (e.g. in terms of causes of possible incomparabilities) using relevant data obtained from this system and properties of the adjustments aimed at obtaining the reliable balance of accounts. For more details see e.g. DESTATIS (2008).

Another interesting tool enabling an efficient assessment of comparability are **mirror statistics**. There are some variables for a given unit (area) that have their counterparts in another region or even country. For example, if employment-related population flows (commuting) from unit A to unit B is observed, then the number of employed persons coming from A to B should be equal to the number of employees coming to B from A. Similarly, in classical migration studies, emigration from Poland to UK and immigration to UK from Poland should coincide. Using such data a flow matrix can be constructed and analysed (see, e.g. T. Józefowski and A. Młodak (2009)). Another example: exports of country A to country B should be equal to imports of country B from country A. Thus, an effort should be made to ensure that relevant two-way statistics coincide. Hence, differences in definitions have to be carefully studied and minimized. As one can see, mirror statistics can be used to verify coherence, geographical comparability and accuracy. Thus, the difference in 'mirror' data can indicate the lack of accuracy in either or both of the outputs and/or may reflect the lack of comparability between regions (or countries) for the same data items. For example (cf. Eurostat (2009)), if the Polish estimate of emigration to UK in a particular year exceeds that of the British estimate of immigration from Poland for the same year by 10%, then this fact may indicate the lack of accuracy due to overestimation in Poland or underestimation in UK. One should then look for causes, e.g. in methodologies, such as the definition of emigrants/immigrants. In general, there are regulations concerning Business Registers (BRs) and statistics, like Structural Business Statistics (SBS) and Short–Term Statistics (STS), regulations about statistical units for the observation and analysis of the production system in the Community, where unit delineation and the BR together form an important part of the basis of statistical output. To overcome the most important problems, one should guarantee co-ordination between statistical surveys, for example in questionnaires, instructions for respondents, data processing, etc. This should, ideally, be done within respective National Statistical Institutes, where it would be much easier than elsewhere. According to M. Bergdahl *et al.* (2001), using the same BR as a frame, constructing the frame at the same time, updating units in the same way at the same time (with regard to business structure, classifications, etc.), addressing questionnaires to the same unit, etc., are further actions influencing coherence and accuracy.

The next problem is **coherence/comparability with other statistics and benchmarking**. It means that there may be other statistical surveys or data sources such that their statistical outputs can be used in combination with the results of a given survey. For example, some data from reporting on financial results of enterprises can be combined with relevant data from the tax register. Any such combination can show discrepancies, which should be (all possible) contained in the quality report. This problem is also connected with the question of benchmarking. According to Iurcovich *et al.* (2006), benchmarking is understood as an improvement process in which a company, organisation or any other (multi-organisational) system carries out three processes:

1) compares its performance against best-in-class systems;
2) determines how these systems have achieved their superior performance; and
3) uses the collected information to improve its own performance. Basically, all processes can be the object of benchmarking.

In the statistical context, benchmarking can be treated as a continuous process in which systems continuously seek to challenge their practices. More precisely, benchmarking is aimed at improving processes based upon the insights on what makes processes effective and efficient. Iurcovich *et al.* (2006) observe that benchmarking stems from the private sector business and has become increasingly popular for political systems over the past years, as nations and regions face increased competition from other competing systems. Statistics also uses benchmarking tools. For example, Barcellan (2005) provides a wide overview of such methodology from the point of view of the national accounts and analyses specific aspects of the European aggregates benchmarking, such as composition and quality of the indicators as well as possible effects of breaks generated by foreseen methodological changes and the impact of the introduction of the new approach to price and volume measures (chain-linking) on the benchmarking-based compilation of the European aggregates. The general benchmarking used in analyzing data comparability takes into account benchmarking of political systems and infrastructures at national level and less so at regional level by adoption and adaptation of modern techniques (like benchmarking) for improvement of data comparability. Statistical benchmarking comprises also the collection, analysis and documentation of good practices and use them to elaborate new solution, as much harmonized across countries and regions as possible.

To enable any comparison, some definitions (units, reference time, variables, sample frame, etc.) should be equivalent. The user should obtain precise information about all differences, problems and their practical consequences. Such information is usually contained in special quality reports. The problems of accuracy should also be mentioned in this context.

4. Measurement and improving degree of coherence and comparability

It is obvious that an efficient and complete quality report should contain basic measures of survey coherence and comparability. These measures should be constructed on the basis of expert opinions on coherence and comparability in terms of various particular aspects described in previous sections of this paper. Such proceeding is motivated by the fact that the degree of coherence and comparability is, in general, not measurable and may depend on subjective assessment of importance of particular factors of them. Our proposal in this respect consists of three main indices, where the third of them is constructed using two previous ones and can be interpreted as total complex index of coherence and comparability.

Let m, p, q and t be natural numbers and $X_1, X_2, \ldots, X_m$ denote variables collected in successive t editions of the analyzed surveys. Assume that assessment is made by p independent experts. Our first index will describe **methodological differences** in definitions of variables used in the aforementioned editions. To define it, it should be assumed that the subjects of assessment are q aspects of definition of analysed variables. It seems to be obvious that various experts can have different hierarchy of importance of the evaluation aspects, i.e. what is crucial for one expert can be of no significance for another. The index can be then computed as a sum of non-negative values expressing evaluations made by these experts:

$$I_{M(r)} \stackrel{def}{=} \frac{1}{pq_0(t-1)} \sum_{i=1}^{p} \sum_{j=1}^{q} w_{ij} \sum_{k=2}^{t} d_{ijkr} \tag{1}$$

where $d_{ijkr} \in [0,1]$ denotes the level of consistency of the definition of variable X_r in k-th edition of the survey in comparison of its $k-1$–th edition in context of j–th aspect according to the opinion expressed by i–th expert (where 0 denotes total inconsistency, 1 – full consistency while an increasing evaluation note denotes a decreasing level of consistency in terms of a given aspect) and $w_{ij} \in [0,1]$ is the weight associated by i–th expert with j–th aspect expressing its potential effect on comparability (0 denotes that relevant aspect is dropped, the larger is the weight, the more important it is for an expert), $i = 1,2,\ldots,p$, $j = 1,2,\ldots,q$, $k = 2,3,\ldots,t$ and $r = 1,2,\ldots m$; $0 < q_0 \leq q$ is the number of aspects which have non–zero weight according to at least one expert. It is also assumed that (according to classical construction of weights)

$$\sum_{j=1}^{q} w_{ij} = 1 \tag{2}$$

for every $i = 1,2,\ldots,p$.

The index I_M takes values from $[0,1]$ – larger value indicates better consistency. It will be next assumed that the variable X_r for which $I_{M(r)} \geq \theta$ is regarded as sufficiently consistent from the point of view of its definition. The parameter $\theta \in (0,1]$ is the arbitrarily assumed threshold of consistency (it can be established, e.g. at the level of 0.6). Let S_M be the set of variables which are sufficiently consistent according to (1), i.e. $S_M \stackrel{\text{def}}{=} \{X_r: r \in \{1,2,\dots,m\} \wedge I_{M(r)} \geq \theta\}$.

The construction (1) can be perceived as an extension of the Borda group evaluation method. That is, the expert opinions are here only one of several factors affecting the final evaluation. On the other hand, the weights w_{ij} can be obtained on the basis of 'hidden' (i.e. subjective and not commonly published) individual matrices of preferences of experts concerning the importance of the investigated aspects for the comparability using the fuzzy Borda approach (see, e.g. García Lapresta and Martínez Panero (2002)). An interesting alternative could be here the use of especially modified Litvak's method (cf. Litvak (1983)). In the classical version of this method such option is preferred by an expert, which is ranked first in any preference order that minimizes the sum of distances between vectors of preferences. In this situation one can propose to set the expert weight of an aspect using the distance (computed according to the aforementioned definition) which is minimal for all preference vectors where this alternative is ranked first. The weights will be set as these minimal distances after normalization to satisfy the condition (2).

If the index I_M yields satisfactory results (i.e. if a sufficient number – e.g. more than a half of n – variables are consistent), a second index, based on similar rules, should be constructed. It is **related to comparability over time and over domains of interest** (i.e. spatial units, branch clusters of economic entities, etc.) – restricted to those variables with similar definitions. It should take into account that for some variables comparability over time can be more important than comparability over domains (e.g. total revenues), for others this expectation can be opposite (e.g. for some newly introduced variable having form of an index, which uses results from new phenomena and needs of users). Formally, the index I_D is constructed in the following way:

$$I_{D(r)} \stackrel{\text{def}}{=} \frac{1}{ap(t-1)} \sum_{i=1}^{p} \left(w_{ir(T)} \sum_{k=2}^{t} d_{ikr(T)} + w_{ir(U)} \sum_{k=2}^{t} d_{ikr(U)} \right) \qquad (3)$$

where $w_{ir(T)}$ and $w_{ir(U)}$ denote weights associated by i–th expert with time and analyzed domains for the variable X_r, respectively ($w_{ir(T)}, w_{ir(U)} \in [0,1]$), $d_{ikr(T)}$ and $d_{ikr(U)}$ are assessments of comparability over time and domains ($d_{ikr(T)}, d_{ikr(U)} \in [0,1]$, where 0 denotes total lack of comparability, 1 – full comparability in a given context), $i = 1,2,\dots,p$, $k = 2,3,\dots,t$ for every $r \in \{1,2,\dots,m\}$ such that $X_r \in S_M$. The parameter a takes values 1 or 2 depending on whether only one or both of these aspects of comparability have non–zero weight in the opinion of at least one expert. Similarly as in the case of (1) also the index

(3) takes values from [0,1] where 1 indicates the ideal comparability and zero – its total lack.

Finally, for each edition of the survey a complex **index of complex coherence and comparability** based on (1) and (2) should be computed. It will be defined as

$$I_C \overset{\text{def}}{=} \frac{\underset{r=1,2,\ldots,m}{\text{med}} I_{M(r)} + b \underset{r\in\{1,2,\ldots,m\}:X_r\in S_M}{\min} I_{D(r)}}{2}.$$

where $b = 1$ if $|S_M| \geq \left[\frac{n}{2}\right] + 1$ and $b = 0$, otherwise[1].

As one can see, the comparability of selected variables has here especial importance. It is motivated by the fact that methodologically consistent survey should have relevant and high level of comparability. Of course, $I_C \in [0,1]$ and higher value inform about better quality of the survey in this context. These indices can significantly help to assess survey quality. It is worth noting that the expert weights have a subjective character. That is, every time when the set of experts is changed a new value of the index I_C (and, of course, previous indices) is obtained. Thus, it is recommended to have stable and competent team of experts, i.e. such that possible change in it will not affect significantly the structure of preferences. On the other hand, the change of the weights may be also a result of importance of actual or foreseen changes in quality of data collection of a given variable of its methodology over time. In this context such elasticity of weights is rather desirable.

A small example illustrating these problems can be as follows. Assume that a survey do journeys to work is conducted. The data following three variables are collected: X_1 – average time of journey to main workplace, X_2 – distance between place of residence of main workplace and X_3 – means of transport from the place of residence to the main workplace. Data on variables X_1 and X_3 are collected from the sample survey of 20% of the population and data on X_2 are determined on the basis of the tax register. The following methodological aspects are considered: 1) frame population, 2) sources of data and sample design, 3) data collection and processing and 4) imputation and estimation. The problem of data comparability on the level of NUTS 2 (voivodships) and NUTS 4 (powiats) Polish regions and over time will be also considered. In case of the index (1) the aspects 1), 3) and 4) will have greater importance for consistency than 2) because most data are collected in sample surveys. Hence, $w_{ij} > w_{i2}$ for any i. The value of d_{ijkr}'s will depend on changes in methodology occurring in comparison with previous edition of the survey. If the changes are very small then d_{ijkr}'s will be close to 1 in any case. Otherwise, e.g. if in the next edition of the survey the data on X_2 are collected also from sample survey then $d_{ij(k+1)2}$ will be smaller,

[1] The symbol $[c]$ denotes the integer part of the real number c, i.e. the greatest integer not greater than c. The symbol $|Z|$ denotes the number of elements of the set Z.

especially for $j = 1,2,4$. As regards the index (3), the comparability over time can be here less important than other regions, and hence one can expect that $w_{ir(T)} < w_{ir(U)}$ for any r. However, taking the character of data collection for particular variables into account, one can observe that the comparability for NUTS 4 regions for X_1 and X_3 will be significantly smaller than for X_2, i.e. $d_{ikr(U)} \ll d_{ik2(U)}$ for $r = 1,3$. For NUTS 2 regions, due to higher quality of estimation, the comparability of X_1 and X_3 can significantly increase.

One more problem, which was not indicated previously due to its practical complexity and importance, is how to treat variables which are expressed using various measurement units and vary from the point of view of the range or significance of information provided. In many cases – despite efforts made to harmonize methodologies – such discrepancies are unavoidable owing to specific needs of users of data produced as a result of particular surveys, diversification of traditions observed between regions or countries, etc. Therefore, some obstacles cannot be overcome methodologically. Hence, numerical methods for the normalization of variables should be applied.

It is worthwhile to note that the normalization should denote also uniformization of characters of particular variables. Each variable has its own status being arbitrarily established, taking into account the relationship between the values and the reality they describe. The following three types of variables can be then distinguished:

- *stimulant* – the higher the value of the feature, the better (e.g. average monthly wage and salary or GDP per capita),
- *destimulant* – smaller values are much more desirable than higher ones (e.g. unemployment rate),
- *nominant* – has an optimum level of value (called also an imbuement point). Thus, below this point the feature is treated as a stimulant and above it – as a destimulant. Or, conversely – increasing values (and simultaneously lower than the optimum) are "worse" whereas decreasing ones (but greater than the optimum) are regarded as "better".

Destimulants and nominants should be converted into stimulants (e.g. by taking respective "bad" values with opposite signs). Next, they are then normalized. The relevant formula can lead either to unification of basic statistical measures (such as arithmetic mean, median, standard deviation or median absolute deviation) or ranges of features. A rich collection of those methods can be found in articles by Zeliaś (2002) or by Młodak (2006 a). They are based on the following formula:

$$z_{ij} = \frac{z_{ij} - a_j}{b_j},$$

where x_{ij} is the value of j–th variable for i–th unit, z_{ij} is its normalized value, a_j and b_j are parameters for j–th variable ($i = 1,2,\dots,n$, $j = 1,2,\dots,m$, where n denotes the number of analysed units) such as mean, standard, median, median

absolute deviation, j–th coordinate of the Weber median or another singular points of the multidimensional space (cf. (A. Młodak (2006 b)), etc. In general, if the arithmetic mean of each of normalized variable amounts to 0 and its variance is equal to 1, then normalization is called standardisation; if the range is constant and equal (e.g. to 1), then such transformation is called *unitarisation*, and if $a_j = 0$ and b_j is assumed to be arithmetic mean, minimum, maximum, median, sum, sum of squares, square root of the sum of squares of the value of the j–th variable ($j = 1,2, \dots, m$), then it is called *quotient transformation*.

Another option in this context is the use of benchmarking. That is, distances of a given object (represented by a vector of data on $\mathbb{R}^m$) and the benchmark are computed, i.e. a best–in–class object (real or artificial, created i.e. by optimization of values of particular variables), which all investigated objects are compared with. To this end, a special measure of distance between structures can be applied (e.g. Gower's distance, Minkowski's metrics and its particular cases, Canberra metrics, Kaziniec's, Gatew's or Jeffrey's and Matusita's formulas or Pearson's correlation coefficient – cf. Kaziniec (1968), Gatew (1977), Młodak (2006), Bruzzone *et al.* (1995)). Hence, the benchmarking normalization can be performed, e.g. as

$$z_{ij} = \frac{\left|x_{ij} - \varphi_j\right|}{d_i}$$

where $\varphi = (\varphi_1, \varphi_2, \dots, \varphi_m)$ is the benchmark, d_i denotes the distance of i–th object from the benchmark, $i = 1,2, \dots, n$, $j = 1,2, \dots, m$.

Data from various questionnaires can also be used to construct complex indices of development in a given domain. It is very comfortable for the user, who obtains one synthetic information instead of many various data that tend to be directly incomparable. Some concepts concerning this issue are contained in the handbook by OECD (2008), but they are based mainly on simple normalizations and do not exploit all interesting properties of the modelled data. Hence, one can recommend here also Polish output combining the aforementioned normalization and benchmarking and taking into account efficient choice of diagnostic variables (cf. A. Młodak (2006 a)). Also some suggestions concerning theoretical methods of establishing the share of each object in the overall development of their whole collection have been formulated (cf. A. Młodak (2002)).

If this analysis is restricted only to the *uniformization* of variables, one could produce desirable *characteristics of accuracy*. Although it may often be of no primary importance, sometimes (e.g. if results of one of two surveys to be compared exhibit bias errors) its characteristics would be desirable. For example, when the same statistics are obtained in two different surveys, measures of their accuracy may be crucial in order to assess the consistency of resulting estimators. The quality report should also include a description of differences between various data producers with the same country, e.g. organizations, agencies,

statistical offices, etc. which can affect the final quality and comparability of collected data as well as differences between quality measures used in other countries. Accuracy can be quantified by measurement errors for units sampled with probability one (that do not contribute to sampling errors) or differences between basic characteristics of respondents in two surveys. If a unit has different respondents in different surveys and their internal structures also differ, the measurement error can be serious, if it was undetected or not reduced. Some secondary indicators (e.g. wage per employee or profitability rate) can be burdened by missing data, if the number of non-response units is relatively large. Imputation precision can also be perceived as a measure of the survey's accuracy. In many cases, however, accuracy should be assessed using several indicators describing various aspects of accuracy (i.e. variability caused by non-response), precision of the sampling frame, random errors due to sampling, etc. Using all components of accuracy statistics (including estimation error), one can determine a symmetric uncertainty interval around a given point estimate, which informs us about the expected precision of this estimate. The smaller this interval is, the more efficient the resulting statistical inference and data comparisons made on the basis of both surveys. To avoid any output distortion, it is very important to take all error sources into account. On the other hand, some negative effects can be mutually reduced. For example, if indices defined as a ratio of two quantity statistics obtained from the same survey or different surveys with similar structures and intensity of errors are considered, such a ratio can reflect them sufficiently, since they are included (with similar values) both in the numerator or denominator of such an index). A similar problem can be observed if profit (the difference between revenues and costs) is investigated. Hence, quality measurement should also be done using the simplest data, expressed in non-negative, absolute values (e.g. revenue in EUR, number of employees, fixed assets in EUR, etc.). In the case of monetary values observed over time, one should rely on the inflation factor using the fixed-base index (fixed prices, if applicable). They can also be normalized.

M. Bergdahl *et al.* (2001) also consider the method of co–ordinating statistical output called also benchmarking (but it is dynamic, whereas the previously described approach can be perceived as static), where one set of estimates is forced to agree with another. Typically, short-term statistics could be benchmarked against annual statistics, if the former (after aggregation to the calendar year) are an indicator of the latter. This procedure can have two benefits for the user: it unifies the two time series (ensuring that the monthly series has the same annual sum as the annual series); and it improves the accuracy of short–term statistics. For this to be meaningful, the two sets of statistics should have the same target parameters for the calendar year. The necessary comparison should be conducted both at the macro and micro–levels.

5. Conclusions

Coherence and comparability of data are one of the key aspects deciding on the quality of statistical output. Their role is underlined in many official documents and recommendations. There are various dimensions of these problems as well as their sources and typologies. In business statistics coherence and comparability have specific formal backgrounds and depend on quality of integration of information obtained from enterprises using different data sources. The methodological solutions differ between countries and sometimes even regions. Moreover, another inconveniences making difficulties in estimation of sufficient quality can occur, e.g. due to the scale of inter–company transfers in some economic entities as well as a dispersion of local units of the same enterprise over various countries. In any case, the user of statistical data should obtain precise information about all such differences, problems and their practical consequences. Hence, some international standards of quality reports were elaborated.

Thus, the construction of some universal methodological solutions guaranteeing at least roughly international coherence and comparability is very difficult. The only way to the improvement of these aspects of data quality leads through harmonisation of sampling frames, survey designs and used tools. However, to do this efficiently a reliable assessment of the level of problems should be performed. The presented original concepts of aggregated indices of coherence and comparability can be a significant support in realisation of this task. Of course, they are based to a large extent on subjective appraisal of the size and importance of key aspects. However, the stable and competent teams of experts and professional use of presented tools (such as, e.g. benchmarking) can ensure reliable results of such recognition which could be a hint how to permanently improve the data quality in the context of the investigated aspects.

REFERENCES

BARCELLAN, R., (2005). *The use of benchmarking techniques in the compilation of the European quarterly national accounts situation and perspectives*, Working Papers and Studies, European Commission, Euro Indicators, Office for Official Publications of the European Communities, Luxembourg, available at http://www.uni-mannheim.de/edz/pdf/eurostat/05/KS-DT-05-026-EN.pdf.

BERGDAHL, M., BLACK, O., BOWATER, R., CHAMBERS, R., DAVIES, P., DRAPER, D., ELVERS, E., FULL, S., HOLMES, D., LUNDQVIST, P., LUNDSTRÖM, S, NORDBERG, L., PERRY, J., PONT, M., PRESTWOOD, M., RICHARDSON, I., SKINNER, CH., SMITH, P., UNDERWOOD, C., WILLIAMS, M., (2001). *Model Quality Report in Business Statistics*, General Editors: P. Davies, P. Smith, http://users.soe.ucsc.edu/~draper/bergdahl-etal-1999-v1.pdf.

BRUZZONE, L., ROLI, F., SERPICO, S. B., (1995). *An extension of the Jeffreys–Matusita distance to multiclass cases for feature selection*, IEEE Transactions on Geoscience and Remote Sensing, vol. 33, pp. 1318–1321.

CASCIANO, M. C., DE GIORGI, V., OROPALLO, F., SIESTO, G., (2012). *Estimation of Structural Business Statistics for Small Firms by Using Administrative Data*, Rivista Di Statistica Ufficiale No. 2–3, Instituto Nazionale Di Statistica, pp. 55–74.

COOK, L., (2007). *International experience in setting up an economic statistics compilation programme*, Regional Workshop for African countries on Compilation of Basic Economic Statistics jointly organized by United Nations Statistics Division (UNSD) and African Centre for Statistics at Economic Commission for Africa (ACS), United Nations, Department of Economic and Social Affairs, Statistics Division, 16th – 19th October 2007, Addis-Ababa, Ethiopia, Doc. No. ESA/STAT/AC.136.3, document available in the Internet at the website http://unstats.un.org/unsd/economic_stat/intl%20coop%20and%20workshops%20(bes)_files/AddisOct2007/UNSD%20documents/WS-BES-ECA-136-3-Intl-experience-Len%20Cook.pdf.

DAVIES, P., (2000). *Assessing the Quality of Business Statistics*, Office for National Statistics, Proceedings from the Second International Conference on Establishment Surveys, June 17th – 21st, 2000, Buffalo, New York, http://www.amstat.org/meetings/ices/2000/proceedings/S38.pdf.

DESTATIS, (2008). *National Accounts Quarterly Calculations of Gross Domestic Product in accordance with ESA 1995 – Methods and Data Sources*. New version following revision 2005, Fachserie 18 Series S. 23, Statistisches Bundesamt (Federal Statistical Office), Wiesbaden, Germany, https://www.destatis.de/EN/Publications/Specialized/Nationalaccounts/QuarterlyCalculationsGrossDomesticProductAccordance.pdf?__blob=publicationFile.

EUROSTAT, (2009). *ESS Handbook for Quality Reports*, Series: Methodologies and Working papers, Office for Official Publications of the European Communities, Luxembourg.

EUROSTAT, (2008). *Quality Control of Urban Audit Variables*, Unit D2, Office for Official Publications of the European Communities, Luxembourg, April 2008.

EUROSTAT, (2003). *Handbook "How To Make A Quality Report"*, Series: Methodological Documents, Working Group "Assessment of quality in statistics", Sixth meeting, Luxembourg, 2nd – 3rd October 2003, Document No. Doc. Eurostat/A4/Quality/03/Handbook, available at the website http://190.25.231.249/aplicativos/sen/aym_document/aym_biblioteca/Documento%20de%20soporte/Methodological%20documents%20handbook%20%20how%20to%20make%20a%20quality%20report%20-%20NACIONES%20UNIDAS.pdf .

EUROSTAT, (2005). *European Statistics Code of Practice for the National and Community Statistical Authorities*, Statistical Office of European Union, Eurostat, Luxembourg, available at http://unstats.un.org/unsd/dnss/docViewer.aspx?docID=2636.

GARCÍA LAPRESTA, J., MARTÍNEZ PANERO, M., (2002). *Borda Count Versus Approval Voting: A Fuzzy Approach*, Public Choice 112(1–2), pp. 167–184.

GATEW, K., (1977). *Статистическо характеризание на структурни измениения*, Трудовэ на Висшия, Икономический Институт - К. Маркс, София, vol. 3, pp. 10–42 (in Russian).

IURCOVICH, L., KOMNINOS, N., REID, A., HEYDEBRECK, P., PIERRAKIS, Y., (2006). *Mutual Learning Platform. Regional Benchmarking Report. Blueprint for Regional Innovation Benchmarking*, European Commission, Committee of the Regions, IRE Innovation Network, available at http://www.rttn.ru/_files/fileslibrary/90.pdf.

JÓZEFOWSKI, T., MŁODAK, A., (2009). *Observation of flows of population in Polish statistics – problems and challenges*, [in:] E. Elsner, H. Michel (eds.) "Assistance for the Younger Generation. Statistics and Planning in Big Agglomerations", Institut für Angewandte Demographie IFAD, Berlin, pp. 61–76.

KAZINIEC, L. S., (1968). *О методах сводной оценки стуктурных сдвигов*, Вестник Статнстики, No. 11 (in Russian).

KÖRNER, T., PUCH, K., (2011). *Statistics and Science. Coherence of German Labour Market Statistics*, Volume 19, Statistisches Bundesamt (Federal Statistical Office), Wiesbaden. Available at https://www.destatis.de/DE/Publikationen/StatistikWissenschaft/Band19_CoherenceLabourMarket1030819119004.pdf?__blob=publicationFile.

LITVAK, B., (1983). *Distances and consensus rankings*, l Cybernetics and systems analysis, 19 (1), 71{81. (Translated from Kibernetika, No. 1, pp. 57–63, January-February, 1983).

MALINA, A., ZELIAŚ, A., (1998). *On Building Taxonometric Measures on Living Conditions*, Statistics in Transition, vol. 3, pp. 523–544.

MŁODAK, A., (2006a). *Taxonomic analysis in regional statistics*, ed. by DIFIN – Advisory and Information Centre, Warszawa, Poland (in Polish).

MŁODAK, A., (2006b). *Multilateral normalisations of diagnostic features*, Statistics in Transition, vol. 7, pp. 1125–1139.

MŁODAK, A., (2002). *An Approach to the Problem of Spatial Differentiation of Multi-feature Objects Using Methods of Game Theory*, Statistics in Transition, Vol. 5, pp. 857–872.

OECD, (2008). *Handbook on Constructing Composite Leading Indicators: Methodology and User Guide*, Global Inventory of Statistical Standards, Organization for Economic Cooperation and Development, link: http://unstats.un.org/unsd/iiss/Handbook-on-Constructing-Composite-Leading-Indicators-Methodology-and-User-Guide.ashx.

STATCAN, (2006). *Metadata to Support the Survey Life Cycle*, Invited Paper, Joint UNECE/Eurostat/OECD work session on statistical metadata (METIS), Topic (iii): Metadata and the Statistical Cycle, Submitted by Statistics Canada for the Conference of European Statisticians, United Nations Statistical Commission and Economic Commission for Europe, European Commission Statistical Office of the European Communities (Eurostat) Organisation For Economic Cooperation and Development (OECD), Statistics Directorate, Geneva, 3rd – 5th April 2006, http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.40/2006/zip.5.e.pdf.

DE WAAL, T., PANNEKOEK, J., SCHOLTUS, S., (2011). *Handbook of Statistical Data Editing and Imputation*, John Wiley & Sons, Inc., Hoboken, New Jersey.

YANCHEVA, D., ISKROVA, K., (2011). *Reducing the administrative burden for the business in Bulgaria: Single Entry Point for Reporting Fiscal and Statistical Information*, [in:] Proceedings from BLUE-ETS Conference on Burden and Motivation in Official Business Surveys Statistics Netherlands, Heerlen, March 22 & 23, 2011, pp. 189–198.

ZELIAŚ, A., (2002). *Some Notes on the Selection of Normalization of Diagnostic Variables*, Statistics in Transition, vol. 5, pp. 787–802.