DIFFERENCES BETWEEN SURVEY AND ADMINISTRATIVE DATA

CASE OF EMPLOYMENT AND WAGES INDEXES

REZARTA MYRTOLLARI, INSTITUTE OF STATISTICS
rmyrtollari@instat.gov.al

Introduction

In the production of official statistics, for a certain phenomenon, the available data may come from statistical surveys and administrative sources. Nowadays, the combination of these two sources is a promising and innovative strategy which affects the quality and quantity of research and increases the potential of data (Künn, 2015). However, this usage is often accompanied by serious challenges, given the simple fact that the purpose of designing the two sources is different. Administrative data are defined as data sets collected by government institutions or agencies for tax, benefit or public administration purposes (UNECE, 2011). On the other hand, data from surveys are collected specifically for statistical purposes. This article examines the differences between administrative sources and surveys focusing specifically on common variables of payroll as an administrative source and those from surveys conducted in INSTAT.

According to Penneck (2007) surveys differ from administrative data in the sense that they are specifically designed for analytical purposes, so coverage of population, definitions, methodology and time can be designed to meet these analytic needs. However, the sample size might be a problem if it is small since large-scale surveys are expensive and small-scale surveys have limited use. Samples are also subject to errors and non-response bias. In addition, mentions Penneck, we cannot be sure of the accuracy of business survey responses, compared, for example, with the administrative data collected for tax purposes. Administrative systems also require data from individuals, but the latter often see this as an indispensable part of the administrative process rather than as an additional statistical burden. The following sections will examine some of these issues in detail based on the work of Johnson and Moore (2008), illustrating them with concrete examples from the case of estimating average wage and employment indicators from two different sources in INSTAT.

1. POPULATION COVERAGE ISSUES

A system of administrative records defines the population covered by legislation based on the scope of the program intended for registration. This population is often limited by specific demographic or economic characteristics. According to Johnson and Moore (2008), in some cases individuals may need to undertake some actions to become part of the administrative system (e.g. registration of farmers in the tax and social security system by obtaining the NIPT to benefit from government support schemes). It is therefore important, say the authors, to consider what encourages individual units to be part of schemes. There may be some favouring factors for some individuals to avoid registration, especially if their circumstances place them close to a threshold that requires obligatory participation or gets associated with financial costs (p.10), such as setting a minimum wage on which contributions are calculated. Another factor is the change of policies that may fluctuate the population taken in study from year to year.
The Federal Committee on Statistical Methodology points out the differences on the unit of interest. The study unit needed for statistical purposes often focuses on the characteristics of groups formed by units (e.g. enterprises operating in a particular activity or large enterprises), while the administrative data focus on identifying specific units so that based on their individual characteristics (e.g. full-time employees or dual employment) certain actions can be undertaken. Thus, the differences in the entity reported in the tax statements limit the usefulness of the data for some types of research. Johnson and Moore explain that the target population of a survey is determined by the purpose of the study, the sampling frame availability, and the cost of the sample. Population for most surveys is derived from existing sources, such as population data based on geography, address lists, or other administrative sources. Often one of the most difficult issues in designing a survey is finding a suitable population (Lessler and Kalsbeek, 1992). If the population frame chosen for sampling is not suitable, it can lead to under coverage problems which may affect the results obtained from the survey data. Another problem arises if a survey targets a population that is difficult to find or measure. Directly related to the availability of a population is the potential cost of receiving population data and the cost of interviewing a sample of the desired size. For target populations that are difficult to find, simply the cost of increasing sample size to provide better coverage may be obstructive to undertake such an initiative (Johnson and Moore, 2008, p.13).

2. CONTENT ISSUES

Johnson and Moore list several content issues that need to be considered while working with administrative and survey data. One of them is the purpose for which administrative data are collected, which may have a significant impact on their usefulness for statistical purposes regarding the amount of available data, data definitions, consistency between different time periods and data quality. The authors argue that many times the usefulness of administrative registry systems is limited because only those variables needed to administer the tax and tax payment program are collected. These variables can only be a small part of the data reported in an administrative form (p.15). In addition, because program requirements are defined by legislation, the concepts and definitions of variables used to meet program needs do not necessarily match those required for social or economic analysis (Brackstone, 1987). For example, one of the problems faced in our administrative data comes as a result of using the concept of working days instead of working hours in the declaration of the taxpayer's payroll. This difference makes it difficult to compare employment data with those of national accounts. Such factors may pose serious limitations on the overall usefulness of the administrative data systems or require that the administrative agency undertake collecting and / or editing additional data, causing financial costs and delaying the availability of data.

An important aspect of the data content is continuity over time of the variables included and their definitions. Coverage and content in administrative data systems may be subject to discontinuities resulting from changes in laws, regulations, administrative practices or the scope of the program (Brackstone, 1987). For example, the revisions of the law on the minimum and maximum wage level made that the minimum monthly base salary for employees required by any legal or natural person, domestic or foreign, is 24.000 ALL from 22.000 ALL that was before this period. Such changes have a significant impact on the statistical uses of data for comparisons between periods.

Administrative data systems also can not ensure perfect data quality. Information that might be important for statisticians, but less important for administrative purposes, is often reported and processed imperfectly, noticed Johnson and Moore (2008). Here we can mention the choice of profession by the person who makes the declaration of salary and wages. The variable that indicates the profession category has a secondary importance for the administrative agencies as long as the person

regularly declares his or her contributions. On the other hand, this variable is of particular importance for the production of statistics about average wage by group of professions. The other variables used mainly as secondary or complementary information may be of low or even incomplete quality (as in the case of working days for which the declaration is usually a standard of 21 or 22 business days). This phenomenon may also occur with data specifically collected for statistical purposes using existing administrative channels such as in the case of enterprise activity classification in Nace Rev.2 collected by the administrative agencies for the account of INSTAT. These variables may be of lower quality if their priority is not too high for the administering authority or the entity providing the information (Jensen, 1987). Another issue pointed out by Johnson and Moore (2008) is data reliability which can be affected if the information provided to the tax entity can cause profits or losses for the declaring subject. Moreover, given that the data collected and processed for administrative purposes are generally given priority over what is required for statistical purposes, the amount of processing required to provide administrative data suitable for statistical purposes may affect the time that these data are made available to statisticians, argue the authors.

Many of the issues raised above are best addressed and resolved through surveys, stated Johnson and Moore. However, the authors notice that other content and validation issues of some kind appear in the survey data. The key issue here is the voluntary nature of responses to surveys versus the legal obligation to participate in administrative data programs. The respondent needs to be persuaded to give their time and the information required despite the fact that there are no consequences if he or she refuses and there are no benefits if the survey is filled. Still, if a respondent agrees to participate in the survey, it is possible that he or she refuses to answer the questions in a "real" manner (p.21).

For respondents who agree to attend and respond to all survey questions, the measurement error is still a concern for the survey data, state Johnson and Moore (2008). Respondents can "think" answers to questions or they may have difficulty to remember past events. Other typical measurement errors include rounding of amounts, misunderstanding of questions and changing responses due to fears about the disclosed data or the desire to protect privacy. Numerous studies exist regarding error measurements and their effects on observation data (Lessler and Kalsbeek, 1992). While it is true that, for administrative data, non-response is not an essential problem, it is not clear whether administrative records are always more accurate than observation data, report the authors. An example would be the number of employees declared by the enterprise; some companies intentionally may declare a lower number of employees into their statements to reduce their tax obligations. The same individuals can report the true value in responding to a questionnaire as there are no legal consequences if true value is stated. Another content issue for the survey data is the timeline of data. While many simple surveys are carried out at frequent frequencies, monthly or quarterly, most of the most complex surveys occur annually or even rarer. Costs and other resource constraints are major factors in timely use of survey data. A final content issue for surveys elaborated by Johnson and Moore (2008) is validation of data. According to the authors sometimes it is possible to conduct validation studies after a survey has been completed and these studies add additional costs to the survey or validation of selected data variables can be carried out with external sources such as censuses or administrative records, but there is often no validation source (p.25).

3. PRIVACY ISSUES

In their work Johnson and Moore (2008) consider data privacy as a very important issue. The authors explain that any use of administrative data for research purposes should take into account the laws that protect the privacy of the data. The research of administrative data is often limited to uses within the scope of an agency mission and should be carried out only by persons working for the agency as employees, contractors or under Memorandum of Understanding that allow employees of various institutions to exchange the data. The way the public percept privacy protection of their data has a direct impact on the continuity of the levels of declarations (p.26). Often, because of these factors, the available data does not contain identifying variables. For example, in the case of individual data from the administrative source, variables which directly identify the subject are missing. Of course in another scenario the availability of these variables could lead to wider statistical use and a combination of data from different sources.

However, the authors emphasize that data confidentiality is of great importance to the current and future success of any administrative observation and registration. If the subjects do not believe that their data is sufficiently protected, response rates and overall data quality will be subject to deformation. Confidentially and privacy laws offer significant protection against potential abuse of data (Johnson and Moore, 2008).

4. EMPLOYMENT AND WAGES AND SALARY DATA

Both administrative sources and surveys (such as Quarterly Short-term Statistics Survey) provide important information at quarterly frequency regarding the number of employees and salary fund. This data is used to calculate wages and salary index and the employment index. Administrative records have richer demographic information about the individual and detailed data on social and health contributions. On the other hand, survey data is more limited, including only the number of employees and the wages and salary fund of the surveyed enterprise.

The most important changes between the two sources, as theoretically are discussed above, relate to the survey unit, population coverage, and sample size. STS is a quarterly survey where the surveyed unit is the enterprise and the main variables required are Net Sales, Industrial Production, Construction Production, Average Number of Employees, Wages and Salaries Fund, Production Prices, Import Prices, Construction Costs (INSTAT, 2017). All produced indicators are expressed in indices, in annual and quarterly changes. The study unit for the administrative source is always the individual, and the average salary indicators are expressed as absolute values. The size of the STS sample is limited due to frequency and cost, and the survey does not cover all economic activities, leaving out the assessment of the agricultural activities (section A), those financial services and insurance (section K), real estate (section L), public administration (section O), education and health (sections P & C), as well as arts, entertainment and entertainment activities, other services and activities of international organizations (i.e. sections R, S, T, U) which are outside its coverage area. This means that the quarterly information from the survey about employment and salaries is missing for these industries. On the other hand, the information from the administrative source includes individuals and enterprises in all economic activities.
The change in methodology has a direct impact on the estimates derived from each source. In addition, STS estimates are not particularly focused on estimating average wages and the lack of detailed employee information (e.g. full-time or part-time employment, dual employment, contribution category etc.) makes it is impossible to apply the same methodology as the one used to estimate the average salary from the administrative source. Furthermore, the data from surveys are subject to the weighing process, while the assessment from the administrative source is straightforward.

5. CONCLUSIONS

Nowadays there is a need to satisfy a growing demand from users for good quality statistics, enabling faster measurement of new phenomena. At the same time, the demands of these users are in line with the needs of today environment that the burden placed on businesses and citizens diminishes (Laux, Baigorri, & Radermacher, 2009). Therefore, the use and combination of administrative or secondary data by statisticians is seen as a necessity in the present days, but it is also accompanied by a number of challenges. Some indicators, such as those discussed above, can be produced with data that can be derived from both administrative and statistical sources, but the fundamental structural differences between these two sources, as well as changes in the applied methodology, result in differences in estimations and, of course, the final results obtained from them. These changes are present in almost all dimensions of quality, such as relevance, accuracy, timeliness, accessibility, comparability, and timing. For this reason, users should be aware of these changes at the time of using the estimations from different sources and should understand the origin of data, their collection and use, in order to avoid mistakes and misunderstandings. This allows them to choose the indicators that fit the best their study goals (Laux, Baigorri, & Radermacher, 2009). More than just competing sources, administrative data and surveys should be seen as complementary sources. As Kapteyn & Ypma (2007) say, the question of whether administrative resources or observations show "truth" is almost a philosophical question.

BIBLIOGRAPHY

Brackstone. (1987). Statistical uses of Administrative Data: Issues and Challenges. Statistical Uses of Administrative Data Proceedings, (fv. 5-26).

INSTAT. (2017). Statistikat Afatshkurtra- Raporti i cilësisë. Tirana, Albania: INSTAT.
Jensen, P. (1987). The Quality of Administrative Data From a Statistical Point of View, Some Danish Experience and Considerations. Statistical Uses of Administrative Data Proceedings, (fv. 291-300).

Johnson, B., & Moore, K. (2008). Comparing Administrative and Survey Data. IRS Statistics of Income Working Paper Series.

Kapteyn, & Ypma. (2007). Measurement error and misclassification: A comparison of survey and administrative data. Journal of Labor Economics, 513-551.

Künn, S. (2015). The challenges of linking survey and administrative data. IZA World of Labour, 214.

Laux, R., Baigorri, A., & Radermacher, W. (2009). Building Confidence in the Use of Administrative Data for Statistical Purposes ., (f. 9).

Lessler, & Kalsbeek. (1992). Nonsampling Error in Surveys. New York : John Wiley & Sons.

Penneck, S. (2007). Using Administrative Data for Statistical Purposes . ICES-III. Montreal, Quebec, Canada.

UNECE. (2011). Using Administrative and Secondary Sources for Official Statistics - A Handbook of Principles and Practices, United Nations Economic Commission for Europe. New York and Geneva: UNITED NATIONS.