WP.12
ENGLISH ONLY

**UNITED NATIONS STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE**

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Vienna, Austria, 21-23 April 2008)

Topic (ii): Editing administrative data and combined sources

QUALITY OF ADMINISTRATIVE DATA – A CHALLENGE FOR THE MAINTENANCE OF THE STATISTICAL BUSINESS REGISTER

Supporting Paper

Prepared by Norbert Rainer, Statistics Austria, Austria

I. INTRODUCTION

1. It is obvious that the business registers run by the National Statistical Institutes (NSIs) are some of the most important instruments for the production of an increasing variety of official statistics. In the past these instruments were primarily important for the traditional core of economic and enterprise statistics. Nowadays there are a lot more statistical domains addressing enterprises than the typical economic statistics requesting data for business cycle and structural analysis. Examples are R&D statistics, innovation statistics, statistics on the use of ICT, statistics on vocational training in enterprises, environment statistics, etc. For all these statistics it is highly desirable that they use the same frame for deriving their samples as for the core of enterprises statistics. These coherence requirements are clearly a quite high challenge within the NSIs, and of course even much more if some of these statistics is elaborated by other institutions than the NSIs.

2. The traditional use of the business register as provider of the sample frame and the corresponding information for the grossing up procedures is more and more supplemented by further uses; the most important ones in our Austrian case are the links that the register provides to administrative records and the use of the business register as a statistical database of its own. The first additional goal supports the use of administrative data to supplement or replace statistical surveys in a consistent way which moreover reduces response burden. The second is that the register is used to derive statistical pictures of the structure of the economy, for instance to replace the traditional censuses of local units. Another example is business demography statistics that is also ideally to be derived from the business register database.

3. The additional uses of the business registers increase also the requirements concerning coverage and data quality. Feedbacks from statistical surveys or special register surveys are no longer the most important sources for the maintenance of the register. Regularly available sources with high coverage and easy access and use become the central focus. The register is primarily not maintained by information coming directly from the enterprises to the NSIs but from indirect sources, such as administrative records. Thus, the availability and quality of administrative data sources become a crucial factor of the register performance.

4. Administrative data sources follow partly different definitions and rules than what would be needed or desired for statistical purposes. This makes it difficult to use these sources in a straightforward way without any additional efforts and data editing procedures. Furthermore, the availability of such administrative data is not always sufficient and the data are of different quality. It is therefore important not only to improve the data editing and maintenance procedures required but also to improve the

cooperation between the NSI and the holders of the administrative data, and to undertake various efforts in order to achieve improvements of the administrative data.

5. This short contribution is non technical; it does not describe the various editing and imputation methods in any detail that are run in the maintenance process of the business register. Focus is rather on the general frame under which administrative data are used, some of the main quality problems encountered and the strategies that are developed to improve the situation. The background is of course the Austrian situation, which in one or the other example may not be different from the experience in other countries.

II. MAIN ADMINISTRATIVE DATA SOURCES FOR THE BUSINESS REGISTER

A. Register requirements

6. The business register of Statistics Austria covers legal units/enterprises, establishments and local units for the market sector (except agriculture), the government and the non-profit sector. Thresholds for size of the units are employment and turnover criteria. All units are classified according to the European activity classification NACE with their main activities and in many cases also with their secondary activities. Currently, the classification is done both for the old version of NACE (NACE Rev. 1.1) as well as for the new version (NACE Rev. 2). The location data of all units are linked to the buildings and dwellings register. The buildings and dwellings register covers also the official addresses as assigned and entered into the system by the municipalities. Each building has its own identification code and the link to it allows using the official address data (post code, municipality, street name, buildings number) for purposes of mailing questionnaires and attribution to geographical areas.

7. The main stratification data for defining statistical reporting obligations are the activity code, the size classes by employment and by turnover. In addition to the activity codes according to NACE also the institutional sector classification is applied to the institutional units in the register. The institutional sector classification is still less important for stratification and economic surveys but its importance will increase. The enlarging coverage of certain surveys into the fields of services which are provided by enterprises as well as government and non-profit institutions needs a clear delineation of the intended survey coverage and the sector classification can serve as the provider of the criterion. Examples of such services are education and health services. Therefore, the new European Regulation on statistical business registers stipulates that the sector classification has to be introduced.

8. The business register should cover active units in the sense that the unit performs economic activities that contribute to GDP. However, the very small enterprises are not covered in the business register: the current threshold is either employing at least one person or an annual turnover of more than € 10.000.-. For certain activities the thresholds are differently defined and measured.

9. The business register is not only used within Statistics Austria, one outside user is the Austrian Central Bank with which a register cooperation has been started some time ago. Currently, Statistics Austria delivers to the Central Bank the activity and the sector classification code. Since beginning of 2008, also the social security systems receives the activity codes in order to enable them to elaborate the official employment statistics on the basis of the same classification of the enterprises as the statistics performed by Statistics Austria.

B. Administrative data sources

10. For the maintenance of the market sector units of the business register four main data sources are regularly used. In addition there are further sources for specific domains, such as registers of the central bank in the areas of financial institutions. For the government sector various other sources have to be used, such as the registers of schools and kinder gardens or the staff information system of central government. The four main sources for the units in the market sector are:

- Register of companies

- Tax register
- Register of employers (social security register)
- Register of the Federal Economic Chamber

11. The reason for using four different data sources in parallel is just that only the “sum” of all sources provide the data necessary. The register of companies provides the legal name and form for all corporate enterprises, the tax register - as the most comprehensive one - ensures full coverage of the market units, the social security register provides information of the number of employees, and the register of the Federal Economic Chamber delivers information on the trade licences issued for a single location.

12. The register of companies covers only corporate enterprises and some of the incorporated enterprises when they are above a certain size class threshold. It is the official company register and publicly accessible. For the business register the company register provides the official company name and the legal form, basic information on the demography and on ownership. It covers all the data that is required under commercial law. It is run by the commercial courts.

13. The tax administration register is the most comprehensive administrative register. It contains basic enterprise profile data such as name, address, legal status; in case of sole proprietorship firms also date of birth and sex of the entrepreneur. The tax register information allows linking to the VAT-declaration data base and deriving from there the turnover data. VAT-declarations have to be made on a monthly basis by enterprises with an annual turnover of more than € 100.000 and for all enterprises on an annual basis above a threshold.

14. The social security register covers data of all units employing persons, it thus covers not only enterprises but also non-profit and government institutions. The social security system in Austria is structured by kind of employment and by regional criterion. The Main Association of the Austrian Social Security System collects the data of the individual social security institutions. The data most relevant for the business register are the employment data which after establishing the employer link are forwarded to the business register on a monthly basis.

15. The last main source is the membership register of the Austrian Federal Economic Chamber. In Austria membership is compulsory. The chamber of commerce is organised under public law. It is also federally structured. Enterprises are members in the respective trade branch of their region. Thus enterprises can hold more than one membership. Trade licences are required for each kind of activity according to the chamber of commerce classification and for each location separately. This source is the only one that provides information for the level of local units. However, holding a trade licence does not mean that economic activities are actually performed. Not all economic activities are covered by the Federal Economic Chamber. For example, the traditional free lance activities (architects, doctors, lawyers, etc.) have to be member in special chambers of commerce.

III. EXAMPLES OF QUALITY ISSUES

A. No unique numerical identifier

16. The various administrative registers described above use their own identification system which leads to the result that each enterprise has to deal with several identifiers for their reporting requirements (and also a different one for the statistical reporting obligations). As long as the administrative databases are not linked, the missing unique numerical identifier is no problem for the administration. However, for the statistical use it has the consequence that extensive data matching processes have to be applied in order to match the records of the sources used. This is principally no methodological problem as we use the quite efficient bigram-method. Of course, parsing, standardisation and similar pre-checks of the databases are required. However, not all records can be matched and thus additional clerical work is necessary which in the past was quite laborious when the matching work had to start from scratch. For a description of our matching procedures see Haslinger, 2004.

17. Using these procedures we were able to link these four main administrative sources more or less exhaustively. However, each month the exercises need to be repeated with the new data deliveries of the administrative registers in order to link the new records. As the administrative data are not updated in the same way and time frame, it has also the consequence that the linking fails in cases where in one source a new record is delivered but not in another source. However, this aspect has to be seen with a view to the problem of the deviating unit definitions and continuity.

B. Different definitions of units

18. The register of companies and also the tax register have the legal units as their register basis. Until recently this was not the case in the social security register as this register is run regionally decentralised. A legal unit with locations in two or more provinces has two or more identification numbers. However, the profile data of the different units in the regional registers do not necessarily coincide so that the same legal unit/enterprise in the central social security register is displayed as more than one unit. Furthermore, in the regional registers a unit can have more than one employers account differentiated by social security legislation. Even then these employers' accounts may not have the same units' profile.

19. The situation is similar in the registers of the chamber of commerce. The only difference is that the federal chamber itself merges the various regional memberships and runs an enterprise register at the central level. This is of much help for the statistical business register; however, their register is not linked to any of the other administrative registers.

C. Different continuity procedures

20. The continuity rules in the administrative registers are also quite different. Changes of the legal form disrupt continuity in most of the administrative registers, however, to different extent. Clearly, a change of the legal form changes the legal and financial environment. Other liability, taxation etc. rules are valid so that for the administration the unit is to be treated as a new one. This situation would create no problems for the statistical use even if the statistical rules may be quite different as long as the information on these changes is recorded in a way so that the respective units can be traced back and the units before and after the change can be linked. This is, however, only the case in the register of companies and even there the relevant information needs to be analysed on the basis of verbal descriptions.

21. Thus, a new identifier code in one of the administrative registers does not necessarily mean that it is a new unit from the statistical point of view. For the units not registered in the company register (which cover only about 40 % of all units) no explicit information is available as to whether it is a newly created unit or not. Thus, data matches and clerical checks are required.

D. Estimation needs to supplement missing data

22. An important example where data editing and imputation methods need to be performed in the business register is the estimation of employment data for the level of local units. The social security register can only provide employment data at the level of legal units/enterprises. However, also employment data for the level of the local units need to be recorded in the business register. The only statistical source for such data is the structural business survey. However, these data are available 18 months after the reference year, provide only annual information, and do not cover all activities and only enterprises above a certain threshold. Therefore, an estimation procedure was elaborated which allocates the employment data of the social security register for each enterprise to the various local units where the data from the structural business survey are used as exogenous variables. For new local units, not covered in the structural business survey, the number of employees is estimated on the basis of average employment ratios calculated for each enterprise. This first estimate is then treated as an exogenous data set for the allocation step. The final result is then that the number of employees of each local unit sums up to the number of employees for the total enterprise as given by the social security data. This procedure is repeated every month using the newest monthly social security information.

23. A similar estimation procedure is currently being developed for the missing turnover data, either because of the quite long time lag in the tax declaration data availability or because enterprises can declare their VAT data for the enterprise group as a whole. This estimation task will probably be more complex as the one for the employment data.

E. Other quality issues

24. The above described quality problems are inherent to the system, and thus only a change of the system will diminish or abolish these deficiencies. In addition, there are the quality problems of the content of the data records. Examples are that changes of enterprise names and addresses may not be updated; address data do not conform to the official address records, etc. Such problems can only be reduced if the daily procedures in the administrative system are improved. However, there is sometimes no real pressure to improve the database. For example, as long as an enterprise transmits its monthly social security contribution, there is no real need to update the name or address of the enterprise in the administrative database.

25. The administrative database for the payslip information has been enlarged recently in order to provide data on local units. Every employer has to report to the tax and to the social security authorities for each of his/her employee the relevant data on the payslip. This has to be done when an employee leaves the enterprise and for all employees end of the year. The data are to a very high percentage delivered in electronic form via a special internet portal. Since last year the employers are obliged to deliver also the address of the local unit where the employee was working on 31 December. This is thus a very important new database as it provides information on local unit addresses and how many employees have been working there at least on 31 December. Furthermore, this data are very quickly available as the employers have to deliver them within two months after the end of the reference year. These data are also quite relevant for wages and salaries indicators. However, these data are suffering from the address quality problem. Addresses, such as name of streets, are written differently and differently abbreviated, etc. The result is that only two thirds of the data records can be merged with the data in the business register automatically. Currently we are analysing these data in order to increase the automatic matches.

IV. IMPROVEMENT STRATEGIES

26. Creating awareness of the data quality problems is a first step in trying to achieve improvements. Compared to the administrative units, the NSIs have a privileged status as they have access to various (all) administrative data and are allowed to match and compare the data for statistical purposes. Thus, the NSIs clearly recognize the quality problems. Any improvement in the quality of the relevant administrative data improves the situation of the NSI. Such an improvement may not just result in an increase in the quality of the statistical data; it will decrease editing needs and certainly lower the workload of the NSI.

27. Quality improvements in the administrative data can be achieved by various single steps without any great costs. The most important example of this kind concerns the address problems. There is an official address register which even is run by Statistics Austria as part of their buildings and dwellings register. However, Statistics Austria is not the data owner and is not allowed to provide these data to other users. Data may be purchased from the cadastral authority. Thus, the use of these data is not widespread. However, address data should be viewed as public good and available online free of charge. This would increase data quality to a very high degree, especially also the new enlarged payslip data. Furthermore, not only the administration would profit from broad and easy availability of the official address data but also the enterprises themselves as addresses - even today in our internet society – are indispensable data.

28. Another example is the NACE coding of the businesses. Both the social security and the tax authorities are classifying the units according to NACE. In addition, for certain tax declarations the enterprises are requested to provide self-coding information. The NACE code is important information for the business register, especially for the new enterprises which needs to be entered into the register on the basis of the administrative information. However, the NACE code provided by the administration is

partly of low quality which is not astonishing as neither the enterprises nor the staffs of the social security and tax administration are classification experts and classifying according to a statistical classification is not viewed as their most important task. It would thus be better if the enterprises would provide information on their activities performed and this information is forwarded to the NSI which transforms this information to an NACE code and delivers back the code to the administration. However, in order that this procedure works it is necessary that the units in the administrative registers are linked and unambiguously identified.

29. The Austrian case is to be characterised as a decentralised system in the sense that each of the administrative registers is run and organised separately. Moreover, some of these registers are run regionally decentralised. Under this situation each administrative system would need to make its own efforts, without achieving much synergy effects. It would therefore be more advisable if these administrative registers would be linked together or really centralised. It would be a big step forwards if the tax and social security register would be linked or run together. The reorganisation of the social security register which came into effect in 2008 would be a good basis now for such integration. For the integration of the units in the decentralised social security registers Statistics Austria provided its matches as a starting basis. However, other administrative registers should be considered also.

30. An integration of the main administrative registers and provisions for its common maintenance would certainly be a great step forward. This would also reduce the workload both of the administration and the enterprises as currently maintaining the administrative registers is done by each of the administrations separately and the enterprises have to report the same data to various administrations. This should create a win-win situation both for the administrative authorities as well as for the enterprises. It would anyway be a win situation for official statistics as any improvement reduces the workload of data matches and clerical data editing in the business register and will increase also the quality of the statistical data. In our view, register integration would be more efficient for all partners, including the businesses, than improvements in the various registers separately.

References:

Alois Haslinger (2004): Data Matching for the Maintenance of the Business Register of Statistics Austria, Austrian Journal of Statistics, Volume 33, Number 1&2, 55-67.

Norbert Rainer, Thomas Karner (2007): Measuring and improving the NACE coding in the business register, Paper presented at the 20th International Roundtable on Business Survey Frames, Wiesbaden, 21 – 26 October 2007.