ESSnet Administrative data, WP9
Use of administrative data in the production of business statistics
Seminar in Rome, March 2010

The integrated system of editing administrative data for STS in Germany

by
Robin Lorenz
Federal Statistical Office of Germany

1 Introduction

When in 2003 the Administrative Data Use Act was passed in Germany, the statistical offices in Germany got monthly access to two administrative data sources: the turnover tax files held by the tax authorities of the German states (*Länder*) and data from the Federal Employment Office on persons in employment liable to pay social insurance contributions and on persons with minor employment. The Administrative Data Use Act obliged the Federal Statistical Office of Germany (FSO) and the statistical offices of the *Länder* to examine the suitability of the data for different purposes of economic statistics. The focus was on tests in the field of short term statistics (STS). Extensive testing pinpointed promising sectors for use of the new data sources. It also turned out that the administrative data does not meet the demands of STS completely, but that the weaknesses of the administrative data could be eliminated in sufficient quality by estimates and additional information from the business register in some fields of economy. In consequence, in the fields of crafts the survey of quarterly 40.000 mainly small enterprises could be fully replaced by the use of administrative data, in the field of other services the response burden of another formerly quarterly surveyed 40.000 enterprises could be drastically reduced by a multiple-source mixed mode design combining a statistical survey of the 4.000 largest enterprises with administrative data for the small and mid-sized enterprises. For the STS-other services and STS-crafts the administrative data is already in live operation since the second quarter of 2007 respectively the first quarter of 2008. In other fields such as building completion and installation, wholesale trade as well as sales and repair of motor vehicles administrative data can be used in various mixed designs that still have to be realised.
The paper in hand gives an overview of the Admin data sources and the methods used for editing and adjusting the data for STS-purposes by integrating additional information from the BR.

2 Data Sources

2.1 VAT Data

2.1.1 Content of the VAT data set

The turnover data is provided by the fiscal authorities of the 16 states *(Länder)*, which get the data from the VAT prepayment notice and payment procedure. Under the current VAT legislation all enterprises above an annual VAT limit of 1.000 Euro in the previous year have to submit their turnover to the tax authorities either 10 days (20 -25% of the total turnover) after the end of the reporting period or - with a permanent extension - 40 days (75-80% of the total turnover) at the latest after the end of the reporting period. The reporting period is for the bigger enterprises (above 7.500 Euro VAT in the previous year) the month and for the smaller ones the quarter. About 50 % of the enterprises are monthly taxpayers representing more than 90 % of the turnover in most NACE -sections.

Every month, around the 20th of a month, each of the 16 fiscal authorities delivers a file to the FSO (altogether about 3.5 million records) containing all the current VAT prepayment notices, which have been processed in the tax agencies since the last delivery. If all enterprises met the legal deadline of 40 days for their VAT notices, the VAT data would be complete for producing STS with a timeliness of t+60. In practice about 10% of the enterprises (usually representing less than 10% of the turnover) are late with their VAT reports. The missing reports are delivered in one of the following deliveries. Nevertheless, the production of quarterly results at t+60 is possible as well as monthly results by applying appropriate imputation methods for the missing values. Moreover first efforts were made to achieve a timeliness of t+30 for quarterly turnover results. Even though 30 days after the end of the quarter the data referring to the third month of the quarter is rather incomplete, the first two months of the quarter are sufficiently covered (t+60 and t+90 respectively). Altogether this leaves a share of missing turnover of about one third of the total quarterly turnover, that could be estimated. But so far the estimations did not lead to satisfying quarterly results at t+30. In any case monthly turnover results at t+30 are out of reach at the moment.

Apart from the turnover values and their reference periods the data set of the tax authorities contains some other variables, which are essential for editing the data: the tax number, the former tax number (in case of a change) and the intra-community trade number as identifiers, the address of the enterprise and an activity code (NACE).

2.1.2 Deficiencies of the VAT Data

The suitability tests of the data for STS-purposes revealed that the VAT data has some weaknesses with regard to the statistical requirements:

- *Allocation of the activity code:* The activity code in the VAT data does not entirely meet the statistical requirements. This is less a problem of standardisation, because the classifications

used are more or less the same, but the tests showed that the allocated codes for a single unit differ depending on the source. A comparison between the statistically surveyed industrial classification codes (NACE) - predominantly from annual statistics - and those from the administrative sources shows huge discrepancies. On the two-digit level of NACE about 20% of the units have different codes in the administrative sources and in the surveys. On the three to five-digit levels the shares of units with deviant codes rise up to 50% and more taking the surveyed codes as a reference.

- *Deviations in definitions:* The definition of turnover within the tax prepayment notice differs in some respects from the statistical definition of turnover. Some extraordinary receipts such as rental income for company-owned machinery, dwelling or land used by third parties or sales of land or used machines are not included in the statistical definition, but they are included in the tax prepayment notice under the same heading ("non-taxable goods and services with no deduction of tax prepayment") as are statistically relevant goods and services such as sales of stamps in the sector of postal activities. In addition, in the tax legislation a number of enterprises can be combined in an integrated VAT group. The internal turnover between the members of a VAT group is not taxable.
- *Deviations from required statistical units:* In the case of the already mentioned VAT groups only the controlling company will report the total turnover to the fiscal agency. Unfortunately, top-selling enterprises are often organised as a VAT group and the data suppliers do not provide any information about the division of the turnover among the different enterprises in the group. The total share of VAT groups varies in different sectors of the economy. At the federal level over all branches VAT groups make up 45% of the turnover, while for example their share on the two-digit level (NACE Rev. 1) in the service sector can reach up to 85-90% in the fields of air transport and communications, or 25% in the crafts sector.

2.2 Employment Data

2.2.1 Content of the employment data set

Data on employees is provided by the Federal Employment Office (Bundesagentur für Arbeit), that receives the data from the integrated reporting procedure for social insurance (health insurance, pension institutions). Every month, around the 15th of a month, a file is delivered to the FSO containing for each local unit (nearly 3 million) the number of persons employed who are liable to pay social insurance contributions and the number of those with marginal employment. The numbers of employees in one file refer to three different reference days. For example the data delivered mid-July refers to end of May, end of April and end of January. The early data for May can be used to produce STS with a timeliness of t+60, the data for April and January for revised results (t+90 respectively t+180). With an increasing time-lag the completeness of the data rises from about 90% of the relevant ins and outs of employees in the t+60 data, to 95% in the t+90

data and almost 100 % in the t+180 data. Apart from the numbers of employees the file from the Federal Employement Office also contains the local unit number as identifier, the address and the NACE-code of the local unit.

2.2.2 Deficiencies of the employment data

The employment data of the Federal Employment Office is quite a reliable data source, but there also some deficiencies in the data:

- *Allocation of the activity code:* The quality of the activity code from the Federal Employment Office seems to be slightly higher than the quality of the activity code from the fiscal agencies. But also in the employment data the activity code does not entirely meet the statistical requirements. The tests revealed big deviations between the activity code from surveys (which is considered to be correct) and the activity code in the employment data.
- *Deviations in definitions:* There is also a different delimitation in the "employees" variable compared to the statistical requirements: The administrative source provides information about those who are liable to pay social insurance contributions and those with marginal employment. But it does not cover the self-employed, (unpaid) family workers, civil servants and slightly short-term employees. Altogether the administrative source covers about four-fifths of the working force.
- *Deviations from required statistical units:* The Federal Employment Office delivers data for local units, whereas in the sectors where the admin data can be used for STS-purposes, it is the enterprise that is asked for. In many sectors on the two digit level more than 30 % of the employment can be found in units where the activity code of the local establishment does not correspond to the one of the enterprise.

2.3 Business Register (BR)

In the system of editing the Admin Data for STS the BR can be considered as an additional data source. The BR is updated regularly with annual data from the tax authorities and the Federal Employment Office as well as other structural data from sources such as Chambers of Commerce and Chambers of Crafts. In addition information from surveys is integrated into the BR.
Once a year an extract from the BR is taken for STS-purposes, which is used for one calendar year and which contains all the additional information required for the editing process. For example the BR-extract taken at the end 2009 is valid for all reporting months and quarters in 2010.The data in this BR-extract refer to 2007, so there is considerable time lag between the current Admin data an the BR-information. The information in the BR-extract is particularly used to partially compensate the deficiencies of the Admin data, but also to control the editing process. The BR-extract contains the following important information:

- *Identifiers:* Since the two data sources for STS, the data tax authorities and the Federal Employment Office, are also the main sources for the BR (even though on an annual basis), there are common identifiers between the BR and each of the monthly data sources: the tax number for the VAT data and the local unit number for the employment data.
  In particular in the mixed mode designs combining a survey and Admin data, it is essential to have the possibility to link the monthly data with the BR-extract and to identify the surveyed enterprises in the Admin data by these identifiers. Otherwise the enterprises that are surveyed could not be excluded when editing the administrative data and the turnover and the number of employees respectively would be counted twice, in the survey and in the Admin data.
- *Activity code:* The insufficient quality of the NACE codes in the administrative data sources can partly be treated by using the NACE Code stored in the BR. Due to the sampling techniques, bigger enterprises are usually covered better by statistical surveys than small and medium-sized enterprises, so that statistically surveyed NACE codes are often available for a high percentage of the turnover and employees. Even in case of enterprises without surveyed information the NACE-Codes in the BR are of higher quality because due to the NACE revision many activity codes were verified in the BR.
- *Information on VAT groups:* With the help of information from other sources stored in the business register the turnover of VAT groups can be broken down to the single enterprise for statistical purposes. In a multiple regression model the business register annually estimates the turnover for enterprises that are part of a VAT group. These estimates are used to form a key to split the monthly submitted turnover of the controlling company.
- *Relation between enterprise and local units:* This information, which can only be provided by the business register, is important in case of enterprises which are active in more than one state (Land). These multi-state enterprises often make a significant share of the total turnover, as for example in the field of air transport (93 %). According to the STS regulation the results must be delivered referring to enterprises. The data on persons employed however is submitted referring to local units. By using the relation between enterprises and local units it is possible to allocate every local unit to the economic activity of the belonging enterprise. In addition the relation between enterprises and local units is used to break down the turnover values to the level of the German states.
- *Flag for crafts enterprises:* For the STS-crafts the information on whether a unit belongs to the craft sector can only be taken from the business register.

3 Editing the Admin Data

The following section gives a description of the most important steps in the editing process of the Admin data for STS.

3.1 Data Linking

When the monthly files from the 16 tax authorities of the states arrive the first step - after a formal check of processability – is the linking to the turnover database, a relational database on a MySQL-platform, where the turnover data from the previous deliveries is stored. The single records in the files are linked to the stock of the database in a complex algorithm using as identifier the tax number and – in case the tax number does not match - the former tax number and the intra-community trade number. Linking problems occur, when e.g. an enterprise moves from one *Land* to another, so that another fiscal agency becomes responsible. The tax number of the enterprise changes in this case and not always the information on the former number is submitted. In this case a second record in the database for the same enterprise is attached, i.e. the current turnover can not be linked to the turnover of the previous months/quarters. This can lead to incorrect results, e.g. when for the current turnover a possibly other activity code is used than for the turnover of the previous periods.

The employment data is stored in the central database (MySQL) on employees, which is updated monthly by the current file from the Federal Employment Office. During the import of the current file the single records are linked to the stock of the database by the local unit number as identifier. As the local unit number is the only identifier there are few linking problems.

3.2 Combining Admin data and BR-information

After the transmission and the import of the data in the turnover database and the employment database respectively extracts out of the data bases for STS-purposes are taken containing all the variables needed for the following editing process. Afterwards both the monthly STS-extracts from the turnover database and those from the employment database are linked to the currently valid BR-extract generated once a year. As in the BR both the identifiers from the tax authorities (tax number, former tax number and intra community trade number) and from the Federal Employment Office (local unit number) are stored, the linking algorithm is quite similar to the one used for updating the data bases. However the matching rates between the BR-extract and the current STS-extracts are quite low (55%of the units in the VAT data, 70% of the units in the employment data, however representing a much lower percentage of the values). One reason for the low matching rate is that very small VAT payers which never exceeded a threshold of 17500 € turnover per year are not included in the BR at all. Furthermore the currently valid BR-extract has - compared to the current admin data - a time lag of more than two years. This means, relatively

young enterprises and local units set up after the reference year of the BR-extract are not contained in the BR.

For all the units where the linking was successful the additional BR-information is used to compensate the weaknesses of the Admin data. The activity code from the Admin source is replaced by the activity code from the BR. In case of local units of multi-state-enterprises the activity code of the enterprise is assigned to every local unit. Moreover the BR-information is used to split the turnover of VAT groups on its members (see subsection 3.5). The remaining units in the Admit data, which could not be linked to the BR-extract, go straight into the results using the activity code from the Admin source. It might be the case that the enterprise is the reporting head of a VAT group or that the activity code is incorrect, but there is no other reasonable way of treating these units.

3.3 Automated plausibility checks

To secure reliable results automated plausibility checks in the VAT data are conducted on a micro level. Implausible turnover values can result from errors while processing tax reports at the agencies (e.g. errors from scanning) or from sales of assets particularly in case of dying enterprises. To avoid genuine data values to be altered and trends in the data being missed the detection of implausible values currently concentrates on extreme outliers. In order to identify these outliers the data is divided into five different classes depending on the median of the last six submitted monthly turnover values (last four quarterly turnover values). The higher the median the stricter is the plausibility check. For example enterprises with a median of more than 1.000.000 € per month are marked as outliers when they tenfold the median turnover. On the other hand small enterprises with less than 10.000 € are allowed to increase their turnover a thousand fold. Once an outlier is detected the same method as for missing values is used to substitute the value.

3.4 Treatment of missing values

The studies showed that missing values occur in the VAT data to an extent where imputations are essential as otherwise the results are unsuitable for economic purposes. As mentioned before, the data at t+60 are incomplete. On average at the time of receiving the data from the fiscal agencies about 90% of the monthly payers have reported. But the proportion of the monthly missing turnover varies strongly since missing data occurs even in the case of large enterprises, sometimes the all dominant enterprises on the relevant level of classification for evaluation. To resolve the problem of missing VAT data the change rate of those enterprises from the same sector (two-digit level) that have reported is taken and applied to the turnover values of the previous months/quarters. Thereby a distinction is made between big enterprises, where the

change rate of all enterprises is used and small and mid-sized enterprises where the change rate of only the small and mid-sized enterprises is used. The method proved to be suitable for nearly all enterprises with the exception of outstanding market leaders.

Missing data can also be found in the employment data. Ins and outs are known after a waiting period of two months to a considerable extent, but only after about six months are the data almost completely updated. Thereby a systematic effect occurs in the way that in the majority of the cases two-month values lie beneath the six-month values meaning that the outs seem to be reported faster than the ins. But at the level of an individual enterprise or a local unit no indication is available in order to determine whether there are changes or not. The impact of missing data can therefore only be reduced by estimates at aggregate level. So far, various methods of adjustment have been tested and the tests are ongoing as a “best practice” has not yet been found. Nevertheless, in some economic sectors reliable results can be produced by using the original data without estimations for missing records. In general the revisions on the aggregation level of the two-digit codes are acceptable. Critical are those sectors where either dominant enterprise determines the figures or where a relevant share of minor employed contribute to the change rate. In particular the last named have proved to be less reliable.

3.5 Splitting the turnover of VAT groups and multi-state enterprises

In order to avoid that the current total turnover of a VAT group is allocated to the economic activity of the controlling company, the turnover is split to the members of the VAT group. For this a key derived from the estimations of the annual turnover of the VAT group members in the BR is used. The key indicates the percentage of the turnover that is allocated to each VAT group member and is applied to the current monthly or quarterly turnover of the VAT group.
Through this method e.g. an increase of the monthly total turnover of the VAT group by a certain percentage leads to a corresponding increase of the estimated monthly turnover value of each member by the same percentage. However, the assumption that all VAT group members behave in a similar way is only realistic, when VAT groups have a high degree of homogeneity on the demanded level for evaluation, i.e. the VAT group members belong to a high extent to the same industrial sector (and state). In this case the risk to transfer business trends in or from other sectors is rather limited. On the federal level the homogeneity on the two-digit level of the activity classification is usually sufficient but on the regional level it is not always the case so that business trends from other sectors and other German states can be reflected in the figures of one state. Another weakness of the splitting method is that the key refers to the reference year of the BR, which is at the time of editing the Admin data already past for two years. Changes in the composition of the VAT group since the end of the reference year are not (or only to a small extent) taken into account.
For the production of regional results the turnover of enterprises that are active in more than one state (multi-state-enterprises) must be broken down. Again a key from the BR is formed which

bases on the number of employees in each local unit of the enterprise and is applied to the current turnover value.

3.6 Calculation of growth rates

The steps described in the previous subsections are conducted separately for the VAT data and the employment data. The results of these steps are the so called enhanced STS-extracts for turnover and persons employed respectively. For both variables the enhanced STS-extracts are generated every month. Covering the whole economy they contain the complete micro data including imputations and the most reliable NACE code available, so they are the basis for all further calculations in the STS production process. For each reporting period the absolute aggregate values for certain economic activities can immediately be calculated out of these enhanced STS-extracts. But when it comes to calculate growth rates with Admin data one has to account for that the two absolute values involved in the calculation of the growth rate have to be computed on the basis of the same BR-information. So the annual growth can not be calculated by taking the ratio of the absolute aggregate value of the reporting period (month or quarter) in a certain sector and the corresponding value of the previous year as it is usually done in Germany when STS are produced with surveyed data. The reason is that these two values were generated by using BR-information from two different years. There might have been changes in the activity codes of the enterprises, in the composition of VAT groups and accordingly the keys used to split their turnover, and in the relations between enterprises and local units. These changes can have substantial effects on the absolute values, so that aggregate values basing on different BR-extracts are not comparable. The same problem occurs when calculating monthly or quarterly growth rates. Even though the growth rates within a calendar year can be computed without any problems, the growth rate at the transition to the following calendar year (i.e. December to January, Quarter 4 to Quarter 1), when the next BR-extract is valid, is not usable as the aggregate values involved base on different BR-information.

To resolve this problem, the data of the last period of a year (month/quarter) is edited twice, one time using the old BR-information and another time using the new BR-information. For the November-to-December (Quarter 3 to Quarter 4) growth rate the data with the old BR-information and for the December-to-January (Quarter 4 to Quarter 1) growth rate the data with the new BR-information is used. Having calculated the monthly/quarterly growth rates by this means the annual growth rate is calculated by forming the series of 12 monthly growth rates or 4 quarterly growth rates respectively.

Example: let $Q2_Y$ be the aggregate value of the second quarter of the current year and also the current reporting quarter, then the annual growth rate GR would be

$$GR = (Q2_Y/Q1_Y * Q1_Y/Q4_{Y-1,\,newBR} * Q4_{Y-1,oldBR}/Q3_{Y-1} * Q3_{Y-1}/Q2_{Y-1} \quad - 1\,) * 100$$

This expression cancelled down and rearranged yields

$$GR = (Q2_{Y}/Q2_{Y-1} * Q4_{Y-1,oldBR} /Q4_{Y-1,\ newBR} * \quad - 1) *100$$

So the annual growth can be seen as the ratio of the absolute value of the reporting period and the corresponding value of the previous year adjusted by a coefficient that indicates the effect of the change of the BR-information. If monthly results were required this BR-coefficient would refer to December instead of quarter 4.

4 Conclusion

The previous sections gave an overview of the Admin data sources and the system of processing the data for STS-purposes by integrating additional information from the BR. As mentioned before the data is at present used for STS -other services and STS-crafts, both statistics with a timeliness of t+60 and the quarter as reference period. For these two purposes the system of editing the Admin data has proved to generate reliable results in a sufficient quality at the federal and with some exceptions at the state level. In order to reduce these exceptions in the results for these two STS an additional safety net has been implemented by means of additional manual checks in case of individual VAT data values strongly affecting the results. But only few units with a weight of 2 % in a state on the level of evaluation and abnormalities in their turnover values are affected by these checks.
Even though the editing system runs quite well at the moment, there are some areas of further development. When the mixed models in the fields of wholesale trade and sales and repair of motor vehicles will be implemented, both STS with a timeliness of t+60 and the month as reference period, the methods to split the turnover of quarterly payers on the months should be reconsidered. Moreover the production of quarterly turnover results at t+30 as well as the treatment of missing values both in the VAT and particularly in the employment data are areas of further research.