# Calculating Business Demography Statistics based on Administrative Data

Alois Haslinger, Norbert Rainer
**Statistics Austria**

Paper prepared for the 2009 European Establishment Statistics Workshop
7 – 9 September, Stockholm, Sweden

**Abstract:** Statistics on business demography comprise data on the population of **active enterprises**, on **births of enterprises**, their **survival**, and on **deaths of enterprises.** The main characteristics refer to number of employment, size class, legal form and of course economic branch. Business demography statistics play a key role in forming a basis for political decisions and analyses: newly born enterprises stimulate the economy by creating new jobs, competition and structural change. **Derived indicators** such as births, deaths and (two-year) survival rates also form part of the EU structural indicators, which are used to monitor the progress made in the Lisbon process aimed at boosting growth and employment.

Austria first took part in the harmonised data collection for the 2005-2006 reporting years. The production of these figures could not be done – as in other countries – solely on the basis of the Austrian business register; additionally, administrative registers had to be used. This paper highlights the **methodological problems** in creating a synthetic data base by combining data from the tax register, the social security register and the business register. Since these 3 registers have **no unique identifier** data had to be merged by **record linkage** procedures. Other problems we had to manage: the **determination of time** of birth and of death, the distinction between **real births** and unreal ones (split-offs, break-ups, changes of legal form,…), the treatment of **different monitoring concepts** in the administrative registers, and so on.

## 1 Basic concept of business demography statistics

Because of the growing demand for statistical information which – by focussing on the dynamic component of the business universe – could supplement traditional data on the structure of the businesses in a given period of time or at a certain point in time Eurostat together with the member states has started to develop a harmonized concept of business demography statistics. After a period of pilot phases with voluntary contributions of member states, a legal basis for business demography statistics was adopted in 2008. In 2009 the member states were for the first time legally obliged to deliver business demography statistics data to Eurostat.

Around 2005 also the OECD joined in the development of business demography statistics. However, their approach started from the development of a set of various indicators to describe the “entrepreneurship” of a society; demography data were just a part of these, even if of course an important one. With the OECD joining, a broader view of comparability came into play. It was recognized that OECD countries outside Europe often exclude the very small enterprises in their demography statistics whereas the European concepts do not exclude them in principle. This has led to a second basic concept of business demography, namely, the “employer enterprise demography”. Under this concept only enterprises with employees are considered. The consequence of this different conceptual approach was a duplication of the basic definitions and a second conceptually different business demography data base. In addition to the definition of an enterprise birth, a definition of an employer enterprise birth had to be added. An enterprise may be born as an employer enterprise already or may be born without any employees. An enterprise born without employees may become an employer after some time and will then be counted as an enterprise birth in the year when it started to engage (an) employee(s). So, such an enterprise is born twice in different periods of time. The same logic applies to survival and death.

Like in other statistical domains there are quite strict definitions of the key variables in business demography also. This can be illustrated by the concept of enterprise birth. The aim is to produce data on the creation of new enterprises that have started from scratch and that have actually started economic activity. The creation of a new legal enterprise is not enough to be considered as an enterprise birth. An enterprise creation can be considered as enterprise birth if new production factors, in particular new jobs, are created. And they have to be created without the involvement of other enterprises. New enterprises created by break-ups, mergers, split-offs, take-overs or restructuring are not considered as newly born enterprises. Also the change of legal form, the change of location or activitiy is not to be counted as newly born enterprises. The same is true in the case of reactivated enterprises if they restart activity within 2 calendar years.

It is therefore necessary to know whether an enterprise is active or not, and whether newly created enterprises were created on their own or not. Active enterprises are defined as having either turnover or employment at any time during the reference period. Provided that appropriate data on turnover and employment are available, this condition can easily be tested. However, the second condition is more difficult to apply as for each newly created enterprise it should be known whether is was not part of a merger, break-up etc., or just a change in the name, legal form, location or activity. In case of two different units which have the same economic activity and business location in common, but have a different name it just could be continuation of the business by the son of the former entrepreneur, in which case the newly created legal unit is not considered as a new enterprise. Analogous concepts are valid for the survival and death data. The challenge of business demography statistics is therefore the knowledge of the predecessor or successor relations of legal units created or closed, respectively.

## 2 The Austrian Business Register (BR)

Since the Federal Statistics Act 2000 has become effective, Statistics Austria receives monthly copies from at least four main administrative data sources or registers (AR) all covering information about Austrian enterprises or subunits of them (company register, tax register, social security register, chamber of commerce register). This information is not only used for the updating and maintenance of the Business Register but also as a partial or total surrogate for censuses and surveys. The idea is to reduce the response burden on businesses as much as possible. If any information which is needed for statistical purposes is already stored somewhere else in the public administration, then Statistics Austria should use that information instead of surveying the enterprises or the citizens again.

The BR of Statistics Austria serves as an instrument for all surveys addressed to enterprises. It has been designed according to the EU requirements and contains about 390.000 enterprises including their establishments and local units. All in all it has about 530.000 active and 450.000 inactive units (i.e. enterprises, establishments and local units) and has been in operation since mid-1995. Currently, the BR does not distinguish between a legal unit and an enterprise, like the registers of most EU member states.

## 3 Reasons why BR alone is not sufficient for business demography statistics

Conceptually, the BR should serve as a data base for demography statistics as the demographic characteristics are part of the variables of the BR system. However, there are some conceptual and practical reasons why demography statistics cannot solely be derived from the BR. One important reason is the threshold that is applied in the BR.

This threshold was lowered in the past years but is yet not fully implemented at the envisaged level. However, in the future the thresholds applied should be the same in both domains.

A second reason is that for demography statistics it is necessary to provide data on employment of active, newly born and death enterprises. Thus employment data from the social security sources have to be used. Similarly, for the decision whether an enterprise is active or not, data on turnover from the tax register are necessary. While both administrative registers are to a very high degree linked to the BR, the original sources still have to be used, also because longer time series information is not directly available in the BR.

However, the most important reason is the problem of data on the relations between legal units over time. From administrative registers information on the status of new units is very scarce. Only from the information in the company register conclusions concerning the status of a unit can be drawn. For such units the demography variables in the BR are normally of good quality. However, most of the newly born enterprises are very small and have the legal form of sole proprietorship and are thus not covered by the company register. Furthermore, more than half of the newly born enterprises have no employees. Thus, a high share of units is only covered by the tax register, without any information on predecessor or successors, if appropriate.

It is evident that administrative data do also not fulfil the requirements of the demography concepts; if that had been the case, the BR would already have integrated these data. Nevertheless, being aware of the insufficiencies of the administrative registers, it was decided to base the compilation of the demography data on administrative data that should supplement the BR data base and to adjust the administrative data to the demography concepts by applying statistical methods.

## 4 Supplementary administrative data used for compiling Business Demography Statistics

The following two administrative registers are used for the compilation of Austrian business demography statistics:

**Social Security Register (SSR)** Each Austrian employer has to register his employees in one of about 20 different social security insurance institutions. It depends on the region where the establishment is situated and on the kind of contract of employment, which insurance institution is responsible for a certain employee. It is possible that the employees of an employer are registered at two or more different insurance institutions (e.g. if an employer has local units in more than one province). For each employment of a certain person with a certain employer, a data record is created in the responsible social security insurance institution containing, among other things, the social security number of the person, an identification code of the employer, a code of the insurance institution, sex of the person and the kind of contract of employment. Because of the federal organisation of the social insurance system most of the institutions are member of an umbrella organisation called Main Association of Austrian Social Security Institutions. The Main Association has access to the employment registers of its members and additionally maintains a register which holds one record for each combination of insurance institution and employer. This register of employer accounts (SSR) contains the name of the employer, postcode, address, place of the enterprise, NUTS-3 and NACE codes, and contains about 350.000 units. The units of this register are not comparable with the units of the BR.

**Tax Register (TR)** The most comprehensive administrative register is the register of the tax authorities. It contains basic information like name and address, date of birth, sex and civil status (the last 3 primarily for persons), legal status and economic

classification according to NACE (primarily for enterprises) for about 6.8 million taxable units (persons, business partnerships, corporations, institutions, associations, etc.). The coverage of this basic tax file is much broader than that of the BR. To get a sub-file from the basic tax-file which is comparable with the BR it has to be merged with the turnover taxation file from the tax authorities containing about 600.000 units. Statistics Austria receives both files of the tax register monthly. Both files include a unique subject identification key which can be used for merging the two files. The turnover taxation file contains all units from the basic file which did a turnover tax return in at least one of the last 3 years. A problem is the lag between a fiscal year and the time, when all units have received their tax assessment and data get available. This lag is about 2-3 years. To get a definite value of total turnover in 2007 one has to wait at least until mid 2010. The merged file of that date covers units which are no longer active at present. On the other hand, in the merged file units of the BR are lacking which are not liable for turnover taxation (e.g. turnover from medical activity). Nevertheless, most of the units of this merged file are in accordance with the enterprises of the BR. From the start in 2003 on, the problem of the time lag of turnover returns has been reduced, because now each enterprise with a turnover above 100.000 € in a year has to provide a monthly turnover tax advance return beginning with January of the following year. Enterprises below that threshold can provide an advance return either monthly or quarterly on a voluntary basis. Therefore, new enterprises are registered earlier than in the past in the basic tax file.

## 5 Basic Problems for compiling business demography based on administrative data

- The units of the AR do not exactly agree with the units of the BR (enterprises, establishments and local units). Usually, one enterprise of the BR consists of 0 to n units of the Social Security register (employer accounts). The tax register consists of legal units which correspond more or less with the enterprises of the BR.
- Some information in the AR is incomplete or occasionally wrong (e.g. the NACE code), also the timeliness of the data in the AR can be different from that in the BR.
- The maintenance of the units of the AR is done differently. If an enterprise changes its legal form, it gets a new identifier in the tax register. In the SSR this was not always the case in the past.
- The biggest problem is the non-existence of a unique numerical identifier for the units in different registers. The matching of the register units has to be done by comparing mainly text fields like name of the company and address. These fields are not standardised and are of different length in different registers. Generally, the most suitable kind of unit of the BR for linkage with the units of an external AR is the enterprise. For each AR a table is loaded in the DB/2-database of the BR where each enterprise gets assigned the identification keys of the corresponding units in the AR.

For record linkage of the units of two registers Statistics Austria uses the bigram method. The bigram method is simple to implement, is usable for each language and robust against permutation of words in a phrase. To achieve satisfying results with the bigram method it is necessary that the compared text fields of the same unit in different registers are not written too differently. Before the text variables of two registers are compared for similarity, they must be standardised and parsed.

The comparison of thousands or even millions of records of one register with all records of another big register would consume very long elapse times even for a modern computer . Fortunately, both the BR and the different AR store the postal and/or municipality code for each unit which can be used for blocking the comparisons which diminishes the number of necessary comparisons highly.

## 6 Methodological aspects of the data compilation

In the following chapter some main methodological issues in compiling the Austrian business demography data are presented. As noted above, the general approach was to use two administrative registers (tax register, social security register) in combination with the BR.

The **creation of the data base** for reporting year 2007 started with the merging of the turnover taxation files for the years 1997 till 2007, all in all about 1 million records. The monthly number of employees for each employer account of most of the taxation units has been stored in the BR just as the relation of the employer account keys and the tax identification keys of the tax records.

Employer accounts and their corresponding employment figures whose relation to a tax unit was not stored in the BR were attached to a tax unit by **record linkage methods**. (Ch. 5)

**Estimation & imputation of missing data:** The employees of all employer accounts belonging to the same enterprise of the tax population were summed up. At the time of compiling the business demography for reporting year 2007 (mid 2009) the turnover taxation file did not contain all enterprises that were active in 2007. Therefore that file was supplemented by employer units of the SSR. Missing non-employer units were added by including the units from the turnover taxation file. The amount of missing units was estimated from the past experience (growth of the enterprise stock for 2006 in the turnover taxation file between mid 2006 and mid 2007). For units with missing turnover on the current margin of the time span 1997-2007 the missing turnover was estimated from the monthly tax advance returns.

**Temporal harmonization of records and variables:** Because of the different maintenance concepts of the social security register and the tax register it would happen that an enterprise which only changed its legal status (e.g. in 2005) would be recorded as two different tax units, one unit with turnover from 1997 until 2005 and another unit with turnover from 2005 till the end of the observation period. If the SSR stored the employees before 2005 under the same employer account key as after 2005, then the employees before and after 2005 were allocated to the record belonging to the legal unit after the change of the legal unit. This happened since the relation between the SSR keys and the tax keys was only available at the time of producing the business demography and not for the whole time series. To fit the number of employees to turnover it was necessary to move employees in years without turnover to another record with an adequate turnover.

**Flagging of real births:** For business demography only real births and deaths are relevant. A real birth occurs if only one enterprise is involved and a combination of production factors – in particular employment – is created. The enterprise is established from scratch, so to speak. Additions to stock, for example, due to merger, break-up or restructuring are not real births; nor is a pure change of economic activity, legal form or location regarded as a birth. Records of the synthetic data set with more than 20 employed persons were checked manually whether they are real births. From the smaller units a sample of some thousand units was selected and also checked manually . From this operation we achieved estimators of the percentages of real births on all births classified by year of birth, size class and legal form. This information was used to flag the appropiate percentages of births as real birth or not.

**Assembling parts of a business biography:** To get unbiased survival rates it was necessary to fit together fragments of the same enterprise during its life. Therefore all births of a given year were compared with all deaths in the same year to find automatically data belonging to the same unit (by using record linkage procedures comparing name, address and NACE for similarity). Two or more similar records were fit together to a new record.

**Creation of variables:** Before tabulating the results, the synthetic data set was enhanced by variables like year of birth or death, NACE Rev. 1.1 code, legal form, size class, region and sex of the entrepreneur. If a record had a relation to a unit in the BR, then the values in the BR were used for creating the variables of the synthetic data set. Otherwise the information was taken from the AR, mainly from the tax register. As a last resort missing values were imputed by hot deck methods. The year of birth was defined as the year in which an enterprise really started activity, measured by turnover or persons employed. The year of death was defined analogously.

Bibliography

Bundesstatistikgesetz 2000. (Federal Statistics Act 2000), BGBL I Nr.163/1999, idF BGBL I Nr.136/2001, I Nr. 71/2003 and I Nr. 92/2007, Vienna.

Eurostat – OECD Manual on Business Demography Statistics, 2007 edition, Eurostat Methodological working papers, Luxembourg 2007.

Haslinger, A., 1997. Automatic Coding and Text Processing using N-grams. Conference of European Statisticians. Statistical Standards and Studies – No. 48. Statistical Data Editing, Volume No. 2, Methods and Techniques, pages 199-209. UNO, New York and Geneva.

Haslinger, A., 2004. Data Matching for the Maintenance of the Business Register of Statistics Austria. Austrian Journal of Statistics, Volume 33, No. 1&2, pp. 55-67. http://www.stat.tugraz.at/AJS/ausg041+2/041+2Haslinger.pdf

Regulation (EC) No 177/2008 of the European Parliament and of the Council of 20 February 2008 establishing a common framework for business registers for statistical purposes and repealing Council Regulation (EEC) No 2186/93, OJ L61 of 5 March 2008.

Regulation (EC) No 295/2008 of the European Parliament and of the Council of 11 March 2008 concerning structural business statistics (recast), OJ L97 of 9 April 2008.