STATISTICS AUSTRIA
The Information Manager

ESSnet: Use of Administrative Data in Business Statistics

Rome, 18-19 March 2010

Session – Methods for editing and controlling quality of administrative data

Model based estimation of enterprises below thresholds in Structural Business Statistics

Martin Haitzmann, Statistics Austria

1 Introduction

The SBS-Council-Regulation[1] is the basis for compilation of Structural Business Statistics (SBS) in Austria from the reference year 1997 onwards. Since 2002, a new national regulation[2] based on the Federal Statistics Act 2000 has foreseen a completely new data collecting and estimation concept for SBS in Austria. To satisfy national users' needs and European requirements the new concept was implemented by conducting a yearly cut-off survey in combination with the use of administrative sources and statistical calculation methods instead of formerly applied stratified random sampling.

By this new concept, the annual primary statistical SBS-survey is only relevant for enterprises exceeding legally defined threshold values comprising about ~37.000 (2007) large and medium sized enterprises which cover about 12.5% of the total population. For the 255.000 small enterprises below legal thresholds instead, an estimation model was developed by using information from administrative sources in combination with model based estimation for all variables not available from administrative sources. In Austria, model based estimation for enterprises below threshold values is carried out by the Methods Division on behalf of Directorate Enterprise Statistics. The model development started in 2003 with a one year lasting period of preliminary analysis, building the basis of the final estimation model.

The results concerning the reporting year 2002 were therefore the first SBS results emerging from model based estimation. The main conceptual changes were

- cut off survey instead of random sampling,
- estimation of micro data instead of free grossing up,
- use of administrative sources and
- changes in the list of variables;

This required a temporary development of model based estimation by using census data from 1995, because data from the SBS survey for reference year 2002 itself was available too late for being integrated in the course of model development.

From the reporting year 2003 onwards, however, it was necessary to develop an estimation model with surveyed enterprises above thresholds from the particular reporting year. The most important reason for this change was that "old" 1995 data would not have allowed to model economic

[1] Council Regulation (EC, EURATOM) No 58/97 of 20 December 1996 concerning structural business statistics, respectively Council regulation (EC, EURATOM) No 295/2008 of 11 March 2008 (SBS-recast).

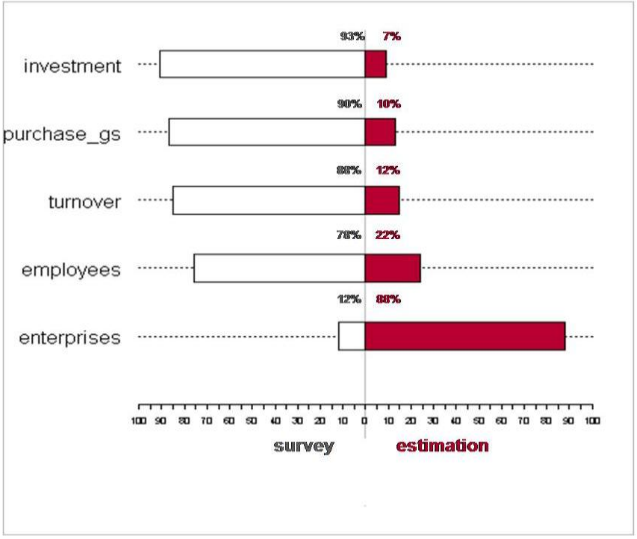
[2] National Regulation for SBS (Leistungs- und Strukturstatistik-Verordnung, BGBl. II 428/2003, changed by BGBl. II 266/2009)

developments which had happened in the meantime. The reason why data from sample surveys 1997 – 2001 had not been taken into account was the reduced catalogue of variables for small enterprises.

As already mentioned, cut-off data information was available for the first time regarding reference year 2002. After having analyzed the data from various aspects[3], experts from Directorate Enterprise Statistics finally decided to develop an estimation model based on primary (surveyed) data only.

In general, the applied strategy is based on the *principle of data concentration* covering a small number of large and medium sized enterprises, where the ”most important information” should be obtained directly from the units, whereas ”less important” information from small enterprises should be gained through model based estimation. Thus, there is no general loss of information and detailed information can be provided for all units on record level. Figure 1 illustrates the “principle of data concentration”.

Figure 1: *The principle of data concentration regarding selected variables*



Because of missing information from enterprises below thresholds the model parameters calculated with enterprises above thresholds have to be applied to data which cannot be considered in the estimation model. The ideal case - the structure of variables for enterprises above thresholds is identical to the structure of variables for enterprises below thresholds – can not be generally assumed. Therefore it is necessary to take the most “similar” enterprises above thresholds as basis for the calculation of the model parameters. Additionally, parameter calculation has to be carried out on the most detailed NACE level possible, provided that enough enterprises at this level are in the cut off sample. This is necessary due to structural differences between distinct economic activities. Insufficiently covered economic activities with structural differences require some special treatment, subsumed under the so-called expert rating (adaptations of estimation model). The following chapters explain model based estimation in detail.

2 Database and data sources for the compilation of micro data

The following data from statistical surveys, business register and administrative sources are matched for model based estimation and for various analysis and time series comparisons:

- **Recent information from business register, administrative sources and SBS survey**
  - Total active enterprises in relevant NACE categories named in the SBS regulations from the Business Register of Statistics Austria at the end of the reporting year
    - Number of enterprises
    - Regional classification
    - NACE classification
    - Legal form

[3] Test calculations taking only primarily surveyed data into account were as well carried out as time series analysis in order to get a picture of influence of economic developments, showing promising results.

    - Self employed persons
  - Results of SBS survey for the particular reference year for enterprises above thresholds (primary data)
  - Administrative data linked on a single enterprise basis with the enterprises in the business register:
    - Turnover from annual tax declarations for the reporting year and the year previous to the reporting year
    - Aggregated monthly tax declarations (VAT advance return)
    - Employment data from social security authority by sex and status level
  - Results from short term statistics in the economic sector “Industries and Construction”
  - Structure of enterprises for unit non response from the year prior to the reference year - provided that primary data for the enterprises in question were available.
- ➢ **Information from previous years (for data analysis and time series comparisons)**
  - Census 1995
  - SBS results between 1997 and 2001 (stratified random sample surveys)
  - SBS results from 2002 to 2007 (cut-off-surveys)

3 Legal threshold values

Until the reference year 2007 the threshold values for “Industries and Construction” were defined in terms of persons employed and for trade and services in terms of turnover. The national SBS Regulation foresaw the following threshold values for defining enterprises above (to be surveyed) and below (enterprises to be estimated) thresholds:

Threshold values for Production (sections C-F of NACE Rev. 1.1)

20 persons employed - **90%** of the **total turnover** of a NACE division shall be reached; if this is not the case the threshold value is reduced step by step till 90% of the turnover in the relevant NACE division is reached (minimum level is 10 persons employed).

Threshold values for Trade & Services (sections G-K except 65, 66[4] of NACE Rev. 1.1)

- ➢ Section G (Wholesale and retail trade, repair motor vehicles, motorcycles and personal and household goods, Group 633 (Activities of travel agencies and tour operators), Group 634 (activities of other transport agencies): yearly turnover of EUR 1,5 Mio.
- ➢ Other Service sectors (NACE H, I except 633 and 634, 67, K: yearly turnover of EUR 750,000.

Threshold values have been adopted in course of implementing SBS recast and NACE Rev.2 and are therefore slightly different to those mentioned above with the beginning of reference year 2008. Most relevant changes mainly concern an increase of threshold values in general and the introduction of a turnover threshold in Industries and Construction as well as a newly introduced threshold concerning employees in Trade and Services

4 Compilation of basic data

For the compilation of basic data, information about number of employees and turnover is derived from administrative sources.

4.1 Basic data from administrative sources

In Austria, the statistical units (enterprises) of the business register are linked to administrative sources on a single enterprises basis. Thus, for each enterprise below threshold as well as for unit

[4] Data for NACE 65 and 66 can be taken from supervisory authorities completely

non responses, information about employment variables can be taken from Social Security Authority and turnover values can be derived from Tax Authority.

As basic variables like turnover and the number of employees by sex and status level are available either from the statistical survey or administrative sources, these variables can be seen as deriving from a census, obeying the restriction that administrative data from the reporting year is not always available for all enterprises at the time of calculation. The treatment of such missing basic data in the estimation modes is explained in chapter 4.2.

In contrast to the quality and completeness of social security data, which is very satisfactory, turnover definitions from tax declarations and SBS do not correspond exactly. Therefore, a variety of analysis regarding these differences in comparison to former SBS surveys has been carried out. Deviations between SBS definitions and administrative data depend on different reasons such as foreign tax accounts, definition differences, structural changes, group company tax declarations, financial year records (for SBS), calendar year (tax) or deficient tax declarations etc. Deviations were observed mainly for large or medium sized enterprises, which were in the primary survey anyway. Differences for small observable enterprises were almost determined as rather negligible. The differences between basic data and administrative data depend mostly on events not possible to be influenced, but incurred analysis have to be continued regularly in order to determine systematic deviations and for constructing “perfect” model estimation in the end.

4.2 Treatment of missing basic data

Because of missing tax declarations or incomplete links of administrative sources with the Business Register, administrative data records for the reporting year are not available for all enterprises below thresholds. As a first step, regarding enterprises without tax data for the reference year, aggregated monthly tax declarations (VAT advance return) are taken into account as an annual value. Missing tax declarations for single months are imputed concerning the development of single enterprises as well as of the whole NACE category.

For quality reasons monthly tax declarations are preferred to yearly tax declarations from the previous year only in the case of at least 10 monthly declarations available. Otherwise an extrapolated yearly tax declaration of the previous year is preferred. (Base: available tax data by NACE category/method: LTS-Regression/Model: turnover$_{\text{ref.year}}$= ß0+ß1*turnover$_{\text{prev.year}}$.

If there is neither a tax declaration of the previous year nor monthly declarations available, a substitution of turnover is carried out by calculating a ratio of turnover/person employed by economic activity, provided that Social Security data is available (Base: Data of Social Security Authority and tax data by NACE category/method: LTS-Regression/model: Turnover$_{\text{ref.year}}$= ß1* employees$_{\text{ref.year}}$)

Enterprises without a link to social security authority are supposed to have none employees if on the other hand tax data from Tax Authority is available. (Assumption: enterprises which have no employees just have self employed persons). If information from short term statistics in “Industries and Construction” is available, turnover and employment information from STS is taken into account.

Enterprises with no information from administrative sources at all, are not taken into account for estimation.

The distribution of population according to availability of administrative data is shown in the following table.

Table 1: Distribution of basic values (reference year 2007)

Turnover			Employees		
	%enterpr.	%turn		%enterpr.	%empl.
Survey	12%	88%	Survey	12%	77%
Estimation	88%	12%	Estimation	88%	23%
thereof			thereof		
Yearly tax declaration for 2007	71,4%	68,5%	Social security authority	61,7%	95,6
VAT advance return	14,6%	19,3%	Zero	37,6%	0,0%
Extrapolated yearly tax declaration	10,9%	4,4%	STS production	0,7%	4,4%
Substituted	2,3%	1,8%			
STS production	0,7%	5,9%			

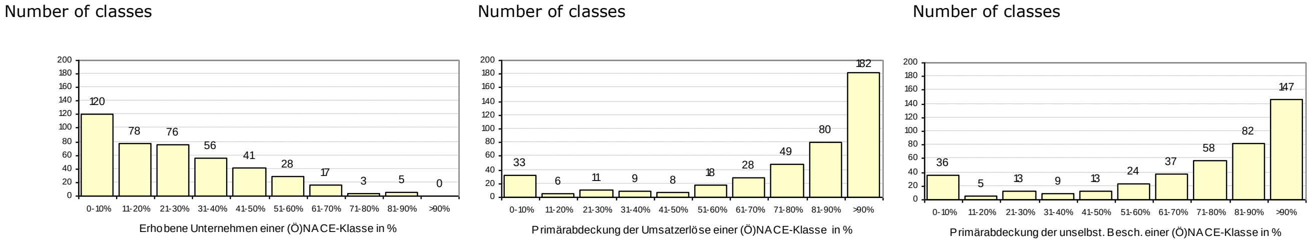
Table 1 depicts that the main part of the turnover of enterprises below thresholds is available either from yearly tax declaration or from aggregated VAT advance return from the reference year.

Regarding social security information the table shows that almost all information concerning number of employees is available from social security authorities.

4.3 Coverage of basic data by enterprises above thresholds (SBS survey)

The coverage of enterprises above thresholds is shown for the basic variables “number of enterprises”, “turnover” and “number of employees”.

Figure 2: Coverage of basic data by enterprises above thresholds (SBS survey)



Percentage coverage of number of enterprise at NACE 4 digit level

Percentage coverage of turnover at NACE 4 digit level

Percentage coverage of number of employees at NACE 4 digit level

The principle of concentration is displayed in figure 2. Generally, for the majority of NACE-classes only a small percentage of enterprises (first chart) is necessary to reach a relatively high coverage of basic variables (second and third chart). More than two third of NACE-classes have at least a coverage of 70% of the basic variables turnover and number of employees.

However, in order to be able to apply concentration principle to all NACE classes, an iterative adaptation of threshold values would be necessary. For NACE classes (without concentration of variables to some large enterprises) consisting of homogenous turnover values additional random sample could be taken into consideration. But this would on the other hand cause a higher number of enterprises in the yearly survey.

A random sample of about 500 enterprises for insufficiently covered NACE activities in production was carried out once for the reference year 2003 on a voluntary basis. As this survey was on a voluntary basis the response was not high enough to have reliable results. Cost and benefit analysis were evaluated and it was decided that random sampling for enterprises below thresholds is not being planned for the future.

5 Estimation of model parameters

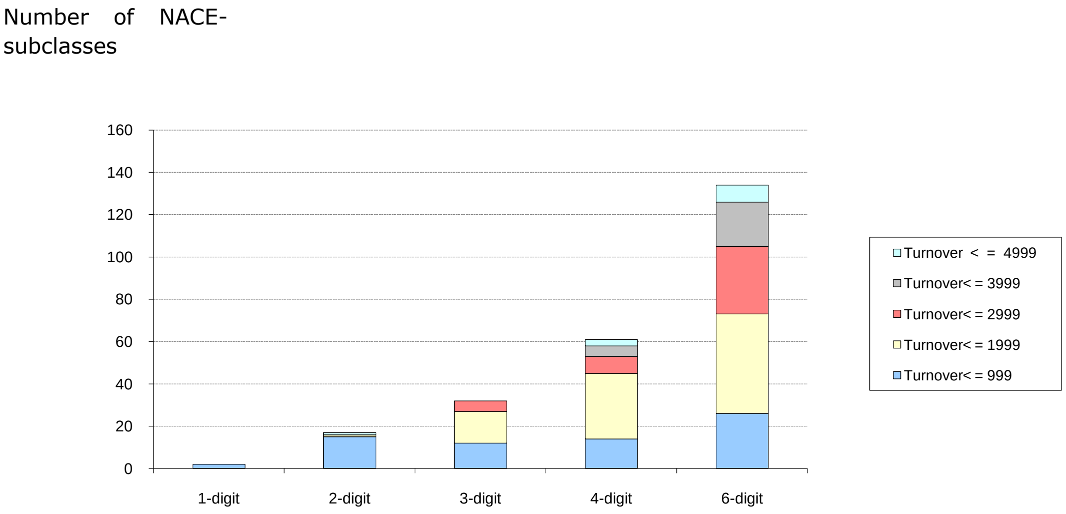
For small enterprises below thresholds an estimation model including all main and detailed variables not available from administrative sources was developed.

5.1 Database for parameter estimation

In order to determine a model basis, various different aspects have to be taken into account. First of all, enterprises which are as similar as possible to enterprises below thresholds have to be taken into consideration for the calculation of model parameters. The calculation itself has to be carried out at the most detailed NACE level to be able to consider different structures of economic activities. Therefore calculations are in a first instance applied to sub classes (6-digit of national version of NACE Rev. 1.1 respectively 5-digit of national version of NACE Rev. 2) for economic activities provided that a sufficient number of enterprises from the survey is available in the relevant size classes.

To construct the model base for about 600 NACE sub classes, the choice of enterprises is started from the enterprises above thresholds by NACE sub classes (homogeneous start position) with an initial upper limit of turnover of EUR 999.000. If there are not enough enterprises (at least 15 enterprises are necessary for regression and at least 10 for ratio estimation) at the related NACE level, the upper limit of turnover is raised until up to EUR 5 Mio in steps of EUR 1 Mio. If there are not enough enterprises with less than EUR 5 Mio turnover in a NACE sub class the model parameters are calculated on a higher NACE level.

Figure 3: Model base construction for the parameter estimation, for e.g. the variable "Purchases of goods and services"



5.2 Linear regression model for estimation of the main variables

5.2.1 Regression Model

For explaining the correlation between main variables and basic data – e.g. turnover derived from administrative data (record of Tax Authority) is necessary for estimation of "purchases of goods and services" – a linear regression model is used. Regressors are the variables "turnover" and "number of employees.

Names:

Total number of enterprises	*n=1,2 , . . . , k*
'Group' (at the deepest possible level of NACE)	*g=1,2, . . . , number of ÖNACE subclasses*
Number of enterprise in the Group g	*ng*
Estimated value of enterprise i in group g	$\hat{y}_{gi}$
Regressor of enterprise i in the group g	*xgi*
Number of the Regressors	*m*
Regressand of the enterprise i in group g	*ygi* ($\bar{y}_g$ *= mean value)*
Regression coefficients in group g	$\hat{b}_{go}$, . . . , $\hat{b}_{gm}$

Regression model: $y = x\beta + \varepsilon$ with ε . . . Error.

$E(\varepsilon) = 0$ and $E(\varepsilon^T \varepsilon) = \sigma^2 I$

Estimation of the regression line: $\hat{y}_g = X_g \hat{b}_g = \hat{b}_{g0} + \hat{b}_{g1} x_{g1} + \ldots + \hat{b}_{gm} x_{gm}$

Design matrix:
$$X_g = \begin{bmatrix} 1\ x_{g11} \ldots x_{g1m} \\ 1\ x_{g21} \ldots x_{g2m} \\ \ldots\ \ldots\ \ldots\ \ldots \\ 1\ x_{gn1} \ldots x_{gnm} \end{bmatrix}$$

Estimated model parameter (Regression coefficients)
$$\hat{b}_g = \begin{pmatrix} \hat{b}_{g0} \\ \ldots \\ \hat{b}_{gm} \end{pmatrix} = (X_g' X_g)^{-1} X_g' y_g$$

by
$$\min_e (Y_g - X_g \beta_g)^T (Y_g - X_g \beta_g)$$

Quality of the adaptation of the data to the model
$$R_g^2 = \frac{\Sigma(\hat{y}_{gi} - \bar{y}_g)^2}{\Sigma(y_{gi} - \bar{y}_g)^2}$$ *... R square*

Deviation of the estimated value from the observed value $e_{gi} = (y_{gi} - \hat{y}_{gi})$ …. *Residual*

Table 2 gives an overview of regressors used for the estimation of main variables:

Main Variable	Regressor (Basic Variable)
Purchases of goods and Services	Turnover
Other expenditures	Turnover
Income (without turnover)	Turnover
Stocks	Turnover
Investments	Turnover
Wages	Wage earners and turnover
Salaries	Salary earners and turnover
Gross compensation for apprentices	Apprentices and turnover
Social security costs	Wages and salaries

5.2.2 Influence of outliers in the regression model

Outliers can have a great influence both on the position of regression line and on the quality of the model adaptation. These outliers can cause serious bias and can furthermore lead to total useless estimations. That's why these outliers were analysed in a first step by means of COOKs' D. COOKs' D measures the shift related with estimated parameter vector without the observation i. If $D_{gi}>4/(n_g-k-1)$ then observation i has a major impact.

$$D_{gi} = \frac{(\hat{b}_g - \hat{b}_{gi})'(X'_g X_g)(\hat{b}_g - \hat{b}_{gi})}{(m+1)\ s_g^{\ 2}}$$

b_{gi} *Regression coefficient without observation i in the group g*

5.2.3 Robust regression (regression estimator with high breakdown point)

In general the OLS (ordinary least square regression) is very outlier- sensitive. Since COOKs Distance is not robust, it is negatively affected by outliers or groups of outliers itself. Thus a robust regression model is used for SBS estimation in Austria. As it is necessary to calculate a regression line for each variable of the NACE sub classes, it is important to find a method for detecting outliers which works automatically.

The most suitable automatic methods are compute-intensive robust regression models respectively models with high breakdown point. The breakdown point is a criterion for the minimal ratio of observations that can influence the estimation arbitrarily. The OLS regression has a breakdown point of zero. (minimal ratio 1/n), therefore e.g. just one very large or small value can lead to meaningless results. (*leverage point*).

On the other hand robust regression models like Least Median of Squares (LMS) or Least Trimmed Squares Regression (LTS) have a maximum breakdown point of nearly 50%.

In the case of LTS regression only a predefined number of the smallest residuals (h) of the minimised function are taken out of all observations (n). Therefore large residuals didn't have any influence on the estimator.

To be minimized[5] $\sum_{i=1}^{h}(e^2)_{i:n}$ whereas $h \in [\frac{n+1}{2}, n]$ is the number of the observations in the function to be minimized and $(e^2)_{1:n} \leq (e^2)_{2:n} \ldots\ldots\ldots\ldots \leq (e^2)_{n:n}$ are the ordered squared residuals.

The smaller the value of h is, the more robust is the LTS-estimator. The larger h is the more information the estimator gets from the data, which means the larger is the data-majority the algorithm follows.

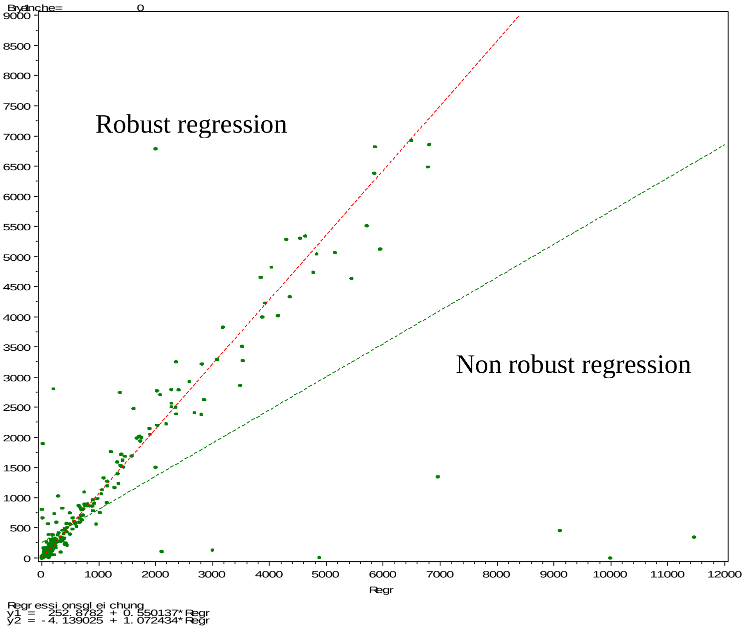
[5] Find the h observations out of n, with smallest sum of squared residuals (SAS Online Doc)

5.2.4 Combined (weighted) regression model

Because of the just stated disadvantages of non robust regression methods, a combined regression model was selected to explain correlation between variables for SBS. An OLS regression is carried out on the results of robust regression estimation for parameter estimation of enterprises below thresholds.

That means in a first step that an LTS regression has to be applied. LTS is a very robust regression estimator but hard to compute, therefore the Fast-LTS algorithm[6] is in use. In order to have a good compromise between robustness and efficiency, the number of observations (h) in a branch is determined with h=0,75*n. Based on the LTS-results, an OLS regression neglecting observations with large residuals[7] is carried out (compare the functions in figure 4).

Figure 4: Comparison of a robust and a non robust regression model (Using not transformed highly skewed data with some leverage points)



5.2.5 Quality of model parameters

R^2 is a parameter for the quality of correlation between independent and dependent variables. The R^2 criterion is defined between 0 and 1 and measures how well the variation of dependent variables can be explained with the model. For example, how well the main variable “gross wages” can be estimated with the regressors “turnover” and “number of employees” by a multiple linear regression model. Higher values indicate that the model fits better (closer to the points). R^2 cannot be taken as main criterion for whether a fit is reasonable – it doesn’t generally mean the fit is well in other ways.

[6] Fast_LTS algorithm for large data sets (Rousseeuw and Van Driessen, 1999)

[7] based on the LST-Regression the stand. Residuals> |2.5| of all n observations are weighted with zero

Figure 5: Distribution of R^2 (quality of the different regression functions to the estimation of a variable), for selected variables at the deepest level of NACE

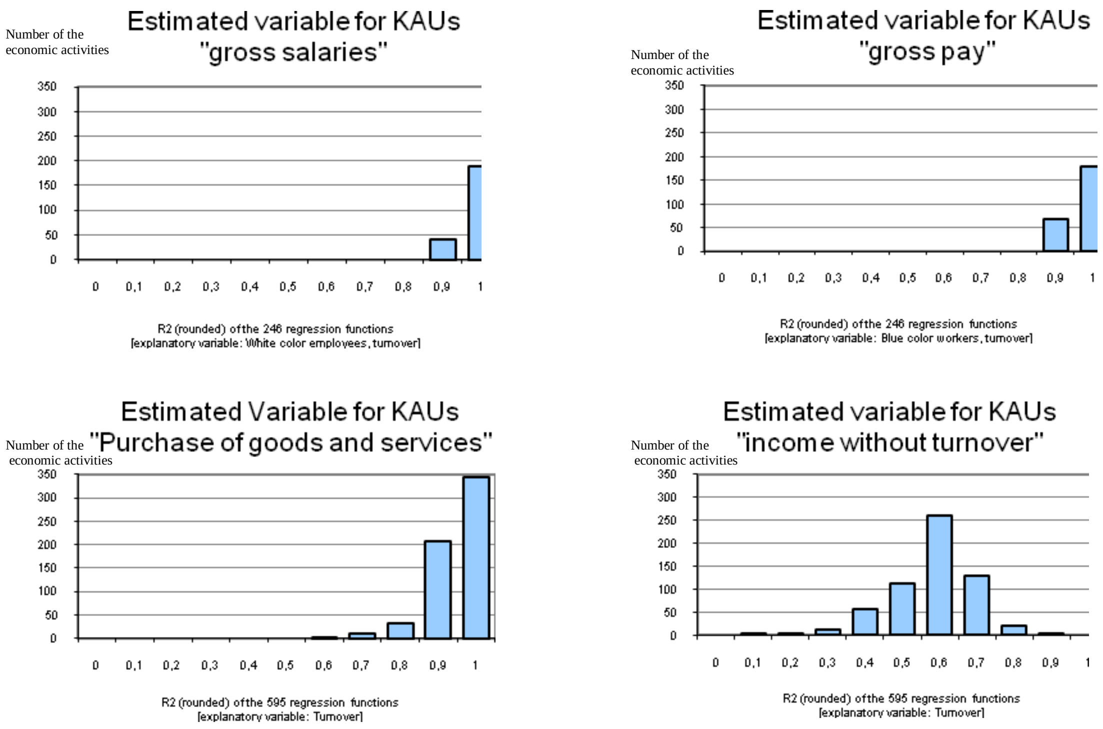


Figure 5 shows that the correlation between e.g. turnover and purchases of goods and services can be described very well by a linear model.

For many branches of NACE classification, however, such a linear model is not the adequate way to explain gross investments, stocks or other income. For partly balancing the quality of results of investments, newly born enterprises got the 80.percentile of the distribution of the relationship between investment/turnover instead of regression parameters originally calculated. (Assumption: newly born enterprises=higher investment and probably fewer turnover).

5.3 Ratios for the estimation of detailed variables

To explain the correlations between main variables and detailed variables (e.g. breakdown of turnover), the main variable (HMM) is broken down to detailed parts.

$$HMM_j = \sum_{i=1}^{r} MM_{ji},$$

with r = number of detailed variables and j=1, . . ., h

$$\sum_{j=1}^{h} HMM_j = \sum_{j=1}^{h}\sum_{i=1}^{r} MM_{ji},$$

with h = Number of main variables

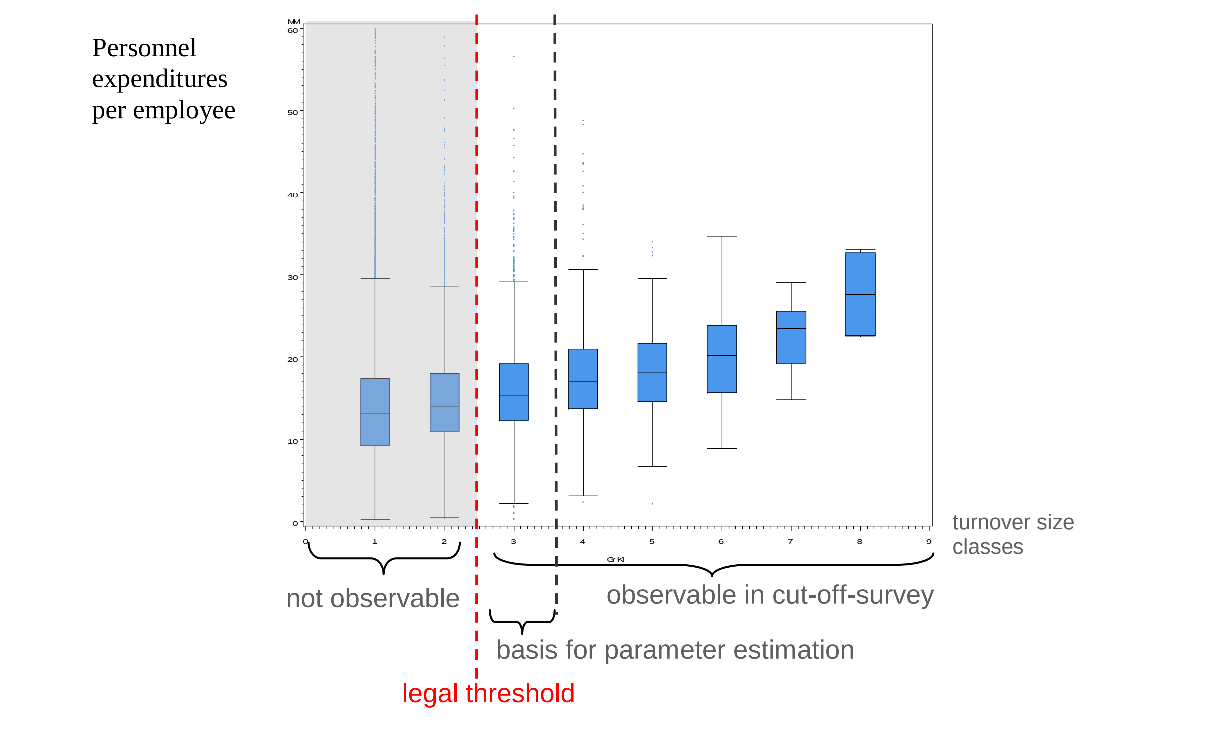
and parameter calculation $G_{ji} = \dfrac{MM_{ji}}{\sum_{i=1}^{r} MM_{ji}}$, j=1, . . ., h

by ratio estimation.

5.4 Model based estimation from the methodological point of view

Because of non availability of random sample information, model based estimation has been applied instead, which quite leads to some difficulties, because the parameters calculated with enterprises above thresholds have to be applied to a data area that is not included in the model. The ideal case – the structure of variables for enterprises above thresholds is the same as for enterprises below – can not generally be assumed. This is shown in figure 6 which shows the box plots for personnel expenditures per employee for a selected NACE subclass and different turnover size classes in the last SBS-Census (1995). Despite such an application is generally not allowed from methodological point of view, due to new basic conditions a compromise was necessary.

Figure 6: Representation of observable and not observable data area



Therefore, based on economic activities, the step by step approach with increasing turnover size classes was used in order to calculate model parameters with enterprises which are most similar to enterprises below thresholds.

For variables without sufficient correlation with the basic variables (such as investments) alternative models (with combining new data sources) are planned in an ongoing process. But for the most cases the correlation of main variables and basic data can be described by a linear model (partly as depend on branch) very well.

For economic activities with a high number of enterprises below thresholds, a variable based optimization of the estimation model and further basic analysis about the use of administrative data as well as the transferability of model parameters to the non observable array would be useful.

Especially for economic activities, with a high dependency on size classes a systematic bias cannot be avoided. Data information getting from the random sample surveys (carried out between 1997 and 2001) and the census 1995 were analysed very well. Because of different limitations ("old" structures of economic activities, reduced catalogue of variables for small enterprises etc.) this source is not usable for further model modifications.

In general the conceptual change from "design based estimation" (1997 – 2001) to "model based estimation" from 2002 onwards would have required a first basic statistical SBS survey for enterprises below thresholds, which was not possible due to lack of resources and legal restrictions. Therefore from our point of view the best possible option was chosen to reach the best possible compromise between quality of results, response burden and resources.

6 Estimation of Micro Data

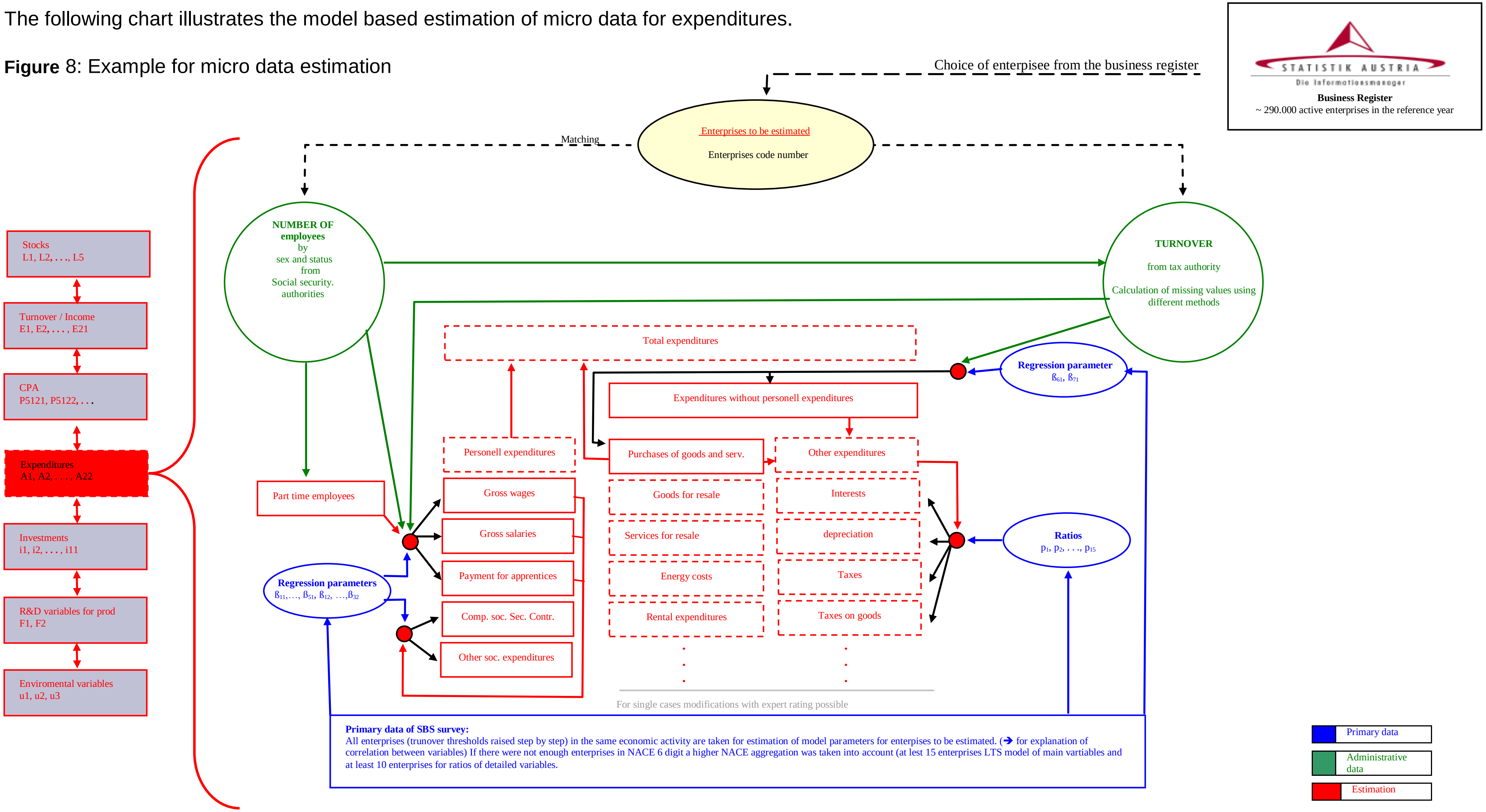
After having derived basic variables (turnover and employees) from administrative sources and SBS survey data, activity-based model parameters are calculated. These model parameters will be used for calculating all main and detailed variables for enterprises below thresholds using model based estimation.

6.1 Compilation of Variables

1. Taking basic variables (***turnover*** and ***number of employees)*** from administrative sources.
2. **Imputation** of missing basic data
3. Calculation of **model parameters** with SBS survey data (primary data)
4. **Estimation of main variables** like personnel cost, purchases of goods and services, stocks etc., with activity based regression parameters (LTS)
5. **Estimation of detailed variables** like purchases of goods for resale, investments on machines, compulsory contribution of employers to social security, part-time employed etc., with main variables and ratios calculated
6. Adaptation of estimation model by "**expert- rating**": "Expert- rating" is mainly used for insufficiently covered inhomogeneous economic activities and economic activities where structure of enterprises above thresholds is very different from that of enterprises below thresholds (e.g. NACE categories in which trade and intermediary activities are put together). Expert-rating is the subjective assessment of results by qualified experts of data editing staff. In the case of known systematic deviations a modification of calculation and parameters can be carried out within the estimation model in order to increase the quality of results.
7. **Calculation of main aggregates** like value added or production value
8. Consideration of **results from STS** in "industries and construction": Takeover of some turnover and employment variables (for unit non response only)
9. Consideration of **variable-structure on enterprise level** from previous years' primary data
10. Estimation of remaining **unit non response** with model based estimation developed for enterprises below thresholds – in contrast to enterprises below thresholds higher turnover thresholds (max. EUR 50 Mio.) are applied for parameter calculation; estimated unit non response enterprises are analysed very carefully in the course of macro plausibility checks.

The following chart illustrates the model based estimation of micro data for expenditures.

Figure 8: Example for micro data estimation



Primary data of SBS survey:
All enterprises (trunover thresholds raised step by step) in the same economic activity are taken for estimation of model parameters for enterpises to be estimated. (➔ for explanation of correlation between variables) If there were not enough enterprises in NACE 6 digit a higher NACE aggregation was taken into account (at lest 15 enterprises LTS model of main vartiables and at least 10 enterprises for ratios of detailed variables.

Variables like income or stocks have to be also estimated in the same way. The logical dependencies between categories of variables have to be modeled. The illustrated estimation of micro data was applied for every single enterprise on the possible detailed NACE level (~600 ÖNACE sub classes).

7 Conclusions

The estimation model was developed with consideration to all legal and technical circumstances by using all data sources available for the moment. Various analysis and test calculations were carried out in advance. Analysis regarding the reference year 2002 have shown qualitatively high results mainly for basic data “turnover” and “number of persons employed” and for SBS results on regional level. Results for all other variables were also very encouraging. With the estimation model from the reference year 2003 onwards, which was based on primary data for the reference year in question, a better consideration of development of personnel expenditure and part time employment or the increasing importance of service sector was possible.

A concrete evaluation of the quality of the resulting data by calculating measurement errors (as done in the former concept) is not possible because of missing random sampling. Therefore other different quality aspects have to be taken into account. Optimizing the following different quality aspects is very important for increasing the quality of results.

In general the following Quality aspects have to be considered:

1. Quality of business register
    a. Control of data quality
    b. Update of business register
    c. Completeness of links to administrative sources
2. Quality of primary data
    a. Data quality: micro and macro plausibility checks in the course of data editing and data analysis
    b. Quality of data for parameter estimation
3. Quality of administrative data (social security authority and tax data)
    a. Data quality: regular control of data quality of administrative sources
    b. Correspondence with definitions: analysis of definition differences between SBS and tax data
4. Quality of imputation methods for missing administrative data: regular controls
5. Accuracy of estimation model used for model basis
    a. A specified number of enterprises is necessary from primary survey for parameter calculation on level of economic activity
6. Suitability of estimation for description of dependency between variables
    a. Use of alternative model for variables with low dependency like investments
    b. Consideration of new data sources (e.g. income tax data) ➔ planned for the future respectively just being implemented
7. Transferability of estimation model to enterprises
    a. Exact evaluation is possible only if a basic statistical survey will be carried out for those enterprises which is not foreseen at the moment
8. Coverage of enterprises above thresholds should be high enough

The predefined aim of developing an estimation model on basis of primary data of the reference year was reached. Consideration of already mentioned quality aspects is very important for future analysis. The model based estimation developed for enterprises below thresholds will be applied for the following reference years also. Adaptations in the estimation model will be carried out in the cases of basic legal changes or if resources are available. Because of changes caused by NACE Revision and SBS Recast affecting reference year 2008 onwards, adaptations of the Austrian concept for compiling SBS data are being implemented. At the moment, the use of additional administrative sources for wages and salaries, is another aspect added to the estimation concept for the first time concerning reference year 2008.

Documentation on the project methodologies

http://ec.europa.eu/comm/eurostat/ramon/nat_methods/SBS/SBS_Meth_AU.pdf (english)
http://www.statistik.at/web_de/wcmsprod/groups/gd/documents/stddok/007205.pdf (german)
http://www.statistik.at/web_de/wcmsprod/groups/gd/documents/stddok/007191.pdf

Bibliography

P. J. Rousseeuw, A. M. Leroy. Robust regression and outlier detection. Publisher: John Wiley & Sons, Inc., August 1987
P. J. Rousseeuw, K. van Driessen. Computing LTS Regression for Large Data Sets, Springer Netherlands 2006 (http://www.springerlink.com/content/06k45m57x01028x6/fulltext.pdf)