WP.9
ENGLISH ONLY

UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Ljubljana, Slovenia, 9-11 May 2011)

Topic (i): Editing of administrative and Census data

Combining administrative and survey data: potential benefits and impact on editing and imputation for a structural business survey

Supporting Paper

Prepared by Casciano C., De Giorgi V., Luzi O., Oropallo F., Seri G., Siesto G. (ISTAT, Italy)

I. Introduction: the re-design of Italian Structural Business Statistics

1. The new European Regulation on Structural Business Statistics (SBS) (March 2008) establishes that, in order to estimate information on the structure of National production systems, Member States (MS) can integrate data available in different information sources, including administrative data. Besides, due to budget restrictions, MS need to reduce statistical production costs while maintaining high data quality levels and reducing the statistical burden on enterprises. The latter problem is especially relevant in the Italian economic system, characterized by a large amount of small and medium enterprises: this fact, together with the high level of detail for SBS required by the European SBS Regulations and the amount of information to be estimated, implies relevant burden on both enterprises and the Italian Statistical Institute (Istat) and high non response rates especially for small and medium enterprises. Finally, the traditional "stove-pipe" model which is currently adopted at Istat for SBS production (one statistical survey for one specific statistical subject matter) implies additional production costs and additional burden on enterprises due to redundancies of questionnaires' contents.

2. Important re-design projects in this direction have been either carried out or are currently ongoing in several European Countries, among others France (Brion et al., 2009), UK (Lewis, 2010, Elliott, 2010), Portugal (Chumbau et al., 2010), Finland (Tolkki, 2007). A number of activities aiming at extending and improving the integrated use of administrative data sources in statistical production processes, are in progress at Istat, too. In effect, the Italian information system is at present characterized by the availability of high quality administrative data on enterprises. The existing administrative sources cover a large amount of business population and may provide direct information on key variables for estimating SBS.

3. Based on these premises, a new project has started aiming at completely revising the SBS production system, moving from the current stove-pipe model (where administrative data are used as complementary sources of information mainly for data validation purposes) to a completely integrated system where administrative data represent the *core* of information on SBS and direct surveys are designed accordingly in order to gather complementary information on specific economic issues or businesses' sub-populations.

4. Istat is currently developing a number of supporting tools in order to guarantee continuous and secure access to external data: besides formal agreements with the Italian Tax Authorities to establish a stable cooperation protocol for business data exchange, some technological tools are under development to facilitate the access to administrative data by the direct electronic transmission of information from

enterprises to Istat (adopting the eXtend Business Reporting Language technology - XBRL - and creating a statistical web portal for the direct electronic acquisition of businesses' financial statements).

5. In general, it is well known that the use of administrative data for statistical purposes implies a deep revision of the statistical production process, especially in terms of data integration and data validation, due to additional problems in terms of data quality and data usability.

6. Concerning data quality, the variety of definitions adopted in literature to describe this concept in case of administrative data (Wallgren et al, 2007; FMI, 2004; Thomas, 2005; ONS, 2005; Daas and Fonville, 2007; Daas et al., 2008) proves the complexity of defining a framework in this area which is agreed at international level. We refer here to the definitions adopted at European level (Eurostat, 2003; Eurostat, 1999) for assessing the quality of administrative data integrated with statistical data. The adopted definitions partly refer to concepts which are similar to those adopted for statistical data (e.g. *relevance*, *accuracy*, *accessibility* and *clarity*, *timeliness*, *coherence* and *consistency*), even if "adapted" to the characteristics of administrative data (as an example, the definition of *relevance* has to take into account the need of consistency of administrative and statistical definitions for both the target population and target parameters; the concept of *accuracy* implies the statistical adequacy of administrative items for estimating target parameters).

7. Concerning data usability, integrating administrative data in statistical production processes implies the assessment of specific requirements like *coverage* (in terms of target population units), *completeness* (in terms of variables which can be directly obtained from the source), *periodicity*, *stability* (in terms of availability of the source in a long period of time), *costs* (for data acquisition and data treatment), *potential reduction of statistical burden* deriving from the use of the source in statistical production. Furthermore, a crucial element to be considered is the availability of meta data about *methodological characteristics* of the data validation and treatment performed by the owner and provider of administrative data.

8. This paper deals with the re-design project involving the Italian sampling survey on *Small and Medium-sized Enterprise* (*SME*), which collects information on balance sheets of enterprises with less than 100 employees. It mainly contains preliminary exploratory analyses aiming at assessing both the accuracy and the usability of administrative data for estimating the survey target parameters for the main SBS. The results will help us to make a first evaluation of the types and amount of quality problems affecting administrative data, hence the potential costs of editing and imputation (E&I) on these data when integrated in the SME's survey process.

9. Part of the results shown in the paper have been obtained in the context of the *ESSNet on the Use of Administrative and Accounts Data for Business Statistics*" (*ESSNet Admin Data*) (http://essnet.admindata.eu/), which aims at developing a quality framework and recommended practices for the use of administrative data for statistical purposes in business statistics (Elliott et al., 2010). The *ESSNet* is one of the ongoing projects in the context of the European *MEETS* program (*Modernisation of European Enterprises and Trade Statistics*), approved by the European Council and Parliament on December 2008.

10. The paper is structured as follows. Section II contains a short summary of quality aspects determining the usability of administrative data, and their potential impact on the structure and costs of a statistical data editing and imputation process. In Section III the main aspects characterizing the Italian SME survey and the available administrative sources on enterprises are illustrated. Section IV contains some preliminary results which can be considered as useful for assessing the potential usability of these external sources for estimation purposes. In Section IV some first evidences in terms of potential impact of integrating administrative data in the current SME production process are discussed.

II. Potential impact of using administrative data on E&I processes

11. It is straightforward to note that the use of administrative information in statistical production processes has a strong impact on data editing and imputation (E&I) strategies from both an organizational and methodological point of view. We assume for the rest of this discussion that no errors are originated from data linking activities.

12. In general, the impact of integrating administrative and survey data on E&I depends on the type of use of the external information in the statistical production process:

(a) administrative information are used as auxiliary information in the statistical survey process, i.e. for improving *accuracy* and *completeness* of the final survey data (e.g. for optimizing error detection, and imputation of the survey item and unit non response);
(b) administrative data can be used as primary source in the SBS production process, complemented by direct sampling surveys to estimate either non covered sub-populations, or target variables which are not directly available from administrative sources; in this case, sample survey design can also exploit the availability of the additional administrative information as additional "frame information", in order to improve the efficiency of the estimation process in terms of trade-off between sample dimension and estimates accuracy.

13. We are especially interested in situation (b). In this case, a first element to consider is that administrative data are collected and validated by external Authorities based on typically non statistical quality requirements. Being the validation process out of the control of the Statistical agency, information about external data treatments are to be collected, and the possible additional data validation activities to be performed in order to guarantee the usability of the external data are to be implemented. In other terms, a first element to be considered is the trade-off between the reduction of costs and burden on enterprises, and the additional costs for external data analysis and validation.

14. When assessing the potential impact of using administrative data for SBS estimation, important elements to be considered are *completeness* and *coverage* of external sources, which directly affect respectively the amount of item and total non responses to be treated once administrative data are directly used in the estimation process:

- total non responses may derive from the under-coverage of administrative sources with respect to specific business sub-populations;
- item non responses may derive from the under-coverage and incompleteness of administrative sources : this implies the need of developing appropriate methodologies to compensate for the partially non available information.

15. Focusing on *accuracy* and *coherence*, in general administrative data suffer from non sampling errors similar to those affecting survey data (Daas, 2008; Hoogland, 2010; Haitzman, 2010) (systematic or random measurement errors, outliers, influential errors). In addition, inconsistencies deriving from the integration of information from different sources may affect the quality of the data used for estimation.

16. As a consequence, in case of administrative data, traditional error detection and non responses imputation methodologies can be adopted: Hoogland (2010) illustrates an overall editing strategy for elementary and aggregate data to check administrative information used in the area of business statistics which is similar to traditional strategies for survey data (e.g. *selective editing*, a specific sistematic errors detection, a *macro-editing* approach for outlier detection). In particular, for outlier detection similar approaches to those adopted for statistical data can be adopted (see among others, Lorenz, 2010; Elliott, 2010; Zach, 2007). Yung and Lys (2007) describe the overall editing and imputation strategy adopted in some economic surveys at Statistics Canada.

III. The SME survey and the available administrative data

17. The Italian sample survey on *Small and Medium-sized Enterprises* (SME) is an annual survey aiming at investigating mainly profit-and-loss accounts of enterprises with less than 100 persons employed, as requested by SBS EU Council Regulations. The sampling units belong to the industrial, construction, trade and services economic activities. The mode of data collection is mixed (both paper questionnaires and web electronic questionnaires) in order to facilitate the data transmission based on each enterprise's characteristics and technological capabilities.

18. The survey frame is represented by the Italian Business Register of active enterprises (BR), which is the logical and physical combination of data from both statistical and administrative sources. The SME

reference population consists of about 4.5 million of enterprises employing approximately 12.9 million people. The SME survey is a multi-purpose and multi-domain survey which produces statistics on mainly economic and employment variables for three types of domains, each defining a partition of the population of interest:

(a) *class of economic activity* (4 digits of Nace code);
(b) *groups of economic activity* (3 digits of Nace code) by size *classes of persons employed*;
(c) *divisions of economic activity* (2 digits of Nace code) by *regions*

19. The SME design is a one stage stratified random sampling with the strata defined by the combination of economic activity, size class in terms of persons employed, and administrative regions. The SME sample consists of about 105,000 enterprises, and for the reference year t it is drawn from the BR updated to the year t-1 and is launched in June of the year t+1. The updated frame is available for the estimation phase only 15 months after the end of year t.

20. In the current E&I procedure, data editing is performed disregarding the availability of external auxiliary information on the investigated phenomena. Imputation of item non-responses is based on nearest-neighbor donor methods using the economic activity (either 2 or 3 Nace code digits), the number of employees and the geographical region as matching variables. Calibration methods are used to compensate for units non responses.

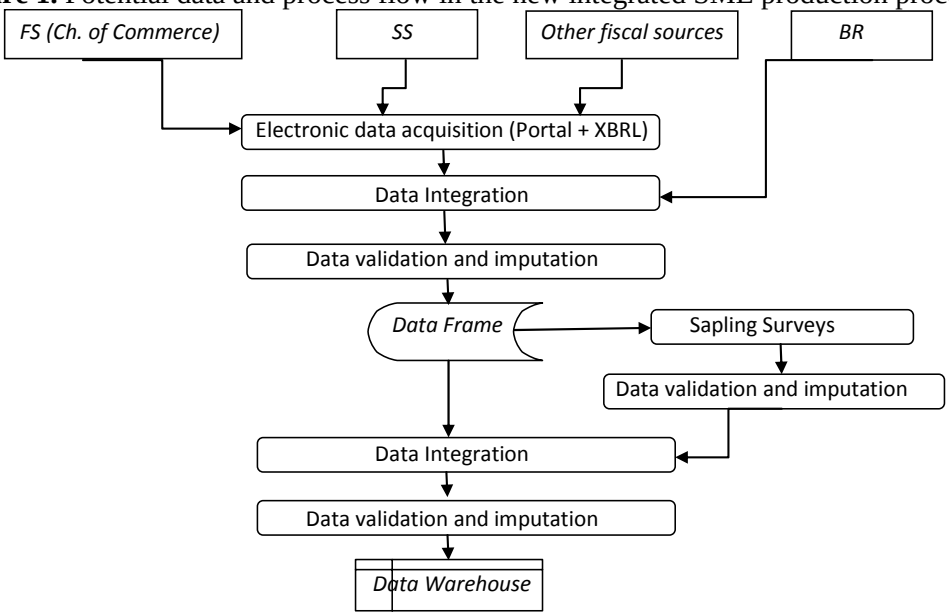
21. Besides BR, the relevant administrative sources considered on the SME survey target population and parameters are *Financial statements* (*FS*) and *Tax Authority sources* (*Tax returns forms* and *Fiscal Authority survey*). Out of the 4.5 million frame enterprises, corporate companies (about 15,7% in terms of enterprises and 37,2% in terms of persons employed) are liable to fill in FS. This is the best harmonized source with the SBS Regulation definitions.

22. All other enterprises are obliged to declare their taxable income to the Fiscal Authority by filling in tax forms. In particular, Istat acquires directly from the Tax Authority the *Sector Studies survey* (*Fiscal Authority Survey*) source: it is a fiscal survey aiming at evaluating the capacity of enterprises to produce income and to know whether they pay taxes correctly. It includes enterprises with a turnover greater than 30 thousand euros and less than 7,5 million euros: almost all enterprises are obliged to fill in the *Sector Studies survey* (*SS*) form together with the tax return one and to declare in detail costs and income items. As the common part of all sector studies questionnaires is like a financial statement, it can be used in a more effective way than tax return data.

23. Based on this information context, in order to reduce survey costs and respondents' burden, Istat plans to implement a new integrated system for the production of SBS for SMEs, where administrative data represent the main source of information, further integrated by direct investigation of either specific sub-populations, or variables which cannot be directly obtained from available administrative sources. This implies a complete re-design of the surveys' E&I strategy to deal with the new information framework and the additional data quality requirements.

24. In Figure 1, a preliminary flow chart describing the potential data and process flow in the new SME production process is delineated. As it can be seen, data validation and imputation appears in different phases of the process, since the data acquisition strategy results further complicated by the need of integrating data characterized by different quality levels and quality requirements.

Figure 1. Potential data and process flow in the new integrated SME production process



IV. Some preliminary results

25. In this section we illustrate some results of preliminary exploratory analyses and first experimental studies aiming at assessing the usability and quality of administrative sources for some SMEs key variables. The objective is to obtain useful information about the potential organizational and methodological impact on E&I deriving from the integration of external sources in the SBS production process.

26. The evaluation is performed on the 2007 SME data, characterized by a response rate of 42,1% (about 37,000 questionnaires), representing the set of final reliable replies (excluding non contacted units, out of coverage and list errors). The following analyses have been performed:

(a) coverage and completeness of external sources;
(b) completeness and accuracy of administrative data;
(c) coverage of administrative data: potential impact on estimates due to source effect;
(d) completeness of administrative data: potential impact on estimates due to item non responses imputation.

A. Coverage and Completeness of external sources

27. In terms of *coverage*, unless list errors, *FS*, *SS* and *Tax return modules*, cover almost all sample enterprises (see Figure 2): what remains uncovered are the largest partnerships with an ordinary accountancy regime and very small sole proprietorships. The large ones are asked to fill the RF form of tax return form which is not comparable with the profit and loss scheme. The very small ones, called *minimum taxpayers*, are liable to compile the special CM form of Tax return form starting from 2008 reference year. In Table 1, the sample coverage figures are showed by administrative source. If we do not take into account list errors and units out of SME frame, the total coverage is about 95%.

28. As for the whole SME target population, in Table 2 the coverage of FS and SS is reported. As it can be seen, about 87% of enterprises and 90% of total employees are covered. As it can be seen, the SS is the most relevant administrative source in terms of sample/population coverage: 67% of the sample, 44% non overlapping with the FS (percentages increase if referred to the population). These results strongly support the actual feasibility of the SME redesign project.

Figure 2: Coverage analysis by legal type and size class

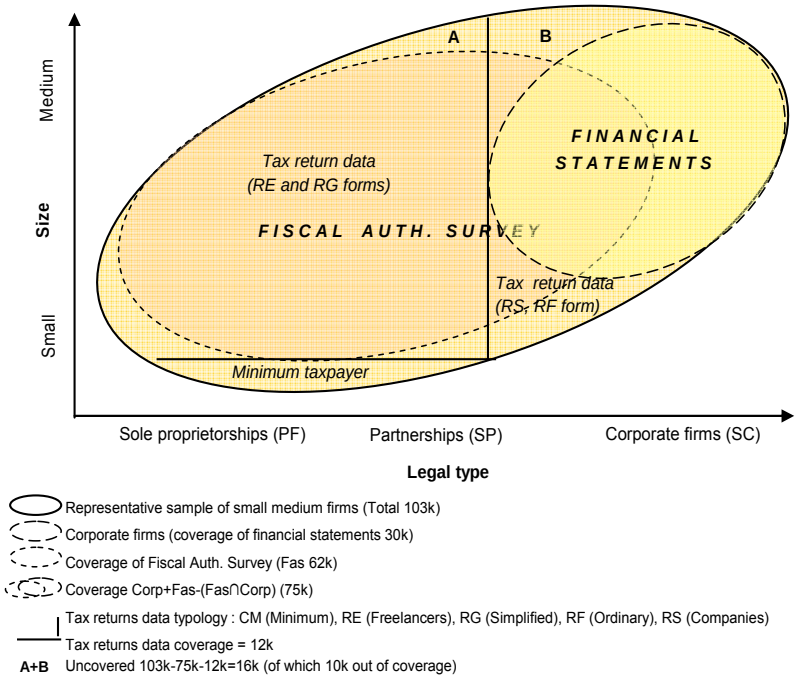


Table 1: Coverage of the initial sample by type of response and administrative data – Year 2007

	Initial theoretical sample		
Source	*Non respondents*	*Respondents*	*Total*
Financial Statements	10.370	19.739	30.109
Fiscal Authority Survey (F)	24.655	17.798	42.453
Fiscal Authority Survey (G)	1.343	1.223	2.566
Tax Return data - PF-RG	2.312	990	3.302
Tax Return data - PF-RE	747	483	1.230
Tax Return data - SP-RG	810	378	1.188
Tax Return data - SC-RS	4.546	1.839	6.385
From survey only	-	1.251	1.251
Total	44.783	43.701	88.484
Out of coverage and list errors			10.218
No sources			4.337
Total sample units			103.039

Table 2: SME target population coverage (percent) of the available administrative sources, in terms of number of enterprises (ENT) and number of employees (EMP) by economic activity - Year 2007.

	FS		*SDS-F*		*SDS-G*		*TOTAL*	
	ENT	*EMP*	*ENT*	*EMP*	*ENT*	*EMP*	*ENT*	*EMP*
C-Mining and quarrying	49.9	69.8	39.5	23.9	0.1	0.0	89.5	93.8
D-Manufacturing	22.5	54.5	64.8	37.9	0.0	0.0	87.3	92.4
E-Electricity, gas and water supply	57.5	81.8	2.1	0.6	0.1	0.0	59.7	82.4
F-Construction	14.3	33.4	72.9	56.5	0.1	0.0	87.2	89.9
G-Wholesale and retail trade; repair of motor vehicles, motorcycles and personal and household goods	11.1	30.7	77.0	60.1	0.1	0.0	88.2	90.9
H-Hotels and restaurants	10.8	24.5	75.1	66.1	0.0	0.0	85.9	90.7
I-Transport, storage and communication	16.3	47.6	68.7	39.9	1.1	0.3	86.1	87.8
J-Financial intermediation	6.1	13.9	72.8	68.0	6.3	4.3	85.2	86.3
K-Real estate, renting and business activities	13.9	31.5	23.9	22.8	49.9	34.6	87.7	88.9
M-Education	19.2	45.6	22.6	15.9	1.5	0.5	43.3	62.0
N-Health and social work	4.6	31.1	2.9	4.2	81.9	55.3	89.4	90.7
O-Other community, social and personal service activities	7.8	26.2	64.5	53.9	3.7	1.8	76.0	81.9
TOTAL	13.2	37.0	56.6	45.1	17.0	7.8	86.8	90.0

B. Completeness and accuracy of administrative data

29. In terms of *completeness* w.r.t. the target statistics of the SME survey, it has to be underlined that the considered administrative sources may contain the same information for a subset of key variables: among others, *Number of employees, Turnover*, *Changes in stocks*, *Changes in contract work in progress, Other income and earnings, Purchases of goods and services, Use of third party assets, Other operating charges, Personnel costs;* moreover two further variables, *Value added* and *Gross operating value,* can be derived from the previous ones with some additional data processing activities. Given this situation, priorities among sources are to be set, based on: (i) source coverage (see previous section), (ii) *number of comparable* variables (with few discrepancies in definitions) which are available in the source, (iii) *coherence* to the SME survey variables. *Coherence* between each SME target variable and the corresponding one in each considered administrative source has been assessed based on the Kolmogorov-Smirnoff (KS) test.

30. For each available source in Table 3 we report the number of available comparable variables, the number variables for which the KS test resulted in acceptance of the null hypothesis (the compared distributions are similar), and the priority assigned to the source.

31. Accordingly to priority rules stated above, FS are considered as primary source as it covers almost all the corporate firms and it supplies with the highest number of comparable (21 variables) and coherent (13 out of 21) variables. Further analyses have been carried out on some of the administrative variables apparently non coherent with the SME survey. In particular, *Changes in stocks stocks of finished and semi-finished products* (*CS*) has been investigated in order to understand the reasons of the KS test failing.

Table 3: Number of comparable variables and KS test on sampling vs administrative data distributions, by type of source - Year 2007

Source	*Number of available variables*	*Number of favourable KS tests*	*Priority assigned*
Financial Statement	21	13	1
Fiscal Authority Survey (F)	15	8	2
Fiscal Authority Survey (G)	13	7	3
Tax Return Data – PF - RG	13	6	4
Tax Return Data – PF – RE	14	6	5
Tax Return Data – SP – RG	14	6	6
Tax Return Data – SC - RS	16	2	7

32. *CS* is defined as the difference between two components: *Changes in stocks of finished and semi-finished products* and *Changes in stocks of raw materials and for resale*, which are currently surveyed as the difference of the stocks at the end and at the beginning of the reference period. A deeper comparison of these variables values (both the components are available in FS) allowed us to detect errors mainly due to mistakes in filling in the SME questionnaires (such as sign inversion). The KS test on the remaining data resulted favourable to the null hypothesis: the two variables are similarly distributed. This further analysis confirm FS to be an accurate and reliable administrative source for accounting data of corporate firms.

33. Based on Table 3, and taking into account coverage and completeness, SS is the second best choice with respect to FS because the coverage and the coherence in terms of variables is lower with respect of FS. Tax return data are considered in case no other administrative source is available.

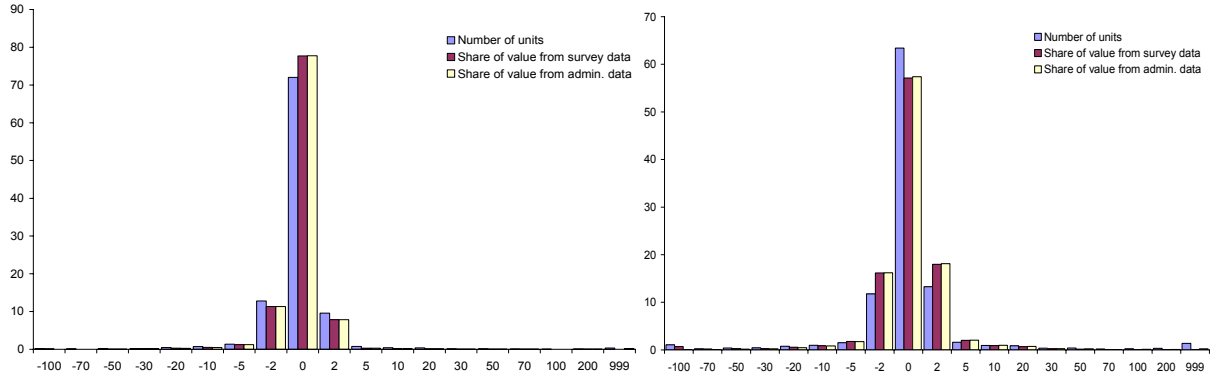
34. Turning to the *accuracy* of the analysed administrative sources, it has bee defined in terms of statistical comparability of variables coming from the different sources. In particular, accuracy has been evaluated by analysing both the definition contents and frequency distributions of the differences between elementary data[1]. Assuming that the variable contents are defined in a comparable way in the two sources, the distributions of the difference ranges between administrative and survey variables have been drawn: in Figure 3 and Figure 4 we report the bar charts obtained by considering FS and SS for variables *Turnover* and *Value Added*, respectively. The horizontal axis represents the classes of relative percentage differences while

[1] Relative differences are computed according to the following: diff(V)/100= [V(Admin Source) – V(Survey)]/V(Survey).

the vertical axis represents the relative frequencies.

35. The left chart shows that around 94% of the surveyed units, linked with FS (19,739 obs.), shows differences for the variable *Turnover* in the range +/- 2% (70% equal to zero: the sources are equivalent). Also for SS the comparison is satisfying: 88% of differences lies between the range of +/- 2%, 63% being equal to zero. In both cases, it can be noticed the distribution of differences is quite symmetric, that supporting the idea of randomness of errors.

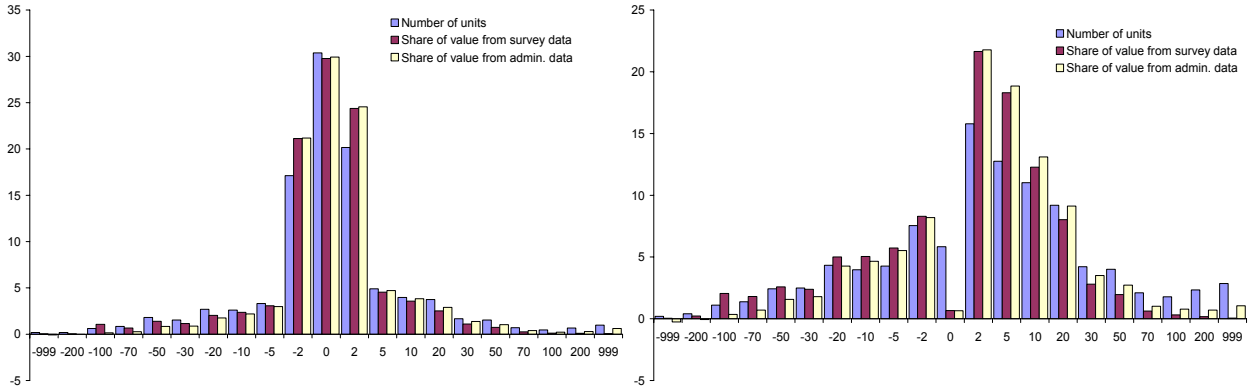
Figure 3: *Turnover*: distribution of sampling and administrative data, by range of differences



Financial Statements (19,739 obs.) **Sector Studies (28,138 obs.)**

36. Variable *Value added* has a more complex situation, since the variable is the result of an algebraic combination of other variables. Therefore, errors in each single components are summed up in the derived variable making the coherence between difference sources more problematic (another example are variables referring to *Changes in stocks* discussed before that contribute to the *Value Added* as well). Anyway, FS (left chart) confirm to be more accurate and reliable for statistical purposes with respect to the fiscal sources[2].

Figure 4: *Value Added:* distribution of sampling and administrative data, by range of differences:



Financial Statements (19,739 obs.) **Sector Studies (28,138 obs.)**

37. It worth notice, *Turnover* and *Value Added* of corporate enterprises (subject to financial statements obligation) are much greater of those reported by enterprises filling in SS. For these last enterprises about 46% of linked observations and 55% of the *Value added* lies between an error of +/- 5%. For both variables, the comparison analysis shows zero-balanced and symmetric distributions of differences that can be interpreted as due to random discrepancies/error.

38. These analyses confirm the priorities assigned to the administrative sources for estimation purposes, privileging FS as the more accurate and reliable followed by SS and then by the other fiscal sources. Under this assumption, it is straightforward to plan a production process where FS and SS represent the major

[2] More details can be found in Casciano et al., 2010.

sources of information for the subset of key SME target variables, on which data validation activities are to be focused using the traditional editing approaches, in particular outlier detection and influential errors.

39. As an example of exploratory data analysis involving FS data, Figure 5 and Figure 6 contain the scatter plot of FS data and SS data for corporate enterprises (logarithmic scale) for Nace divisions 17 (*Textile Industry*) and 52 (*Retail Trade*) for variables *Turnover* and *Changes in Stocks*, respectively. As it can be seen, information coming from the two sources is highly coherent, with few "outlying" values which may be due to either errors in one of the two sources, or not completely harmonised variables definitions.

Figure 4: Nace Divisions 17 and 52: scatter plot of FS *Turnover* vs SS *Turnover* for corporate enterprises (log scale)

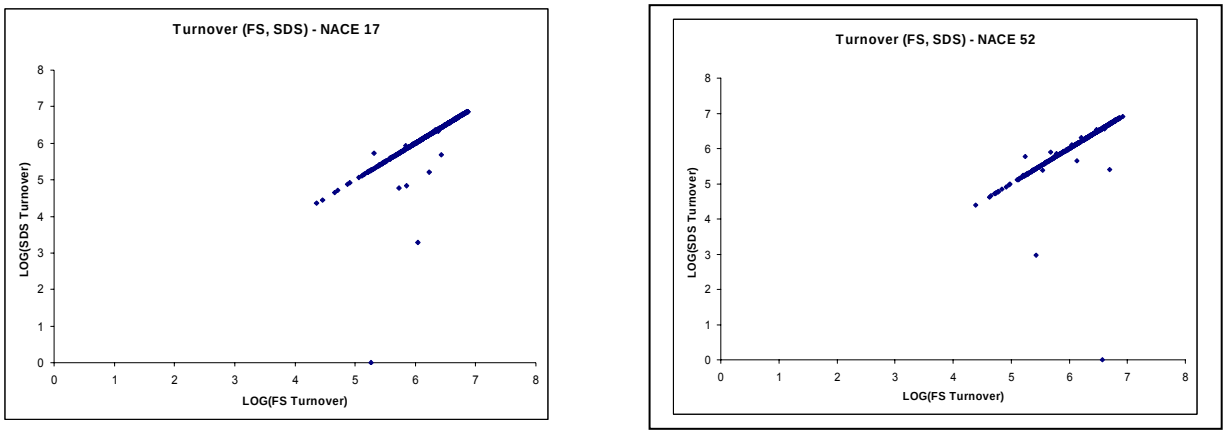
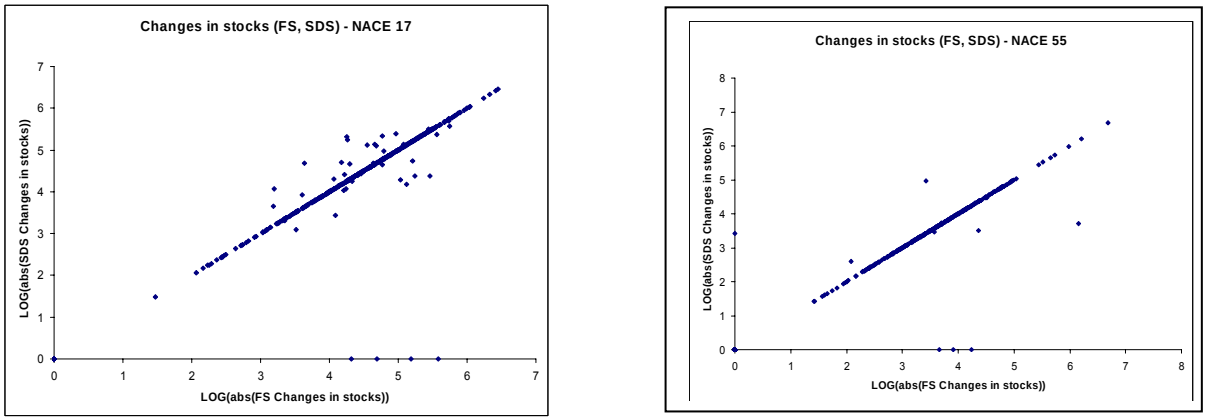


Figure 5: Nace Divisions 17 and 52: scatter plot of FS *Changes in Stocks* vs SS *Changes in Stocks* for corporate enterprises (log scale)



40. For our purposes, these preliminary results can be considered as promising in terms of potential use of multiple (possibly overlapping) sources of (administrative) information in the SBS validation process for SME. The specific methodologies to be adopted need additional evaluation and experimental studies.

C. Coverage of administrative data: potential impact on estimates due to source effect

41. As seen in previous subsections, administrative sources cover almost all the SME sample for a subset of key variables: this allows us to evaluate the potential effect of estimating key SBS variables for SME population only based on the use of administrative information on sampled enterprises. The methodological developments and results resumed in this section are drawn from Casciano et al. (2010).

42. Let S1 be the set of 43.701 SME responding units, w_k the associated sampling weights (referred to as "old" weights) ($k \subset S1$), S2 the set of 88.484 sample units covered by administrative sources, w_k^* the associated sampling weights (referred to as "new" weights) ($k \subset S1 \oplus S2$), y_k and y_k^* the elementary survey and the administrative data for a variable *Y*, respectively. The SME "original" estimate of *Y Total* is:

$$\tilde{Y}_\alpha = \sum_{S1} y_k \, w_k$$

while the corresponding estimate computed by using only administrative date is given by:

$$\tilde{Y}_\alpha^* = \sum_{S2} y_k^* \, w_k^*$$

43. We are interested in measuring the difference between $\tilde{Y}_\alpha$ and $\tilde{Y}_\alpha^*$ due to the use of administrative data as a substitute of sampled data ("source" effect). In order to isolate the source effect from the non response effect, the following estimate is also considered:

$$\tilde{Y}_\alpha^{**} = \sum_{S1} y_k \, w_k^* + \sum_{S2-S1} y_k^* \, w_k^*$$

which represents the estimate of *Y*'s total based on the integrated set of survey data (on S1) and administrative data (on S2). We can write:

$$\tilde{Y}_\alpha^{**} = \sum_{S2} y_k^* \, w_k^* + \sum_{S1} y_k \, w_k^* - \sum_{S1} y_k^* \, w_k^* = \tilde{Y}_\alpha^* - \sum_{S1} (y_k^* - y_k \,) w_k^*$$

44. Given the difference between $\tilde{Y}_\alpha^*$ and $\tilde{Y}$:

$$\tilde{Y}_\alpha^* - \tilde{Y}_\alpha = \sum_{S2} y_k^* \, w_k^* - \sum_{S1} y_k \, w_k$$

it can be proven that: $$\tilde{Y}_\alpha^* - \tilde{Y}_\alpha = \sum_{S1} (y_k^* - y_k) w_k + \sum_{S2} y_k^* \, (w_k^* - w_k)$$

45. Remembering that $\tilde{Y}_\alpha^{**} = \tilde{Y}_\alpha^* - \sum_{S1} (\, y_k^* - y_k) w_k^*$, by subtracting $\tilde{Y}_\alpha$ from both the first and the second member we obtain that:

$$\tilde{Y}_\alpha^{**} - \tilde{Y}_\alpha = \tilde{Y}_\alpha^* - \tilde{Y}_\alpha - \sum_{S1} (\, y_k^* - y_k) w_k^*$$

or in other terms that:

$$DIFF = \tilde{Y}_\alpha^{**} - \tilde{Y}_\alpha = \sum_{S1} (y_k^* - y_k) w_k - \sum_{S1} (\, y_k^* - y_k) w_k^* + \sum_{S2} y_k^* \, (w_k^* - w_k)$$

46. From the last formula it results that, as expected, in order to evaluate the potential effect on SME estimates due to the use of administrative data, three elements are to be considered:

- a source effect, due to differences among survey and administrative data (given the weights structure), which can be further divided in two components depending on the specific weights system (either old or new);
- a non response effect, due to the fact that the survey is affected by a given amount of unit non responses.

47. The differences (*DIFF*) between estimates have been computed for variables *Turnover*, *Total costs* (purchases of goods and services, use of third party assets, other operating charges), *Personnel cost* and *Value Added*. Table 4 shows the decomposition of the differences between the final mixed estimate Y** obtained combining survey and administrative data and the initial (official) estimate Y of the same variables.

48. For variable *Turnover* the total difference is very close (+0,03%) to the official estimate produced with a high variability in results when we breakdown economic activities.. The source substitution effects are -1.07% (old weights) and 0.66% (new weights) respect to the previous estimate. The difference due to the TMR is higher in construction activities and lower in service sector.

49. For variable *Total costs* the mixed estimate is higher than the official estimate produced (+1,26%) with differences in results by economic activities. The source substitution effects are -1.72% (old weights) and 0.86% (new weights) respect to the previous estimate. The difference due to the TMR is 2,12% and is quite similar in industry and services while is reduced in construction.

50. For variable *Personnel cost* the mixed estimate is lower than the official estimate produced (-0,91%) with a lower difference in industry. The source substitution effects are 0.51% (old weights) and -0.21% (new weights) respect to the previous estimate. The difference due to the TMR is -1,21% and is higher in services activities.

51. For variable *Value added* the mixed estimate is lower than the official estimate produced (-4,50%) in special way in construction and service activities. The source substitution effects are -0.15% (old weights) and 0.34% (new weights) respect to the previous estimate. The difference due to the TMR is higher than other variables (-4,68%).

Table 4: Differences between sources by sectors of activity and size class – Year 2007

%DIFF - Total differences in finale estimates (Y**-Y)/Y

Activities	Turnover	Costs	Personnel cost	Vaule added
Industry	1.36	2.15	0.58	-1.88
Construction	-3.53	-1.31	2.30	-6.35
Servics	0.06	1.32	-2.91	-5.32
Total	**0.03**	**1.26**	**-0.91**	**-4.50**
SSw - Source substitution effect for S1 (wiht old weights)				
Industry	-1.16	-1.32	0.30	-1.33
Construction	-0.76	-1.52	0.80	0.51
Servics	-1.09	-1.94	0.57	0.26
Total	-1.07	-1.72	0.51	-0.15
SSw* - Source substitution effect for S2 (wiht new weights)				
Industry	0.73	0.75	-0.13	0.89
Construction	0.82	1.02	-0.16	0.38
Servics	0.60	0.87	-0.28	0.03
Total	**0.66**	**0.86**	**-0.21**	**0.34**
NRD - differences due to TMR				
Industry	1.80	2.72	0.41	-1.43
Construction	-3.59	-0.82	1.66	-7.24
Servics	0.55	2.39	-3.21	-5.62
Total	**0.44**	**2.12**	**-1.21**	**-4.68**

52. In general, the source effect is essentially due to the following reasons: first, administrative data have not been checked, in particular w.r.t the presence of outliers and/or influential errors; furthermore, there exists some discrepancies between the definitions adopted for statistical and fiscal purposes that unavoidably influence the comparison and which need to be carefully analysed and reconciled.

53. As general conclusion, the analysis illustrated in this subsection shows that the parameters estimates differences are more affected by TMR than the substitution of data sources: for the analysed variables, administrative data can be then fully considered as appropriate for estimation purposes. Note that, since the non response mechanism is not random, the use of administrative data for integrating the SME survey TMR also reduces the possible biases caused by the non-respondents self-selection.

D. Completeness of administrative data: potential impact on estimates due to item non responses imputation.

54. As discussed before, administrative data directly provide information which is useful for estimating a subset of SME target variables. The remaining variables are to be estimated by direct surveys, based on the *core* information provided by administrative sources. Anyway, even for variables which can be directly estimated from external sources, a problem of item non response may exist, e.g. for those variables which can be obtained by external sources only for specific SME subpopulations. In this cases, appropriate strategies are to be adopted to compensate for the non available information, depending on the specific variable/information context.

55. As an example of this type of situation, in this subsection we illustrate some preliminary experimental results relating to the estimation of the two components of *Changes in stocks of goods and services* (*CS*), *Changes in stocks of finished products and work in progress* (*Csfp*) and *Changes in stocks of raw materials and for resale* (*Csrm*), as they are involved in the computation of the *Production Value* and *Gross margin on goods for resale*. The following relation holds:

$$CS = Csfp - Csrm \qquad (1)$$

56. For corporate enterprises, all the variables involved in (1) are available from FS. For other enterprises (un-incorporate and sole proprietorship enterprises, which are not covered by FS), only variable *CS* is directly available from the other administrative sources, the problem being to estimate the *Csfp* and *Csrm* values. We assume that MRPs are Missing At Random (Little et al., 1987), so that we are allowed to treat them as "similar" to the fully observed units inside appropriate domains.

57. An experimental MonteCarlo simulation study has been performed to assess the potential biasing effects on *Csfp* estimates[3] due to direct imputation of its (partial) non responses. Target parameter is the *Csfp Total* by domain *D*, defined as $\hat{T}^D_{Csfp} = \sum_{k=1}^{n_D} w_k Csfp_k$, where n_D is the number of units in domain *D* ($\sum_D n_D = n$ is the sample size), and w_k are the sampling weights.

58. The simulation consisted in I=100 iterations of the following steps: 1) simulation of 5% MRPs on *Csfp* and *Csrm* on a random sample of un-incorporated and sole proprietorship enterprises; 2) imputation and estimation of *Csfp* total; 3) evaluation.

59. Evaluation is based on the (*Mean) Relative Estimation Error due to Imputation/Estimation* (*REEI*), and on the *Relative Root Mean Squared Error* (*RMSE*) of parameter estimates (by domain):

$$REEI^D = \frac{1}{100}\sum_{I=1}^{100} \frac{\left|\hat{T}^D_{Csfp,ori} - {}_I\hat{T}^D_{Csfp,imp}\right|}{\hat{T}^D_{Csfp,ori}} \qquad (2)$$

where $\hat{T}^D_{Csfp,ori}$ is the total of *Csfp* in each domain *D* estimated on original data, and ${}_I\hat{T}^D_{Csfp,imp}$ is the corresponding estimate obtained from imputed data at iteration *I*.

60. Within cells nearest-neighbour donor (NND) imputation, within domains robust regression imputation (RR), and within cells median imputation have been comparatively evaluated. In NND, the imputed value at unit level is the proportion $p_i = \frac{Csfp_i}{CS_i}$ observed in the closest complete unit in the imputation cell. In RR, estimates of model parameters β_j in the simple model $Csfp_i = \alpha + \beta_j X_j + \varepsilon_i$, are obtained based on the Least Trimmed Squares algorithm (Rousseew at al., 1987) as implemented in SAS language. In within cells median imputation, the imputed value at unit level is the p_i cell median (zero values are excluded from calculations). In both RR and median imputation, a preliminary logistic regression is used to classify units based on their probability of having either zero or non zero changes in stocks components, by domain.

61. Auxiliary information used in all methods are *Economic activity* (either 2- or 3-digits Nace code), *Number of employees*, *CS*, *Turnover*, *Purchases of goods and services for resale in the same condition as received*. Different experiments have been performed by changing the criteria used to form imputation cells for each evaluated method. We report in Tables 5, 6 and 7 the results at 2-digits Nace code level for divisions 17, 52 and 55 (*Hotels and Restaurants*), corresponding to the "best" criteria in terms of *REEI* and *RMSE* for each imputation method.

62. These preliminary results show that methods performing better seem to be robust approaches (within-cells regression models and median imputation), which explicitly take advantage of the correlations existing between *Csfp* and the used auxiliary information, while reducing the influence of anomalous behaviours on missing data predictions. These outcomes are also confirmed by the corresponding results at 3-digits Nace code level for the same divisions (which are not reported here for shortness).

[3] *Csrm* will be deductively derived from relation (1), as *CS* is assumed to be known from the available administrative sources for the considered business domains

63. As a general conclusion, for the considered variable, imputation seem to be appropriate to compensate for partially not available data in administrative sources. Additional experiments and analyses are needed, especially for specific Nace divisions and enterprise subpopulations. Furthermore, if we extend this type of evaluation of approach to other SME variables which could be potentially affected by partially not available data in administrative sources, it is evident that an additional modelling effort will be needed to define and implement an overall imputation strategy in the new SBS production process for SME.

Table 5: quality indicators by domain (2 digits Nace) and imputation method (form of imputation/model estimation cells) – Year 2007

Method	*NACE*	*REEI*	*RMSE*
NND (3 digits Nace+Legal form+CS sign)	**17**	0,050	0,060
	52	0,060	0,100
	55	0,110	0,160
RR (2 digits Nace+Legal form)	**17**	0.020	0.040
	52	0.030	0.030
	55	0.010	0.020
Median (3 digits Nace+Legal form+Size+CS sign)	**17**	0.024	0.032
	52	0.013	0.020
	55	0.001	0.003

V. Conclusions and future work

64. In this paper the re-design project of the SBS production process for the Small and Medium Enterprises (SME) survey ongoing at Istat is briefly illustrated. The project aims at reducing statistical burden on enterprises and statistical production costs by exploiting as much as possible information on Italian enterprises which is already available from administrative sources of information (Financial Statements and fiscal information collected by the Italian Tax Authorities). A summary of quality problems relating to the integrated use of administrative and survey data is provided, and a potential structure of a new SME survey integrated strategy is delineated.

65. From the proposed scheme, it is evident that the re-design project will have a strong impact on the overall E&I strategy in terms of process organization, adopted methodologies and technologies and data management. The aim of the paper is to illustrate, on the basis of some preliminary data analyses and experimental studies, which is the potential impact of the new integrated production process in terms of additional data validation, data modelling and non responses imputation.

66. In effect, besides the traditional E&I activities to be designed and implemented on sampling survey data, additional validation activities with respect to the traditional ones are needed to guarantee accessibility, usability and quality of administrative data, especially in terms of completeness, coherence and accuracy.

67. The results obtained so far, partially presented in the paper, are very promising and encourage us in proceeding with further analyses and developments, in order to achieve the planned objective in the very next future. Next steps in this direction will consist in analysing more in depth the characteristics of SME population and target variables, better specifying the data processing flow of the new SBS production process fro SME, and identifying the most appropriate E&I approaches for the different data problems and data types involved in the new integrated process, at the different phases of the data processing flow.

References

Brion P., Gros E. (2009), Methodological issues related to the reengineering of the French structural business statistics, *Proceedings of the European Establishment Statistics Workshop (EESW09)*, Stockholm.

Casciano M.C., De Giorgi V., Oropallo F., Siesto G. (2010), *Experimental Analysis in the estimation of SBS variables for small firms by using administrative data*, paper presented at the *Seminar on Using Administrative Data in the Production of Business Statistics - Member States Experiences*, Rome, 16-18 March.

Chumbau A., Pereira H. J., Rodrigues S. (2010), Simplified Business Information (IES): Impact of Admin Data in the production of Business Statistics, presented at the *Seminar on Using Administrative Data in the Production of Business Statistics - Member States Experiences*, Rome, 18-19 March, http://www.ine.pt/filme_inst/essnet/papers/Session3/Paper3.6.pdf.

Daas P.J.H., Arends-Tòth J., Schouten B., Kuijvenhoven L. (2008), Quality Framework for the evaluation of Adminsitrative data, *Proceedings of the European Conference on Quality in Official Statistics (Q2008)*, Rome, 8-11 July.

Daas P.J.H., and Fonville T.C. (2007), *Quality control of Duch Dutch Administrative Registers:An Inventory of Quality Aspects*, paper presented at the Seminar on Registers in Statistics - methodology and quality, Helsinky.

Elliott, D., (2010), The potential use of additional VAT data in ONS business surveys, *Proceedings of the European Conference on Quality in Official Statistics (Q2010)*, Helsinki, 4-6 May.

Elliott, D., van Elswijk D., Luzi O., Siesto G., Redling B., Kavaliauskiene D. (2010), Methods of estimation for business statistics which cannot be obtained from administrative data sources, *Proceedings of the European Conference on Quality in Official Statistics (Q2010)*, Helsinki, 4-6 May.

Eurostat, European Commission (1999), Use of Administrative Sources for Business Statistics Purpose, *Handbook on good practices*, 1999 Edition.

Eurostat (2003), *Item6 - Quality Assessment of Administrative Data for Statistical Purposes*, Luxemburg.

Haitzmann M. (2010), Model based estimation of enterprises below thresholds in Structural Business Statistics, paper presented at the *Seminar on Using Administrative Data in the Production of Business Statistics - Member States Experiences*, Rome, 16-18 March.

Hoogland J. (2010), Editing Strategies for VAT Data, paper presented at the *Seminar on Using Administrative Data in the Production of Business Statistics - Member States Experiences*, Rome, 16-18 March.

Lewis D. (2010), Integrating data from different sources, in the production of business statistics (WP5), *Proceedings of the European Conference on Quality in Official Statistics, (Q2010)*, Helsinki, 4-6 May.

Little, R. and D. Rubin (1987), *Statistical Analysis with Missing Data*. Wiley & Sons, New York.

Lorenz R. (2010), The integrated System of Editing Administrative Data for STS in Germany, paper presented at the Seminar on *Using Administrative Data in the Production of Business Statistics - Member States Experiences*, Rome, 16-18 March.

ONS (2005) *Guidelines for measuring statistical Quality*, version 3.0, London: Office of National Statistics.

Rousseew P.J., Leroy A.M. (1987), *Robust Regression and Outlier Detection*. Wiley & Sons, New York.

Tolkki V. (2007), Finnish SBS System: use of administrative data, methods and process, presented at the *Seminar on Reengineering of Business Statistics*, Lisbon, 11-12 October.

Yung, W., Lys P. (2008), Use of Administrative Data in Business Surveys - The Way Forward, Statistics Canada - *IAOS Conference on Reshaping Official Statistics* - Shanghai, 14-16.

Wallgren A., Wallgren B. (2006), Register-Based economic statistics on enterprises – Editing Issues, *UNECE Work Session on Statistical Data Editing*, Bonn (Germany).

Wallgren A., Wallgren B. (2007), *Register-based Statistics: Administrative Data for Statistical Purposes*, John Wiley & Sons.

Zach S. (2007), Model based estimation of enterprises below thresholds in Austria, presented at the *Seminar on Reengineering of Business Statistics*, Lisbon, 11-12 October 2007.