WP.23
ENGLISH ONLY

UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Oslo, Norway, 24-26 September 2012)

Topic (iii): Editing and Imputation in the context of data integration from multiple sources and mixed modes

Imputing Missing Values When Using Administrative Data for Short-Term Enterprise Statistics

Prepared by
Pieter Vlag, Statistics Netherlands, the Netherlands

I. Introduction

1. National Statistical Institutions (NSIs) are often faced with the problem that administrative data are not available at the time when they are needed for enterprise statistics. This is particular the case when using Value Added Tax (VAT) and Social Security data for Short Term Statistics (STS). Turnover and employment and wages may be derived from VAT and Social Security data, respectively. However these administrative data are still incomplete when the monthly or quarterly turnover estimates have to be provided. This incompleteness might be temporal (e.g. to late response of single units because the deadline for reporting is later than that for the statistical estimates) or structural (e.g. because units below certain thresholds are only available for a different periodicity).

2. As this estimation problem is common, an European ESSnet-project was started. In a first step, existing practices to impute and check incomplete admindata for monthly or quarterly estimates have been investigated. In a second step, the most promising practices were tested and compared.

3. The project distinguishes two situations. Situation I is that the available admin data provide good coverage when the estimates have to be made. This situation applies in general to regularly produced quarterly estimates, because almost all commercial enterprises in continental Europe have to declare their VAT and Social Security data on a monthly or quarterly base and the deadline for reporting is– depending on the country – between 30 to 40 days after the reporting period. The other practices are used when no or limited and non-representative admindata are available (situation II). This situation mostly applies for monthly estimates, because part of the enterprises declare per quarter and some deadlines for monthly statistical publications are early, e.g. before the deadline of reporting to the tax office.

4. Many National Statistical Institutes (NSIs) use a coverage of about 80% of the total turnover by the large enterprises survey and available admin data, before they feel that reliable figures of turnover can be published in a certain publication cell (see, e.g., Vlag, 2012). This percentage can, of course, be decreased or increased, depending on the coverage of the large enterprise survey and how much risk a NSI is willing to take. Pragmatically we have used this previously mentioned 80 % coverage as the lower limit for situation I “good coverage of admin data”. In practice, we have noticed that that the coverage of the large enterprise survey plus available admin data often exceeds 90 or 95 % when making STS-estimates.

5. This paper deals with situation I only. The main administrative data sources for producing the STS indicators are VAT data (for turnover indicators) and Social Security data (for employment and wages indicators). In the remainder of this document we focus on VAT data, but unless noted otherwise, VAT-data will in the context of this document refer to administrative data in general. For specific information on using the social security data for STS-estimates we refer to Baldi et al. (2011).

II. The statistical process for producing short term statistics with administrative data

A. Overview

6. The general set-up when utilizing VAT data for producing STS is that a combination of a survey and VAT data is used (see, e.g., Orchard, Moore and Langford, 2011; Karus, 2012; Kavaliauskiene, 2011; Lorenz, 2011; and Šličkutė-Šeštokienė, 2011, for some current approaches to using VAT data for STS in EU countries). In the survey the large enterprises are generally completely enumerated. Large enterprises often have a complex structure. Correct observations from those large enterprises are considered crucial for producing reliable STS figures. For the remaining small and medium enterprises VAT data are used instead of direct observations by the NSI. In some specific cases only a small sample may be observed directly for small and medium enterprises In other words, the general system of admin data based STS-estimate is a survey under the large enterprises and use administrative data, i.e. VAT data in our case, for nearly all remaining smaller enterprises.

7. When administrative data are used for economic and STS-statistics, a number of steps in the statistical process can be distinguished that require special attention (Teneva, 2012):

- Data transfer from the tax office to the NSI;
- Matching administrative units to statistical units;
- Dealing with differences in definitions;
- Administrative data editing and outlier detection;
- Determining the active population;
- Estimation/imputation methods;
- Determining the active population.

This paper covers the last two points. For further information about the other steps, we refer to De Waal et al, 2012.

B. Growth rate estimates versus level estimates for STS

8. When using VAT data to produce the STS, one has to choose whether the aim is to produce estimates for population totals (and implicitly also for growth rates) or for growth rates only. At first sight there may not seem to be a big difference between the two choices. After all, by comparing the population total of the current period to that of a previous period, one can estimate the growth rate between these two periods. Conversely, given the growth rate between two periods and the population total in the first period, one can estimate the population total in the second period. However, there is a big difference between the two choices.

9. To estimate the growth rate it is often sufficient to draw a sample for both periods and use the growth rate of the respondents common to both periods as an estimate for the growth rate in the population. Population characteristics, such as the size of the population in terms of the exact number of active units (enterprises) are often not required. One may even reduce the sample to those common

responders which do not present outlying or suspicious values, or neglect differences in definition which may affect levels but – on average – not growth rates. This does not only the simplify the data-editing and data-checking process but also the estimation process.

10. However, the option generally chosen by Statistical Institutes in Europe in situation I is producing both levels and growth rates. This implies that more information can be published, but that the population of the current period should be known or estimated. In practice, the latter is the case. Depending on the construction of the statistical business registers (SBR) time-lags between the SBR and VAT-data concerning starting, stopping and merging enterprises may arise. This implies that in case of missing VAT data for current period the question should raised whether the VAT-data is missing due to late reporting or because the enterprise is not active anymore. Similarly, the question raises whether VAT records new enterprises timely.

C. Dealing with missing data: Estimates for growth rates versus estimates for population totals

11. Another fundamental issue is whether estimates are produced on a micro level, i.e. for individual enterprises, or on a macro level, i.e. combinations of branch of industry and size class. Both approaches can be used, but producing estimates on a micro level has some practical advantages over producing estimates on a macro level.

12. An important practical advantage is that changes in the structure of enterprises and the population, for instance enterprises moving from one branch of industry or size class to another, can easily be dealt with when estimates are made on the micro level. Another advantage is that with estimates on the micro level one can easily take special characteristics of individual enterprises, such as enterprises with a limited number of employees but with a relatively high turnover, into consideration, resulting into a more thorough control of problems arising in the estimation process due to measurement errors or missing values. This is especially important in case it is decided that enterprises above a certain threshold (or with a certain activity) are surveyed. A third advantage is that estimates at the micro level make further processing rather easy: one simply has to aggregate over all enterprises in a publication cell to obtain an estimate for that publication cell. This also allows higher flexibility in deciding the breakdown of the released estimates and allows one to provide estimates at more detailed domains.

13. We have observed that estimates in situation I are generally produced on a micro level while for Situation II, where the administrative data are incomplete or unavailable, many NSIs do produce estimates on a macro level instead. Main reason for producing micro level estimates in case of situation I is that by far most data are already available (and the datasets enormous), allowing statistical institutes to construct an enriched dataset with turnover data of almost all enterprises to which other survey might be connected. In this document we therefore focus on producing estimates on a micro level. This means that estimation of missing data means in practice imputation of missing data. Therefore, we discuss imputation of missing data only in this paper.

D. Dealing with missing data: Summary

14. The most important conclusions are

a) Missing VAT-data are imputed at micro level.

b) Statistical offices produce both levels as growth rates for quarterly turnover estimates. As a result the population has to be estimated . When facing the problem of missing VAT-data the question should be raised whether the data are missing because of

   i) late reporting or reporting for another periodicity. In this case the missing VAT has to be imputed.

   ii) the enterprise has stopped. In this case the missing VAT should not be imputed.

In the remainder of the document, we first discuss imputation techniques (point a)), before addressing the question which unit should be imputed

III. Imputation of missing VAT: methods

A. Methods to impute missing data by assuming that the available VAT is representative for the target population

15. The most common practice for imputing missing values is using the turnover of units with complete and plausible data. These units are divided into several groups by size classes, NACE-codes and in some cases the period of VAT declaration (monthly/quarterly payers). For each of these groups the group-specific growth rate of the total turnover is calculated. Some countries – like Germany - use as growth rate O_t/O_{t-1} (current period versus previous period), other countries – like the Netherlands - use as growth rate O_t/O_{t-12} (current period versus same period last year). The stratification level for which the group-specific growth rate are calculated may also differ per country. Germany, for example, uses two size classes (Threshold: 15.000.000 € turnover per year), NACE-two-digits and the period of VAT declaration (monthly/quarterly payers). The Netherlands, for example, uses more detailed level: nine size classes (defined by persons employed), NACE-three/four-digits. The advantage of more detailed groups is that groups are theoretically more homogeneous. Disadvantage is that the number of (donor)units becomes too small, which increases the effect of outliers ect.

16. The main assumption behind these imputations rules is that the available VAT is representative for the units without data. This assumption seems realistic, because of the completeness of the data (in practice 90 % of the estimated VAT-turnover is observed when the estimates have to be made). In case of very large enterprises, in particular outstanding market leaders, this assumption may not be correct. However, these large enterprises are for this and other reasons surveyed.

17. Using of O_t/O_{t-12} instead of O_t/O_{t-1} ratios for imputation has advantages and disadvantages. The same is true for the choosing high instead of low aggregation levels at which these ratios are calculated. However, as these techniques are used when at least 80 and generally more than 90 % of the estimated VAT-turnover is available, it is a matter of debate whether these theoretical pro's and con's lead in practice to differences in publication (Lorenz, 2011). Especially because the large enterprises with a large impact on publication are surveyed To test this hypothesis, Statistics Estonia has tested the following imputation rules on enterprises with NACE-code 47 (retail trade):

1) imputation with O_t/O_{t-12} ratios of available VAT; calculated at NACE-2digit level.
2) imputation with O_t/O_{t-12} ratios of available VAT; calculated at (a more detailed) STS-output level.
3) imputation with O_t/O_{t-1} ratios of available VAT; calculated at NACE-2digit level.
4) imputation with O_t/O_{t-1} ratios of available VAT; calculated at (a more detailed) STS-output level.

18. These results have been analyzed for NACE 47 and its underlying publication levels as defined by the STS-regulation. The selected period was 2010-2011. On a quarterly base, the largest differences between the four methods in terms of growth rate are smaller than 0,3 %. No method produces significantly higher or lower growth rates. The same conclusion can be drawn for levels. Revisions compared to final estimates – when all VAT are available - are lower than 0,5 % under the assumption that the population is fixed within a year. The results were confirmed by tests on VAT-data in Finland. Therefore we have concluded that if 80 % or more of estimated turnover is available, the available information is representative and covers such high part of the population that the STS-estimates are insensitive with respect to which exactly chosen imputation method despite the theoretical differences ! This conclusion applies to imputation technique. It is not valid for the question which units have to be imputed.

B. Imputing missing values in case of only publishing growth rates?

19. The current imputation method for Statistics Finland's differs fundamentally from those in other countries. It is based on data from previous periods instead of stratum-imputations. Moreover, it is only valid for calculating growth rates. Estimation of missing data is based on five different imputation rules calculated at micro-level. Imputations are made for the latest month and the second latest month to the pairs of values thus obtained. If an enterprise has missing data for three or more months, imputation is not performed and the imputations performed earlier are deleted since business is assumed to have stopped.

20.Five different imputation rules are used for enterprises with missing data (t = month to be estimated):

$$\hat{x}_t = \frac{x_{t-1} + x_{t-2} + x_{t-3}}{x_{t-13} + x_{t-14} + x_{t-15}} \times x_{t-12}$$ Mean annual change

$$\hat{x}_t = x_{t-1}\sqrt{\frac{x_{t-1}}{x_{t-2}}\frac{x_{t-2}}{x_{t-3}}}$$ Geometric mean of monthly changes

$$\hat{x}_t = x_{t-1},$$ Previous turnover

$$\hat{x}_t = avg(x_{t-1}, x_{t-2}, x_{t-3}),$$ Mean turnover

$$\hat{x}_t = x_{t-12},$$ Turnover of comparison month

21. These five imputation rules are tested by comparing the results of each rule with 'real' (already arrived) data for the five to seven previous periods. A model with less than 20% maximum proportional forecast error and the same direction of change as in the last month and in the median change is admissible. A model with greater than 20% and less than 50% maximum proportional error or a different direction of change than in the last month and in the median change is non-admissible, but can be accepted to the calculation by the statistician. The imputation rule in which the largest prediction error is the smallest is chosen. Due to strict criteria, the proportion of missing observations for which a value is imputed is fairly low and can this method only be used to calculate growth.

22. Kiema and Remes (2012) compared the results of this approach with the results of the more common O_t/O_{t-12} and O_t/O_{t-1} imputation rules[1]. These authors also combined the O_t/O_{t-12} and O_t/O_{t-1} imputation rules with the Finnish approach: comparing O_t/O_{t-12} and O_t/O_{t-1} imputations with 'real' (already arrived) data for the five to seven previous periods and admitting imputations only when the maximum proportional forecast error is less than 20 %.

23. The results of the analyses of Kiema and Remes are summarised in figure 1. It shows that for a large cell with many enterprises and VAT-data like NACE 47 (retail trade), all imputations methods provide similar results and that the added values of imputing missing value when publishing growth rates only can be questioned. For the other smaller cells – five have been selected for this figure – the current Finish method slightly improves the estimates in general. The O_t/O_{t-12} and O_t/O_{t-1} imputation rules, however, may for the smaller cell slightly increase the revisions compared to the 'non imputation' approach. Hence, the added value of applying these such imputations rules is questionable when publishing growth rates only even after adding restrictions by not admitting all imputations.

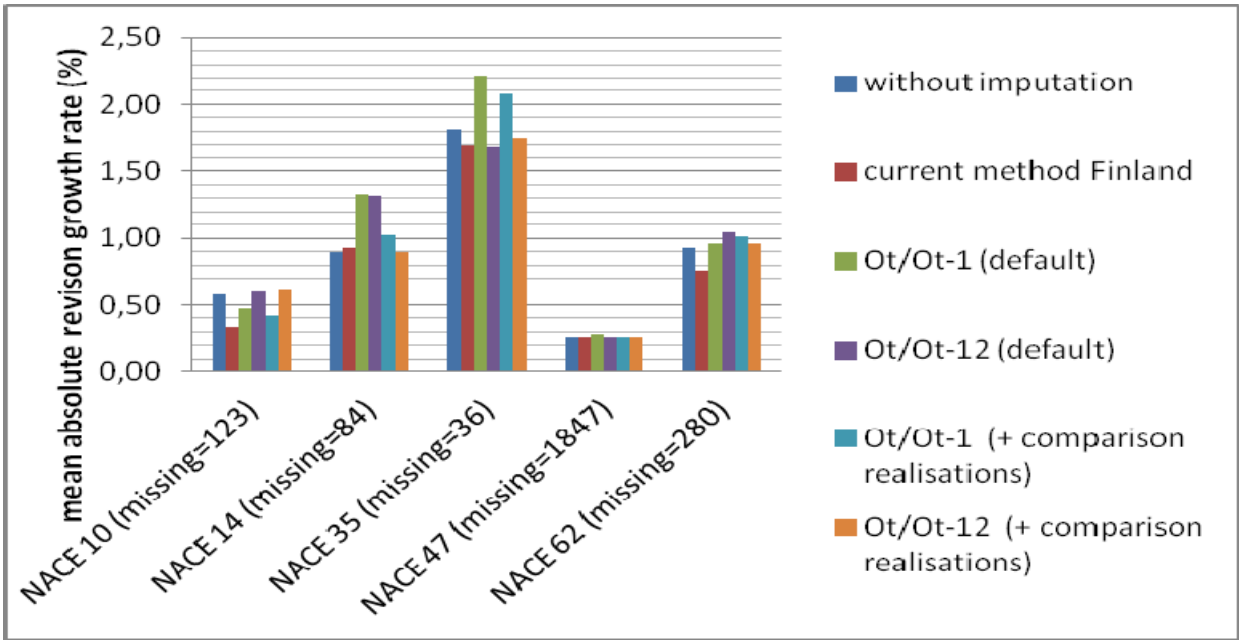
24. Possible explanations that standard O_t/O_{t-12} and O_t/O_{t-1} imputation rules might slightly deteriorate the results for small groups are:

a) Although an outlier filter has been applied when calculating O_t/O_{t-12} and O_t/O_{t-1} ratios, effects of heterogeneity on these ratios cannot excluded in small cells.

[1] When calculating Ot/Ot-1 and Ot/Ot-12 imputations, these authors used as stratification level: NACE-2 digit + 4 turnover size classes.

b) For large groups with many VAT-data available and a high share of VAT to the total estimate, the assumption that available VAT is representative can be made (chapter IIIa). However, it is probably not entirely correct. In countries with high automatic 'fining' rules of the tax office in case of late VAT-reporting (like Finland) one might assume that 'missing VAT' due to late reporting might have specific reasons (delays due to merging/splining or special projects, problems with systems/invoices etc.). This effect may become slightly visible at the smallest publication levels, because in contrast to the large groups the varying effect of these specific reasons on the later reported VAT-turnover may not average out.

Figure 1. Comparison of the mean absolute revision in growth rate between regular STS-estimates with incomplete VAT and final STS-estimates with complete VAT for the period 06/2008–09/2011. The better the regular STS-estimate, the lower the mean absolute revision. This figure shows five selected NACE-groups (x-as) in the case of a) no imputations and b) five different imputation methods. Note that missing units are relatively constant over the NACE-codes. Hence, a large (small) number of missing units within a NACE-group can be related to many(few) enterprises with this activity.



25. On the other hand we note that the differences between all different imputation models and the 'no imputation' scenario are very limited. This suggest that the results are fairly insensitive with respect to which estimation/imputation method is used, and that determining the active population (i.e. which units have to be imputed) might be more important. In this perspective, we recommend that one should aim for an optimal trade-off between benefits and costs when developing imputation rules for missing VAT-data (known) to be imputed rather than aiming for the "optimal" quality when choosing an imputation method. The exception might be small groups with persisting late reporting (Baldi et al., 2012). Pending on the desired output quality for these groups and knowing the possibility that the major assumption of the O_t/O_{t-12} and O_t/O_{t-1} imputation rules (available VAT is representative for the population) might not be entirely correct , one might consider to publish only growth for these groups and apply an approach like Statistics Finland by using historical data with comparison with realisations.

IV. Imputation of missing VAT: which units to be imputed

A. Active population in an administrative data system

26. A distinctive feature of STS based on administrative data derives from the fact that the administrative data for a reference period normally represents the population of enterprises active in that

very period. In other words, the administrative data provides a representation of the currently active population of enterprises alternative to that of the SBR.

27. To illustrate the problem, let us suppose for the moment that the administrative data sent to the NSI for the STS deadline is not affected by late reporting, i.e. all the enterprises subject to reporting are included in the data delivered. Depending on how up to date the SBR is, the representation of the currently active population by the administrative data may have some advantages over that of the SBR[2]. In fact, since the obligation of sending the VAT declaration arises from the presence of non-zero turnover, the administrative data include even enterprises starting in the reference period and do not include enterprises that have stopped in the previous period or before. This may not hold true for the SBR available for the STS deadline due to the reception schedule of the base sources of the SBR and/or to the time necessary to process these sources. In practice, this implies that the match with the SBR will result in a fraction of enterprises found in the SBR and not in the administrative data and, vice versa, a fraction of enterprises found in the administrative data and not in the SBR. Apart from matching errors these two fractions consist, respectively, of the enterprises for which the end (or suspension) of activity has not yet been registered in the SBR and the enterprises whose new activity has yet not been taken into account in the SBR. The definition of a list of enterprises currently active (i.e. the currently active population) may thus require to update the SBR with information coming from the administrative source.

28. This determination of the currently active population poses no problem for the final estimate for which all the active units have reported.

B. Active population and STS-estimates

29. Since not all units report in due time there is uncertainty for STS-estimates on which units are active and which are not. On one hand, if a unit reporting for a previous period is missing, it might be only late (late reporter) or might have become inactive (stopping enterprise). On the other hand, while the list of reporters will generally include starting enterprises (or enterprises recovering from a suspension) in the reference period, some of these starting enterprises will report only afterwards. In other words, for the preliminary estimate a *provisional active population* has to be compiled. This is composed by the set of early reporters for that month, which are obviously active, and the set of expected late reporters, i.e. the enterprises that are assumed to be still active. This latter set consists of the enterprises for which the values of the target variables have to be imputed. The quality of the *provisional active population* can at later stage be checked with the *final population* consisting of all VAT-units (Baldi et al., 2012).

30. For the final estimate, when the administrative data will include also the late reporters and a more up-to-date SBR is available, it is possible to verify the actual status of an enterprise (active of non-active) versus the estimated status (see Table 1). More in depth it is possible to evaluate which units assumed active are actually active (cell *b*) and which are not (cell *e*). Among the units which have not been assumed active it is possible to evaluate which are (actually) active (cell *c*) and which are not (cell *f*). Cell *d* is logically empty.

31. The quality of the estimates strongly depends on the magnitude of the sets represented in Table 1. In particular it is important to remark that even if the first best solution would be to minimize (in terms of units and turnover – actual or imputed) the magnitude of the sets in cells *c* and *e* and maximize the magnitude of the set in cells *b* and *f*, for aggregate estimates it is sufficient that the sum of (imputed) turnovers of units incorrectly assumed active (the set in cell *e*) roughly compensates the sum of the (actual) turnovers of the units incorrectly not assumed active (the set in cell *c*).

32. The most important issue is how to estimate whether an unit is assumed active (group b) or assumed inactive (group d) in case active or not. This will be dealt in next chapter.

[2] At least with respect to the STS indicators. More generally, the relation between the two populations depends also on how well the definition of activity according the SBR matches with the VAT reporting of non-zero turnover.

Table 1. Units (expected) status in the provisional population compared to the (actual) status in the final population.

		Status in Provisional Population		
		Active = early reporter	**Assumed active = expected late reporter**	**Assumed NON active**
Status in Final Population	**active = reporter**	Active (a)	Correctly assumed active (b)	Incorrectly assumed non-active (c)
	non active = non reporter	- (d)	Incorrectly assumed active (e)	Correctly assumed non-active (f)

33. Detecting starting enterprises is less an issue, even for small enterprises. Starting enterprises are reported by the Chambers of Commerce and/or the tax office. Subsequently they are, generally with some delay after they started, included in the SBR. If this delay is small, it should be present in the SBR and this list is sufficient to cover all starting enterprises. If this delay is longer, a part of the starting enterprises might not be included in the SBR but in the VAT-data. However even in this case these enterprises can be included in the population (and in the estimates) provided that a reliable NACE code can be obtained from the administrative data itself or from another data source (de Waal et al., 2012).

C. Estimation of assumed active and assumed inactive enterprises

34. The most important issue with respect to determining the active population is how to detect whether an enterprise has stopped its economic activity. This is especially a problem for small enterprises. Larger enterprises are generally well-recorded and are usually timely updated in the SBR. This problem is enhanced by the possibility that enterprises that stop do not always report this to the Chambers of Commerce and/or the tax office. Therefore the SBR might include them improperly for a long time after their closure. Alternatively, the tax register may apply different rules than the NSI for declaring the administrative unit (enterprise) dead. For example, the tax authority may need to keep the enterprise alive until all outstanding transactions between it and the tax office are completed. Hence, whatever the quality of the SBR and/or VAT-register in several countries a method to determine whether a missing unit is active or not has to be developed.

35. A common method is used at several NSIs for determining whether enterprises still are active: NSIs simply check whether the enterprise has reported any turnover to the tax office for the last few months. When the enterprise has not reported any turnover to the tax office for the last x months in the case of monthly VAT declarations, or the last x quarters in the case of quarterly declarations, the enterprise is considered inactive, otherwise it is considered to be still active. The common method differs across various NSIs with respect to the value of x, and with respect if one looks only at the previous months or also at the next month(s)[3]. For instance, for monthly turnover estimates, Lorenz (2012) suggested that considering an enterprise for which the turnover value is missing for more than one month as stopped produces results that are sufficiently reliable. The value of x may even differ for different

[3] The choice of x strongly depends on the timing of the reception of the administrative data at the NSI and the way these data accumulate over subsequent data deliveries for the same period.

branches of industry within the same NSI. For monthly retail trade estimation at Statistics Estonia, for example, enterprises are regarded as stopping enterprises when they have not submitted declarations to the Tax and Customs Board for more than 1 month after the reference month, whereas for other branches of industry at Statistics Estonia the value of *x* is set to 6 months. At Statistics Finland, if no input has been received from an enterprise for 3 months, it is considered as stopped.

36. An alternative approach is to compare the VAT data with another data source, such as social security data. This approach is, for instance, used at Statistics Estonia, besides the common approach sketched above. In this alternative approach an enterprise that does not report turnover to the tax office, but that does pay social security fees is considered to be active. It depends on the timely availability of social security data, or similar data sources, if this approach can be used.

D. Impact of the determination of the active population on imputations

37. Lorenz (2011) results suggested that, depending on the NACE-code and period, 5- 10 % of the missing units should not be imputed for the STS-estimate. This because they are incorrectly assumed to be active (group e in table 1). Although their impact on the revision between the first and final estimate is partly compensated by late VAT-reporters which were not imputed in the first estimate (group c in table 1)[4], Lorenz (2011) suggested that the uncertainty of population is a major source of revision in an admin data based STS-estimate. Its contribution to the total revision is much larger than how units are imputed.

38. Starting from the German experience on VAT data, Italy has experimented a method to impute the missing values of employment due to late reporting in the Social Security data. From a methodological point of view, there is no fundamental difference in testing these experiments on VAT or Social Security data. The Italian experiments consist of two stages:

a) the determination of the *provisional active population.* Defined by the available admin data plus a *list* of missing units to impute (according to table 1 in chapter IVb).

b) the *imputation stage* consisting in the assignation of values to the list of units to impute.

39. The quality of this approach was tested by comparison with data of t+12month (a year later). With this comparison both the quality of the provisional active population and the imputation itself have been tested. As imputation models O_t/O_{t-12}, O_t/O_{t-1} and a regression model have been tested (Baldi et al., 2012). The data have been tested on quarterly estimates in the period: 1st quarter 2008 – 3rd quarter 2010. The results are summarised in table 2.

40. Table 2 shows that the coverage of admin data when the estimates have to be made is > 95 % (column *a*). Therefore, the results of the imputation methods them self are adequate if it is correctly assumed that the unit has to be imputed. This can be seen by same values in columns *b* and *c*[5]. As already indicated by Lorenz (2011), table 2 show that the revisions are mainly caused by uncertainty in the *provisional active population*. The values in columns *d* (wrongly imputed values as incorrectly assumed active) and *e* (wrongly not imputed values as incorrectly assumed not active, in practice late reporting starting enterprises) are relatively high. They do not compensate leading to a small bias in the estimates. The latter is revealed by lower values in column *h* (=comparative with STS-estimates) and *i* (= final estimates one year afterwards).

41. Summarising these results confirm the previous conclusion of chapter IIIa that the method of imputation for missing VAT-data is not crucial, if the population is known. The challenge is to determine adequately, which units have to be reported (Baldi et al., 2012).

[4] This are enterprises incorrectly to be assumed nonactive (group c in table 1). This are mainly starting enterprises.

[5] Table 2 shows the results of the Y_t/Y_{t-1} imputation, but similar results are obtained by choosing other imputation techniques (Baldi et al., 2012).

Table 2. Summary of estimation. Values relative to total reported values. Average over January 2008 – September 2010. The column "total imputed value" represents the situation when the STS-estimates have to be made. The column "total reported value" represents the final estimate, the estimate a year after the STS-estimates when all admin data are available. Imputation method is Y_t/Y_{t-1}

Nace Division	Early reporters	Units with imputed missing			Units without imputed values at t but reporting at t+12			Total imputed value	Total reported value
		with later reporting		without later reporting		Imputed value	Reported values		
		imputed values at t	reported values at t+12	imputed values a t					
	a	b	c	d	e	f=(b+d)	g=(c+e)	h=(a+f)	i=(a+g)
10	98.2	1.2	1.2	0.4	0.7	1.6	1.8	99.8	100.0
15	98.4	1.0	1.0	0.5	0.7	1.5	1.6	99.8	100.0
25	98.6	0.9	0.9	0.3	0.4	1.3	1.4	99.9	100.0
28	98.5	1.0	1.0	0.3	0.5	1.3	1.5	99.8	100.0
30	98.1	1.2	1.2	0.5	0.7	1.7	1.9	99.8	100.0
41	97.7	1.4	1.4	0.8	1.0	2.2	2.3	99.9	100.0
47	98.0	1.2	1.2	0.6	0.8	1.8	2.0	99.8	100.0
64	98.2	1.2	1.2	0.3	0.6	1.5	1.8	99.7	100.0
71	98.3	1.1	1.1	0.3	0.6	1.5	1.7	99.8	100.0
81	96.3	1.8	1.8	0.7	2.0	2.5	3.7	98.8	100.0

V. Conclusions

42. When making STS turnover and employment estimates in an admin data based system with the largest enterprises being surveyed two situations can be distinguished. Situation I is that the available admin data provide good coverage when the estimates have to be made. As a rule of thumb, we consider a coverage of about 80% of the total turnover by the large enterprises survey and available admin data as a good coverage. This is generally the case for quarterly estimates and in some countries for regular monthly estimates. The other situation II (admin are limited or not yet available) often applies to (early) monthly estimates and is not discussed in this paper.

43. The option generally chosen by Statistical Institutes in Europe in Situation I is producing levels and growth rates. In this case, the estimates are generally produced on a micro level implying that missing data are imputed. As the share of the available data is very large, our results indicate that the results of the imputations methods are fairly insensitive with respect to the exactly chosen method. In the case of only publishing growth rate, one might even question the added value of imputation. Therefore, we recommend that when choosing an imputation method one should aim for an optimal trade-off between benefits and costs rather than aiming for the "optimal" theoretical quality.

44. For the STS estimates, there is uncertainty on which enterprises are active and which are not. As a consequence, the question arises whether a admin data unit is missing due to late reporting or because it is not active (anymore). In the latter case the missing unit needs not to be imputed. To deal with this challenge, a two step imputation model might be considered:

a) determination of a *provisional active population:* Define a *list* of missing units to be imputed

b) the imputation technique itself.

45. Analyses in Italy, Germany and Finland show that main revisions in STS-admin data estimates are caused by the uncertainty in population. Therefore, it is recommended that research and development should be devoted on choosing the most adequate estimation method for active population (i.e. units to be imputed).

VI. Literature

Baldi, C., M.C. Congia, S. Pacini and D. Tuzi (2011), *The Quarterly Employment Estimates in Italy Based on the Employment Register*. Deliverable of work package 4., http://essnet.admindata.eu

Baldi, C., F.Ceccato, S. Pacini and D. Tuzi (2011), Imputation of employment admin data in Italy, Interim report of the ESSnet AdminData http://essnet.admindata.eu (upon request)

Orchard, C, K. Moore and A. Langford (2011), *Practices for Using VAT Turnover Data within the UK to Produce Estimates of Growth and Monthly Turnover*. Deliverable of work package 4. http://essnet.admindata.eu

Karus, E. (2012), *How to be Convinced as a User to use Admin Data for STS*. Presentation for work package 4. http://essnet.admindata.eu

Kavaliauskiene, D. (2011), *Application of Ratio and GREG-Estimator to VAT for Monthly Turnover Estimates*. Deliverable of work package 4. http://essnet.admindata.eu

Kiema, S., Remes, T. (2012) Testing of imputation methods on Finnish VAT data, Interim report of the ESSnet AdminData http://essnet.admindata.eu (upon request)

Lorenz, R. (2011), *Current Results and Future Improvements in Respect of Estimates for Missing Values in the VAT Registration*. Deliverable of work package 4. http://essnet.admindata.eu

Maasing, E. (2012), *Testing Imputations on Estonian Retail Trade Data*. Paper Q-2012 conference.

Šličkutė-Šeštokienė, M. (2011), *Application of GREG Estimators for (Administrative Data Based) Short Term Lithuanian Labour Statistics*. Deliverable of work package 4. http://essnet.admindata.eu

Teneva, M. (2012), *Use of VAT data in Monthly Business Survey*. Presentation for work package 4. http://essnet.admindata.eu

De Waal, A.G., Vlag, P.A., Baldi, C., The use of administrative data for STS. Situation I: Good coverage provided by administrative data Milestone of work package 4. http://essnet.admindata.eu