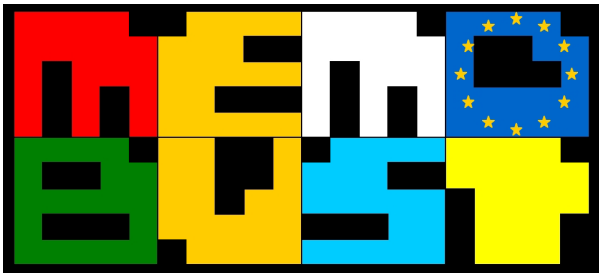

This module is part of the

Memobust Handbook

on Methodology of Modern Business Statistics

26 March 2014

Theme: Estimation with Administrative Data

Contents

General section

1. Summary

Official statistics produced by national statistical institutes (NSIs) can be based on primary or secondary data. Primary data are collected by the organisation also responsible for the statistical estimates, i.e., in this case the NSI. Secondary data are collected by another organisation or individual other than those responsible for the collection and aggregation of data from their initial source. An important secondary data source are administrative data. Administrative data are defined as data collected by another organisation for implementing an administrative regulation (or group of regulations). This module describes estimation techniques in case administrative data are used as replacement for a survey when estimating statistical variables. To keep the paper concise and illustrate the challenges with concrete examples, it is focussed on the use of Value Added Tax (VAT) data for estimating turnover.

2. General description

2.1 Introduction

Official statistics produced by national statistical institutes (NSIs) can be based on primary or secondary data. Primary data are collected by the organisation also responsible for the statistical estimates, i.e., in this case the NSI. Secondary data are collected by another organisation or individual other than those responsible for the collection and aggregation of data from their initial source. An important secondary data source are administrative data. Administrative data are defined as data collected by another organisation for implementing an administrative regulation (or group of regulations). Some of these administrative data can be used for statistical purposes. Tax data, i.e., data collected by tax authorities, can be considered as administrative data source. VATdata (Value Added Tax) of the tax office are the most widely used administrative data for enterprise statistics. For a complete overview of existing administrative data sources and their use for statistical purposes in Europe, we refer to Constanzo (2013) and the general results of the ESSnet project on the use of Administrative and Accounting Data ("ESSnet AdminData").

The use of administrative data for statistical purposes has increased considerably during the last decade. Administrative data can be used in two ways:

- as auxiliary information in the statistical process.
- to replace survey data in the statistical process.

Examples of using administrative data as auxiliary information are:

- checking the validity of outlying survey values with administrative data,
- benchmarking the validity of survey estimates with administrative data,.
- weighting survey results with GREGtype estimators (Kavaliauskiene et al., 2013) and administrative data as auxiliary information.

In this module, we focus on methodological issues arising when administrative data are used to replace survey data. Examples of using administrative data as replacement for survey data are:

- The use of the VATdata of the tax office for turnover estimates.
- The use of social security data from the tax office or social security agencies for employment estimates.
- The use of corporate tax data or building permits to estimate specific variables for annual statistics.
- The use of accountancy data to estimate specific variables for annual statistics.

2.2 Factors determining whether administrative data can be used to replace surveys

The first question which needs to be addressed is whether the used administrative data are suitable to replace variables from surveys. The answer on this question depends on several factors and affects the estimation technique. The most important factors to decide whether administrative data are suitable to replace survey variables are:

1. Does the NSI have legal access to these admin data?
2. Is the data transfer to the NSI guaranteed?
3. Is the NSI able to process large amounts of (administrative) data in a short time?
4. Can the administrative data be linked to the population frame derived from the statistical business register (SBR)?
5. Do the administrative data provide information about almost all enterprises within the target population when complete (i.e., completeness)?
6. Are the administrative data timely available (i.e., timeliness)?
7. Do differences in definition between administrative variables and statistical variables exist? Are these differences substantial and do they lead to biases in level and/or growth rate estimates. Are definition differences constant in time or not. Can the impact of differences in definition be monitored and corrected if required? (I.e., accuracy.)

If the answer on these seven questions do not reveal insuperable barriers – which is the case in several northern and north-western European countries – one may consider to use admin data for thereplacement of a survey. However, a methodology and a process needs to developed if the use of administrative data for replacing (variables in) surveys is considered. Guidelines for such a methodology, with is of course related to the statistical processes, are provided in the remaining part of this module.

VAT is the most commonly used administrative datasource in business statistics. Therefore, the term VAT instead of administrative data is used in concrete examples in the next modules of this theme. This choice has been made for sake of readability and concreteness. Methodologically the guidelines for VAT are also valid when using other admin data for estimating other statistical variables.

2.3 Using administrative data to replace surveys: general considerations

The general set-up when utilising VATdata for producing turnover estimates is that a combination of a survey and VATdata is used. In the survey, the large enterprises are generally completely enumerated. Since large enterprises often have a complex structure and their impact on the estimates is high,

correct surveyed observations from those large enterprises are considered crucial for producing reliable turnover estimates.

For the remaining small and medium enterprises VATdata are used instead of direct observations by the NSI. In other words, the general system of admin data based STSestimates consists of two parts:

1. use of a survey for the large enterprises (LEsurvey);
2. use of administrative data, for the remaining smaller enterprises.

The coverage of the large enterprise survey is a matter of debate. In order to define a more objective method to determine the coverage of the LEsurvey, Langford and Teneva (2012) have developed a method to calculate the impact of the 'incompleteness' factor on the boundary of the LEsurvey and the VATpart in an administrative data based STSsystem. This method is based on calculating revisions between the first estimates (with incomplete administrative data) and final estimates (with complete administrative data), by experimenting with different boundaries between the LEsurvey and the VATparts. More information about revisions in official statistics can be found in the theme module "Quality Aspects – Revisions of Economic Official Statistics".

Two fundamental issues arise for the population part which is estimated with administrative data:

1. Is the aim is to produce estimates for population totals (and implicitly also for growth rates) or is the aim to produce growth rates only?
2. Are estimates produced at the microlevel, i.e., for individual enterprises, or at the macrolevel, i.e., using combinations of activities and size classes?

At first sight, there may not seem to be a big difference between estimates for population totals or for growth rates only. 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 which affects the methodology. Estimates for population totals require better information about population characteristics and suspicious values than estimates ofgrowth alone.

Concerning micro- versus macro-level estimates, both approaches can be used. The main advantage of micro-level estimates is that further processing is then easy: one simply has to aggregate over all enterprises in a publication cell to obtain an estimate for that publication cell. On the other hand, the choice of macro-level estimates can also be justified, because weighting with the Horvitz-Thompson estimator provides the same results as imputing missing values at microlevel using stratum averages.

The project ESSnet Admin Data has observed that most European NSIs produce:

- both levels and growth rates, and
- data at the microlevel.

The main reason for producing a) both levels and growth rates and b) micro-level estimates is that the great majority of the data are already available, allowing NSIs to construct an enriched dataset with (VAT-)turnover data of almost all enterprises, to which other survey data may be linked. Publishing growth rates only may provide slightly more stable results, as the link with the population frame is less critical. Moreover, outliers or non-predictable enterprises can be excluded more easily. However, level

estimates for year, quarter and (if possible) month can be used as auxiliary information and reference for (early) STSestimates with no or few admin data available (see section 2.9).

2.4 The statistical process

When administrative data are used for economic and STSstatistics, a number of steps in the statistical process can be distinguished:

1. Data transfer from the tax office to the NSI;
2. Checking the data when entering the NSI (completeness, obvious errors, etc.);
3. Linking administrative units to statistical units;
4. Combining the result of the large-enterprise (LE) survey with the administrative data used for small and medium sized enterprises (SME);
5. Dealing with differences in definitions;
6. Editing of influential errors and outlier detection;
7. Determining the active population;
8. Estimation/imputation methods.

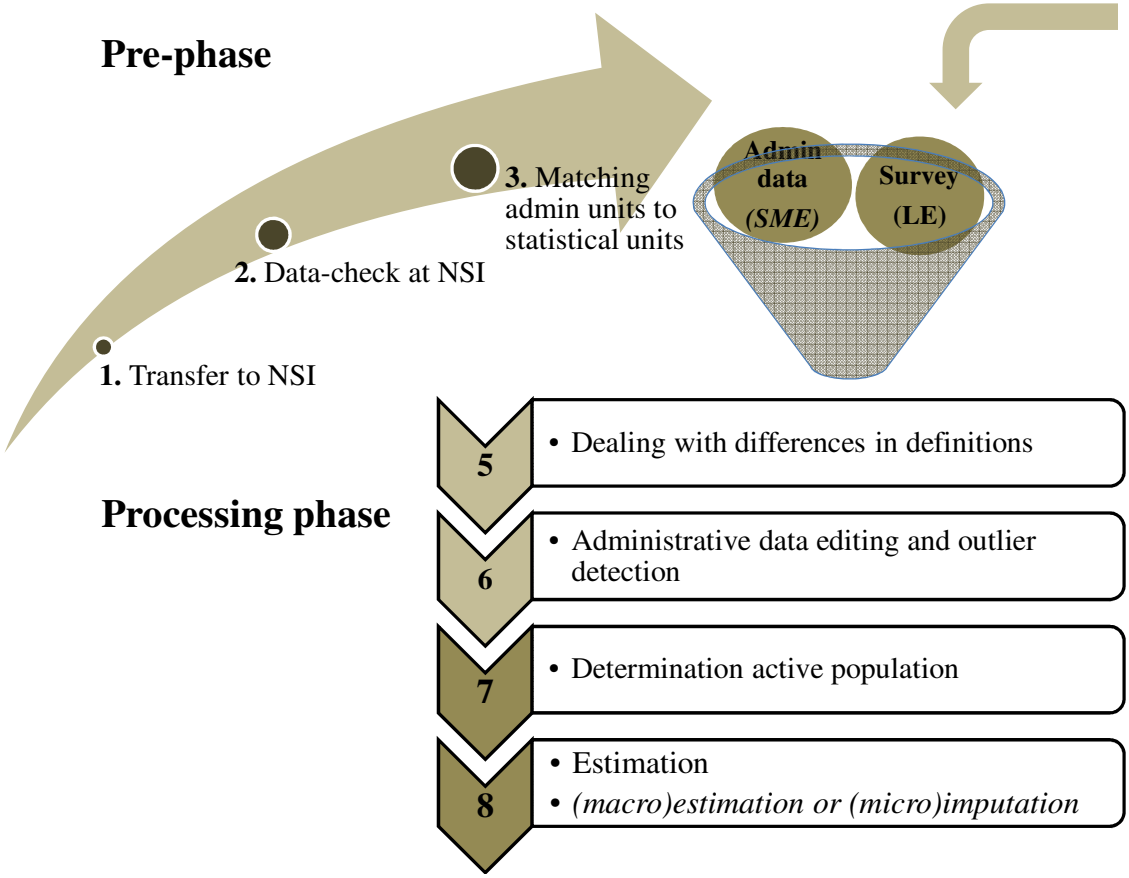

Figure 1. Overview of the statistical process of STSstatistics using admin data.

This process is visualised in Figure 1. Note that admin data have to be matched with the business register first (step 3), before combining the results of the LEsurvey with the administrative data used for SME (step 4). Steps 1 until 6 are summarised in this theme. The theme module "Data Collection – Collection and Use of Secondary Data" provides more information about these steps. The remainder of this document is focussed on the determination of the active enterprises and estimation procedures, because both are closely related.

2.5 Findings per process step

Step 1: Data transfer from the tax office to the NSI

To produce annual and quarterly and monthly short-term statistics (STS) with administrative data, the transfer of admin data to a NSI should be guaranteed. Furthermore the NSI must decide whether it opts for only one transfer per month or quarter, or several data transfers per month. The latter allows more flexibility, especially for STSestimates used for internal use (e.g., for National Accounts).

Step 2: Checking the data (completeness, obvious errors etc.)

It is common practice for NSIs to perform elementary checks on the administrative data as soon as they arrive at the NSI. This is in order to check whether there is anything wrong with a specific admin data delivery (e.g., less/more admin data than usual, different distribution than usual, large errors).

Step 3: Matching administrative units to statistical units

This step consists of linking the administrative data to the population frame which is generallydefined by the Statistical Business Register (SBR). Theoretically, a 100% match between two frequently used administrative sources (VAT and social security data) and the population for enterprise statistics should exist. In practice, this 100% match is not achieved due to:

- different enterprise units in the SBR and the admin data;
- time-lags, which may cause different timings of starting, stopping, merging and splitting enterprises in admin data and the SBR;
- maintenance peculiarities, which result in differences between administrative data and the SBR; and
- a (slightly) different population coverage because the smallest enterprises are exempted from VATreporting in some countries.

However, the large majority of enterprises in the SBR should be matched with administrativedata on an annual base. If this is not the case, it is recommended to improve this before proceeding further, because the added value of using VAT for turnover estimates is that these tax data are available for all enterprises when the data are complete. Furthermore, if not all administrative data can be matched with the SBR, it may be well possible that the non-linkable units represent specific parts of the populations. When this issue of non-matchable units is not resolved at this processing step, it may lead to estimation problems at a later stage.

When linking admin data to the SBR for STS another important issue is added;the incompleteness of the administrative data. Due to time-lags between the SBR and the administrative data source, late reporting starting enterprises are missed in the first estimates (because they are not yet included in the

SBR). In the case of apparently missing admin data, it is difficult to determine whether this is due to a) late reporting or b) because the enterprise has stopped. This issue is particularly important if the SBR population corresponds with a previous period (e.g., all active enterprises at the end of the previous year). However, the problem also arises if administrative data for the current month or quarter are linked to an up-to-date SBR. This situation is sketched in Figure 2. It will be described in more details in the next section of this module.

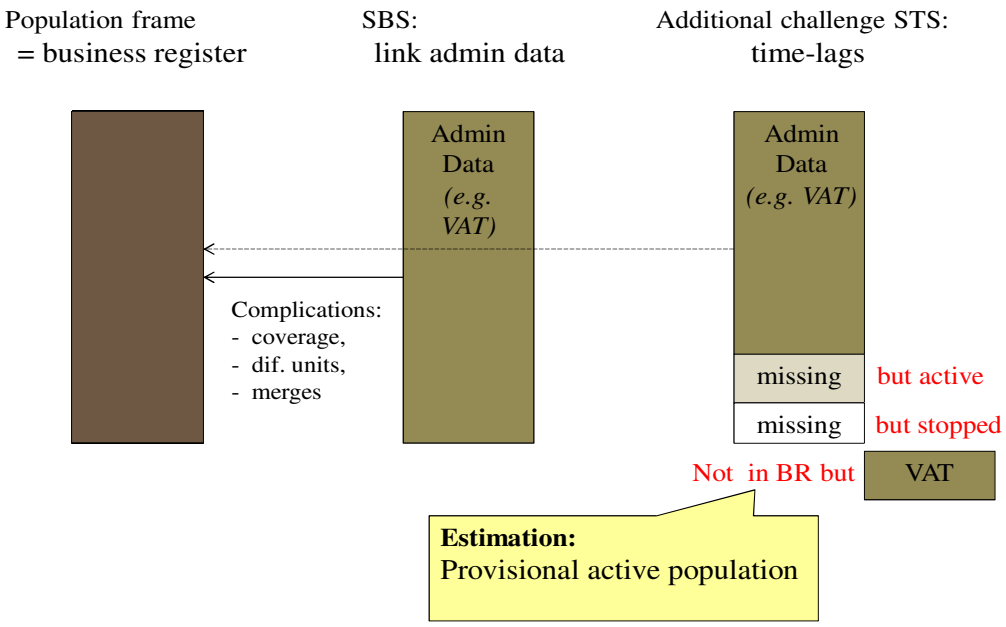

Figure 2. Schematic sketch of a) general challenges when linking admin data to the SBR (middle column) and b) specific challenges for STS when linking incomplete admin data to the SBR.

Step 4: Combining the large-enterprise (LE) survey data with the administrative data for small and medium-sized enterprises (SME)

To ensure stable timeseries, it is recommended to use a 'frozen' LEsurvey within a reference year, i.e., the LEsurvey remains unchanged within a year unless an enterprise stops. The use of a 'frozen' LEsurvey also prevents enterprises switching from survey to admin data (and vice versa) within a year. The LEsurvey can be updated at the beginning of a new calendar year, at which time new enterprises can be added to LEsurvey and other enterprises may be removed. It is recommend to keep the 'to be removed' enterprises within a survey for one extra period in order to maintain the stability of the crucial LEsurvey timeseries.

Step 5 Combining the large-enterprise (LE) survey data with the administrative data for small and medium-sized enterprises (SME)

Definitional differences between VATturnover and STS turnover and administrative variables and statistical variables in general do exist. More information about this topic can be found on the Information Centre of the ESSnet AdminData (http://essnet.admindata.eu/).

The most common approach is the use of linear regression analyses to correct for definition differences between administrative data variables and survey variables. These analyses are carried out on observed survey and administrative data of enterprises with similar activities in a reference period, i.e.,*x* years before current year. More specifically, the factor β describing the linear relationship between administrative variables in this reference period like VATturnover and statistical variables like turnover are used to correct the "VATturnovers". Note, however, that this technique is applicable only if:

- the relationship between administrative data and survey variables is linear;
- the relationship between administrative and survey data is constant in time;
- the relationship between administrative and survey data is not dominated by errors or other sources of noise.

A review by the ESSnet AdminData project of current practices in the use of VAT for annual and short-term statistics showed that differences in definition have little impact on level and growth rate estimates for most industrial activity sections. Therefore, several NSIs do not carry out corrections for definition differences for all variables.

Step 6: Administrative data editingand detection of outlier detection

The topic "Statistical Data Editing"in the Memobust handbook describes several methods for statistical data editing and detecting outliers. Specific for administrative data is that many 'suspicious values' may be caused by uncertainties in the link between admin data and the SBR. Some of these suspicious values may be more easily (and reliably) resolved at a later stage when more 'confirmed' adminand SBRdata are available. Especially when using administrative data for STS, it is advisable not to correct too many suspicious values at the first estimates, but to exclude these suspicious values in the first estimates (and consider them as missing) and correct them when more information becomes available.

Furthermore, it can be recommended that a relationship is established between the stratification level used for administrative based estimates and the stratification level used for detecting:

- influential erroneous values which need to be corrected (= dataediting);
- influential correct values which need to be considered as 'unique' cases when estimating aggregates (= outliers).

If these two stratification levels do differ, it is hard to determine whether a suspicious and outlying values is influential on the estimates or not. Current practices in the use of VATdata for turnover estimates differ in respect of the stratification level at which missing values with group-specific Y_t/Y_{t-1} or Y_t/Y_{t-12} growth rates are calculated.

Some NSIs use detailed groups. Detailed groups have the advantage that they are theoretically more homogeneous, because growth rates may differ for:

- enterprises with (slightly) different activities; and
- enterprises of different sizes.

The disadvantage of using small groups is that the number of available (donor) units may become too small, which increases the effect of outliers etc. Hence, a good outlier filter to detect anomalous growth rates in the available VATdata should be developed if detailed groups are used for imputation.

Disadvantage of this approach is that it is – in practice – for most NSIs impossible to check VATdata structurally at microlevel, due to the enormous dataset and the generally short production time. As a result, the cause for outlying values (errors, change in reporting unit between the reference period and the current period, or a valid economic explanation for a deviating growth rate) remains often unclear. This implies that if too many outliers are detected in small groups and a valid (economic) explanation for these values does not exist, some selectivity in the remaining values used for imputation should be introduced by filtering out all these outliers. To prevent this, and to keep the process more transparent, other NSIs use higher stratification levels (= bigger groups). This has the advantage that the impact of outliers on the imputed ratios is generally smaller because more (donor) admin data are available. For this reason, some NSIs use higher stratification levels (= bigger groups). This has the advantage that the impact of outliers on the imputed ratios is generally smaller because more (donor) admin data are available.

2.6 Determining the active population

Statistical registers and frames are described in the topic "Statistical Registers and Frames" of the handbook. Specific for the use of administrative data in enterprise statistics, and especially STS, is the determination of the active population. For example in STS, the most important issue with respect to determining the active population is to detect whether VATdata are missing because:

- the enterprise has stopped (or changed) its economic activity; or
- the enterprise is a late reporter.

This is especially a problem for small enterprises. Larger enterprises are generally well-recorded and are usually quickly updated in the SBR. This problem is enhanced by the possibility that enterprises do not always report their closure immediately 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 to 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.

A common method for determining whether enterprises are still active is simply to 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 reporting) or the last x quarters (in the case of quarterly reporting), the enterprise is considered inactive, otherwise it is considered to be still active. The ESSnet Admin Data has tested different rules and suggests that x should be chosen to be larger than 1, in order to minimise errors in the active population estimate due to late reporting. The most suitable values for x seem to be 2 or 3.

Detecting starting enterprises in time is also an issue. Starting enterprises are reported by the Chambers of Commerce and/or the tax office. Subsequently they are included in the SBR, with some delay after they started. If this delay is small, the starting enterprises should be present in the SBR. If this delay is longer, some of the starting enterprises might not be included in the SBR but in the

VATdata. However these enterprises, when present in the VATdata source, can be included in the population (and in the estimates) provided that a reliable NACE code can be obtained from the administrative data source or elsewhere.

This ESSnet has analysed the impact of starters and stoppers on the estimates. More specifically, it has analysed the impact of:

- incorrectly assumed active enterprises; and
- missing starters

on revisions in growth rate between the preliminary and final estimate. These analyses have been performed on VATdata in Estonia, Finland and Germany and on social security data in Italy. The conclusions are similar:

1. The contributions of starting and stopping enterprises do not average out. This may lead to a systematic over- or under-estimation (bias) in the preliminary STSestimate.
2. The relative contribution of starting and stopping enterprises to the total revisions is large, compared with the small share of starting and stopping enterprises to the total estimate.

These conclusions seemed to be independent of the exact estimation methodology, because the bias effect differs per period. Therefore, it is recommended that NSIs invest in the development of a suitable and efficient approach to determine the active population, before using administrative data for STSestimates. For annual statistics, this challenge has less impact because the administrative data are complete. For these statistics, it is however still crucial that all admin data can be linked to the SBR to prevent estimation challenges due to selectivity problems of the 'matched' administrative data which may vary per year.

2.7 Estimation: available administrative data when the estimates have to be made

As previously mentioned, the most frequently used administrative data sources to replace survey are VATdata for monthly, quarterly and annual turnover estimates. VAT may be reported on a monthly, quarterly or annual basis to the tax office. It has to be reported between 30 and 40 days after these reporting periods. The thresholds between these obligatory reporting periods differ per country, but as a rule of thumb it can be stated that:

- Large enterprises report VAT on a monthly base. Monthly reporting is also common for enterprises expecting a VATrefund;
- Most enterprises report VAT on a quarterly basis;
- Only very small enterprises (those having a turnover less than a few thousand Euros) are annual VAT reporters. In some countries, these enterprises do not need to declare VAT at all.

Therefore, it can be concluded that:

A. VATdata are complete when the annual estimates are produced. This is because all monthly, quarterly and annual VAThas been reported. VATdata are only missed due to matching problems with the SBR.

A. VATdataprovide generally very good coverage for quarterly estimates which are produced (more than) 45 days after the quarter.

This is because all monthly and quarterly VAThas been reported and the share of the missing annual reporters is very limited (in general < 5%). Hence, the available VAT should provide information about almost all enterprises apart from the smallest ones and some matching problems with the SBR.

Altogether is the implications that estimation is only needed for relatively few missing enterprises when producing annual and quarterly statistics with VAT.

B. No or few (and selective) VATdata are available at the time the monthly estimates have to published in most countries (30-45 days after the end of the month).

This is because many enterprises report quarterly or annually and publication deadlines for STSestimates are often earlier than deadlines for VAT reporting to the tax office.

Obviously under these circumstances, other estimation techniques are required than in case almost all admin data are available. This is illustrated in Figure 3.

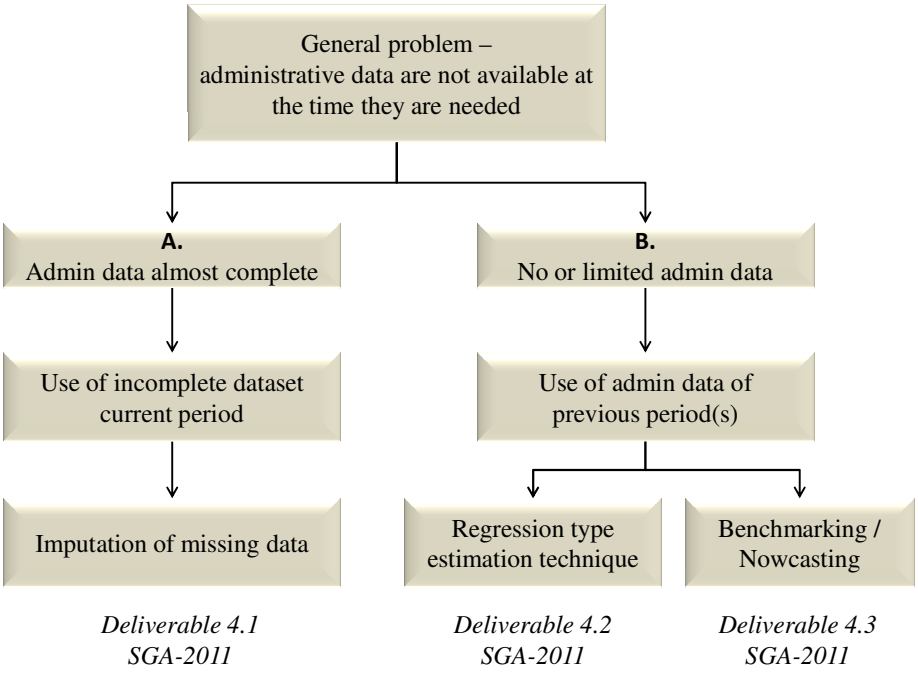

Figure 3. Relationships between estimation techniques when using administrative data as replacement for survey data and available administrative data. This figure also shows the relationship between the techniques and the documentation of the ESSnet Admin Data project (deliverable 4.1 = Maasing et al., 2013; deliverable 4.2 = Kavaliauskiene et al., 2013; deliverable 4.3 – Vlag et al., 2013).

Note VAT reporting periods differ from standard month, quarter and year in a few countries (United Kingdom, Ireland and Iceland). In this case, the VATdata have to be calendarised into values that cover the standard intervals such as month, quarter and year (Maasing et al., 2013). As a consequence, the timeliness of calendarised VAT is worse and a complete set of calendarised VAT is only available for the annual turnover estimates (Vlag et al., 2013).

Pragmatically the ESSnet Admin Data has used 80% coverage as the lower limit for "good coverage of admin data".This threshold has been chosen because many NSIs require that the large enterprises survey and the available admin data together cover about 80% of the total turnover beforereliable turnover estimates can be published. This 80% threshold is arbitrary. Revision analyses, as performed by Langford and Teneva (2013) or by Baldi et al. (2013), can provide more objective criteria to determine whether or not a dataset is almost complete and representative.

In the remaining part of this module, estimation techniques will be discussed in case of:

- almost complete coverage of the available administrative data;
- few administrative data available.

2.8 Estimation in case of almost complete coverage of administrative data

As mentioned in section 2.3, the most common practice is to produce level estimates and micro-imputations when using VAT for STS. This practice is recommended when the VAT (or other administrative data sources) are almost complete. The consequence is that the few missing VATdata need to be imputed. The most common practice for imputing missing values is using the available VATturnover data from enterprises with plausible data for current period. To impute missing values, enterprises are divided into several groups by size class, activity (e.g., NACEcode) and in some cases the period of VAT declaration (monthly/quarterly payers). These groups are the so-called stratification groups. The main assumption behind this stratification is that within these groups the available 'average' VATdata are representative for the enterprises without data.

Current practices differ with respect to the exact imputation technique, but the two most common imputation techniques are:

1. Average growth rates between current period and previous period of available VATdata. The imputation of the variable *y* for unit (enterprise) *i* at month *t* belonging to a generic stratification group is in formula:

$$\hat{y}_{it} = y_{it-1}\frac{\sum y_{jt}}{\sum y_{jt-1}} = y_{it-1}\frac{Y_t}{Y_{t-1}}$$

   where the summation is on the units reporting in both *t* and *t*-1 belonging to the same stratification group. These are the so-called Y_t/Y_{t-1} imputations.

2. Average growth rates between current period and the corresponding period of the previous year of available VATdata. In formula:

$$\hat{y}_{it} = y_{it-12}\frac{\sum y_{jt}}{\sum y_{jt-12}} = y_{it-12}\frac{Y_t}{Y_{t-12}}$$

   These are called Y_t/Y_{t-12} imputations.

The main advantage for choosing Y_t/Y_{t-12} imputations is that these ratios should provide more robust estimates in case of strong seasonality patterns. The disadvantage of using Y_t/Y_{t-12} growth rates for imputation is that these growth rates are affected by changes in activity between *t* and *t-12*. This is especially true when using detailed stratification levels and therefore relatively few *t;t-12* growth rates of enterprises with change of activity are available. Another disadvantage is that enterprises which

started during the last 12 months cannot be taken into account when imputing missing data, which may lower the quality. If these disadvantages are dominant, Y_t/Y_{t-1} imputations are preferred over Y_t/Y_{t-12} imputations

The fundamental question is whether the theoretical pros and cons of Y_t/Y_{t-12} versus Y_t/Y_{t-1} lead in practice to differences in publications (Vlag et al., 2012), This question was raised because these techniques are used when at least 80% and generally more than 90% of the estimated VATturnover is available. The same is true for the choice of aggregation levels at which these ratios are calculated.

Testing by the ESSnet Admin Data on VATdata by Statistics Estonia and Statistics Finland and testing on social security data by ISTAT has demonstrated that the impact of the different imputation methods on the published results is negligible, due to the high coverage of the available administrative data combined with the use of a LEsurvey. Hence, when choosing an imputation method one should aim for an optimal trade-off between benefits and costs, rather than aiming for the "best" theoretical quality. The testing of the ESSnet Admin Data also revealed that the impact of different imputation rules on the STSestimates is less than the impact of the uncertain active population on the estimates (i.e., **which** units are to be imputed). Hence, when developing a statistical production system for admin data based STSestimates, it is recommended that research and development should be concentrated on choosing the best method for determining the active population (i.e., **which** units are to be imputed).

2.9 Estimation in case of few administrative data available

The work of the ESSnet Admin Data has demonstrated that if few VAT (or other admin data) are available due to timeliness issues, this VAT cannot be used to replace a survey. Main reason is that, analyses in Finland, the Netherlands and the United Kingdom show that the few available VAT is selective and that this selectivity varies in time. As the extend of the selectivity can only be determined afterwards, it is not straightforward to correct for this selectivity with weighting techniques at the time the estimates are needed. Hence, provided that turnover levels and growth rates can be estimated with a later stage, the challenge is to find estimation methods for (early) month when no or only a few VATdata are available. This as alternative for a costly standard monthly survey among all enterprises within a branch.

The application of possible alternatives for a standard survey depends on the observation whether the long-term trends and short-term movements of the timeseries are similar for the larger enterprises, covered by a LEsurvey, and smaller enterprises, covered by administrative data like VAT for quarterly and annual estimates. Depending on the outcome, this information can be used to decide whether for first monthly estimates:

- a small survey under small medium-sized and small enterprises should be added to the LEsurvey which is also used for quarter and annual. This option is called alternative I in the remainder of this module.
- a survey under the largest enterprises (a LEsurvey) only is sufficient for the (first) monthly estimates, knowing that VAT covering the entire population becomes available at a later stage or for the quarters. This option is called alternative II in the remainder of this module.

- the LEsurvey should be combinedwith a separate estimate for the smallest enterprises based on extrapolation of the VATseries. This option is called alternative III in the remainder of this module.

When one of these three alternatives have been chosen, all alternatives do have in common that VAT of previous periods is used as auxiliary information for the estimates. Note that alternative I uses a small survey. Alternatives II and III are im- or explicitly model-based. Basically, the latter two alternatives provide implicitly a temporal estimation for growth of small and medium sized enterprises. This temporal estimation is 'overwritten' as soon as sufficient VATdata become available.

Alternative I is described by Kavaliauskiene et al. (2013). The basic idea is that VAT is too late to produce turnover estimates for the current month. As alternative a mini-survey for current month *t* is weighted by using VAT of previous period *t-1* as auxiliary information. This can be done by using GREGtype estimators. The use of GREGtype estimators in combination increases the precision of the estimates. Hence, a smaller survey can be used compared to the Horvitz-Thompson estimator. The use of GREGtype estimators is an established technique, which provides acceptable results. Disadvantage of the method is that it is elaborative in terms of detection and handling of outlying values. Furthermore, the reduction of the survey might be limited as this method requires a minimum amount of data.

As a result, the decision whether the smaller enterprises should still be sampled for estimates until VAT becomes available (alternative I) or whether temporal estimations for small enterprises can be considered (alternatives II and III) basically depends on five factors:

- the target of a National Statistical Institute (NSI) to reduce production costs;
- the target of a NSI to reduce administrative burden;
- the desired quality;
- the output level;
- the risk factor of using temporal estimations in case of unforeseen circumstances.

In general it can be stated that higher targets of reducing production costs and administrative lead to a lower survey coverage. Quality and output level may generally lead to a larger coverage of the surveyed part. However, this is not necessarily correct because if the sample size becomes too small, a survey estimate may have a large imprecision and a temporal estimation may have a better quality. The Standard Mean Error (SME), a combination of bias and imprecision, may help to compare the quality of a (small) survey estimate with temporal estimation (Vlag et al., 2013).

Alternatives II and alternatives III provide a temporal estimation for the smaller enterprises for current period *t* without using survey and VATdata (as the latter are not available yet).

Alternative II is based on the assumptions that:

- the short-term movement of the growth of the non-surveyed small enterprises is similar to the short-term movement of the surveyed large enterprises;
- changes in the business cycle and sudden events are simultaneously registered in the surveyed and non-surveyed part.

Alternative III is based on the assumptions that:

- the short-term movement of the growth of the non-surveyed small enterprises differ from the short-term movement of the surveyed large enterprises due to time-lags. Time-lags may occur if the small enterprises within a branch supply goods and services to larger enterprises or a subcontractors of larger enterprises;
- changes in the business cycle are differently recorded in the surveyed and non-surveyed part.

Note that available VAT of previous periods are needed to test whether the above mentioned assumptions are valid or not. Hence, although is not directly used in alternatives II and III, VAT is implicitly used for estimation.

If the growth of the larger enterprises is related to the growth of the smaller enterprises (= alternative II), the total estimation can be determined by

$$G_{t,t-1} = \frac{Y_{MLE}.G_{t.t-1;MLE} + Y_{SE}.G_{t,t-1;MLE}.C}{Y_{MLE} + Y_{SE}}$$

with:

$G_{t,t-1}$, $G_{t,t-1;MLE}$ the growth rates of the entire target population and the surveyed medium and large enterprises (MLE) respectively;

Y_{MLE}, Y_{SE} the(extrapolated) turnover level for MLE and small enterprises (SE), respectively;

C factor to correct for systematic differences in growth between MLE and SE.

The most simple model is assuming that C = 1. In more sophisticated models C is basically based on bias corrections or by benchmark nowcasting (Fortier et al., 2007; Brown et al., 2012; Vlag et al., 2013).

If the growth of the larger enterprises is not related to the growth of the smaller enterprises (=alternative III), the total estimation can be determined by

$$G_{t,t-1} = \frac{Y_{MLE}.G_{t.t-1;MLE} + Y_{SE}.G_{t-x,t-x-p;SE}.C}{Y_{MLE} + Y_{SE}}$$

with:

$G_{t-x,t-x-p}$, the growth rates of the SE of a previous period;

C correction factor, in this case basically a nowcasting factor.

The most simple model is assuming that p=1 and C=1. In this case the growth rate of previous period is applied to the current period. Several timeseries models exist to determine C, including Holt-Winters, ARIMA and SSA nowcasting techniques.

The approaches and underlying models are sketched in Figure 4.

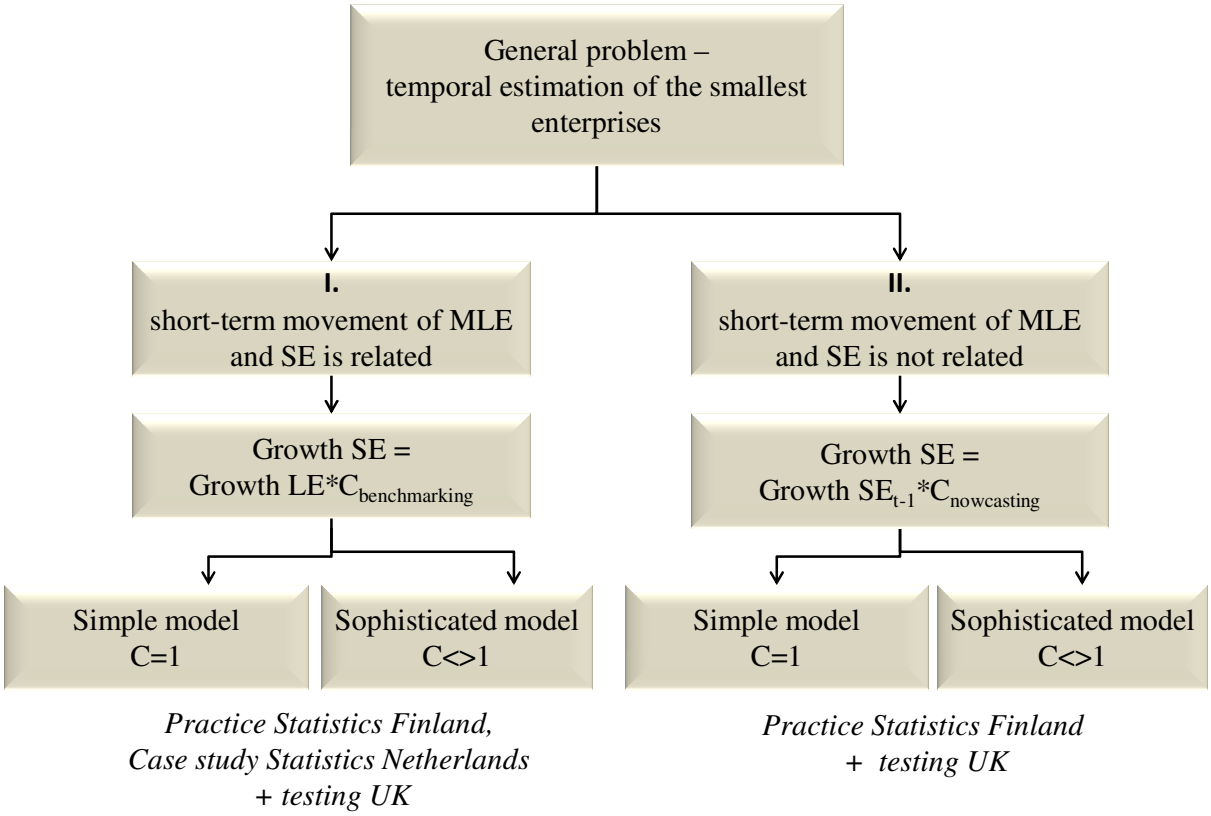

Figure 4. Simplified sketch of temporal estimation methods for small enterprises depending on the relationship between growth rate of large enterprises versus small and medium sized enterprises. Note that the ESSnet Admin Data project has tested these alternatives on data in Finland, the Netherlands and the United Kingdom.

The ESSnet Admin Data has analysed and described three cases (on real data) in which alternative II is used and two cases (on real data) in which alternative III is used. It is beyond the scope of this theme to describe the findings in details. Therefore, only the most important considerations are summarised.

In general these methods provide acceptable results. This especially the case for alternative II, which is also more data-based. However, even for alternative II, it can never be excluded that the underlying assumptions for the temporal estimations are invalid in case of unexpected changes in the business cycle or sudden events. If the surveyed part of the enterprise population (LEsurvey) covers 70-80% of the turnover and analysis on longseries of historical survey or VATdata demonstrate that maximum difference in growth rates between the surveyed part and the non-surveyed part is limited (e.g., a few per cent points), the potential impact of an incidental less performing temporal estimation for small enterprises is limited on the published total estimate for a branch. In this case a National Statistical Institute may accept the risk of an incidental less performing temporal estimation. However, the risks may be considered as unacceptable if the potential impact of an incidental less performing temporal estimation on the total estimate is larger. Risks may be high if the coverage of the LEsurvey is small and historical data suggest that the maximum difference in growth rates between the surveyed part and

the non-surveyed part might be large. Therefore, it is recommended to perform risk analysis before determining the size of small enterprise part which is temporal estimated.

Another finding that whatever alternative is used, artefacts in the (implicitly) extrapolated VATdataset can easily be magnified, leading to erratic temporal estimates of small enterprises. Therefore, is recommended to:

- consider the use of index series, which are panel based series, rather than level based series, i.e., that is using all available data;
- to spend time for correcting 'previous' VATdata for outliers, level shifts and other irregularities when implicitly using these data of previous periods for 1st estimates.

3. Design issues

4. Available software tools

Several productions system do exist for producing statistical estimates with VAT and/or social security administrative data. For more information, we refer to Maasing et al. (2013) and references herein.

Statistics Canada has developed SASprocedures for benchmarking and benchmark-nowcasting. The module "tempdisagg" in R can also be used for benchmarking, benchmark-nowcasting and other nowcasting techniques.

5. Decision tree of methods

The first decision to be made is whether administrative data can be used to replace surveys. This decision may depends whether:

1. the NSI has legal access to these admin data;
2. the data transfer from the tax authorities to the NSI is guaranteed;
3. the NSI is able to process large amounts of (administrative) data in a short time;
4. the administrative data can be linked to statistical business register (SBR).

Then decisions have to be taken whether the aim is:

1. to produce estimates for population totals (and implicitly also for growth rates) or to produce growth rates only;
2. the estimates produced at the microlevel, i.e., for individual enterprises, or at the macrolevel, i.e., using combinations of activities and size classes.

In the next step decisions have to be made about:

1. matching the administrative data to the SBR;
2. dataediting and outliers detection;
3. the stratification level at which estimation and detection of influential erroneous or outlying values takes place;

4. the determination of the target population, i.e., active enterprises.

In the next step one has to analyse the completeness and selectivity of the available administrative data when the estimates have to be made, because it determines:

1. whether imputation techniques using growth rates (or levels) of available VAT of current period can be used to impute few missing VATdata need to be imputed;
2. whether available VAT for current period cannot be used for estimations and estimation have to be based on indirect and implicit use of VAT of previous periods.

6. Glossary

For definitions of terms used in this module, please refer to the separate "Glossary" provided as part of the handbook.

7. References

Baldi, C., Tuzi, D., Ceccato, F., Pacini, S., Karus, E., and Vlag, P.A. (2013), STS-estimates based on admin data: dealing with revisions. *Deliverable 4.3 of the ESSnet on AdminData – SGA2011*, http://essnet.admindata.eu.

Brown, I. (2012),An Empirical Comparison of Benchmarking Methods for Economic Stock Time Series. *Proceedings ICES-IV conference- June 18-21, 2012, Montreal, Quebec, Canada*, 399–412.

Constanzo, L. (2013), Report to Eurostat on the "Overview of Existing Practices". *Deliverable 1.2 of the ESSnet on AdminData -SGA2011*, http://essnet.admindata.eu.

Fortier, S. and Quenneville, B. (2007),Theory and application of benchmarking in Business Surveys. *Proceedings ICES-III conference- June 18-21, 2007, Montreal, Quebec, Canada*, 422–434.

Kavaliauskiene, D., Slickute-Sestokiene, M., and Vlag, P.A. (2013), The use of regression estimators for admin data based STS estimates. *Deliverable 4.2 of ESSnet AdminData – SGA2011*, http://essnet.admindata.eu.

Langford, A. and Teneva, M. (2012), Analysis of revisions of admin data based short term statistics. Application to UK retail sales data and implications for the definition of the boundary between survey and administrative data coverage. Internal report ESSnet AdminData (upon request).

Maasing, E., Remes, T., Baldi, C., and Vlag, P.A. (2013), STS estimates based solely on administrative data: final results and recommendations. *Deliverable 4.1 of the ESSnet on AdminData*, http://essnet.admindata.eu.

Vlag, P.A. (2012), Imputing missing values when using administrative data for short-term enterprise statistics. Paper for UNECE work session on Statistical Data Editing, Oslo.

Vlag, P.A., Bikker, R., de Waal, T., Toivanen, E., and Teneva, M. (2013), Extrapolating admin data for early estimation: some findings and recommendations for the ESS. *Deliverable 4.3 of the ESSnet on AdminData*, http://essnet.admindata.eu.

Waal, A.G. de, Vlag, P.A., Baldi, C., and Tuzi, D. (2012), The use of administrative data for STS. Situation I: Good coverage provided by administrative data. *Milestone of work package 4*, http://essnet.admindata.eu.

Interconnections with other modules

8. Related themes described in other modules

1. Overall Design – Overall Design
2. Statistical Registers and Frames – Main Module
3. Statistical Registers and Frames – The Populations, Frames, and Units of Business Surveys
4. Dynamics of the Business Population – Business Demography
5. Data Collection – Collection and Use of Secondary Data
6. Statistical Data Editing – Main Module
7. Statistical Data Editing – Editing Administrative Data
8. Weighting and Estimation – Main Module
9. Quality Aspects – Revisions of Economic Official Statistics

9. Methods explicitly referred to in this module

1.

10. Mathematical techniques explicitly referred to in this module

1.

11. GSBPM phases explicitly referred to in this module

1. Phase 5 - Process

12. Tools explicitly referred to in this module

1.

13. Process steps explicitly referred to in this module

1. GSBPM Sub-process 5.7: Calculate aggregates

Administrative **section**

14. Module code

Weighting and Estimation-T-Estimation with Administrative Data

15. Version history

Version	Date	Description of changes	Author	Institute
0.2	23-10-2013	draft version	Pieter Vlag	CBS
0.3	07-02-2014	revised after comments by Editorial Board	Pieter Vlag	CBS
0.3.1	12-02-2014	revised after comments Leon Willenborg and Sander Scholtus	Pieter Vlag	CBS
0.3.2	12-02-2014	preliminary release		
1.0	26-03-2014	final version within the Memobust project		

16. Template version and print date

Template version used	1.0 p 4 d.d. 22-11-2012
Print date	26-3-2014 13:36