# The Use of Administrative Data for Short Term Business Statistics: Lessons from a Cross-Country Experience

Ciro Baldi, Francesca Ceccato, Silvia Pacini, Donatella Tuzi[1]

**Abstract**
In order to reduce the burden on respondents and the cost of statistics, in 2009 Eurostat launched the ESSnet project "The Use Of Administrative And Accounts Data For Business Statistics" whose aim is to disseminate the best practices already in use and to provide recommendations to interested member states. Within this project, the Work Package 4 (WP4) deals with the problems related to the production of short term business indicators. The scope of this paper is to present the work made so far by the WP4, starting from the mapping of methods that can be applied in common situations to their application in specific country contexts.

## 1 Introduction[2]

The use of admin data for Short Term Statistics (STS) has to cope with the problem that this data might not be complete to comply with the dissemination deadlines. The reasons for incompleteness may depend on the rules of reporting, implying availability of data for a different periodicity than that required (*periodicity* issue), or because of the late response of single units (*timeliness* issue). The main goal of the WP4 in the ESSnet Project is to provide practical recommendations on the methods that can be used in order to produce reliable estimates despite of these problems. The focus is on the estimation of two main variables of the EU STS regulation: the turnover and the number of employees that can be estimated respectively from the Value Added Tax (VAT) registers and the Social Security registers.

A preliminary work of the group was to make an inventory of the applications already in production or being developed by the National Statistical Institutes (NSIs) in order to map the most common data situations and the appropriate methods to be used. A good deal of effort has been devoted to establish a common language and share the knowledge base. In order to pass from best practices to recommendations, methodological improvements have been discussed and, to ensure that recommendations were not based only on theoretical grounds, cross country comparisons and tests have been programmed.

Moreover, attention has been also paid to the issue of revisions between estimates based on incomplete admin data (preliminary estimates) and estimates based on complete admin data (final estimates) as a tool to test the performance between the

[1] Istat, Short Term Business Statistics Directorate.

[2] This work synthesizes the WP4 group results. To all members of the group, starting from the coordinator Pieter Vlag (CBS-Netherlands), goes the gratitude of the authors. However for the sake of brevity some aspects have been simplified and re-elaborated  by the authors. All errors and misrepresentations of the WP4 work are attributable only to the authors.

different approaches and also evaluate costs and benefits of methods at implementation levels.

## 2 A map of data situations and methods

**Figure 1**. ESSnet WP4 Framework: Data sources and model assumptions

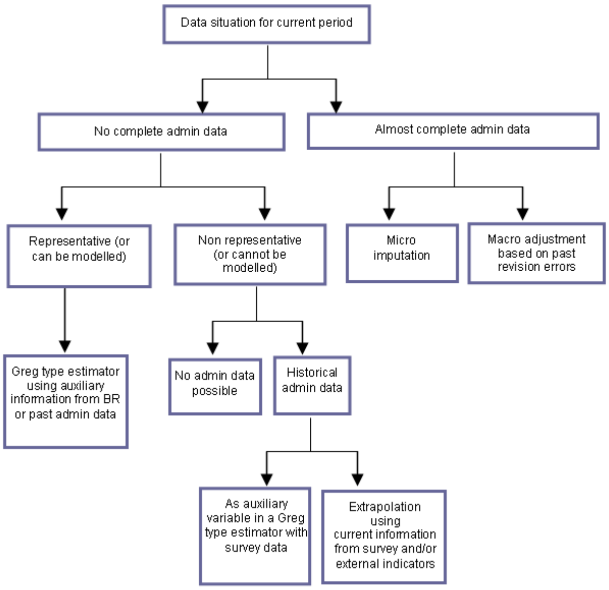
Data situation for current period
No complete admin data
Almost complete admin data
Representative (or can be modelled)
Non representative (or cannot be modelled)
Micro imputation
Macro adjustment based on past revision errors
Greg type estimator using auxiliary information from BR or past admin data
No admin data possible
Historical admin data
As auxiliary variable in a Greg type estimator with survey data
Extrapolation using current information from survey and/or external indicators

Figure 1 reports the framework map that relates the available data to the appropriate methods. Two main situations were distinguished: 1) the administrative data for the reference period are far from completeness when the STS estimates have to be compiled; 2) the administrative data are almost complete [5].

In the first case the admin data, albeit being incomplete, might be representative or not of the target population. When they are representative, or can be modelled to become representative, a model assisted estimation (such as a Greg type) can be used along with auxiliary variables from the business register or administrative datasets of previous periods to produce an unbiased estimate. When they are not representative, the admin data of previous periods (when they are complete) can be used as auxiliary variables either in a regression or calibration estimator for a small statistical survey or in a context of extrapolation techniques on aggregated data using current information from external correlated indicators or from a survey (see §3).

In case of almost complete admin data two possibilities arise. In one case, available data are corrected at an aggregate level with an adjustment procedure that exploits the revision errors of previous periods. A second possibility is to identify the missing values and impute them at micro level (see §4).

# 3 Administrative data not complete and not representative

When the admin data are far from complete and resist the attempt of the statistician to model them, in order to produce unbiased estimates a survey collecting information on the latest period is necessary. Apart from the unavailability of admin data on the current period, there may be other reasons to keep a survey. The main one relates to differences between the definition that has to be measured and the content of administrative information. In this situation, the administrative data from previous periods can (should) still be used to improve the estimate and/or reduce the sample size of the survey.

Two main situations have been envisaged. In the first, the sample survey covers the whole target population. In this case, the estimates can be obtained through a regression or calibration estimator that uses the administrative data of previous periods as auxiliary variables. This case is well represented by the recent methodological improvements introduced by Statistics Lithuania for the index of turnover [2]. The choice made by SL is a regression estimator exploiting the correlation between the statistical turnover measured at month *t* by the survey and the administrative turnover of month *t*-1. The regression estimator is used for the middle sized enterprises, since for the largest ones the sample strata is *take all* and for the smallest one, for which no administrative data are available, it is still used an Horwitz Thompson estimator. The frame for sample drawing and population definition is represented by the Business Register. The sample design is not very different from the design under the previous situation when the Administrative data were not used. However the passage from an Horwitz Thompson estimator to a regression estimator has allowed a significant reduction of the sample size.

In the second case only a survey on a subpopulation is available or reputed acceptable in terms of budget. Typically the survey covers the largest enterprises due to their influence on the estimates. For instance, Netherlands keeps on running a monthly survey on the top companies in terms of turnover [6]. The case of Netherlands is also interesting since a recent legislation change has made the monthly reporting of VAT data voluntary for the companies below a certain threshold and thus only quarterly admin data are complete and reliable. In this case, the monthly estimates can be produced using the survey as a concurrent indicator and exploiting the time series correlation of survey and admin data to produce monthly estimates that cover the whole target population.

A similar situation is faced by the ONS in UK, where there is the additional complication that the reporting periods, overlapping staggers of quarters, are different from the estimation periods (months) [4]. There, interpolation techniques are used to transform the staggered data into monthly data.

# 4 Administrative data almost complete

A typical situation occurs when a relatively small share of reporters is not available for the deadline of the preliminary estimate. This may happen either because the time when data have to be available to the NSI is before the end of the administrative reporting deadline (or just few days after it) or because some data have not yet passed the checks of the administrative institution. The consequence, in both cases, is that the preliminary estimate has to be based on the early reporters while the revised estimate will include also the late ones. The percentage of late reporters may vary according to a number of circumstances. For instance in Italy the employment estimate released at 60 days from the reference quarter is based, in normal conditions, on a base of reporters that accounts for 95-98% in terms of employment. However, this completeness rate, is variable among economic activities and is subject to sudden anomalous drops due to particular events.

Two solutions are possible in this case: a macro adjustment and a micro imputation procedure, with this latter being a more flexible solution and allowing the production of more disaggregated estimates. The imputation methodology in this framework has to solve two main issues: 1. definition of the list of active units for the reference period *t*, and of the associated list of units to be imputed (non reporters) as complement to the reporters; 2. model/rules to impute a value to non reporters.

The first issue is peculiarly associated with the use of admin data. At the time when the preliminary estimate is performed, unless the Business Register is up to date with respect to the reference period (which is rarely the case), there is uncertainty on which non reporting units are active and which are not. On one hand a missing unit might be only a late reporter or have become inactive (stopping enterprise). On the other hand, while the list of reporters generally includes businesses born (or recovered by a suspension) in the reference period, some of these starting enterprises will report only afterward. This implies that the definition of rules of activity has to be based on assumptions, usually related to the pattern of reporting in previous periods and that, this list cannot include the late reporters which are born in the very last period. Consequently the performance of the imputation method depends, basically, on the ability of the method of balancing the over-imputation of stopping enterprises (considered active) and the under-imputation of starting enterprises (not included in the list of active units).

The second issue is more traditional as it implies setting up a model to impute the values for the units included in the list defined above. While many methods have been reviewed within the WP4, it has been decided to focus on methods that use the reporters growth rates (either year on year, or month on month). This framework represents the practice of Destatis [3] for turnover and the procedure currently in development at Statistics Estonia for turnover and Istat for employment [1].

## References

The following documents are available on: http://essnet.admindata.eu/WikiEntity?objectId=5276

1. Baldi C., Congia M.C., Pacini S., Tuzi D., The STS-employment estimates in Italy based on administrative data, ESSnet WP4 (2011).
2. Kavaliauskiene D., Use of regression type estimation technique as a tool for solving timeliness problem in short-term statistics, ESSnet WP4 (2011).
3. Lorenz R., Imputation of missing values in the VAT data in Germany, ESSnet WP4 (2011).
4. Orchard C., Langford A., Moore K., National practices of the use of administrative and accounts data in UK short-term business statistics, ESSnet WP4 (2011).
5. Vlag P., Ortega Azurduy S., Karus E., The use of admin data for monthly and quarterly estimates: common issues and challenges in Estonia, Finland, Germany, Italy, Lithuania, The Netherlands and the United Kingdom, ESSnet WP4 (2011).
6. Vlag P., Ortega Azurduy S., Van Loon A., Scholtus S., Monthly turnover estimates with VAT: challenges in the Netherlands, ESSnet WP4 (2011).