DGINS conference 2014:
Towards global business statistics

25 September 2014

Integration of business and trade statistics: limitations and opportunities

Giorgio Alleva

President of the Italian National Institute of Statistics

1. Introduction

In a time of increasing uncertainty on the future perspectives of growth for the EU area, national statistical authorities can play an active role by complementing already existing economic data with new indicators finalised to shed some lights on the sources of competitiveness of European businesses.

The availability of new information on business behaviour is crucial to assess the competitiveness and evaluate the potential growth of the EU economic system in the future. It also plays a crucial role to set up or fine tune policy measures oriented to boost productivity growth and create new jobs. The presence of a great deal of heterogeneity in business performance can only be partially explained by standard classification variables such as enterprise size, sector of economic activity or geographical location. Performance gap across firms should be analysed according to business relevant taxonomies, such as exporting versus non exporting firms and/or innovative versus non innovative companies.

EU and national statistical authorities can further strengthen their role as data providers of official and business relevant data as long as they are able to clearly identify and consistently track the business characteristics and performance of sub-populations of businesses that are of a primary interest for business analysts and policy makers. This requires to further enhance the internal and external integration and harmonisation of the key components of a statistical production process: data collection, estimation methodologies, and dissemination. Data collection includes both data from surveys and data derived from the statistical processing of administrative data sources. Estimation methodologies encompass all the statistical processes finalised to produce and validate different statistical outputs from data collection. Dissemination includes the definition of outputs (data, indicators and analyses), formats and tools for communication.

As is now well understood, the traditional stovepipe approach has remarkably limited the range of outputs that can be obtained from original data sources. As suggested by the MEETS results, a radical change of perspective in the organisation of statistical production can dramatically boost the production and dissemination of new outputs with limited impact on the statistical burden on respondents. According to current proposals for the realisation of *Vision 2020*, focusing on the adoption of a Business Architecture (BA) model, the challenge is to build and maintain common infrastructures of data and metadata - by linking matrices for the integration of information regarding the same units, or group of units - and to develop generalized principles, rules and

methods for data collection and processing, crossing the single traditional production processes and sharing IT services and infrastructures. This allows producing and disseminating information on the variables of interest integrated with others, fostering the interpretation of the phenomena and their relationships with others.

Furthermore, the state of the art of statistical tools and methods for the measurement of business phenomena make feasible the development of new indicators on the business structure and performance of specific sub-populations of businesses, consistent with the Business Register (BR) frame and Structural Business Statistics (SBS) figures, such as enterprises engaged in international activities, with limited costs and in a relatively short time span.

More in general, the development of new methodologies finalised to the statistical processing and quality improvement of administrative data sources has opened the floor to substantial information gains in the structural business domain. In particular, the applied and theoretical methodological research in this area increasingly focuses on the exploitation of micro-level data from already available administrative data sources in a way that is consistent with statistical standards and procedures. The increasing availability of business data from administrative data sources also leads to reconsider and upgrade the use of direct reporting for the compilation of business statistics. While direct gains in terms of a consistent reduction of the statistical burden on the respondents is the most obvious outcome of this approach, the redefinition of the direct reporting approach should also consider the need to focus on variables, statistical units and emerging phenomena those information are not available from standard administrative data sources.

The aim of this paper is to highlight the crucial features of the limitations and opportunities stemming from the integration of business and trade statistics. While opportunities, especially for data users and no additional burden on respondents, are quite self-evident, limitations appear sometimes more difficult to be perceived and correctly assessed given the limited integration so far achieved in this area of economic statistics.

The second paragraph describes the advantages as well as the drawbacks associated with the use of administrative data in statistical production. It also provides an overview of the key methodological issues involved in the data integration between different data sources. In particular, it focuses on the peculiar features of the integration process between foreign trade and structural business data (SBS). The third paragraph provides a EU level overview on the exploitation of administrative data for the compilation of SBS data. Despite the widespread consensus on the relevance and usefulness of administrative data, there is still a great deal of heterogeneity across EU countries as far as their effective implementation in the statistical processes is concerned. The fourth paragraph highlights the main methodological and data processing innovations recently introduced in the compilation of Italian SBS data through a more intensive and systematic use of administrative data. In particular, it introduces the Frame-SBS recently developed by Istat to support the compilation of key national account figures and SBS data (the first benchmark will be SBS revised data to be transmitted to Eurostat by the end of 2014 for reference year 2012). Paragraph five further expands these results by showing their implications for the production and dissemination of new data and indicators focusing on business sub-populations selected from the overall population of all resident enterprises in Italy included in the Business register. In particular, it provides some concrete examples of structural business indicators with respect to the target population of exporting enterprises in Italy. The new figures included in this paragraph substantially expand business data already made available by TEC data (trade by business characteristics) being consistent with SBS official estimates. In particular, they provide some preliminary answers to questions that rank high in the policy

makers agenda: Which is the impact of exporting enterprises on value added and turnover across industries and size classes? Do exporting enterprises outperform non exporting ones in terms of productivity and profitability? How much enterprises benefit from their engagement in global value chains?. These questions are relevant also in a short term perspective, as in many countries the recent crisis has been characterized by a significant growth gap between domestic and foreign demand. In this context, the ability of enterprises to fulfill the demand for imports from other countries seems to be a key factor in supporting the economy, and a powerful factor in enterprise selection, redirecting the overall industrial system towards more profitable and high growth activities. Paragraph six draws some conclusions. In particular, it emphases that new information gains stemming from the use of combined primary or secondary data sources represent both an opportunity and a potential threat for the overall consistency of economic statistics. The design of a new system for the compilation of economic statistics that takes into account all potential sources and biases and maximizes the mix between direct reporting and use of administrative data is foreseen as the only possible approach to expand the outputs in a consistent and efficient way.

2. Methodological issues involved in the integration between foreign trade and business data

Methodological solutions proposed in literature to deal with integration between (primary and/or secondary[1]) data sources usually discriminate between sources composed of different units (which typically happens when integration involves two or more sample surveys) and sources composed of (almost) the same units (this case is typical of integration between two or more statistical registers, or statistical registers and sample surveys).

In this paper we focus on the latter case, as we are interested in the integration between foreign trade in goods data and SBS data. Since foreign trade data encompass all trade operators[2], thus representing a census like population, methodological issues to be managed in the integration process mostly depend upon the characteristics of the SBS data, being this latter either a sample or a complete data set covering the whole target population.

The choice between the sampling or the census like nature of SBS data crucially relies on the methodological approach adopted to combine direct reporting with administrative data.

Traditionally, SBS estimates are based on direct sample surveys. In this context, administrative data (AD hereafter) are normally used in an indirect way as a source of auxiliary information to either support processing of primary data (e.g., for data editing, imputation, or calibration of estimates), or to increase the accuracy of parameters' estimates generally based on the use of model-based estimation approaches at both micro and aggregate level. At micro-data level, the proposed estimation approaches (see Essnet AdminData, 2010, for an overview of such methods) consist in fitting a model on the sample, and applying it to predict values for units on the unobserved part of the sample itself using information obtained from AD. At aggregate (estimates) level, methods such as small area/domain estimation and post-stratification/calibration can be

[1] Primary data are statistical data collected by the statistical organization itself, while secondary data correspond to data that are already available in the outside information system. Secondary data sources can be either statistical, consisting of statistical objects and statistical data collected by other (survey-oriented) organizations, or administrative sources, i.e., secondary sources that have an administrative purpose (Wallgren and Wallgren, 2007).

[2] Foreign trade data used in the integration process are firstly reclassified by trade operator code and then fully integrated with the Business Register. This data standardisation process is well known by EU countries since it represents the underlying data source for the compilation of TEC statistics. These data are mandatory under the EU Regulation on foreign trade statistics.

adopted in order to increase the parameters' estimates accuracy. In particular, small area estimation (Rao, 2003) allows to combine the strength of the survey frame (assumed to have little bias) with the greater information detail of AD to produce a coherent and richer set of estimates.

Alternatively, when highly reliable AD sources are available on the SBS target variables/population, they can be directly used in the statistical estimation process as a replacement for primary data at unit level. This approach very much expands the availability of information across different dimensions: spatial-demographic and longitudinal. In effect, since AD usually covers an almost complete business population[3], they also provide very exhaustive information to expand existing domains (such as regional statistics) or to develop new publications targeted to specific dissemination events. Another positive effect of data integration is the ability to estimate parameters of the relationship between phenomena not jointly collected in any surveys or AD sources. In addition, since AD usually cover multiple time periods and maintain a stable composition over a relatively long period of time, they are also very suited for detailed longitudinal studies. However, in these situations additional costs and methodological issues are to be managed, mainly relating to the need for a deeper analysis of the quality of AD. In fact, it has to be reminded that AD are normally affected by different types of errors similar to those normally emerging in statistical surveys[4], and which can be treated (at least partially) using classical data editing methods. Furthermore, a process of *harmonization* is required, consisting in the treatment of errors on a "conceptual" level, for instance matching classifications, concepts and definitions with respect to the target units and the statistics of interest. In addition, as any AD source is gathered for other purposes than statistical ones, it usually does not cover all the variables/population of interest: for this reason, normally information from a given AD source needs to be complemented by additional data, which can be obtained from other AD sources, and/or based on statistical prediction processes (*imputation*), and/or from direct survey data. Concerning the former solution, it has to be underlined that when different AD sources are combined, the benefits deriving from higher levels of coverage are balanced by additional issues deriving from the integration process itself, which is an important source of other types of errors.

If integration is based on record linkage, regardless of whether the sources to be integrated are surveys or AD, there should be a framework of rules on decisions related to (1) how to conduct the match, (2) how to measure the errors, and (3) how to take them into account in making inference on the population (Istat, 2008). When the record linkage represents the central process of data integration, its quality should be assessed according to a specific quality framework. The quality of linking will depend on the quality of information of the individual datasets, and in particular the quality of the common unique identifier (UID) and of the linking variables.

Actually, when statistical units are identified and linked, variables are combined from different sources and derived variables are created, different types of integration errors may be originated, like identification errors, consistency errors, and missing values. Unit identification errors can determine bias in final estimates, especially in a longitudinal perspective. If a common unit identifier does not exist in the matched archives, methodologies referred to as *record linkage* (see Cibella *et al.*, 2009 for a complete overview) can be adopted, consisting in identifying pairs of units coming from the different archives, which belong to the same entity, on the basis of the agreement between common variables.

[3] Differences between statistical and administrative target populations need to be carefully taken into account in order not to overestimate the informative potential of administrative data and to correctly consider problems of coverage.

[4] For an overview of error sources in AD, see Bakker (2011), Zhang (2012) and Wallgren and Wallgren (2007, p. 177).

As far as consistency is concerned, information from multiple archives can result to be incoherent due not only to measurement errors, but also to inconsistent definitions deriving from the different source purposes, or to the fact that units may change their structural characteristics over time. Depending on the causes of the inconsistencies (Bakker, 2011), methods referred to as *micro-integration* (see ESSnet DI, 2011) can be adopted in order to search for and eliminate "errors" in combined data sets at unit level. Besides *harmonization,* these methods also include *reconciliation* (consisting in the correction for those measurement errors which cause the failure of logical or mathematical relations among the integrated micro-data). Reconciliation approaches are essentially based on editing and imputation techniques (see Pannekoek J., 2011; MEMOBUST, 2014 for an overview of these methods): for instance, in *Minimum Adjustment* methods a constrained minimization problem is solved on the linked data in order to find the final imputed values that differ as little as possible from the observed data and satisfy the edit rules.

Micro-integration also involves methods (referred to as *completion*) for the treatment of under-coverage (or *representation*) errors, which are due to the fact that the target population is not completely described by the integrated AD sources. Missing values may also be due to the fact that not all topics of interest may be covered by the integrated AD sources, in this case they appear as item non-responses in the final data. Not covered information in multi-source datasets can be treated by *mass imputation* (Essnet DI, 2011*;* Whitridge *et al.*, 1990*)*, consisting in the prediction of missing variables' values at unit level in order to obtain a rectangular data set: note that this approach ensures naturally consistent estimates at any level of detail, however it is feasible only when the amount of data to be predicted is reasonably limited, and the imputation models can be estimated in order to preserve as many relationships as possible in the final data. Nevertheless, methods for evaluating the additional uncertainty due to the imputation process in this context need to be applied (see for example Da Silva, 2014 for a recent discussion on this issue).

A further alternative for managing the incompleteness of multi-source statistical registers consists in designing specific sample surveys: in this case, supplementary information is observed to be used for instance to collect variables that do not appear in AD sources, or to investigate complex or large enterprises, or to compensate for under-coverage (see Kloek *et al*., 2013, and MEMOBUST 2014).

Finally, it is worthwhile to remind that, when parameters estimates are directly produced based on multi-source statistical registers involving the direct usage of AD, the published statistics are rarely accompanied with measures of data quality: the problem of constructing confidence intervals based on register-based statistics has been recently treated in Laitila (2014).

3. Exploiting the information potential of administrative data for the compilation of SBS data: a EU level overview

The extent to which AD sources are exploited in EU and EFTA Countries to estimate SBS has been recently explored by the ESSnet Admin Data[5], together with an analysis of the reasons that in each Country hamper the full utilization of this type of secondary data sources. In particular, an overview of MS's existing practices in the area of SBS estimation has been provided (see Costanzo, 2011 and the reference library available at: http://essnet.admindata.eu/ReferenceLibrary/List), and methods for estimating SBS which cannot be directly derived from administrative sources have been investigated.

[5] ESSnet on "The Use Of Administrative And Accounts Data For Business Statistics", http://essnet.admindata.eu/.

These analyses highlighted that, despite the availability of relevant, high quality and suitable AD (mostly tax-related and Social Security data) in most MSs, their potential is not yet fully exploited by the large majority of NSIs. Only France (Chami, 2010; INSEE, 2008) and Nordic Countries (Sweden, Denmark, the Netherlands and Norway, see Wallgren *et al.*, 2007) have implemented efficient integrated systems for SBS which are based on the *primary* use of AD sources, complemented where necessary by sampling surveys, focusing either on specific sub-populations or on sub-sets of target variables.

A different production strategy has been adopted by Portugal (Chumbau *et al.,* 2010), where the so-called "Simplified Business Information" (IES) system has led to the complete replacement of SBS surveys with administrative sources directly provided by the major owners of data on businesses. In other countries, like Germany, the production strategy for SBS is actually based on a combination of survey and BR data (Essnet AdminData, 2010), where usually the surveyed units are those above a given threshold. In Spain, SBS estimation is essentially based on direct sample surveys and AD are used only as an indirect source, e.g. for data editing and imputation activities (Eurostat, 2010) or for model-assisted estimation aiming at increasing the efficiency of the sample survey strategy (Saralegui *et al.*, 2013).

It is evident that in those countries where a complete SBS archive is available, integration with other statistical registers can be performed at micro-data level: this allows information to be accessible in dimensions not earlier available, enabling for instance more robust analyses of relationships among variables observed in the different sources, as well as the publication of new and more detailed statistics to respond to new user's needs with no additional response burden and/or additional surveys (Wallgren *et al.*, 2007). On the contrary, when combining a statistical register with SBS sample data, specific methodological issues arise. It is well known that when one or more sample surveys are involved in a linking process, as survey designs rarely target multiple purposes, potential selection bias in linked datasets can theoretically result.

Actually, linking and aggregation of information from two or more sources including sample surveys may affect the representativeness and usefulness of indicators from the integrated data set, unless appropriate methodologies are adopted ensuring data consistency at aggregate (estimates) level. As mentioned before, a direct way of obtaining consistent estimates at any level of detail in case of incomplete multi-source databases is by *mass imputation*. Other approaches proposed in literature to obtain consistent parameters' estimates are based on *re-weighting*, consisting in properly modifying the units' weights (see for example Iancu, 2013). However, it has to be stressed that the traditional calibration procedures may ensure consistency only for those variables which are directly used to determine sampling weights. However, when consistency is required for a high number of variables of the statistical registers which are not used in the sampling strategy, different approaches are necessary, since - as well known - calibration becomes unfeasible for more than two variables. In this situation, the *consistent repeated weighting* developed by Statistics Netherlands (Renssen *et al.*, 2001; Bakker, 2011) can be used. It consists in an iterative procedure based on the repeated application of the regression estimator which generates new sets of weights, one set for each output (for instance a frequency table) to be published, in such a way that later estimates are consistent with the previous ones.

4. The new approach in the compilation of the Italian SBS data

Italian SBS data are currently affected by a transition period that will generate in the near future a major shift from direct reporting based on large samples and limited use of AD, to an intensive use of AD coupled by the development of a new system of business surveys based on limited samples and oriented to collect new business relevant data.

In Italy, SBS estimation has been traditionally based on data collected through two direct annual surveys: the sample survey on Small and Medium Enterprises (SME) (enterprises with less than 99 persons employed, about 4.3 million of units as reference target population), and the total survey on Large Enterprises (LE) (about 11,000 enterprises representing the census target population for all enterprises with 100 or more persons employed). Both surveys collect information according to EU harmonised statistical definitions on profit and loss accounts, as well as on employment, investments etc. in the industrial, construction, trade and non-financial services sectors. The SME sample consists of about 100,000 enterprises each year. In this setting, AD were essentially used as indirect auxiliary information.

The increasing stability, timeliness, coverage and accuracy of firm-level information available in some AD sources on businesses' economic accounts[6] has made it possible to develop a new SBS estimation system mainly based on the direct use of AD as primary source of information, integrated with the SME survey data (Luzi *et al*. 2014). In the resulting Statistical Data Warehouse (called "frame SBS"), firm-level data for the main economic aggregates[7] are directly obtained from the integrated sources, as they cover about 95% of the whole target population.

Actually, the new estimation strategy consists in a *mixed* procedure, composed of two main phases. In the first phase, a predictive approach based on *imputation* has naturally allowed to build a complete micro-data file for the main economic aggregates, which are extensively covered by the integrated sources: in this phase, non- available information is predicted on the basis of AD using a combination of different imputation techniques, which have been applied to separate groups of variables taking into account their distributional characteristics and their relations with other variables (see Di Zio *et al*. 2014 for more details). This approach ensures final estimates which are free of the traditional levels of sampling errors.

In the second estimation phase, the remaining economic account's items, which are characterized by inadequate coverage rates, are estimated on the basis of data observed in the SME survey, using the main economic aggregates as auxiliary information. A *design based/model assisted* approach has been adopted, consisting in the use of a *projection estimator* (Kim *et al*., 2011): in this approach, a weighted regression model for each component based on the sample responding units is estimated, after adjustment of the sample weights for nonresponse and using the main economic aggregates as covariates: the unobserved variables values are imputed (*projected*) using the estimated regression parameters, so that the final estimates can be obtained as the sum of the model-based projected values (see Luzi *et al*. 2014 for more details).

The *projection estimator* has some important properties: estimates at high level of detail can be obtained, having a better precision with respect to traditional regression estimators; in addition, the estimator is unbiased at the domain levels (or at more aggregate levels) were the regression models are estimated. Note that, under specific conditions, this estimator corresponds to a modified GREG estimator, in this sense it can be considered a design-based small area estimator (Kim *et al*., 2011).

[6] Financial Statements, which annually provide information on about 75,000 units for limited liability companies; the fiscal Sector Studies survey, which annually includes about 3.5 million of SMEs, the Tax returns form data, containing economic information for different legal forms for about 4.5 million of units each year, and the Social Security Data, which includes firm data and individual (employees) data on occupation and labor cost for all enterprises.

[7] Income from Sales and Services (Turnover), Changes in stocks of finished and semi-finished products, Changes in contract work in progress, Changes in internal work capitalized under fixed assets, Other income and earnings (neither financial, nor extraordinary), Purchases of goods, Purchases of services, Use of third party assets, Changes in stocks of raw materials and for resale, Other operating charges, Personnel Costs.

It should be pointed out that the development of the new system has implied high initial costs not only for the methodological design, but also for the analysis and treatment of the AD sources. It has to be remarked that in the Italian context, each combined source actually covers different - yet partially overlapping - sub-populations of enterprises, and that some sources provide information on (partially overlapping) variables: for each source, this "common" information has been used for assessing the quality of input data, for harmonizing classifications and definitions with SBS concepts described by the SBS regulation, and for editing micro-data (identification of logical inconsistencies/measurement errors, removal of duplicated units, etc.).

Specific analyses have been devoted to manage inconsistencies among data from different sources (for instance in the case of labor cost and number of employees w.r.t. other economic accounts' variables). It has to be remarked that in the context of the frame SBS the unit identification is not an issue, as in each administrative archive enterprises are uniquely identified and classified based on a complex procedure performed at the Business Register construction stage.

The Frame-SBS is now the pillar of the new system of economic statistics in Italy, according to the innovation strategy launched in 2011 (Monducci, 2010).

It is worthwhile to underline that, based on the frame SBS, estimates for key economic account variables can be calculated at a very refined level of detail, thus facilitating the dissemination of larger, more detailed and better focused data to end-users. Furthermore, the frame SBS represents an advanced "intermediate output" which is expected to ensure higher levels of consistency between annual statistics on enterprises and National Accounts (NA). Improvements are also expected in terms of coherence of SBS estimates over time, as well as accuracy of cross-sectional SBS estimates within and across statistical domains.

By the end of 2014 Italy will transmit to Eurostat a new set of SBS data (2011-2012) coherent with the new methodological approach. The Frame-SBS 2011 has been already included in the NA estimates produced for the new benchmark according to the new regulation ESA2010. The results will be published by Istat in September 2014.

In a medium-term perspective, the Frame-SBS will be the reference information system for the revision of the design of structural economic surveys (for instance R&D, CIS, ICT) in an integrated way, for a more coordinated, coherent and non-redundant production system.

5. An example of new indicators based on the integration between SBS and foreign trade data: the business performance of Italian exporting enterprises

The availability of census like data on key SBS variables for all resident enterprises in Italy, included in the Frame-SBS introduced in the previous section, allows expanding the production and dissemination of variables and indicators for the target sub-population of Italian exporting enterprises far beyond what is already made available by standard TEC figures. In particular, the integration between foreign trade and business data is realised at the micro level through the Italian business register with no loss of business relevant information. As a result, data can be explored across different dimensions in a very flexible way (standard SBS estimation domains do not necessarily represent a constraint) and aggregated in a consistent way with SBS official figures. The possibility to link exporting enterprises characteristics with their business performance and to compare those results with non-exporting firms represents the key informative advantage of this approach.

For the sake of simplicity, in this paper new figures are provided for the manufacturing sector only where this sector is responsible for the largest share of Italian exports of goods (80%), with reference year 2011.

There are 83 thousand exporting enterprises in manufacturing, representing 21% of all active firms in this sector (28% of all enterprises with a corporation status) and covering 81% of national value added in manufacturing.

The breakdown of these data jointly by enterprise size and export propensity classes (Table 1) shows that exporting enterprises with more than 20% of their turnover realised with foreign sales explain more than a half of national valued added in manufacturing, while more export intensive firms (more than 50% of their turnover come from exports) account for more than 1/4 of national valued added.

The analysis of the same data by enterprise size classes clearly shows that the export propensity is highly correlated with business size. Indeed, export intensive firms (50% and more of foreign sales over total turnover) in micro businesses represent only 3.2% of the total value added of this size class, while this percentage reaches 36% for medium and 42% for large businesses.

Table 1 – Contribution of Italian firms to national value added in manufacturing by enterprise size and export propensity classes – 2011 (Percentage values)

Export propensity classes (percentage shares of export over turnover)	*Enterprise size classes (persons employed)*				
	1-9	*10-49*	*50-249*	*250 and more*	*Total*
Non exporting	8.9	8.1	1.8	0.3	19.0
0-5	1.5	6.1	4.0	4.2	15.7
5-20	0.9	4.0	4.3	4.7	13.9
20-50	0.6	4.3	7.2	11.0	23.0
50-80	0.3	2.9	7.4	12.1	22.7
80+	0.1	1.1	2.2	2.3	5.7
Total	12.3	26.4	26.9	34.5	100.0

The comparison between exporting and total firms in terms of export intensity by size class clearly shows a substantial divergence for micro businesses only (Table 2).

Table 2 – Export propensity for exporting and all firms in manufacturing by enterprise size class – 2011 (percentage shares of export over turnover)

Type of firms	*Enterprise size classes (persons employed)*				
	1-9	*10-49*	*50-249*	*250 and more*	*Total*
Exporting firms	23,1	27,1	38,1	36,9	34,7
All firms	8,2	20,6	36,0	36,7	30,0

Enterprises that export 50% or more of their turnover contribute for about 60% to the total exports of manufacturing firms (Tab. 3). In particular, firms highly oriented to foreign markets (more than 80% of their turnover is sold abroad) activate national exports for less than 20%.

Enterprises with little export orientation (less than 5%) represent less than 1% of national exports although their contribution to value added in manufacturing is quite important (almost 16%).

Table 3 – Contribution of Italian firms to national exports in manufacturing by enterprise size and export propensity classes – 2011 (Percentage values)

Export propensity classes (percentage shares of export over turnover)	*Enterprise size classes (persons employed)*				
	1-9	*10-49*	*50-249*	*250 and more*	*Total*
0-5	0,1	0,3	0,2	0,4	0,9
5-20	0,3	1,5	1,9	2,7	6,4
20-50	0,7	4,7	8,5	17,9	31,8
50-80	0,8	6,3	14,4	21,7	43,1
80 and more	0,8	3,6	7,2	6,2	17,7
Total	2,6	16,4	32,1	48,9	100,0

The breakdown by manufacturing industries shows that the contribution of enterprises more oriented to exports (more than 50% of turnover sold on foreign markets) remarkably varies across industries. In four sectors, such as “Other manufacturing industries”, “Manufacture of other transport equipment”, “Manufacture of machinery and equipment” and “Manufacture of leather and related products”, these enterprises account for more than 70% of industry exports. On the contrary, they represent a minority share of other industries, such as food processing where they account for about 1/3 of industry exports.

The enterprise level link between those data and the Italian Business register on resident enterprises groups represents another way to expand new figures with already available data. Exporting enterprises that are part of an enterprise group represent only 21,5% of the total but they also explain 82% of national exports and 64% of national value added in manufacturing. Based on this link, further analysis can be developed to assess, for instance, the role played by individual enterprise within an enterprise group (trader, supplier, specialised producer).

The integration between foreign trade and business data is also important to assess the impact on enterprise performance related to different firm internationalisation profiles (Table 4 and 5).

Tab. 4 – Labour productivity for exporting and non-exporting firms in manufacturing by enterprise size class – – 2011 (Value added per person employed, EUR thousands)

Enterprise size classes (number of persons employed)	*Non-exporting firms*	*Exporting firms*	*All firms*
1-9	28.0	41.7	31.1
10-49	38.3	54.6	48.3
50-249	52.2	71.1	69.4
250+	52.1	82.7	82.3
Total	34.1	68.6	58.2

Exporting firms largely outperform non exporting ones in productivity across all enterprise size classes. In addition, labor productivity increases more sharply for exporting enterprises as compared to non-exporting ones as firm size grows. In a similar way, exporting firms are more profitable than non-exporting ones.

Tab. 5 – Profitability for exporting and non-exporting firms in manufacturing by enterprise size class – – 2011 (Gross profit over value added, percentage values)

Enterprise size classes (number of persons employed)	*Non-exporting firms*	*Exporting firms*	*All firms*
1-9	17.0	30.3	21.0
10-49	22.6	33.0	29.8
50-249	27.8	37.9	37.3
250+	20.5	39.0	38.9
Total	20.7	37.0	34.1

Conclusions

The integration between trade and business data represents a very promising opportunity in a context where EU and national statistical authorities are struggling to provide high quality and business relevant data given decreasing financial and human resources.

As has already been demonstrated by the Italian experience, new data and indicators allow to address policy relevant questions, such as the role played by exporting enterprises in boosting productivity and profitability, and creating a large portion of value added in manufacturing. This approach can also be extended to the analysis of Global Value Chains (GVC) by considering import of intermediate goods and FATS variables as key information to develop a new taxonomy of firm internationalisation profiles. In addition, the integration between trade and business statistics only represents a "business case" of what could be achieved in other relevant areas of structural business statistics by considering, for instance, new indicators related to other policy relevant business sub-populations such as innovative versus non innovative firms, ICT intensive versus non ICT intensive firms.

The exploitation of new indicators and variables based upon the integration of already existing primary or secondary data sources, however, requires the adoption of a sound methodological approach and a proper reorganisation of statistical production lines in order to successfully meet the quality standards and the financial and resources constraints that nowadays characterise the production of official statistics.

An efficient and cost saving approach for the development of a new generation of structural business statistics and for the sustainability of current data requirements necessarily calls for an in-depth reengineering of current statistical production pipelines based on a dualistic approach.

The first component of this approach relies on the massive and intensive use of administrative data to feed the need of standard SBS variables for large populations of businesses. The statistical processing of those data according to appropriate methodologies must guarantee a high degree of accuracy for micro-level figures, as well as the full consistency with the overall data frame

provided by the business register and SBS key figures. The presence of an enterprise identification code in the database and the use of record linkage procedures then open the door to further integration with other variables coming from direct reporting or administrative data.

The second component encompasses a dedicated system of direct reporting surveys focused on limited and highly qualified samples of well-targeted business populations. This system is finalised to monitor and improve the quality of core variables in the basic frame for the compilation of SBS key variables as well as to collect quantitative and qualitative information on business units, and especially large statistical units, that exhibit a complex organization and economic behavior.

A road map should be designed and agreed at the European level, including not only the "Vision" but also a possible "industrial plan" to concretely support EU member countries in the development of this approach. The MEETS project has already provided some basic elements in terms of data infrastructures and standardisation tools, but more information and support should be provided by Eurostat in strong coordination with Member States. Given the presence of substantial differences across EU countries in terms of data frame characteristics (availability and quality of administrative data, size of the business target population, etc.) a range of alternative solutions should be explored. Successful country based experiences, related to a specific package of concrete solutions (data sources, estimation methodologies and outputs), should be then extended to the cluster of closest countries in terms of similar data frame characteristics.

References

Bakker, B. F. M. (2011), Micro-Integration: State of the art. In: *Report WP1: State-of-the-art on Statistical Methodologies for Data Integration*, ESSNET on Data Integration, available at http://www.cros-portal.eu/content/wp1-state-art.

Chambers, R., van den Brakel, J. A., Hedlin, D., Lehtonen, R., and Zhang, L.-C. (2006), Future Challenges of Small Area Estimation. *Statistics in Transition* **7**, 759–769.

Chami S. (2010). Reengineering French structural business statistics - an extended use of administrative data. European Conference on Quality in Official Statistics (Q2010), Helsinki.

Cibella N., Fernandez G.-L., Fortini M., Guigó M., Hernandez F., Scannapieco M., Tosco L., Tuoto T. (2009), Sharing Solutions for Record Linkage: the RELAIS Software and the Italian and Spanish Experiences, In Proc. Of the NTTS (New Techniques and Technologies for Statistics) Conference, Bruxelles, Belgium, 2009.

Costanzo L. (2011). An Overview of the Use of Administrative Data for Business Statistics in Europe. Int. Statistical Inst.: Proc. 58th World Statistical Congress, 2011, Dublin (Session CPS035).

Chumbau A., Pereira H.J., Rodrigues S. (2010). Simplified Business Information (IES): Impact of Admin Data in the production of Business Statistics, presented at the Admin Data ESSnet Seminar "Using administrative data in the production of business statistics. Member states experiences" (Rome, March 2010).

Da Silva, D. N., and Zhang, L.-C. (2014). Estimation of the variance due to imputation in the 2011 UK Census. UNECE Work Session on Statistical Data Editing. Paris, 28-30 April.

Di Zio, M., Guarnera, U., Varriale R. (2014). *Imputation with multi-source data: the case of Italian SBS*. Paper presented at United Nations Economic Commission for Europe, conference of European statisticians.

Essnet AdminData, (2010). Main findings of the Information Collection on the Use of Admin Data for Business Statistics in EU and EFTA Countries. Deliverable 1.1., Admin Data ESSnet, Work Package 1: Overview of MSs' Existing Practices in the Uses of Administrative Data for Business Statistics, available at: http://essnet.admindata.eu/WikiEntity?objectId=4774

ESSnet DI (2011). Final Report of the Essnet on Data Integration, available at: http://www.cros-portal.eu/content/data-integration-finished.

Iancu D., Hagsten E., and Kotnik P. (2013). Quality of Linked Firm-Level and Micro-Aggregated Datasets: The Example of the ESSLait Micro Moments Database". Report of the ESSnet on Linking of Microdata to Analyse ICT Impact, available at http://www.cros-portal.eu/content/final-reporting-esslait-project.

INSEE (2008). The French Business Register: from a quality approach to a statistical register, paper presented at the 21st Meeting of the Wiesbaden Group on Business Registers - International Roundtable on Business Survey Frames (Paris, November 2008)

Istat (2008), Metodi statistici per il record linkage, Mauro Scanu editor, Istat

Laitila T. (2014). Constructing Confidence Intervals based on Register Statistics. European Conference on Quality in Official Statistics (Q2014). Vienna, 3-5 June

Luzi O., Guarnera U., Righi P. (2014). The new multiple-source system for Italian Structural Business Statistics based on administrative and survey data. European Conference on Quality in Official Statistics (Q2014). Vienna, 3-5 June.

Monducci R. (2010). Statistiche ufficiali e analisi della competitività del sistema delle imprese: aspetti concettuali, problemi di misurazione, strategie di miglioramento della qualità. Atti della X Conferenza nazionale di statistica, Roma, dicembre 2010.

Pannekoej K. (2011). Models and algorithms for micro-integration. In: Report on WP2: Methodological developments. Essnet on Data Integration, available at http://www.cros-portal.eu/content/wp2-development-methods.

Rao, J. N. K. (2003). Small Area Estimation. New York: John Wiley and Sons.

Saralegui J., Gonzalez C. and Arbués I. (2013). Use Of Administrative Sources To Reduce Statistical Burden And Costs In Structural Business Surveys (Ufaes). NTTS 2013. Brussels, 5-7 March.

Renssen, R. H., Kroese, A. H., and Willeboordse, A. (2001). Aligning estimates by repeated weighting. Research paper 491-01-TMO, Statistics Netherlands, Voorburg/Heerlen.

Kim, J. K. K., Rao, J. N. K. (2011). *Combining data from two independent surveys: a model-assisted approach*. Biometrika. No.8, pp. 1–16.

Kloek W., and Vâju S. (2013). The use of administrative data in integrated statistics. NTTS - Conferences on New Techniques and Technologies for Statistics. Brussels, 5-7 March

Wallgren, A. and Wallgren, B. (2007). Register-based statistics – Administrative data for statistical purposes. John Wiley and Sons, Chichester.

Whitridge P., and Kovar J. G. (1990). Use mass imputation to estimate for subsample variables. In: *Proc. Bus. Econ. Statist. Sect.,* American Statistical Association, Washington DC, 132-137.

Zhang, L.-C. (2012). Topics of statistical theory for register-based statistics and data integration. Statistica Neerlandica 66, 41–63.