Estimating Business Statistics by integrating administrative and survey data: an experimental study on small and medium enterprises

Orietta Luzi[1], *Giovanni Seri*[2], *Viviana De Giorgi*[3], *Giampiero Siesto*[4]

Sommario

Il lavoro affronta il problema della stima di statistiche strutturali sulle imprese, sfruttando informazioni disponibili da fonti amministrative in modo integrato con dati di indagine. In particolare, l'obiettivo è quello di verificare la possibilità di stimare alcune delle principali variabili strutturali che non sono direttamente disponibili dalle fonti: ciò implica la necessità di utilizzare modelli di stima o di imputazione per derivare le stime richieste. In questo lavoro, l'attenzione è focalizzata sulle variabili relative alle variazioni delle scorte di beni e servizi rilevate nell'indagine annuale sulle Piccole e Medie Imprese: diverse strategie di imputazione sono valutate sperimentalmente a seconda dei diversi scenari corrispondenti ai diversi "pattern" di risposta determinati dalla disponibilità delle variabili analizzate in uno, più di uno o nessuno degli archivi amministrativi considerati.

Parole chiave: statistiche strutturali, dati amministrativi, integrazione dati, imputazione

Abstract

The paper deals with the problem of estimating structural business statistics by exploiting already existing administrative information integrated with survey data. In particular, the aim of the study is to verify the possibility of estimating key structural variables which are not directly available from administrative sources: this implies the need of using either estimation or imputation models to derive the required estimates. In the present paper, the attention is focused on the variables relating to changes in stocks of goods and services investigated in the annual survey on small and medium enterprises (Small and medium enterprise survey -SME): different imputation strategies are experimentally evaluated depending on the different scenarios corresponding to the various response patterns determined by the availability of the analysed variables in one, more or none of the considered administrative archives.

Keywords: structural business statistics, administrative data, data integration, imputation

[1] Head of Research (Istat), e-mail: luzi@istat.it.
[2] Researcher (Istat), e-mail: seri@istat.it.
[3] Researcher (Istat), e-mail: degiorgi@istat.it.
[4] Senior Researcher (Istat), e-mail: siesto@istat.it.

The views expressed in this paper are solely those of the authors and do not involve the responsibility of Istat.

1. Introduction

In the area of business statistics, administrative data currently represent a key element for European National Statistical Institutes (NSIs) to reduce production costs and statistical burden on enterprises while maintaining high data quality levels. The new European Regulation on Structural Business Statistics (SBS in the following)[5] establishes that, in order to estimate information on the structure of National production systems, NSIs can integrate data available in different information sources, including administrative ones. A number of re-design projects in this context have been carried out or are currently ongoing in several European Countries[6].

Costs and response burden are especially relevant in the Italian economic system, which is characterized by a large amount of small and medium enterprises (out of about 4.5 million of enterprises, companies with less than 10 number of persons employed are about 95% and sum up about 47% of number of persons employed): this fact, together with the high level of detail required by the European Regulations on SBS and the amount of information to be estimated, imply relevant costs on the Italian Statistical Institute (Istat) and significant burden on enterprises (with high non response rates).

As known, using administrative data for statistical purposes poses a number of additional problems w.r.t. traditional survey processes (Wallgren and Wallgren, 2007) in terms of data integration, data quality (Eurostat, 1999) and assessment of data usability, including coverage and suitability of information contents (in terms of comparability of statistical and administrative definitions). As a consequence, integrating external data in statistical processes implies a deep revision of the overall production strategies.

In this paper we focus the attention on the Italian *Surveys on Business Size and Competitiveness* (SBSC) (cfr. Istat, 2011). The SBSC consists of two different surveys: 1) the total Annual Survey on the Economic Accounts of Enterprises (SCI in the following), involving enterprises with 100 or more persons employed, and 2) the sample survey on small and medium enterprises upon enterprises with less than 100 persons employed. Both surveys contribute to the estimation of SBS. In the context of SBSC, a large amount of high quality administrative information is at present available in the Italian economy: the existing sources, in particular Balance Sheets and Fiscal Authority sources, cover an extensive amount of business population and may provide both direct and indirect information for estimating SBS.

In both surveys, the available external information is essentially used to compensate for non responses on a subset of key variables, by directly replacing missing values with the corresponding administrative data (Casciano et al., 2011). This situation has encouraged Istat in setting up a number of activities aiming at supporting a more extensive and rigorous use of administrative data in this area, by proceeding in two main directions. From one side, a number of supporting tools are at present under development in order to guarantee continuous and secure access to external data: besides formal agreements with the Italian Tax Authorities to establish a stable cooperation protocol for business data exchange, some

[5] March 2008.

[6] among others: France (Brion et al., 2009), UK (Lewis, 2010, Elliott, 2010), Portugal (Chumbau et al., 2010), Finland (Tolkki, 2007).

technological tools are under development to facilitate the access to administrative data by the direct electronic transmission of information from enterprises to Istat[7]. From the other side, a number of experimental studies and data analyses are in progress for evaluating the potential benefits and the statistical impact on quality of results due to the integration of administrative and statistical survey data for estimating the key SBS variables in the SBSC. In this context, important elements to be first considered are completeness and coverage of the external sources, i.e. the sources coverage in terms of items (variables) and units, respectively. These two quality dimensions, in effect, are related to the amount of not available information to be recovered (e.g. by direct surveys or by model estimation) once administrative data are used in the statistical production process:

- under-coverage of administrative sources with respect to specific business sub-populations can be viewed as a "total non response" problem;
- incompleteness of administrative sources in terms of target variables which are not directly available from them[8] can be viewed as an "item non response" problem.

In this paper, we deal with sources incompleteness, with particular attention to the situation where administrative data cannot be used to directly "replace" survey data, but appropriate methodologies can be used to compensate for some of the (partially) unavailable information.

In particular, we illustrate the results of some experiments aiming at evaluating the possibility of estimating the components of the variable Changes in stocks of goods and services (CS in the following) based on related information available in the external archives, and to identify the "best" class of estimation methods (at unit level) that could be used to this purpose. Concerning CS, under the framework of the Eurostat Regulation Ce 295/2008 SBS, details are required for the following variables: Changes in stocks of finished products and work in progress and Changes in stocks of goods and services purchased for resale as they are involved in the computation of the Production Value and Gross margin on goods for resale.

While CS and its components are directly available from administrative archives for large enterprises, the same does not hold for SMEs. For this reason, experimental analyses have been restricted to this latter area.

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

The paper is structured as follows. In Section 2 we briefly describe the current SME survey and the available external sources of information on SMEs. An experimental study to evaluate the performance of alternative imputation methods (both parametric and non-parametric) for estimating components of CS is illustrated in Section 3. To this aim, different scenarios to represent the possible information frameworks to deal with are

[7] adopting the eXtend Business Reporting Language technology - XBRL - and creating a statistical web portal for the direct electronic acquisition of businesses' balance sheets.

[8] Assuming that the definitions of statistical and administrative variables are coherent or can be reconciled.

identified. According to the scenarios, in the same section the specific approaches for unit level data prediction are introduced. Experimental results are shown in Section 4. Final remarks are reported in Section 5.

2. The SME survey and the available administrative data

The sample survey on SMEs is carried out annually with the general purpose of investigating profit-and-loss accounts of enterprises with less than 100 persons employed, as well as information regarding employment, investment, personnel costs and the regional breakdown of some variables, as requested by the SBS EU Council Regulation n. 58/97 and 295/2008 (Eurostat 1999). The survey involves units belonging to the industrial, construction, trade and services economic activities. The survey's frame is represented by the Italian Business Register of active enterprises (BR in the following), resulting from the combination of both statistical and administrative information (Tax Register, Social Security Register, Register of the Electric Power Board, etc.). The BR contains variables such as Economic activity, Turnover and Number of persons employed. It counts about 4.5 million enterprises which employ approximately 17.6 million persons. The 2007 SME target population counts about 4 million enterprises (about 94% of the BR enterprises). Target parameters are estimated by publication domains in accordance with the SBS Regulation[9].

The sampling design is a one stage stratified random sample with strata defined by economic activity, size class and administrative region. In 2007, about 103,000 enterprises were included in the sample. The response rate was close to 40% (varying according to size classes and economic activity sectors) in terms of reliable replies.

Besides BR, the relevant administrative sources available on the SME survey target population and parameters are Balance Sheets (BS) and Tax Authority sources (Tax returns forms and Fiscal Authority survey).

The most accurate and reliable administrative source for SBS is represented by the BS of the corporate enterprises collected by the Chambers of Commerce. Companies liable to fill in the balance sheet are about 650,000 covering less than 20% of the BR, although they are about 57% in terms of persons employed. This source is the best harmonized with the SBS Regulation definitions.

All other enterprises are obliged to declare their taxable income to the Fiscal Authority by filling in tax forms. In particular, Istat acquires data from the Sector Studies survey (Fiscal Authority Survey, SS in the following), that is a survey carried out by the Italian Fiscal Authority to evaluate the capacity of enterprises to produce income and to know whether they pay taxes correctly. The Fiscal Authority allows the SS data to be available at Istat for statistical purposes. In spite of some exclusion and non-enforceability principles, almost all enterprises are obliged to fill in the SS survey form (together with the tax return form) and to declare in detail costs and income items. It involves about 4 million enterprises with the Turnover lying in the interval (30,000 - 7,5 million) euro. The common part of SS forms is a sort of balance sheet providing an important set of key variables, for this reason we selected this source in addition to BS.

[9] The data domains are: 1) class of economic activity (4 Nace-code digits); 2) economic activity (3 Nace-code digits) by size (classes of persons employed); 3) economic activity (2 Nace code digits) by regions (Nuts2 level).

Concerning the coverage of BS and SS with respect to the theoretical sample of the SME survey is graphically described, some specific businesses sub-populations are not covered at all by either BS nor SS. These are the so called minimum tax payers[10] and the sole proprietorships with Turnover>7,5 millions euro. Estimating the target variables for this sub-population requires the adoption of appropriate approaches. Figures about coverage analysis of the actual sample for the 2007 SME survey are reported in Table 1. The proportion of respondents covered by BS is around 45%, but the percentage reduces to 11% when weighted. This fact suggests that the response rate for companies is much higher with respect to the other enterprises.

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

Table 1 - Coverage analysis of the sample of the SME survey by administrative data - Year 2007

SOURCE	Coverage (non overlapping BS)	Coverage % (non overlapping BS)	Weighted Coverage % (non overlapping BS)
Balance Sheets(BS)	19739	~45%	~11%
Sector Studies Survey	29406 (19021)	~67% (~43%)	~91% (~82%)
SME Survey (respondents)	43701		

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

ECONOMIC ACTIVITY	BS		SS-F[11]		SS-G		TOTAL	
	ENT	EMP	ENT	EMP	ENT	EMP	ENT	EMP
C-Mining and quarrying	49.9	69.8	39.5	23.9	0.1	0.0	89.5	93.8
D-Manufacturing	22.5	54.5	64.8	37.9	0.0	0.0	87.3	92.4
E-Electricity, gas and water supply	57.5	81.8	2.1	0.6	0.1	0.0	59.7	82.4
F-Construction	14.3	33.4	72.9	56.5	0.1	0.0	87.2	89.9
G-Wholesale and retail trade; repair of motor vehicles, motorcycles and personal and household goods	11.1	30.7	77.0	60.1	0.1	0.0	88.2	90.9
H-Hotels and restaurants	10.8	24.5	75.1	66.1	0.0	0.0	85.9	90.7
I-Transport, storage and communication	16.3	47.6	68.7	39.9	1.1	0.3	86.1	87.8
J-Financial intermediation	6.1	13.9	72.8	68.0	6.3	4.3	85.2	86.3
K-Real estate, renting and business activities	13.9	31.5	23.9	22.8	49.9	34.6	87.7	88.9
M-Education	19.2	45.6	22.6	15.9	1.5	0.5	43.3	62.0
N-Health and social work	4.6	31.1	2.9	4.2	81.9	55.3	89.4	90.7
O-Other community, social and personal service activities	7.8	26.2	64.5	53.9	3.7	1.8	76.0	81.9
TOTAL	13.2	37.0	56.6	45.1	17.0	7.8	86.8	90.0

[10] Minimun tax payers: sole proprietorships with turnover <= 30.000 euros, with no employees, no exportations, no external personnel, with no gross investments in capital goods or rents of capital goods in the last three years

[11] With SS-F and SS-G two separate sub section of the SS model are indicated, containing each information on specific items in the context of SBS.

3. The experimental application

In this section we describe an experimental application of some imputation methods to deal with the problem of estimating the two main components of the variable CS for SMEs by exploiting the related information which is available from BS and SS. The purpose is both to assess the potential suitability of data prediction at unit level for this specific variables, and to have first indications on the most appropriate imputation approaches to be used. The reference year is 2007.

3.1. The target variables

The CS' main components relate to the following items: *Changes in stocks of finished products and work in progress* (Csfp in the following), defined as the difference between the value of the stocks at the end and the beginning of the reference period; *Changes in stocks of raw materials and consumables* and *Changes in stocks of goods and services purchased for resale*, on the contrary, are defined as the difference between the value of the stocks at the beginning and at the end of the reference period. The sum of the last two variables results in the derived variable *Changes in stocks of raw materials and goods and services for resale* (Csrm in the following). CS is defined as follows:

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

The CS distributions is characterized by the presence of high percentage of zero values (37% of the observed variable values in the sample, representing more than half the whole population), in other words CS is characterized by a semi-continuous distribution which has to be taken into account in modelling data for estimation purposes.

Cs, Csfp and Csrm are currently obtained by direct collection. In particular, CS components belong to the set of variables that cannot be directly obtained from the available administrative archives for the entire SBSC population. In effect, even if all of them are directly available from BS for the sub-population of corporate companies, for the remaining enterprises only the variable CS is directly available from SS. The informative situation supplied by the administrative sources is reported in Table 3, in which we can distinguish different informative scenarios according to the variables availability in the external data sources. It is worthwhile noting that for the three variables under study, a preliminary harmonization of definitions has been performed in order to obtain items which were comparable from a statistical point of view: therefore, possible discrepancies among items from different sources (the survey, the BS and the SS) can be due only to the different measurement processes.

Primarily, for the subpopulation of enterprises subject to fill in the BS a “fully informative” scenario can be defined as all the variables involved in (1) are available from that administrative source. This kind of scenario has been considered to assess the quality of the administrative source BS, that is prioritised as the most important one.

Two other scenarios can be distinguished. The first one, which we will refer to as “partially informative”, is defined by the availability, for a given sub-population of enterprises, of the only variable CS, the problem being to estimate the components Csfp and Csrm. In SMEs, this situation involves the 43% of the sample units covered by the SS survey. It is worth noticing that the 19% of the observed units have CS=0 and it can be assumed that this implies Csfp= Csrm=0 too. The second scenario, which we will refer to as “non informative”, is defined by the non availability of any of the target variables in the administrative sources for a given sub-population of enterprises. In SMEs, this is the situation referred to about the 12% of the sample units not covered by either BS or SS (half of them with CS=0).

Table 3 - Availability and coverage of the variables Changes in stocks of goods and services (CS), Changes in stocks of finished and semi-finished products (Csfp) and Changes in stocks of raw materials and for resale (Csrm), by administrative data: year 2007

	Financial Statements	Fiscal Authority Survey	Sample Coverage %	Weighted Sample Coverage %
Available Variables	*CS*, *Csfp*, *Csrm*	*CS*		
CS ≠0	Available	Available	17%	6%
	Available	---	16%	1%
	---	Available	24%	38%
	---	---	6%	3%
CS = 0	Available	Available	7%	3%
	Available	---	5%	1%
	---	Available	19%	44%
	---	---	6%	5%

3.2 The Imputation Methods

Imputation is a commonly applied approach to compensate for item non response in sample surveys (Kalton and Kasprzyk 1986; Schafer,1997). Single imputation has some desirable properties: 1) complete data can be obtained in order to allow for the use of standard estimation and data analyses methodologies, and 2) under specific assumptions, joint data distributions and information coherence at micro and estimation level are preserved. The main drawback deriving from the use of imputation consists of the additional uncertainty due to the prediction of missing information, which has to be properly considered at the estimation stage in order to obtain valid inferences on final data.

In order to estimate the CS’s main components, both parametric and non parametric imputation methods are considered. Parametric methods have the advantage of exploiting the explicit relationships between the target variables and the set of auxiliary variables. The main disadvantages relate to the need of assessing the underlying model and model fitting at the different data domains. Furthermore, relating to our specific estimation objective, the semi-continuous nature of the target variables (high frequencies of zeros in all domains, and low dispersion of non zero values around the domains’ modal values), suggests to consider non parametric approaches too.

In our study, the target parameters are totals of CS, Csfp and Csrm for *j* specific publication domains (D) defined as:

$$\hat{T}_{Var}^{D} = \sum_{i=1}^{n_D} \omega_i Var_i \qquad D=1,...,j \qquad (2)$$

where *Var*=CS or Csfp or Csrm; n_D is the number of units in domain D (where $\sum_{D=1}^{j} n_D = n$ is the sample size); ωi are the sampling weights.

3.2.1 Scenario 1: partially informative administrative data

Under this scenario, we are in the situation where for some SMEs only the variable CS is directly available from SS, while its two components Csfp and Csrm are to be estimated (about 43% of the SME respondents, about 19,000 sample units). We can treat this case as if we were in presence of partial non responses (MRP) on Csfp and Csrm for a portion of units.

We assume that MRPs are Missing At Random (Little et al., 1987) inside appropriate data domains (corresponding in general to the SME survey estimation domains), so that we are allowed to treat them as "similar" to the fully observed units inside domains.

In the following we show results for only the variable Csfp, since Csrm can be deductively derived from relation (1)[12.]. In order to assess the potential biasing effects on the Csfp total estimates due to the imputation of MRPs, a Monte Carlo simulation study has been performed based on k iterations (k=100) of the following steps:

- for a selected set of economic divisions, simulating pre-defined percentages of non responses on Csfp and Csrm on a sample of responding un-incorporated enterprises randomly chosen (MAR assumption w.r.t. some known auxiliary information);
- on test data, imputation of artificial non responses and estimation of Csfp and Csrm totals;
- evaluation of the impact of imputations on estimates.

Evaluation is based on Relative Bias (RB) and Relative Root Mean Squared Error (RMSE) of parameter estimates (by domain).

A) Nearest-Neighbour Donor

In this class of non parametric approaches, one of the methods traditionally used to predict variables values at unit level is hot-deck. Hot-deck is especially useful when strong explicit relations cannot be envisaged between the target and the auxiliary variables, as well as to deal with semi-continuous variables like the ones investigated in our research. In this case, in order to split CS into its two components in a given unit having the only CS available from administrative sources, a within cells Nearest-Neighbour Donor (NND) method is applied, where the imputed value at unit level is the proportion $p_i = \frac{CSFP_i}{CS_i}$ observed in the closest complete unit in the cell. Imputation cells are defined in terms of Economic activity (either 2 or 3 Nace rev.2 digits), Legal form (corporate, un-incorporate, sole proprietorship), and CS's sign. Also in this case, auxiliary variables used as matching items include information from both BR (Number of persons employed) and administrative sources (CS, Turnover, Purchases of goods and services for resale in the same condition as received).

[12] CS is in effect assumed to be known from the available administrative sources in the considered domains.

B) Robust regression

In robust regression, elementary values of the variable Csfp are predicted based on the simple regression model:

$$CSFP = \alpha + X\beta + \varepsilon \qquad (4)$$

where: **X** is the vector of *m* auxiliary variables available for the whole SME population from either the Italian BR or administrative sources; residuals **ε** are subject to usual theoretical assumptions; **β** is the vector of the regression coefficients to be estimated from observed data, by domain. Robust estimates of **β** are obtained based on the Least Trimmed Squares (LTS) algorithm (Rousseew et al., 1987) in order to obtain predictions for missing data which are not influenced by anomalous behaviors within domains. The auxiliary information explored in model estimation are Economic activity (either 2 or 3 Nace rev.2 digits), Number of persons employed, CS, Turnover, Purchases of goods and services for resale in the same condition as received.

As known, model estimation can be cost and time consuming from both a theoretical and operational point of view. In addition this approach requires that, due to the high frequencies of zeroes characterizing changes in stocks items, a preliminary probabilistic data modelling is performed (logistic regression is adopted here) to classify units based on their probability of having either zero or non zero changes in stocks components, depending on each specific domain.

C) Other parametric models

Other forms of simple (robust and non robust) model-based prediction at unit level by separate domains can be considered, again based on the same assumptions as above.

In particular, the following unit level within cells imputation methods have been tested:

1) *Mean imputation*: the imputed value at unit level is the mean proportion p_{mean}.=$Mean_{i \in D}(p_i)$, where $p_i = \frac{CSFP_i}{CS_i}$, and *D* is the imputation cell (*D*=1,...,*j*).
2) *Median imputation:* the imputed value at unit level is the within cell median of the p_i.

In both methods, zero values are excluded from calculations to avoid high frequencies of null means and medians.

As for robust regression, also in this case, a preliminary probabilistic modeling step is performed (logistic regression) in order to classify units based on their probability of having either zero or non zero changes in stocks components, depending on each specific domain.

3.2.2 Scenario 2: non informative administrative data

Under this scenario, we assume that for some specific SME sub-populations, information on neither CS nor its components is available from administrative sources. This is the case of the already mentioned Minimum Tax Payers and sole proprietorships.

In particular, we focus the attention on Minimum Tax Payers, which for the year 2007 are estimated to consists of 3,304 SME responding units (about 7,5% of the observed sample units). Our aim is to verify the statistical effects of excluding this piece of the SME population (for statistical burden and costs reasons[13]) from direct investigation. For this reason, we simulate the non availability of information on the three target variables for all the SME units belonging to the Minimum Tax Payers sub-population. Under this scenario, different approaches at either unit or estimation level could be considered.

Cut-off sampling (Benedetti et al., 2010; Knaub, 2008) could be a potential method to explore the possibility of obtaining parameters estimates without performing direct data collection on the "critical" sub-populations.

Alternatively, imputation methods at unit level can be adopted to derive the total estimates of CS, Csfp and Csrm based on completed elementary data matrices. In this case, non parametric Mass Imputation (Statistics Canada, 1998) and parametric regression methods have been considered and experimentally evaluated. The assumption is that in this case it is not possible to use neither administrative data nor responding units in the sub-population itself to estimate no one of the three variables of interest. For this reason, under this scenario, in the experimental application no variability is associated to nonresponse, as the sub-population of responding Minimum Tax Payers is wholly determined and variables CS, Csfp and Csrm are simultaneously cancelled in it. Furthermore, as the selected imputation methods do not include "random" elements, no iterations of them are needed.

In order to assess the potential biasing effects on CS, Csfp and Csrm parameter estimates due to unit imputation, the following steps have been performed:

- on all the units of the sub-population of SME enterprises classified as Minimum Tax Payers, artificial deletion of observed values of CS, Csfp and Csrm, to simulating the unavailability of observed information on target variables for this sub-population;
- unit level imputation of missing values obtained in the previous step and estimation of CS, Csfp and Csrm totals;
- evaluation of the impact of imputations on totals' estimates.

Evaluation is based on the distance between the estimates of CS, Csfp and Csrm totals derived from "true" original survey data ($\hat{T}^{D}_{j,ori}$) and data after imputation ($\hat{T}^{D}_{j,imp}$) (by domains):

$$Diff\_Var^{D} = \frac{\left| \hat{T}^{D}_{Var,ori} - {}_{k}\hat{T}^{D}_{Var,imp} \right|}{\hat{T}^{D}_{Var,ori}} \qquad D=1,...,j;\ Var = \text{CS, Csfp, Csrm.} \qquad (5)$$

D) Mass imputation

In this paper we refer to *Mass imputation* as a special case of NND imputation (see case A above) where variables to be imputed are always the same for each record. Under scenario 2 of our experimental application, variables to impute are always CS, Csfp and Csrm, simultaneously considered, while matching variables consist of the common

[13] On the contrary, given their potential impact on estimates, large sole proprietorships are typical candidates for re-contacts in case of non response on key surveyed variables.

information directly available from the external sources: *Economic activity, Legal Form, Number of persons employed, Turnover, Purchases of goods and services for resale in the same condition as received.*

Imputation cells are defined on the basis of Economic activity (3 digits of Nace rev.2), Legal form and Size (size classes: [1-9], [10-19], [20-49], [50-99]). Not available values of CS, Csfp and Csrm are jointly replaced at unit level by the corresponding ones observed in the closest complete unit (donor) in the same imputation cell.

E) Robust Regression

Robust regression modelling at domains level is used to predict unit values of variable CS:

$$CS_i = \alpha + \beta_m X_m + \varepsilon_I \qquad (6)$$

where: X_m indicates the m^{th} auxiliary variable available for the whole SME population from either the BR or an administrative source; residuals ε_i are subject to usual theoretical assumptions; the regression coefficients β_m are to be estimated on observed data, by imputation cells. As for Scenario 1, estimates of β_m should be obtained robustly. Parameter estimation is performed inside imputation cells defined by *Economic activity* (2 digits Nace rev.2), *Legal form*; auxiliary information used in the model is given by *Number of persons employed* and *Turnover*. Once CS is predicted, its components Csfp and Csrm can be derived as done in Scenario 1.

4. The experimental results

In this section we report the results obtained for the scenarios illustrated above for three Nace divisions: 17 (Textile Industry), 52 (Retail Trade) and 55 (Hotels and Restaurants). These divisions have been chosen in order to explore different economic contexts in the areas of Industry and Services.

4.1 Scenario 1: partially informative administrative data

Under scenario 1, the correlations between CS (available from administrative sources) and variables Csfp and Csrm (assumed as unknown) have been evaluated in each Nace division. To this purpose, the Pearson's correlation indexes, reported in Table 4, are computed on the BS complete dataset (year 2007). High values of the index are observed particularly in the Services division (Nace codes 52 and 55). As stated before, it has to be taken into account the presence of zero values for the variables relative to changes in stocks. In Table 5 the percentages of zero and non-zero values for the variables Csfp and Csrm combined are reported (it is assumed that, if CS=0 then Csfp=Csrm=0). It is interesting to note that in Nace divisions 52 and 55 (in the Services sector) it happens frequently that Csfp=0 while Csrm=CS. This information, of course, may be used in estimation/imputation strategies for the unknown variables.

Table 4 - Correlations between CS and Csfp, Csrm by Nace code - Year 2007; source BS

NACE	Variable	2007
17	*Csfp*	0,74
	Csrm	-0,77
52	*Csfp*	0,24
	Csrm	-0,91
55	*Csfp*	0,95
	Csrm	-0,26

Table 5 - Percentages of "0" values for the variables Csfp, Csrm by Nace - Year 2007; source BS

NACE	*Csfp*=*Csrm*=0	*Csfp*≠0;*Csrm*=0	*Csfp*=0;*Csrm*≠0	*Csfp*≠0;*Csrm*≠0
17	11%	10%	29%	50%
52	4%	15%	76%	5%
55	18%	11%	67%	4%

In order to evaluate the performance of the methods under this scenario, a percentage of 5% non-responses (corresponding to the percentage of missing values for variable CS resulting from the SME raw survey data) is simulated in each Nace division 17, 52 and 55 (consisting of 1,653, 3,202 and 1,306 units, respectively): as a consequence, target variables are "cancelled" in about 83 responding units for division 17, in about160 responding units in division 52, and in about 65 responding units in division 55. Moreover, as the methods have been tested by comparing estimates stemming from the survey data before and after replacing the simulated non responses with imputed values, we would like to have an adequate number of unit level comparisons and at the same time to alter slightly the donors population.

In Tables from 6 to 11, the results obtained by applying unit level imputation methods (NND, robust regression, within-cell mean, within-cell median) are shown. Different experiments have been performed by changing the criteria used to form imputation cells for each evaluated method.

The results corresponding to the "best" criteria in terms of RB and RMSE are shown: in Tables 6 and 7 for the NND method; in Tables 8 and 9 for the method based on Robust regression; in Tables 10 and 11 for the methods based on within-cells mean and median imputation.

Imputation cells have been defined by combining the Economic activity (2 or 3 digits), a variable representing the sign of CS, and the Legal form. The analyses have been conducted for each Nace division independently and, if where possible, also detailed for the Nace groups (3 digits level).

Based on these results, we can observe that the methods performing better are robust approaches (within-cells regression and within-cells median imputation), which explicitly take advantage of the correlation between Csfp and CS while reducing the influence of anomalous behaviours on missing data predictions. Note that the regression parameter estimate $\hat{\beta}_{CS}$ is significant at p=0,001 in all the imputation domains.

Unsatisfactory results can be observed for robust regression for a number of Nace groups (e.g. 527 and 552), while within-cells median imputation shows a good performance in almost all domains.

As a general conclusion, these preliminary results show that, with particular reference to within-cells median imputation, under this scenario encouraging results can be obtained in terms of possibility of estimating changes in stocks components using partially available administrative data. However, further investigations and analyses are needed to verify the actual usability of unit-level imputation, both in the problematic domains in the selected divisions (especially when the estimation detail increases), and in other divisions. Taking into account the complex nature of the investigated variables, deep analyses are necessary in order to assess the possible influence on the level of the discrepancies of non-statistical reasons, like legal issues and specific economic behaviours.

Table 6 - NND quality indicators by domain (2 digits Nace) and imputation cells

2 DIGITS NACE	3 digits Nace + CS sign		3 digits Nace+Legal form+CS sign	
	RB	RMSE	RB	RMSE
17	0.05	0.06	0.05	0.06
52	0.09	0.16	0.06	0.10
55	0.08	0.11	0.11	0.16

Table 7 - NND quality indicators by domain (3 digits Nace) and imputation cells

3 DIGITS NACE	3 digits Nace + CS sign		3 digits Nace+Legal form+ CS sign	
	RB	RMSE	RB	RMSE
171	-0,11	0.13	-0.26	0.45
172	0,09	0.12	0.06	0.09
173	0.03	0.06	0.07	0.19
174	0.08	0.10	0.07	0.11
175	0,06	0.09	0.05	0.10
176	1,06	2.10	0.48	0.70
177	0,14	0.27	0.08	0.13
521	0,11	0.15	0.12	0.16
522	0.08	0.15	0.14	0.23
523	0.12	0.17	0.11	0.16
524	0,11	0.21	0.08	0.12
525	0.28	0.66	0.34	0.48
526	0.27	0.64	0.10	0.14
527	0.12	0.20	0.18	0.31
551	-0.21	0.45	-0.36	1.06
552	-0.19	0.44	-0.15	0.36
553	0.13	0.21	0.21	0.43
554	0.10	0.13	0.11	0.17
555	0.50	0.96	0.22	0.37

Table 8 - Robust regression quality indicators by domain (2 digits Nace) and imputation cells

2 DIGITS NACE	2 digits Nace		2 digits Nace+Legal form	
	RB	RMSE	RB	RMSE
17	0.02	0.04	0.02	0.04
52	0.03	0.04	0.03	0.03
55	0.04	0.04	0.01	0.02

Table 9 - Robust regression quality indicators by domain (3 digits Nace) and imputation cells

3 DIGITS NACE	2 digits Nace		2 digits Nace+Legal form	
	RB	RMSE	RB	RMSE
171	-0.07	0.10	-0.04	0.09
172	0.04	0.06	0.02	0.04
173	-0.00	1.21	0.22	0.77
174	0.03	0.06	0.01	0.03
175	0.02	0.03	0.002	0.004
176	-0.17	5.05	-0.21	1.12
177	0.04	0.07	0.02	0.04
521	0.04	0.05	0.03	0.05
522	0.16	0.35	0.14	0.52
523	0.05	0.06	0.04	0.06
524	0.03	0.03	0.02	0.03
525	0.21	1.63	0.13	1.49
526	0.14	0.18	0.07	0.09
527	3.33	15.72	2.44	20.62
551	-0.01	0.38	0.07	0.69
552	2.65	5.55	2.49	14.11
553	0.05	0.08	0.02	0.03
554	0.02	0.04	0.01	0.02
555	0.12	0.79	0.10	0.49

Table 10 - Mean/Median quality indicators by domain (2 digits Nace) and imputation cells

METHOD	2 digits Nace	3 digits Nace+Legal form+Size+*CS* sign	
		RB	RMSE
Mean	17	0.089	0.122
	52	0.400	0.722
	55	0.032	0.056
Median	17	0.024	0.032
	52	0.013	0.020
	55	0.001	0.003

Table 11 - Mean/Median quality indicators by domain (3 digits Nace) and imputation cells (3 digits Nace+Legal form+Size+CS sign)

3 DIGITS NACE	Method			
	Mean		Median	
	RB	RMSE	RB	RMSE
171	-0.151	0.281	-0.040	0.076
172	0.062	0.098	0.026	0.059
173	0.023	0.163	0.004	0.017
174	0.123	0.202	0.037	0.068
175	0.101	0.223	0.021	0.039
176	0.055	2.232	0.328	5.107
177	0.075	0.127	0.050	0.106
521	0.122	0.193	0.010	0.061
522	0.275	0.817	0.123	0.492
523	0.399	0.800	0.050	0.521
524	0.148	0.183	0.022	0.054

Table 11 Continued - **Mean/Median quality indicators by domain (3 digits Nace) and imputation cells (3 digits Nace+Legal form+Size+CS sign)**

3 DIGITS NACE	Method			
	Mean		Median	
	RB	RMSE	RB	RMSE
525	0.023	7.243	0.040	0.117
526	4.698	13.770	0.093	0.400
527	0.224	1.130	0.009	0.084
551	0.020	0.159	-0.042	2.433
552	-0.001	0.013	-0.000	0.558
553	0.051	0.121	0.014	0.050
554	0.039	0.088	0.000	0.068
555	0.015	0.697	-0.034	0.102

4.2 Scenario 2: non informative administrative data

In this case, sub-populations which are not covered by any of the available administrative sources are considered: under this scenario, our aim is to assess the statistical effects of estimating the target variables by exploiting the auxiliary information coming from units belonging to other SME subpopulations. We focus the attention on the sub-set of Minimum Tax Payers belonging to the three selected divisions 17 (86 units, 6% of the responding units in the division), 52 (307 units, 11% of the responding units in the division), and 55 (84 units, 8% of the responding units in the division). The imputation models introduced in section 3.2.2 have been. In the following, some of the obtained results are shown. In order to define a model based estimation framework for Minimum Tax Payers, possible relationships with potential covariates have been investigated first. In Table 12, correlations between the changes in stocks target variables and the assumed most promising covariates (using the complete BS database) are reported for 2007. It is expected that the level of correlations found is very poor, nevertheless for the highest values of the index, correlations have been investigated for a more detailed Nace code (see Table 13). Correlations do not appear to be stable over the two years, 2006 and 2007, and therefore it seems that a reliable relationship does not exist. Afterwards, some Δ-variables (computed as the difference between the value observed in 2007 and in 2006) have been investigated as potential covariate, the underlying idea being that, for example, increasing the Turnover results in reducing the stocks of finished products and vice versa. Even in this case, the values of the correlation index (see Table 14) do not support the hypothesis of a good predictive model for the target variables.

Table 12 - Correlations between the variables CS, Csfp, Csrm and some possibly influential variables by Nace code. Year 2007; source BS

NACE		Production Value	Turnover	Costs	Purchase
17	*CS*	0,16	0,14	0,14	0,21
	Csfp	0,26	0,22	0,24	0,23
	Csrm	0,00	0,01	0,02	-0,08
52	*CS*	0,42	0,43	0,42	0,43
	Csfp	-0,06	-0,08	-0,06	-0,07
	Csrm	-0,45	-0,46	-0,45	-0,47
55	*CS*	0,02	0,02	0,02	0,02
	Csfp	-0,01	-0,01	0,00	0,00
	Csrm	-0,08	-0,08	-0,08	-0,08

Table 13 - Correlations between the variables CS, Csfp, Csrm and some possibly influential variables for Nace code 521. Year 2007; source BS

NACE		Production Value	Turnover	Costs	Purchase
521 (YEAR 2007)	*CS*	0,60	0,60	0,59	0,61
	Csfp	-0,13	-0,15	-0,14	-0,14
	Csrm	-0,59	-0,60	-0,59	-0,60
521 (YEAR 2006)	*CS*	0,19	0,21	0,18	0,20
	Csfp	0,13	0,14	0,13	0,13
	Csrm	-0,15	-0,16	-0,13	-0,16

Table 14 - Correlations between the variables CS, Csfp, Csrm and some possibly influential Δ-variables (2007-2006) by Nace: source BS

NACE		Δ(Prod Value)	Δ(Turnover)	Δ(Costs)	Δ(Purchase)
17	*CS*	0,07	-0,07	-0,01	0,28
	Csfp	0,34	0,14	0,28	0,26
	Csrm	0,20	0,22	0,26	-0,16
52	*CS*	0,42	0,37	0,43	0,54
	Csfp	-0,02	-0,13	-0,02	-0,03
	Csrm	-0,43	-0,42	-0,44	-0,55
55	*CS*	0,08	0,00	0,09	0,08
	Csfp	0,04	-0,05	0,05	0,00
	Csrm	-0,16	-0,15	-0,15	-0,26

In Tables 15 and 16, the results obtained by applying Mass Imputation to the subpopulations of Minimum Tax Payers of the selected divisions are shown (the domain index "D" is omitted in the quality indicator's name). As for scenario 1, different experiments have been performed by changing the criteria used to form imputation cells for each evaluated method. In Tables 15 and 16, the "best" results in terms of the distance $Diff\_Var^D$ (Var=CS, Csfp, Csrm) introduced in section 3.2.2. are shown (the domain *D* is omitted for simplicity). Imputation cells correspond to the combination of 3 digits Nace, Legal form, and Size (size classes: [1-9], [10-19], [20-49], [50-99]).

As it can be seen, Mass Imputation provides encouraging results at both 2 and 3 Nace code digits for the three considered divisions. Exceptions are represented by some Nace groups (e.g. 525, 526, 527 and 554), however it is worth further analysing this kind of approach to better investigate its actual usability.

On the contrary, the application of robust regression in this case has provided very unsatisfactory results at both 2 and 3 digits Nace code levels, and for all the investigated forms of imputation cells. This fact can be considered highly depending on the low correlations existing among the variables on changes in stocks and the potential auxiliary variables available in the considered administrative data sources, which make difficult obtain statistically significant estimates of the (robust) regression models.

These preliminary results can be viewed in any case as a starting point encouraging further analysis of the problem and additional investigations involving alternative estimation approaches.

Table 15 - Mass Imputation: quality indicators by domain (2 digits Nace) and imputation cells

2 DIGITS NACE	3 digits Nace+Legal form+ *Size*		
	Diff_CS	Diff_Csfp	Diff_Csrm
17	0,076	0,008	0,001
52	0,001	0,005	0,008
55	0,035	0,148	0,102

Table 16 - Mass Imputation: quality indicators by domain (3 digits Nace) and imputation cells

3 DIGITS NACE	3 digits Nace+Legal form+ Size		
	Diff_CS	Diff_Csfp	Diff_Csrm
171	0,000	0,000	0,000
172	0,000	0,000	0,000
173	0,000	0,000	0,000
174	0,093	0,016	0,011
175	0,013	0,007	0,012
176	0,000	0,000	0,000
177	0,000	0,000	0,000
521	0,000	0,000	0,000
522	0,004	0,008	0,060
523	0,000	0,000	0,000
524	0,008	0,003	0,000
525	0,093	0,246	0,017
526	0,286	0,019	0,063
527	0,634	0,924	0,719
551	0,000	0,000	0,000
552	0,025	0,032	0,032
553	0,000	0,000	0,000
554	0,116	0,177	0,148
555	0,000	0,000	0,000

5. Final remarks

The aim of this paper is to illustrate the results of experimental studies aiming at investigating the possibility of estimating variables related to changes in stocks of goods and services which are not directly available from administrative sources. The target variables, despite possible differences in definitions (Eurostat, 1999), can be sometimes derived from the available administrative source, maybe not for all the target variables and not for the whole population depending on the country specific administrative rules. In this respect, Balance Sheets (BS) are the most common and prioritized as the 'best' administrative source for the target variables, although their information content it is not standardized across Countries.

Different 'informative' scenarios have been considered according to the coverage of the administrative sources in terms of both population units and variables.

A 'fully informative' scenario can be realistic for (subpopulations of) enterprises subject to filling in a BS, as in Italy. In this scenario the key target variables required by the SBS regulation are available from the BS, this holds at least for the variables: Changes in stocks of goods and services (CS), Changes in stocks of finished products and work in progress (Csfp) and Changes in stocks of raw materials and goods and services for resale (Csrm).

In addition, a 'partially informative' scenario has been considered, in which the variable Cs is available from administrative sources Csfp s and Csrm are not. Different strategies based on estimation/imputation methods can be followed. Tests have been performed on some representative Nace divisions. Results have been presented in section 4 for the methods introduced in section 3. As expected, given the good correlations existing among the available information on CS and the variables to be estimated, robust approaches using economic activity, legal form, turnover and size as auxiliary information seem to be appropriate in the most domains. In particular, within-cell median imputation results to be the best performing method with respect to the quality indicators used. However, further analyses are needed, taking into account the complex nature and behavior of the investigated variables.

Finally, a 'non informative' scenario is considered, in which any of the target variables are available from administrative sources for some specific subpopulations (in particular, for the so-called Minimum Tax Payers). In this situation, imputation models have been tested, too. Regression based models cannot be considered appropriate, especially because of the fact that suitable covariates to be effectively used in this kind of models cannot be found. However, methods which do not require an explicit modeling of data relationships, like donor-based Mass Imputation, have shown a better performance in terms of potential effects on estimates in the considered domains.

References

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

Casciano M.C., Cirianni A., De Giorgi V., Di Francescantonio T., Mazzilli A., Luzi O., Oropallo F., Rinaldi M., Santi E., Seri G., Siesto G. 2011. Utilizzo delle fonti amministrative nella rilevazione sulle piccole e medie imprese e sull'esercizio di arti e professioni. *Working Papers Istat N.7/2011*.

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

R. Benedetti, M. Bee, and G. Espa. 2010. A framework for cut-off sampling in business survey design. *Journal of Official Statistics*, 26(4):651–671.

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

Elswijk D. van, Elliott D., Redling B., Kavaliauskiene D., Luzi O., Seri G., Siesto G. 2010. Methods of estimation for business statistics variables that cannot be obtained from administrative data sources. *European Conference on Quality in Official Statistics (Q2010),* Helsinky, May 2010.

Eurostat. 1999. *Structural Business Statistics Regulation (SBSR) report on matching the definitions of SBSR variables with the definitions of the International Financial Reporting Standards*. http://circa.europa.eu/irc/dsis/accstat/info/data/en/SBSR.pdf.

Eurostat. 2007. *Recommended Practices for Editing and Imputation in Cross-Sectional Business Surveys*. A cura di Luzi O., Di Zio M., Guarnera U., Manzari A., De Waal T., Pannekoek J., Hoogland J., Templeman C., Hulliger B., Kilchman D. http://epp.eurostat.ec.europa.eu/portal/page/portal/quality/documents/RPM_EDIMBUS.pdf.

Kalton G., Kasprzyk D. 1986. The treatment of missing survey data. *Survey methodology*, 12, 1, Statistics Canada.

Knaub, J.R., Jr. 2008. Cutoff Sampling. In "*Encyclopedia of Survey Research Methods*", P.J. Lavrakas (ed.). London: Sage.

Istat (2011), Struttura e competitività delle imprese. http://www.istat.it/it/archivio/43673.

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

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

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

Schafer J. L. (1997). *Analysis of Incomplete Multivariate Data*. Chapman & Hall.

Statistics Canada. 1998. *Functional Description of the Generalized Edit and Imputation System*. Statistics Canada Technical Report.

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

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