rivista di statistica ufficiale

n. 3
2016

In this issue:

The EUREDIT project: activities and results (*Reprint*)
Giulio Barcaroli

A prediction approach for the estimation of hours worked using integrated register and survey data
Fabiana Rocci, Silvia Pacini, Laura Serbassi, Marina Sorrentino, Maria Carla Congia

Exploiting the integration of businesses micro-data sources
Giovanni Seri, Daniela Ichim, Valeria Mastrostefano, Alessandra Nurra

Istat
Istituto Nazionale di Statistica

rivista di statistica ufficiale

n. 3
2016

In this issue:

The EUREDIT project: activities and results (*Reprint*)
Giulio Barcaroli 7

A prediction approach for the estimation of hours worked using integrated register and survey data
Fabiana Rocci, Silvia Pacini, Laura Serbassi, Marina Sorrentino, Maria Carla Congia 41

Exploiting the integration of businesses micro-data sources
Giovanni Seri, Daniela Ichim, Valeria Mastrostefano, Alessandra Nurra 73

Istat

rivista di statistica ufficiale
n. 3/2016

ISSN 1828-1982

Istituto nazionale di statistica
Via Cesare Balbo, 16 – Roma

Editorial Preface

Starting from this issue of the *Rivista di statistica ufficiale* we take the opportunity to occasionally retrieve scientific articles of particular importance and extraordinarily topical that regularly underwent a reviewing process and were already released in the past.

This in order to make them available online again as they are no longer easily accessible.

For this reason, the present publication of the *Rivista di statistica ufficiale*, that includes two original scientific works, opens with the reprint of an article published in 2002 (n. 2/2002).

Written by Giulio Barcaroli, the article reports on the activities and results of the European project titled *Euredit*. The project was carried out between 2000 and 2003 within the *5th Framework Programme of European Research*, which aimed at developing and evaluating new methods for the editing and imputation of statistical data.

The main output of the project consists in a set of reports containing the description and the evaluation of these methods, together with guidelines illustrating the conditions for an optimal use of each of them. As additional results, a number of software tools were produced and made available to all the project contributors.

These new methods taken into consideration - dealing with neural networks, support vector machines, self-organising maps, etc. - are fully in line with the branch known today as *Machine Learning*.

This approach is at the basis of *Data Science* on which official statistics are investing a lot nowadays, with the purpose of systematically introducing it into their production processes.

Data Science represents now a real change of paradigm and a choice not taken for granted as far back as nearly 20 years ago.

The participation of the Italian National Institute of Statistics – Istat in the shared commitment of the *Euredit* project made it possible to create the basis for the enhancement of specific and needed professional skills. Later on these allowed, among other things, to carry out research and development activities in the use of the innovative sources represented by *Big Data*, with Istat assuming a leading role both at national and international level.

The other two articles contained in this issue of the *Rivista di statistica ufficiale* deal with the integration of different sources in the area of the pillar information system for the production of estimates on the economic accounts of businesses, called *FRAME-SBS*. The system was developed by Istat to be compliant with the European Commission Regulation on Structural Business Statistics.

More specifically, the article by Rocci, Pacini, Serbassi, Sorrentino and Congia focusses on an estimate of the hours worked for all businesses with employees in industry and services. The methodology applied is based on joint information from statistical registers and business surveys. The results of the analyses and the outline of the final methodology are well illustrated, identifying subpopulations of companies with significant characteristics.

Finally, the paper by Seri, Ichim, Mastrostefano and Nurra describes the methodological approach and the analyses performed in order to integrate the *FRAME-SBS* main information with the data of two sample surveys on structural business statistics, aiming at producing economic indicators by the exploitation of different interactions.

Nadia Mignolli
Coordinator of the Editorial board

The EUREDIT project: activities and results

Giulio Barcaroli [1]

Abstract

The EUREDIT project was carried out from 2000 to 2003 under the 5th Framework Program of European Research, with the aim of developing and evaluating new methods for data editing in official statistics, in particular with respect to the phases of (i) error localisation and (ii) imputation of errors and missing values. Multi-Layer Perceptrons, Self-Organising Maps, Correlation Matrix Memories and Support Vector Machines have been considered as new methods for error localisation and imputation, together with robust methods for outlier identification and treatment. In addition, standard methods (edit rules based and model-based) have been taken into consideration as a benchmark to evaluate the performance of the new methods. For this purpose, a set of performance indicators were defined, together with an experimental plan making use of different datasets, covering the different typologies of data currently treated in official statistical production processes. Results of experiments are reported and discussed, together with methodological indications on how to optimally conduct evaluation tasks of this kind.

Keywords: statistical data editing, error localisation, imputation, outliers detection, quality evaluation.

1 Istituto Nazionale di Statistica.

Introduction

The EUREDIT project ("The development and evaluation of new methods for editing and imputation") was carried out under the 5th Framework Program of European Research from March 2000 to February 2003. To the project participated six national institutes, four universities and two private firms: ISTAT[2], the UK Office for National Statistics (as coordinator of the project), Statistics Denmark, Statistics Netherlands, the Swiss Federal Statistical Office, Statistics Finland, the universities of Jyvaeskylae (Finland), Southampton, York, and Royal Holloway College (UK), Qantaris GmbH (Germany) and Numerical Algorithms Group (UK). Objective of the project was the application and evaluation of new approaches and algorithms to the problems of (i) error localisation in data, and (ii) imputation of errors and missing values. The most important new approach to be evaluated was identified basically in the family of techniques developed for pattern recognition and based, strictly or loosely, on artificial neural networks paradigm. But not only that: for particular problems, as robust estimation and treatment of time series, also ad hoc methods were ideated and evaluated.

For the evaluation of these new methods, it was necessary to consider and make available the following elements:

1. a set of standard methods, already defined and currently in use, whose performance might be considered as a benchmark for the evaluation of the new ones;
2. a conceptual framework (a set of indicators) for the compared evaluation of the quality of the different methods;
3. a set of different datasets, each of them representing a given typology of data, so as to cover the range of possible situations for a statistical user;
4. an experimental plan for the compared evaluation of all methods.

The basic output of the project is given by a set of reports containing the description and the evaluation of the methods, together with guidelines indicating the conditions for the optimal use of each of them. As additional output, a number of software tools were produced and made available to the project partners[3].

In paragraph 1, a description of *new methods* that were investigated is given. They have been subdivided in those belonging to the class of neural network methods, and those classifiable as robust methods. Methods peculiar for time series are directly dealt with in paragraph 5.

2 Members of the EUREDIT ISTAT project group are: Giulio Barcaroli, Giorgio Della Rocca, Marco Di Zio, Ugo Guarnera, Orietta Luzi, Antonia Manzari, Emanuela Scavalli, Angela Seeber.

3 The results of the project are fully described in deliverable 6.1 ("Methods and Experimental Results from the Euredit Project") and deliverable 6.2 ("Towards effective statistical editing and imputation strategies - Finding of the Euredit Project").

In paragraph 2, *standard methods* are described. Also these are distinguished in two classes, accordingly to their belonging to the Fellegi-Holt family (rules based methods), or to the model-based methods family.

In paragraph 3, the *criteria for the evaluation* of error localisation and imputation performance are introduced.

In paragraph 4 a description is given of the six *datasets* chosen for experimenting the various methods, together with the planned experiments.

In paragraph 5 *results of experiments* are analysed and evaluated for each dataset and, more synthetically, for best methods.

1. The new methods

As already said, we can distinguish two main classes of new methods: those belonging to the wide class of pattern recognition computer intensive methods (mainly *neural networks* methods), and the others related to the particular problem of *outlier* detection and *robust imputation*.

1.1 Neural network methods

In general, a neural network is composed by a set of elementary units (neurones) linked by weighted connections (Bishop 1995, Ripley 1996). Neurons are organised in layers: at least one input layer and one output layer must be present in a net. One or more hidden layers are optional. Weights are determined by using training datasets, in case of *supervised* methods, or on the basis of available data in case of *unsupervised* methods.

Apart from any other possible characterisations, an important feature of neural networks is that they are *non-parametric* methods that can capture not only linear relationships between variables, but also *non-linear*. Another crucial element, very important in the phase of error localisation, is that they do not require the explicit knowledge represented by edit rules, but they rather need a set of cases (training datasets or available data) from which implicit knowledge required to operate is acquired: in other terms, these methods can *learn*.

This is the general approach. Actually, a variety of methods were considered in this class, each of them with relevant peculiarities. In the following, a synthetic description for each will be given.

1.1.1 Multi-Layer Perceptrons (MLPs)

A Multi-Layer Perceptron is a neural network characterised by at least one hidden *layer*. One layer is composed by elementary units called *neurons*, each neuron is linked to others neurons by weighted *connections*. For any given neuron x_j, the input is given by the weighted sum of the outputs of linked neurons, while its output is the result of the application of a non-linear function $f(x_j) = f(a_j + \sum_{i=1}^{k} w_{ij} y_i)$, where f is typically the *sigmoid* function (logistic or tangent hyperbolic).

Weights of the MLP, initially defined on a random basis, are sequentially adjusted by submitting a set of individual cases, with known values, for any of which predictions are made. The adjustment of weights, carried out so as to obtain best possible predictions, is based on different possible algorithms (all of the type *feed forward*), and proceeds until convergence, i.e. when the accuracy of predictions can no longer be significantly increased. To prevent over-fitting[4] and ensure generalisation, during this process a validation set is also used.

A very important aspect of MLP construction is in the choice of input variables to the network. Redundant information may produce noise that limits predictive capability of the net. A number of techniques to select relevant variables were defined and tested in EUREDIT project (Scavalli,2002).

The application of MLP's to the error localisation task can occur in two basic ways (Nordbotten 1995 and 1996):

- in a subset of cases where it is known when an error occurs in a given variable, and is therefore possible to define an *error flag*, a neural network is trained to predict the value (0 or 1) of the error flag. When applied to the complete set of cases, the neural net outputs (or in other terms, *predicts*) the values of the error flag. Values closest to '1' indicate the presence of errors. It is necessary to define a threshold value above which corresponding values can be judged as erroneous: this is generally done by minimising the total amount of misclassifications (false positive and false negatives);
- in a subset of cases that can be reasonably judged as "error free", a neural network is trained for each variable, so as to predict its values. When the neural net is applied to the complete set of cases, predicted values of each variable become available. The distance between current value and predicted value is an indicator of the presence of errors. Also in this case, a threshold value should be defined in order to assess when a value is erroneous or not.

4 When over-fitting occurs, a solution is found that minimises errors in prediction or classification of training data, but does not perform well on other datasets, i.e. it lacks in generalisation.

With regard to the imputation task, a straightforward solution is the following: for each variable, the subset of cases with no missing errors are considered, and in this subset a neural network is trained to predict values for that variable. The neural net is then applied to the subset of cases with missing values for that variable, and predicted values are imputed to the variable.

This solution is acceptable when the missing mechanism is judged to be MCAR (missing completely at random) or at least MAR (missing at random). In case of NMAR (not missing at random) a different solution should be followed, based on the availability of a subset of cases in which, in correspondence of each missing value for a given variable, also its true value is available.

1.1.2 Tree-Structured Self-Organising Maps (TS-SOMs)

A Self-Organising Map (SOM) is a neural network that approximates a first principal curve, that is a low-dimensional representation (typically one or two dimensions) of a multivariate distribution (Kohonen1997).

The Tree-Structured Self-Organising Map (TS-SOM) algorithm combines the representation capability of the SOM and a tree-search of the best matching unit (Koikkalainen 1999).

When training TS-SOM, several SOMs with different resolution (i.e. with a different number of nodes or neurons or data clusters) are trained and are organised in a tree structure, starting from the simplest SOM at the root, ending to the most complex SOMs at the leaves. The more complex is a SOM, the higher is its capability to represent non-linear relationships in data.

As SOMs are unsupervised neural networks, the training does not require the availability of a subset of true data (as in the case of MLPs). To train a TS-SOM it is necessary to define the following parameters:

- the number of layers in the tree: this parameter defines the complexity of the net: the higher the number of the layers, the higher the complexity of the SOMs in the final layers;
- the robustness of the training algorithm: for continuous variables, the observations that are k times the value of the standard deviation in the nodes are considered to be outliers, while for categorical variables a “cut probability” is defined in order to mark the observation as an outlier.

Once a TS-SOM has been trained, it is possible to use it to localise errors in data. This is done by (i) searching in the tree the best matching SOM for the current unit, and (ii) by considering the differences between the SOM model projections and the observed values. Potential errors are those that show the largest differences. A threshold is defined in order to choose actual errors.

To perform imputation, TS-SOM can be used in a similar way. For any observation with missing value, the best SOM is searched in the tree, and a conditional distribution is therefore available for the missing values. There are different possible imputation procedures:

- the mean in the cluster;
- a random draw from a probability density function;
- a random donor;
- a nearest neighbour donor;
- a MLP regression model specific for the node.

The differences between observations and predicted values are computed in terms of Euclidean distance between vectors of values. This requires that data are preventively pre-processed in order to perform equalisation of variable ranges, normalisation of scales, log-transformations and dummy coding of categorical variables.

1.1.3 Correlation Matrix Memories (CMMs)

A Correlation Matrix Memory is a particular type of neural network that is trained to associate pairs of patterns (an input pattern and an output pattern). It requires only a single cycle through the training data in order to learn the association of a pair of patterns, while the majority of neural networks require many training cycles, necessary to fit a non linear-regression model to data, where also implicit relationships between variables are represented. CMM create an explicit associative mapping between input and output patterns, instead of regression-type models.

The use of CMM for error localisation involves the following steps. First, a pre-processing of data provides to convert them into binary format. A training of a CMM using the resulting binary representation of every record of data is performed. Then, for each record in the dataset, the trained CMM is applied to find the j best matches. Similarities between records (or "patterns") are determined by considering their Hamming distance. For these j best matches, the k-NN ("k nearest neighbours") subset is considered, and a DKN (the distance from the record from its k-th neighbour) value is computed in the following way:

- for each of the j matched records (neighbours), the Euclidean distance between them and the current record is computed;
- neighbours are sorted accordingly to their Euclidean distance;
- the DKN is retained for the current record.

All records are sorted accordingly to their DKN values. Given a threshold cut-off distance, all records exceeding this cut-off are considered as erroneous. The error-status of each variable in an erroneous record is determined on the basis of individual contribution to the DKN value.

The use of CMM for imputation is quite straightforward: once the k-NN subset for a given record with missing values (or variables flagged for imputation), has been determined the values to be imputed are determined by using one of five possible methods: nearest-neighbour, random donor, median, mean and weighted mean.

In other words, both for error localisation and imputation, CMM is used only in a first step in order to find a set of closest records, that are used differently accordingly to the specific imputation method.

1.1.4 Support Vector Machines (SVMs)

A Support Vector Machine is an algorithm for defining a smoothing function that predicts the values of a set of target variables from a set of explanatory variables (Vapnik 1995). There are two forms of SVM, one for the prediction of continuous variables (SVM for regression, or SVR), the other for binary categorical variables (SVM for classification), both able to learn non-linear functions from data. SVM, originated in the so-called "machine-learning community", can be grouped with other semi-parametric approaches like Multi-Layer Perceptrons and Radial-Basis Functions: semi-parametric in the sense that they offer the efficient training characteristics of parametric techniques, but have the capability to learn non-linear dependencies as non-parametric methods do.

Another possible definition of SVM is the following: a non-linear generalisation of linear techniques (Cristianini and Shawe-Taylor 2000). Covariate data is projected onto a higher dimensional space ("features space"), and then inserted in a linear algorithm: the parameters of the linear model learned in the higher-dimensional space describe a non-linear model in the original space. The advantage of this approach is that the objective function minimised during training is convex quadratic, and therefore the problem of local minima is avoided. SVM learning also avoids over-fitting by introducing a penalisation factor (regularisation) of over-complex models.

1.2 Robust methods

The problem of outliers identification and subsequent treatment is very important, especially in business data. Unidentified outliers can seriously compromise the accuracy of estimates and the validity of standard analyses of data. Not all outliers are errors: they can be characterised as *representative* outliers (corrects values) or *non representative* outliers (errors) (Chambers, 1986). The treatment subsequent to the identification should take into account this distinction, that conversely is not important in the phase of their detection: even if an outlier is not an error, it is nonetheless crucial to detect it in order to give it a special treatment.

Detection requires first of all a metric able to measure the "outlyingness" of a value. Metrics are usually derived by the adoption of models and measures of the discrepancy between real and predicted values (Barnett and Lewis 1994). A very common metric for continuous data is the Mahalanobis distance.

A problem to be dealt with is that the estimation of model parameters can be influenced by those outliers that should be detected by using the model. Robust methods for outlier detection are based in turn on robust estimation of models and distances.

In the following, a number of methods for outlier detection are synthetically illustrated. As for imputation, most of them make use of a particular software, POEM (*imPutation for Outliers, Edit failures and Missing values*), that is a robust nearest neighbour imputation algorithm, while the last one, WAID, has an embedded function that allows not only to detect but also to impute outliers.

1.2.1 Outlier detection: Robust distance via Transformed Rank Correlation (TRC)

The basic idea of Transformed Rank Correlations (Gnanadesikan and Kettenring 1972) is to compose a pseudo covariance matrix $\widetilde{\mathbf{S}}$ using robust bivariate covariances.

This matrix is built by using the standardised Spearman rank correlation, multiplied by the standardised median absolute deviation of the variables involved. Data are then transformed into the space of principle axis derived from the pseudo covariance matrix: the transformation matrix **B** is defined by the equation $\widetilde{\mathbf{S}} = \mathbf{B}\,\mathbf{\Lambda}\,\mathbf{B}^{T}$, with **B** orthogonal and Λ diagonal. The matrix of data **X** is transformed into **Y=XB**, and medians m' and median absolute deviations s' are transformed back to m=**B**m' and **B** diag(s') $\mathbf{B}^{T}$. Finally, the Mahalanobis distance $d^2 = \mathbf{(x_i - m)^T S(x_i - m)}$ is computed for each point, and its ratio to the median of Mahalanobis distances is compared to an F distribution to determine outliers.

1.2.2 Outlier detection: Forward Search Algorithms (BACON)

Forward Search Algorithms start from an initial subset of data that is judged as being free of outliers (Hadi and Simonoff 1993; Riani and Atkinson 2000). In the case of BACON algorithm ("*Blocked Adaptive Computationally Efficient Outlier Nominators*") (Billor et al 2000), this subset of data, of dimension cp, where p is the dimension of data (number of variables) and c is a constant chosen by the statistician (usually c=3), can be built in two different ways: by considering (i) observations with the smallest Mahalanobis distance to mean, or (ii) observations with the smallest Euclidean distance to the median point in the space; of course, the second alternative offers a more robust starting point. This initial subset is used to estimate a regression

model for the variables of interest. The algorithm then calculates Mahalanobis distances for all observations, based on mean and covariance estimated by the model. The next step is to redefine the "clean" subset by considering these new Mahalanobis distances. The procedure is iterated until (i) distances of observation outside the clean set are too large and the clean set does not vary anymore, or (ii) all observations are inside the clean set.

The BACON algorithm has been adapted in order to take into account missing values: in each iteration, the EM algorithm is applied before BACON, under the assumption of a multivariate normal distribution.

1.2.3 Outlier detection: Epidemic Algorithm (EA)

The Epidemic Algorithm (Beguin and Hulliger 2003) simulates an epidemic whose starting point is in the centre of a data scatter and spreads from it stepwise. As first step, Euclidean distances between all observations are calculated. The centre of the distribution (*sample spatial median*) is the point that has the least sum of distances from all other points. The epidemic is such that the probability that an "infected" point i transmits the "disease" to a non infected point j at the next step is inversely proportional to the distance:

$$P(j|i) = h(d_{ij})$$

where d_{ij} is the distance between observations i and j, and h varies from 1 to 0). The form of the transmission function determines the behaviour of the algorithm: for the EUREDIT project, the *inverse power function* has been chosen:

$$h(d) = ((\alpha^{-\frac{1}{p}} - 1)\frac{d}{d_0 + 1})^{-p} \quad \{d \le d_0\}$$

where the constant d_0 is called *reach of the transmission*, and can be determined over the observations as the maximum of the distances to the nearest neighbour. Given a subset I of infected points, the *total infection probability* of an observation j is $P(j|I) = 1 - \prod_{i=1}^{I} (1 - h(d_{ij}))$.

The algorithm starts at the sample spatial median, and at each step the total infection probability of each uninfected point is evaluated. The expected number of new infected points is calculated, and points with the largest total infection probabilities are infected. The algorithm stops when no new infection occurs, or there are no more uninfected points. Infection times of observations are taken as a measure of outlyingness, and simple univariate decisions can be applied to identify outliers. Doubtless outliers are those observations that have never been infected.

1.2.4 Outlier Reverse Calibration Imputation

We assume that a reliable population total estimate $\hat{t}_y = \sum_{i \in s} w_i^* Y_i$ can be obtained by some outlier-resistant procedure: here, weights w_i^* are weights w_i (inverses of inclusion probabilities, or calibration weights) that have been corrected in such a way so to ensure this resistance. We can also say that $\hat{t}_y = \sum_{i \in s} w_i^* Y_i = \sum_{i \in s} w_i Y_i^*$.

In other terms, we can obtain the same outlier-resistant estimate by imputing values instead of modifying weights. Once outliers have been detected, let s_2 be the subsample of outliers, and s_1 the subsample of inliers. Then, the problem is to define a set of imputed values $(Y_i^*; i \in s_2)$ under the constraint:

$$\hat{t}_y - \hat{t}_{1y} = \hat{t}_y - \sum_{i \in s1} w_i Y_i = \sum_{i \in s2} w_i Y_i^*$$

The imputed values $Y_i^*; i \in {}_2$ should remain as close as possible to the original ones, subject to this constraints. This problem is equivalent to the calibration problem, where the survey variable Y plays the role of the sample weight and the sample weight plays the role of the survey variable. The distance here considered is:

$$d(Y^*, Y) = \sum_{i \in s2} (Y_i^* - Y_i)^2 / 2 q_i Y_i$$

where the q_i 's are constants that are chosen by the statistician. So, it follows (Deville and Sarndal, 1992):

$$Y_i^* = Y_i \left[1 + q_i w_i \frac{\hat{t}_{2y} - \sum_{j \in s2} w_j Y_j}{\sum_{j \in s2} q_j w_j^2 Y_j} \right]$$

1.2.5 Outlier detection and imputation: Robust Tree Modelling (WAID)

In the EUREDIT project, the regression tree modelling software WAID has been used[5]. The basic idea of regression tree models (Breiman *et al* 1984) is to sequentially divide the dataset into subgroups (nodes) that are more and more homogenous with respect to the values of a response variable.

5 WAID regression tree modelling software operates under R (a public domain statistical software) and is an extension of WAID software for missing data imputation developed under the AUTIMP project (Chambers *et al* 2001).

The *univariate* version of WAID allows only one continuous response variable Y and p covariates X_1, ..., X_p, all categorical. The tree modelling is robust in the sense that outliers are down-weighted when the measure of each internal node heterogeneity is calculated: weights are in this case based on *outlier robust influence functions*.

At any iteration, each node k is evaluated in order to decide if it should be split, on the basis of a measure of the heterogeneity given by the weighted sum of square residuals:

$$WSSR_k = \sum_{i=1}^{n_k} w_i (y_i - \bar{y}_{wk})^2$$

where $\bar{y}_{wk}$ is the weighted mean of Y in node k, and the weight w_i is calculated as the ratio $w_i = \frac{\psi(y_i - \bar{y}_{wk})}{(y - \bar{y})}$, where $\psi(x)$ is a given influence function, whose default is the *bi-weight influence function*:

$$\psi(t) = t(1 - (t^2/c^2))^2$$

So, each time a current node is split to create two children nodes, a new set of robust weights is created: outliers receive weights close to zero, while inliers receive weights close to one.

The algorithm defines as outliers those observations that in the overall process of nodes splitting are characterised by an average weight below a specified threshold. The optimal threshold is the one that allows to identify successfully outliers minimising the number of misclassifications.

The only difference of the *multivariate* version of WAID is in the evaluation of the heterogeneity. In particular, one of the possible options defines the weight associated to observation i in candidate node h at stage k as

$$w_i^{(k)} = \frac{\psi\left(\left\| \mathbf{y}_i - \bar{\mathbf{y}}_{wh}^{(k)} \right\|_{wh}\right)}{\left\| \mathbf{y}_i - \bar{\mathbf{y}}_{wh}^{(k)} \right\|_{wh}}$$

where $\mathbf{y}_i$ is the p-vector of response values, $\bar{\mathbf{y}}_{wh}^{(k)}$ is the p-vector of means, and

$$\left\| \mathbf{y}_i - \bar{\mathbf{y}}_{wh}^{(k)} \right\|_{wh} = \sqrt{\sum_{j=1}^{p} s_{whj}^{-2} (y_{ij} - \bar{y}_{whj}^{(k)})^2}$$

where s_{whj}^{-2} is the variance within the candidate node h. Of course, in this case weights

have to be calculated iteratively.

Once a subset of observations have been declared as outliers, the robust tree structure generated by WAID can be used to impute them. There are two possible alternatives: (i) the outlier value is replaced by the weighted mean of the terminal node to which the observation belongs, or (ii) a random donor inside the terminal node is searched for.

2. Standard methods

Standard methods have been considered in the EUREDIT project so as to offer a benchmark for the evaluation of new methods.

These standard methods can be grouped in two different classes:

- methods that are *edit-rule based* and, more specifically, follow the optimal editing approach defined by Fellegi and Holt (Fellegi and Holt, 1976);
- methods that are *model-based*.

2.1 Fellegi-Holt methods (F-H)

These methods are currently being used by a variety of National Statistical Institutes. The set of edit rules is used both for error localisation and for imputation. For error localisation, the subset of edits activated by a given record is processed in order to individuate the subset of variables most likely to contain the errors that caused the activation of those edits. The F-H error localisation algorithm is based on the *minimum change principle*, i.e. the number of variables judged to be erroneous must be the minimum under the constraint to explain all edit failures. A variant to this approach is given by the Nearest-neighbour Imputation Methodology (NIM). Accordingly to NIM, the error localisation is no longer based on the minimum change principle, but on the consideration of the differences between the current record (with edit failures) and a potential donor (a neighbour with no edit failures): this approach can be defined as *data driven*, while the F-H methodology is purely *edits driven*.

For imputation, a range of possible values to impute is first determined in order to avoid values that might cause additional failures of edit rules; then, actual values can be assigned by using a number of different methods, from nearest neighbour to regression imputation. In particular, we include in this category the imputation methods based on the *donor search*, as opposite to regression imputation considered in the model-based methods.

A number of systems incorporating F-H methods have been developed by Statistics Canada, Statistics Netherlands, ISTAT and ONS, and applied in the

EUREDIT project. In the following, a short description for each will be given.

2.1.1 CANCEIS and SCIA for editing and imputation of categorical variables

The CANadian Census Edit and Imputation System (CANCEIS) has been developed by Statistics Canada to be applied to the last Population Census. It fully incorporates the Nearest-neighbour Imputation Methodology (NIM) (Bankier *et al* 2000).

The basic steps of NIM is (i) to search, for each record with edit failures, a set of *nearest neighbours* and, (ii) for each couple recipient-donor, to calculate the minimum number of *imputation actions*, so as to let the recipient failing no edits. As already said, this approach is not strictly adherent to the minimum change principle that characterises the Fellegi-Holt methodology, but has a number of advantages that makes it preferable in some applications. One of them is the editing and imputations of complex hierarchical structures, such as households. NIM allows to consider an entire household as the record to be edited, and experiences carried out made it clear that its performance is higher than that of pure F-H systems or other systems. The NIM approach also allows to handle contemporarily both continuous and categorical data, but so far the only applications we know refer to households categorical data, namely the variables that are linked by constraints that involve more than one member of the household.

On the contrary, SCIA (*Sistema per il Controllo e l'Imputazione Automatici*), developed by the Italian Statistical Institute, is a pure Fellegi-Holt system for the edit and imputation of categorical data (Riccini, 2002). Initially, the set of edit rules defined by the user is analysed and checked for contradictions and redundancies, and the complete set of rules, including implicit edits, is generated. These are applied to each record, and for those failing at least one edit, the minimal set of variables to be changed is determined, on the basis of the coverage of failed edits. Range of acceptable values are also determined for each variable. Then, the imputation step is performed, by searching first a unique donor for all imputations, on the basis of the values of the matching variables. If no such donor can be found, a sequential imputation is tried (one donor for each variable to be imputed). The final option is to impute values on the basis of the marginal distributions.

Experience showed that SCIA performs well for variables that are not subject to hierarchical constraints. Then, a typical edit and imputation application concerning a survey on households will consist firstly of an application of CANCEIS to variables whose edit rules mainly refer to the household composition and constraints (relation to head, sex, marital status and age), and secondly of an application of SCIA involving only individual variables (for instance, level of instruction, social condition, etc.) (Manzari, 2002).

2.1.2 GEIS for editing and imputation of continuous variables

The Generalised Edit and Imputation System (GEIS), developed by Statistics Canada (Kovar *et al* 1988), allows to apply the Fellegi-Holt methodology to continuous data. Only linear edits on non-negative variables are admissible. GEIS enables the user to analyse initial edits, identifying inconsistencies and redundancies. Error localisation is carried out on the basis of the minimum change principle: as in the case of categorical variables, for each record with edit failures the minimum set of variables covering all failed edits is identified and flagged for imputation. It is also possible to apply methods, as the Hidiroglou-Berthelot procedure (Hidiroglou and Berthelot, 1986), for outlier detection (Di Zio *et al* 2002a).

Imputation can be carried out in three different ways (Di Zio *et al* 2002b):

i. *deterministic* imputation, when for a given variable there exists one and only value that once assigned to the variable allows the record to pass the edits;
ii. *nearest neighbour* imputation: among all the units passing the edits, a potential donor with minimum distance is searched and its values, if acceptable, assigned to the recipient variables that require imputation;
iii. *estimated value imputation*: variables are imputed sequentially by using estimates based on different functions (means, ratios, historical trends).

2.1.3 CHERRY-PIE and E-C system for editing and imputation of continuous variables

CHERRY-PIE is another implementation of Fellegi-Holt methodology, that allows the user to handle jointly both categorical and continuous data (De Waal 2002). The output of CHERRY-PIE for each record that fails at least one edit is the list of variables that must be imputed as they have been flagged as erroneous.

The user can adopt whatever imputation method. In EUREDIT experiments a number of them have been used:

- *deductive* imputation (analogous to the GEIS deterministic imputation);
- *multivariate simultaneous regression* imputation: a multivariate regression model is estimated using fully observed predictors, and its predicted values assigned to missing/erroneous values;
- *ratio hot-deck* imputation: in case of balance edits, where many variables are sub-totals referred to a total, regression imputation is not adequate, since imputed variables are never zero and can be also negative; it is therefore better first to impute (by regression or deductively) the total, and then to search a donor (nearest neighbour with respect to the total), and allocate the differences between the variable total and the computed total (as sum of subtotals), by using ratios of subtotals to total in the donor (Pannekoek 2002).

The imputations carried out as outlined above, can lead to additional edit failures, because these imputation methods do not take into account edits. A particular procedure is available, the EC System, that allows to adjust the final values in order to satisfy all rules. Adjustments are made by using the *simplex method*, so as to minimise the distance between imputed and final values, under the constraint that final values satisfy edits.

2.1.4 DIS for imputation of continuous and categorical variables

The Donor Imputation System (DIS) has been developed by the Office for National Statistics to be used in the 2001 UK Censuses. It implements the joint imputation method proposed by Fellegi and Holt in 1976. DIS searches for a donor in three different stages. First, a donor is searched having the same values of the recipient on a set of matching variables (exact match). If no such donor can be found, then categories of each categorical matching variable are collapsed, and the search is repeated. If a donor still cannot be found, less significant matching variables are removed until at least one donor is found. If more than one donor is found, a random selection can be performed. A penalty function is applied in order to avoid imputations of the same donor to many recipients (Yar 1988).

2.2 Model-based methods

The basic idea is to define and fit a (parametric and linear) model for every variable involved in the process of edit and imputation. This model will be used both for error localisation and imputation.

Error localisation is carried out with the following steps:

1. for each variable, an expected value is calculated, conditional on a set of covariates;
2. the actual value is compared to the expected value, and if the two values diverge too much, the actual value can be considered erroneous.

Obviously, problems arise when adopting this approach. Firstly, also covariates can contain errors (or missing values). Secondly, what metric should be adopted in evaluating closeness of actual and expected values, and how to define thresholds beyond which data have to be considered as errors?

As for the imputation, on the basis of a given model the expected value is assigned to missing and erroneous data. Also in this case we have to deal with some problems. First, as in the case of error localisation, we should consider the possibility that covariates may contain errors: if so, also the predicted value will be different from the true one. Second, imputation can be *deterministic* (the predicted value is directly imputed), and in this case first order estimates are generally best preserved,

but further data analysis can be biased by a reduced variability; or imputation can be *stochastic* (the imputed value is drawn by from a conditional distribution), with a reduced preservation of means and totals. Third, imputations carried out in this way generally do not take into account the coherence of imputed values with other values in the record, and edit failures are therefore possible after imputation.

2.2.1 Expectation-Maximisation Algorithm (EM)

EM algorithm is a method for estimating distribution parameters in the presence of missing data, under a specified super-population model and an ignorable non-response mechanism (Dempster *et al* 1977) .

In the presence of missing data the complete data score function, i.e. the first derivative of the logarithm of $L(\theta | Y)$, is not easily computable, so an iterative algorithm is preferred to the analytical solution. The algorithm consists in repeatedly applying standard complete data methods to incomplete data, by iterating the following steps:

1. impute missing data Y_{miss} using current estimates of unknown parameter θ (*expectation* step);
2. re-estimate θ using Y_{obs} and imputed Y_{miss} (*maximisation* step).

The procedure is iterated until convergence to the unique maximum-likelihood estimate of θ.

Two methods of imputation can be used:

- each missing value is imputed with its best prediction $E(Y_{miss}|Y_{obs}, \hat{\theta})$ (the conditional expectation given the observed data and the current estimates of the model parameters);
- the imputation is carried out by drawing randomly from the conditional distribution of missing data given the observed data $P(Y_{miss}|Y_{obs}, \hat{\theta})$.

The first method should be chosen if primary estimates of interest are total or means, while the second is preferable to preserve variability in data.

The convergence of EM algorithm is not ensured if the assumption of multi-normality does not hold, and also imputation is performed on the basis of a multi-normal model. So, real applications do require (i) analysis of data to individuate strata in which multi-normality assumption holds and (ii) transformations of variables (usually, logarithmic transformations).

2.2.2 Integrated Modelling Approach to Imputation (IMAI)

The IMAI approach has been developed at Statistics Finland, and can be used both for error localisation and for missing/erroneous data imputation. It is based on the following different steps:

1. selection of training data and auxiliary variables for any given variable of interest;
2. construction of an error localisation model for the prediction of an error indicator for any given variable of interest, and/or an imputation model for the direct prediction of variables of interest;
3. choice of the criteria for error localisation: in particular, it is necessary to decide a proper cut-off probability for errors;
4. choice of the criteria for data imputation: if the predicted value (with or without an error term) is directly used to impute, then the imputation method is *model-donor*; on the contrary, if the predicted value is used to find a nearest neighbour, the method is *real-donor* (Regression Based Nearest Neighbour, RBNN, see Laaksonen 2000).

3. The evaluation criteria

One of the most important objectives of the EUREDIT project was to individuate best methods for given typologies of data and errors. So, the determination of the evaluation criteria was a crucial task that engaged the first phase of the project.

Different sets of evaluation criteria were defined for error localisation and for imputation. All of them imply that knowledge concerning true values is entirely available. In other words, quality indicators, to be calculated, need to know the true value Y_{ij}^{*} of the j-th variable in the i-th unit in the dataset, the corresponding observed (or raw) value Y_{ij}, and the possibly imputed value. In the following we will introduce separately indicators for the evaluation of error localisation methods and indicators for the evaluation of imputation methods.

3.1 The evaluation criteria for error localisation

When considering an error localisation method, we are interested in evaluating two different performances, namely:

- the *efficient error detection*, i.e. the capability of a method to correctly classify errors and true values in data, or, conversely, its capability to minimise misclassifications (*false negatives*, errors judged as true values, and *false positives*, true values judged as errors);
- the *influential error detection*, i.e. the ability to detect the most influential errors, those with the highest impact on final estimates.

3.1.1 Efficient error detection

After the application of a given method for error localisation, for every variable j of interest in the dataset, the following table can be defined:

	$E_{ij} = 1$ (value judged as correct)	$E_{ij} = 0$ (value judged as erroneous)
$Y_{ij} = Y_{ij}^*$ (correct value)	n_{aj}	n_{bj}
$Y_{ij} \neq Y_{ij}^*$ (erroneous value)	n_{cj}	n_{dj}

It is evident that frequencies on the main diagonal refer to correct classifications, while in the other two cells misclassifications are contained.

We can define the following indicators:

$$\alpha_j = \frac{n_{cj}}{n_{cj} + n_{dj}} \qquad (1)$$

that is the *false negative rate*, i.e. the proportion of errors that have not been recognised as such by the method, and

$$\beta_j = \frac{n_{bj}}{n_{aj} + n_{bj}} \qquad (2)$$

that is the *false positive rate*, i.e. the proportion of true values that have been erroneously recognised as errors by the method.

Finally,

$$\delta_j = \frac{n_{bj} + n_{cj}}{n} \qquad (3)$$

is the *total misclassification rate*, i.e. an estimate of the probability of an incorrect outcome from the error localisation method.

3.1.2 Influential error detection

It is worth while to measure not only the efficiency of the error localisation method in finding errors, but also its capability to find *influential* errors, in other words the errors that more than others could influence the estimates of interest.

To measure this capability, we introduce the concept of *post-edited* value $\hat{Y}_{ij} = E_{ij} Y_{ij} + (1 - E_{ij}) Y^*_{ij}$. If the measured value Y_{ij} is erroneous, and the method can recognise it as an error, then the post-edited value is assumed to be set to the true value. On the contrary, if the method fails in recognising the error, the post-edited value remain erroneous.

For continuous variables an important quantity is $D_{ij} = \hat{Y}_{ij} - Y^*_{ij} = E_{ij}(Y_{ij} - Y^*_{ij})$, i.e. the difference between the post-edited value and the true value. A desirable property of an error localisation method is that the two distributions of true values and post-edited values are as close as possible.

To measure this closeness, we can define the *relative average error*:

$$RAE_j = \sum_{i=1}^{n} w_i D_{ij} / \sum_{i=1}^{n} w_i Y^*_{ij} \qquad (4)$$

that indicates the mean difference between undetected errors and true values.Values w_i indicates sampling weights, and are obviously used only in case of sample surveys. If variable j can assume also negative values, a more suitable indicator is the relative *root average square error*:

$$RRASE_j = \sqrt{\sum_{i=1}^{n} w_i D^2_{ij}} \Big/ \sum_{i=1}^{n} w_i Y^*_{ij} \qquad (5)$$

A useful measure of how much differences between undetected errors and true values are spread, is given by the *relative error range*:

$$RER_j = R_j (D)/IQ_j(Y^*) \qquad (6)$$

where $R_j(D)$ is the range (maximum - minimum) of the non-zero D_{ij} values, and $IQ_j (Y^*)$ is the inter-quartile distance of the true values.

For categorical (nominal or ordinal) variables, a different indicator has to be defined. Considering the joint distribution of post-edited and true values, we have to take into account the number of cases not lying in the principal diagonal (where $\hat{Y}_{ij} = a$ and $Y^*_{ij} = b$, with $a \neq b$), each of them with an associated distance d(a,b). In case of nominal variables, d(a,a)=0, and d(a,b)=1 for any a,b. In case of ordinal variables, d(a,b) is given by the number of categories that lie between a and b, plus one. So, we can define the *influential error detection performance for a categorical variable*:

$$DCAT_j = \frac{1}{n} \sum_{a=1}^{p_j} \sum_{b \neq a} d(a,b) \sum_{i \in j(ab)} w_i \qquad (7)$$

Another useful measure of the performance of an error localisation method refers to the impact of remaining errors in post-edited data to the variance of the estimator in a sample survey. We can estimate this variance by means of the jackknife formula:

$$v_w(Y) = \frac{1}{n(n-1)} \sum_{i=1}^{n} \left(\left\{ n \sum_{i=1}^{n} w_i Y_i - (n-1) \sum_{k \neq i}^{n} w_k^{(i)} Y_k \right\} - \sum_{i=1}^{n} w_i Y_i \right)^2 \quad (7)$$

where $w_k^{(i)} = w_k \left(\sum_{q=1}^{n} w_q / \sum_{q \neq i}^{n} w_q \right)$. In other words, variance is calculated from survey data each time excluding the i-th unit, and rescaling weights to take into account this exclusion.

Then the indicator

$$t_j = \sum_{i=1}^{n} w_i D_{ij} / \sqrt{v_w(D_j)} \quad (8)$$

is a standardised measure of the *effect of error localisation method on the variance of the estimator*. Values of t_j greater than 2 indicate a significant failure of the error localisation method.

Finally, we can compare the moments and the distributions of the *outlier-free* data value is retained, and values with corresponding moments and distributions of the true values, in order to evaluate the capability of a method to detect outliers. Remembering that $E_{ij} = 1$ if the $E_{ij} = 0$ otherwise, for positive continuous variables we can define the *absolute relative error for the k-Mean*:

$$AREmk_j = \left| \frac{\sum_{i=1}^{n} w_i E_{ij} Y_{ij}^k / \sum_{i=1}^{n} w_i E_{ij}}{\sum_{i=1}^{n} w_i Y_{ij}^{*k} / \sum_{i=1}^{n} w_i} \right| \quad (9)$$

where this indicator is typically calculated for k=1 and k=2 (to compare first and second moments of the two distributions).

3.2 The evaluation criteria for imputation

An imputation procedure should be evaluated with respect to the following properties:

i. *predictive accuracy*: an imputation method should preserve single values, i.e. imputed values should be the same than true values (for categorical variables), or as close as possible to the true values (for continuous variables);
ii. *distributional accuracy*: the imputation procedure should preserve the distribution of true data;

iii. *estimation accuracy*: the imputation method should reproduce as much as possible the lower order moments of the distribution of true data (at least first and second moments).

An additional desirable property is that imputed values should be "plausible", i.e. coherent with other data and not failing any edit rule.

3.2.1 Performance measures for the preservation of true values (predictive accuracy)

Given a *categorical nominal* variable Y with c+1 categories, and be Y_i^* its true value and $\hat{Y}_i$ its imputed value, both in i-th observation, a measure of how well an imputation method preserves true vales is given by

$$D = \frac{1}{n}\sum_{i=1}^{n} I(\hat{Y}_i - Y_i^*) \qquad (10)$$

that is the proportion of off-diagonal entries in the square table of order c+1 obtained by cross-classifying true and imputed values.

In case of a *categorical ordinal* variable, we can define a more general version of D to take into account the distance between true and imputed values.

We can test whether D is significantly greater than a small positive constant s that is an acceptable proportion of incorrect imputation. If $D > \varepsilon + 2\sqrt{\hat{V}(D)}$, where $\hat{V}$ (D) is an estimate of the variance of D, we can say that the imputation method does not preserve true values. We can set

$$\varepsilon^* = \max(0, D - 2\sqrt{\hat{V}(D)}) \qquad (11)$$

The smaller this value, the better is the performance of the method in preserving true values.

In case of a *categorical ordinal* variable, we can define a more general version of D to take into account the distanced $(\hat{Y}_i, Y_i^*)$ between true and imputed values:

$$Dgen = \frac{1}{n}\sum_{i=1}^{n} d(\hat{Y}_i, Y_i^*) \qquad (12)$$

In case of *continuous* variables, a completely different approach is followed. If an imputation methods preserves true values, $\hat{Y}_i$ should be close to Y_i^* for all cases where imputations have been made. A first measure of this closeness can be *weighted Pearson moment correlation* r between $\hat{Y}_i$ and Y_i^*. This measure is not recommended for highly skewed data.

Another approach is based on regression: first, a linear model of the form

$$Y^* = \beta \hat{Y} + \varepsilon \qquad (13)$$

is fitted to the subset of imputed data, and then a test is made whether the *slope* β is equal to 1. If the test does not reveal a significant difference (non significant p-value), then a measure of the *regression mean square error* can be computed:

$$\mathrm{MSE} = \frac{1}{n-1}\sum_{i=1}^{n} w_i (Y_i^* - \beta \hat{Y}_i)^2 \qquad (14)$$

Another regression-based measure is the value R^2, the *proportion of the variance in* Y^* *explained by the variation in* $\hat{Y}$.

The preservation of values can also be directly evaluated by calculating the distance $d(\hat{Y},Y^*)$ between the vector of imputed values and the vector of corresponding true values:

$$d_{L\alpha}(\hat{\mathbf{Y}}, \mathbf{Y}^*) = \left\{ \sum_{i=1}^{n} w_i \left| \hat{Y}_i - Y_i^* \right|^{\alpha} / \sum_{i=1}^{n} w_i \right\}^{1/\alpha} \qquad (15)$$

where typical values of α are 1 and 2.

3.2.2 Performance measures for the preservation of distributions (distributional accuracy)

For a *categorical* variable with c+1 categories, the distributional preservation capability of an imputation can be evaluated by calculating the following Wald-type statistic:

$$W=(\mathbf{R}-\mathbf{S})^t[\mathrm{diag}(\mathbf{R}+\mathbf{S})-\mathbf{T}-\mathbf{T}^t]^{-1}(\mathbf{R}-\mathbf{S}) \qquad (16)$$

where **R** is the c-vector of frequencies of imputed values for the first c categories, **S** is the c-vector of frequencies of true values for these categories, and **T** is the square matrix of order c corresponding to the cross-classification of true and imputed values for these categories. Distribution of W is chi-square with c degrees of freedom, and statistical tests concerning distributional preservation can be carried out.

For *continuous* variables, we introduce the weighted empirical functions for true and imputed values:

$$F_{Y^*}(t) = \sum_{i=1}^{n} w_i \, I(Y_i^* < t) / \sum_{i=1}^{n} w_i$$

$$F_{\hat{Y}}(t) = \sum_{i=1}^{n} w_i \, I(\hat{Y}_i < t) / \sum_{i=1}^{n} w_i$$

We can now measure the distance between the two functions using the *Kolmogorov-Smirnoff distance*:

$$KS = d_{KS}(F_{Y^*}, F_{\hat{Y}}) = \max_t \left(\left| F_{Y^*}(t) - F_{\hat{Y}}(t) \right| \right) \qquad (17)$$

An alternative is the integrated distance

$$KS(\alpha) = d_{\alpha}(F_{Y^{*}n}, F_{\hat{Y}n}) = \frac{1}{t_{2n} - t_0} \sum_{j=1}^{2n} (t_j - t_{j-1}) \left| F_{Y^{*}n}(t_j) - F_{\hat{Y}n}(t_j) \right|^{\alpha} \quad (17b)$$

where t_0 is the largest integer smaller than or equal to t_1. Larger values of α give more importance to larger differences. Usual values of α are α = 1 and α = 2.

3.2.3 Performance measures for the preservation of aggregates (estimation accuracy)

For *continuous* variables, we consider the problem of preserving the raw moments of the distribution. We can measure this preservation by using the indicator:

$$m_k = \left| \sum_{i=1}^{n} w_i (Y_i^{*k} - \hat{Y}_i^{k}) / \sum_{i=1}^{n} w_i \right| = \left| m(Y_i^{*k}) - m(\hat{Y}_i^{k}) \right| \quad (18)$$

with typical assignments of 1 and 2 to k.

4. Datasets and planned experiments

Six different datasets were chosen in order to represent a variety of data (continuous and categorical) and surveys (census and sample surveys; enterprises and households; cross-section and panel) typologies. The characteristics of the datasets have been reported in the following table.

Dataset	Type of dataset	Type of variables	Number of variables	Number of records
Danish Labour Force Survey (DLFS)	Administrative records	Continuous, nominal, ordinal	14	15,579
UK Annual Business Inquiry (ABI)	Quarterly Sample Survey	Continuous	26	6,233
Sample (1%) of Anonymised Records of UK 1991 Population Census (SAR)	Population Census	Nominal, ordinal	35	494,024
Swiss Environment Protection Expenditures (SEPE)	Yearly Sample Survey	Continuous	54	1,039
German Socio-EconomicPanel	Panel Sample Survey	Nominal, ordinal,	30	5,383
Survey (GSOEP)		Continuous		
Time Series for Financial Instruments (Shares and bonds, Options)	Time Series	Continuous	87 daily time series from 1995 to 1999	

For ABI, SARS, SEPE and time series , three different evaluation versions have been made available:

- Y^* containing true data, i.e. the dataset assumed to be complete and without errors;
- Y_2 containing data with missing values, but without errors;
- Y_3 containing data with both missing values and errors.

For DLFS and GSOEP only Y^* and Y_2 were produced.

Only versions Y_2 and Y_3 were given to partners for carrying out experiments.

Versions Y^*, considered as the "target" data, were not distributed by the project coordinator (ONS), with the exception of small subsets of data (near 10% of each dataset), necessary for some methods, as neural networks, that require "training" datasets to estimate internal parameters.

Together with datasets, also edit rules currently used by owners were disseminated to partners.

Versions Y_2 and Y_3 were produced by perturbing original Y^* in the following way:

a. missing values were generated by adopting a missing completely at random (MCAR) non-response mechanism;
b. errors were generated trying to simulate the way they occur during the compilation of the questionnaire or the data entry operations.

The percentages of missing values and errors for each variable were determined as much as possible on the basis of real situations verified in previous experiences.

Also *development* datasets were given to partners, in order to let them produce by themselves perturbed versions (also perturbation software was available), apply methods and evaluate their performance, to get valuable experience before the application to the evaluation datasets.

Partners applied each suitable method to different dataset according to the following rules:

1. each *error localisation method* was to be applied only to versionsY_3 of datasets (with both errors and missing values), while *imputation methods* were to be applied to both versions Y_2 and Y_3[6];
2. each partner could use the available subset of Y^* to train neural networks, or to estimate the parameters of a statistical model; for imputation methods, partners could use the complete subset of Y_2;
3. edit rules, given together datasets, could be (i) used by partners without modifications, (ii) with modifications, (iii) not used at all.

Once the different methods for error localisation and/or imputation of data were applied to the datasets, the corresponding outputs were given back to ONS, that provided to calculate the set of performance indicators illustrated in paragraph 3.

6 The rationale for the application of imputation methods to both versions was to test their robustness in the presence of errors.

5. Evaluation results

The experiments that were carried out are analysed from a double point of view: (i) for each dataset, the performance of the various methods that were applied to it are compared, and (ii) methods that revealed to be the best are highlighted.

5.1 Evaluation results bydataset

For any dataset, the different performance indicators will be grouped so as to analyse the following quality indicators:

a. "pure" *error localisation* performance: indicators from (1) to (3);
b. *influential error detection* performance: indicators from (4) to (8);
c. *difference between moments* of true and edited data: indicator (9);
d. *predictive accuracy*: indicators from (10) to (15);
e. *distributional accuracy*: indicators from (16) to (17b);
f. *estimation accuracy*: indicator (18).

5.1.1 Evaluation results in Annual Business Inquiry (ABI)

ABI dataset contain 26 variables organised in a three-level hierarchy: at the top level there are six economic variables and one employment variable. Each of these variables breaks down in a number of elements; for some of the latter there is another level with component variables. Most of the analysis that was carried out refers to the first level, including the six most important economic variables. A high number of error localisation and/or imputation methods were applied to this dataset. Up to 33 experiments involving Y_3 version, and 24 related to Y_2 version, were conducted by applying:

1. CHERRY-PIE plus multivariate regression and hot-deck imputation to Y_3 (CP-MRH);
2. multivariate regression and hot-deck imputation (MRH) to Y_2;
3. MLP to both Y2 and Y3;
4. Integrated Modelling Approach to Imputation (IMAI) to both Y_2 and Y_3;
5. Generalised Edit and Imputation System (GEIS) to both Y_2 and Y_3;
6. Self-Organising Maps (SOM) plus random draw form normal Probability Density Function (PDF), or MLP, or nearest neighbour (NN), or mean (MEAN) to both Y_2 and Y_3;
7. Donor Imputation System (DIS) to both Y_2 and Y_3;
8. Epidemic Algorithm plus POEM (EA-POEM) to Y_3;
9. Bacon plus EM algorithm plus POEM (BEM-POEM) to Y_3;
10. Transformed Rank Correlation plus POEM (TRC-EM-POEM) to Y_3;

11. Univariate robust tree modelling (UWAID with node mean or node nearest neighbour imputation) to Y_3;
12. Multivariate robust tree modelling (MWAID) to Y_3;
13. Univariate Forward Search plus Reverse Calibration Imputation (UFS-RCI), or Nearest Neighbour Imputation (UFS-NNI), or Linear and Log-linear Imputation (UFS-REG and UFS-LREG) to Y_3;
14. Correlation Matrix Memory (CMM) plus weighted mean or median to Y_3;
15. Support Vector Machines (SVM) to Y_2.

Once having standardised the quality indicators, if we consider the six most important economic variables in Y_3, and the *only error localisation* experiments, three methods obtained good values in all the three error localisation groups of indicators (a), (b) and (c), namely MLP, GEIS and SOM. In particular, MLP experiments obtained best results in groups (a) and (c) (pure error localisation performance and differences between moments), while SOM was the best in group (c) (influential error detection). If we consider the *only imputation* experiments, best performance in groups (d) and (e), i.e. predictive and distributional accuracy, was revealed by MLP. Finally, considering *both error localisation and imputation experiments*, good values in all the five groups (a)-(e) were shown by CP-MRH, SOM, UWAID and the set of UFS methods with the various imputation methods (RCI, NNI, REG and LREG). This latest set seems to achieve the absolute best values.

If we consider the Y_2 dataset, again for the six upper level economic variables, two methods rank above the others, namely MLP and MHR. In particular, MLP is the only method that achieves good results for all the considered indicators.

5.1.2 Evaluation results in UK Sample of Anonymised Records (SARs)

The evaluation here concentrated on six key variables, four concerning *individuals* (relation to head, marital status, sex and age), and two the *households* (number of rooms and presence of bathroom).

The methods that were applied to both Y_2 and Y_3 are CANCEIS-SCIA, MLP, SVM and, SOM, while DIS, CMM and IMAI were applied only to Y_2.

Starting with Y_3, if we consider individual variables and the first group of indicators related to the pure error localisation capability, for *alpha* values the best performance is shown by CANCEIS-SCIA and SOM; for *beta* values the best are CANCEIS-SCIA and SVM, while for the overall *delta* the best is always CANCEIS-SCIA.

If we consider the other indicators for the only continuous variable (age), MLP is the best for the influential error detection (root average error, RAE), while CANCEIS-SCIA shows the best performance for estimation accuracy (m_1 and m_2). Instead, Support Vector Machine (SVM) is the best for the preservation of true values (R^2 and d_{L2}), followed again by MLP and CANCEIS-SCIA.

Considering now Y_2, for variable age and indicator R^2 the best method is SOM (with random draws from normal PDF), while for d_{L2} is SVM. Again, CANCEIS-SCIA shows the best performance for estimation accuracy, together with IMAI.

5.1.3 Evaluation results in the Danish Labour Forces Survey (DLFS)

The peculiarity of the Danish Labour Forces Survey (15,579 observations) is that only the variable "income" contains missing values. The distributions of all other variables, categorical, are complete. This reflects a real situation, in which 27% of interviewees refused to respond to this question. The corresponding true values of non respondents can be found in administrative registers, so this is the only non simulated situation, in which it is possible to evaluate the imputation performance in the presence of a real non-response mechanism. The following methods have been applied:

1. MLP;
2. CMM (with different imputation methods: nearest neighbour, random neighbour, mean, weighted mean and median);
3. SOM (with nearest neighbour or random neighbour);
4. SVM (greedy or stratified);
5. IMAI (Regression Based Nearest Neighbour linear or log-linear, with or without noise);
6. Linear Regression;
7. Random Hot Decking;
8. DIS.

As for the *predictive accuracy*, MLP (with 20 neurons) shows the best values for slope (together with CMM and SVM), R^2, d_{L1}, d_{L2} and the MSE, followed by the Linear Regression.

In the *distributional accuracy* group of indicators, MLP is still among the best for KS(2), but SOM is the absolute winner for KS, KS(1) and KS(2).

As for *aggregate preservation*, SOM reveals to be the best for the preservation of the first moment (indicator m1), followed by MLP, while IMAI (log-linear without noise) is the best for the preservation of the second moment (indicator m 2).

5.1.4 Evaluation results in Swiss Environmental Protection Expenditures Survey (SEPE)

EPE data contains 1,039 observations and 54 variables. As in the case of ABI, there is a three-level hierarchy, where at the top level we can find the 4 most important key variables, that are totals of 20 variables, some of which are in turn totals of other 30 variables. Evaluation was carried out concerning the four highest level variables. These methods were applied to the Y_3 version:

1. CHERRYPIE plus multivariate regression plus ratio hot deck method (CP-MRH);
2. DIS;
3. Epidemic Algorithm plus POEM (EA-POEM);
4. Transformed Rank Correlation plus POEM (TRC-POEM);
5. Univariate WAID plus node mean imputation (UWAID);
6. CMM,

and these others to Y_2:

1. Multivariate regression plus ratio hot deck method (MRH);
2. Censoring;
3. SOM plus deterministic imputation or mean or random draw from normal PDF;
4. DIS;
5. CMM.

Considering methods applied to Y_3, there is no evidence of a method clearly doing better than the others in error localisation. On the contrary, the CP-MRH method ranks first with respect to the majority of imputation indicators.

Considering the Y_2 version of dataset, the overall good performances belong to methods MHR and SOM.

5.1.5 Evaluation results in German Socio-Economic Household Panel (GSOEP)

The GSOEP is a panel survey with six different waves, from 1991 to 1996. The dataset contains 30 variables, of which two can present missing values. Both are related to income: personal and household income. Because of the waves, we have up to 12 different variables to be imputed, six for personal income (from 91 to 96) and six for household income (again from 91 to 96). Imputation has been carried out by means of the following methods: SOM (with random draw from normal PDF), CMM (with 5 different imputation options: 2 real donor and 3 model donor), DIS and IMAI (using RBNN imputation method with a log-linear regression model without noise term).

For all quality indicators, IMAI always results to outperform the other methods. To explain this, it is worth while to remark that IMAI is the only method that made use of the panel characteristics of the survey. In fact, while all the other methods modelled auxiliary information on a cross-section basis, wave by wave, IMAI did so only for the first wave (1991): for next waves, information on previous values of income (individual and household), actual and imputed, was considered as auxiliary information, and added to the set of explicative variables in the models. In any case, even if we consider only the first wave, where this advantage for IMAI is not present, values of indicators still are in favour of the method, though less markedly. CMM is the second best, at least for personal income, while SOM is better for household income.

5.1.6 Evaluation results in Financial Time Series

Two datasets have been considered: one containing information concerning *shares and bonds* (daily prices for 51 time series from 1995 to 1999), and the other one related to *options* (36 time series of daily prices over the same period). These are the methods used for imputation:

1. Last Value Carried Forward (LVCF);
2. Multivariate regression imputation (R1) using stock market indicators and exchange rats as covariates;
3. Non-parametric multivariate regression imputation using a moving window of length 100 (NP100), with the same covariates than R1;
4. Multivariate autoregression imputation of lag1 (MARX1), with the same covariates than R1;
5. Univariate autoregression imputation of lag1-lag5 (ARX5), with the same covariates than R1;
6. Univariate multi-layer perceptron (MLP) imputation, with the same input considered in R1;
7. Black-Scholes pricing with cross sectional average imputation of missing volatilities (BSBASE);
8. Black-Scholes pricing with LVCF imputation of missing volatilities (BSLVCF);
9. Black-Scholes pricing with EM imputation of missing volatilities (BSEM);
10. Black-Scholes pricing with MLP imputation of missing volatilities (BSMLP).

The first six were applied to bonds and shares dataset, while the last four were experimented on options dataset.

LVCF is a somehow naïve method consisting in replying for a missing value in the series the more recent value observed for the same unit.

Methods (2) and (6) are not peculiar of time series context. Methods (3) to (5), on the contrary, are based on time relationships among observations.

Black-Scholes is a pricing formula, well known and widely used in financial institutions. The price of an asset at time t is dependent on a set of entities: all of them are usually available, with the exception of the so called *volatility*. When a price is missing in a time series, also volatility is: so, to be able to use the Black- Scholes formula, it is necessary first to estimate volatility. This can be done by using a variety of imputation methods: cross-sectional averaging, last value carried forward, EM algorithm imputation, univariate MLP imputation.

For each dataset, also in this case two versions were considered: one with only missing, and one with missing and errors.

As for shares and bonds, considering the dataset with only missing, LVCF is the worst method (essentially in terms of predictive accuracy), while NP100 is slightly better than the others. But if we consider the dataset version with also errors, we have exactly the opposite situation: LVCF becomes the best method (followed by ARX5), while NP100 results to be the worst.

Considering the options dataset, BSLVCF and BSMLP are best methods for imputing missing data. This is true for both versions of this dataset.

As a general conclusion, it can be said that methods that work on lagged variables are better than those exploiting cross-sectional information.

5.2 Best methods

On the basis of previous analyses, we tried to individuate best methods inside those selected to be investigated in the EUREDIT project. It is important to underline the fact that the concept of *best* is sometimes very relative, as performance for a given method may vary accordingly to the considered (groups of) indicators and subsets of variables. Very seldom a method outperforms all others in all possible situations.

Among *standard methods*, we can say that CANCEIS-SCIA revealed the best performance for categorical data, both for error localisation and imputation. CHERRY PIE was the best for error localisation in continuous data; for imputation, multivariate regression plus hot deck method showed the best results, followed by IMAI predictive mean matching method.

Among *neural network based methods*, MLP applications always stand in the first positions, both for error localisation and imputation, followed by TS-SOM.

In the class of *robust methods*, univariate forward search (BACON) for outlier detection outstands as the best. In association with imputation methods as reverse calibration and nearest neighbour, this method is the best also for imputation of both missing data and errors in continuous data. As second best, univariate WAID obtains comparable results in this class.

6. Conclusions

Experience made in the EUREDIT project led us to say that there is no "best" method, in the sense that no method works best in all situations. In addition, for a given situation, i.e. for a given typology of survey, the procedure for error localisation and imputation can hardly be constructed by utilising a simple method: very often, it will be a *complex* procedure, composed by different steps, possibly involving various methods, accordingly to the various nature of errors to be dealt with, and the different non-response mechanism.

Therefore, the value added of the EUREDIT project is not only (and not prevailingly) in the final indications concerning best methods to be used for different typologies of data (the *winners*). It is rather in the methodological path that was followed in its activities, that can be replied by anyone in order to *continuously* improve editing and imputation procedure. This path can be summarised as follows:

i. for any typology of data of interest, individuate candidate methods for error localisation and imputation;
ii. define a set of indicators useful to evaluate the performance of selected methods;
iii. adopt a simulation approach, by introducing missing values and errors in data in a controlled way so to replicate real situations;
iv. develop procedures containing selected methods and apply to data;
v. evaluate and compare results in order to choose best methods.

Another lesson learnt is in the fact that the more *information* related to (i) data structure, (ii) error nature and (iii) missing data patterns you can introduce in the procedure for error localisation and imputation, the more you can obtain in terms of accuracy of the results. This means that a lot of analysis of these three elements is needed. This job can be done only by expert statisticians, and cannot be delegated to naïve users: it is not just a matter of applying software to data.

Nevertheless, the availability of software is a crucial aspect: some of the investigated methods are so complex that a corresponding software is very costly to develop. So, a value added is also in the software that will be made available to EUREDIT partners and to external users: a software incorporating all robust methods and some of the neural network methods; and also a software useful for the evaluation process, to simulate missing and errors in data, and to produce evaluation indicators. Other software, especially rule-based standard software developed by national statistical institutes, is already available on demand.

The activity of the EUREDIT project will be hopefully continued in the VI Framework European Research Programme. One of the first objective of future work will be the creation of a *knowledge base* containing all the information related to the different methods and tools: methodological and operational aspects, suitable typologies of data, performance.

References

Bankier, M., M. Lachance, and P. Poirier. 2000. *2001 Canadian Census Minimum Change Donor lmputation Methodology*. Work Session on Statistical Data Editing, UNECE Cardiff UK.

Barnett, V., and T. Lewis. 1994. *Outliers in Statistical Data*. New York: John Wiley.

Beugin, C., and B. Hulliger. 2001. *Detection of Multivariate Outliers by a Simulated Epidemic*. Proceedings of ETK/NTTS 2001 Conference, EUROSTAT, pp. 667-676.

Billor, N., A.S. Hadis, and P.F. Velleman. 2000. *BACON: Blocked Adaptive Computationally Efficient Outlier Nominators*. Computational Statistics and Data Analysis.

Bishop, M.C. 1995. *Neural Network for Pattern Recognition*. Oxford: Oxford Clarendon Press.

Breiman, L., J.H. Friedman, R.A. Olshen, and C.J. Stone. 1984. *Classification and Regression Trees*. Belmont, CA, U.S.: Wadsworth International Group.

Chambers, R. 1986. Outlier Robust Finite Population Estimation. *Journal of the American Statistical Association*, N. 81, pp. 1063-1069.

Chambers, R., J. Hoogland, S. Laaksonen, D.M. Mesa, J. Pannekoek, P. Piela, P. Tsai, and T. De Waal. 2001. *The AUTlMP Project: Evaluation of lmputation Software*. Research Paper 0122, Statistics Netherlands.

Cristianini, N., and J. Shawe-Taylor. 2000. *An lntroduction to Support Vector Machines*. Cambridge: Cambridge University Press.

De Waal, T. 2002. *An Algorithm for Consistent lmputation in Mixed Data*. EUREDIT Deliverable 5.1.1. Statistics Netherlands.

Dempster, A.P., N.M. Laird, and D.B. Rubin. 1977. Maximum Likelihood from lncomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society,* B 39, pp. 1-38.

Deville, J.C., and C.E. Sarndal. 1992. Calibration estimators in survey sampling. *Journal of the American Statistical Association,* N. 87, pp. 376-382.

Di Zio, M., U. Guarnera, and O. Luzi. 2002. *GElS application on ABl data - Description of the applied editing methods*. EUREDIT Deliverable 5.1.1, Istat.

Di Zio, M., U. Guarnera, and O. Luzi. 2002. *GElS application on ABl data - Description of the applied editing methods*. EUREDIT Deliverable 4.1.1, Istat.

Gnanadesikan, R., and J.R. Kettenring. 1972. Robust Estimates, Residuals and Outlier Detection with Multiresponse Data. *Biometrics*, N. 28, pp. 81-124.

Fellegi, I.P., and D. Holt. 1976. A Systematic Approach to Automatic Edit and lmputation. *Journal of the American Statistical Association*, N. 71, pp. 17-35.

Hadi, A.S., and J.F. Simonoff. 1993. Procedure for the ldentification of Multiple Outliers in Linear Models. *Journal of the Royal Statistical Society*, B 56, pp. 393- 396.

Hidiroglou, M.A., and J.M. Berthelot. 1986. Statistical Editing and lmputation for Periodic Business Surveys. *Survey Methodology*, vol.12, N. 1, pp. 73-83.

Kohonen, T. 1997. *Self-Organising Maps*. Heidelberg: Springer-Verlag.

Koikkalainen, P. 1999. *Tree-Structured Self-Organising Maps. In Kohonen Maps* pp. 121-130. Amsterdam: Elsevier Science.

Kovar, J.G., J.H. MacMillian, and P. Whitridge. 1988. *Overview and Strategy for the Generalised Edit and lmputation System*. Report, Statistics Canada, Methodology Branch (updated February 1991).

Laaksonen, S. 2000. Regression-based Nearest Neighbour Hot Decking. *Computational Statistics*, N. 15, pp. 165-171.

Manzari, A. 2002. *Application of CANCElS and SClA to the UK SARs data. Description of the application.* EUREDIT Deliverables 4.1.1-5.1.1, Istat.

Nordbotten, S. 1995. Editing Statistical Records by Neural Networks. *Journal of Official Statistics*, Vol. 11, N. 4, pp. 391-411.

Nordbotten, S. 1995. Editing and lmputation by means of Neural Networks. *Statistical Journal of UNECE*, Vol. l3, N. 2, pp. 119-129.

Pannekoek, J. 2002. *(Multivariate) Regression and Hot-deck lmputation Methods*. EUREDIT Deliverable 5.1.1. Statistics Netherlands.

Riani, M., and A.C. Atkinson. 2000. Robust Diagnostic Data Analysis: Transformations in Regressions. *Technometrics*, N. 42, pp. 384-398.

Riccini, E. 2002. *CONCORD User Guide*. Istat Internal Document.

Ripley, B.D. 1996. *Pattern Recognition and Neural Networks*. Cambridge: Cambridge University Press.

Scavalli, E. 2002. *Edit and lmputation Using MLP Neural Networks in SARs data*. EUREDIT Deliverable 4.3-5.3, Istat.

Vapnik, V. N. 1995. *The Nature of Statistical Learning Theory*. New York: Springer.

Yar, M. 1988. *The Development of the Donor lmputation System*. Technical report. Office for National Statistics, U.K.

A prediction approach for the estimation of hours worked using integrated register and survey data

Fabiana Rocci [1], Silvia Pacini [2], Laura Serbassi [3], Marina Sorrentino [4], Maria Carla Congia [5]

Abstract[6]

Istat has released, starting from the reference year 2014, a new estimate of the variable 'hours worked per employee' to complete the integrated system of estimates FRAME, designed to meet the European regulation SBS. The result is an estimate of hours worked for all businesses with employees in industry and services. The methodology is based on joint information from statistical registers and business surveys, both structural and short term. The used method is based on a predictive approach that estimates on observed data the relationship between hours worked and hours paid, as register auxiliary variable, and imputes the same relationship on the rest of the population. Through the integrated use of information from various sources, and in particular those from the census register, it has been possible to identify subpopulations of companies with significant characteristics from the input of work and the relationship being valued. This paper describes the results of the analysis and outline of the final methodology.

Keywords: hours worked, prediction approach, data integration.

1 Istat, email: rocci@istat.it
2 Istat, email: pacini@istat.it
3 Istat, email: laserbas@istat.it
4 Istat, e-mail: mrsorren@istat.it.
5 Istat, e-mail: congia@istat.it.
6 The authors thank Fabrizio Solari as support for the model specification. Although the paper is the result of the combined work of the authors, the Sections are to be awarded as follows: Sections 1 and 4 to Fabiana Rocci; the introduction to Section 2 and Sections 2.1 and 3.1 to Marina Sorrentino; Sections 2.2 and 2.3 to Silvia Pacini and M.Carla Congia; Section 3.2 to M.Carla Congia; Section 3.3 to Silvia Pacini; Sections 3.4 and 5.1 to Laura Serbassi. Section 5.2 to all the authors.

1. Introduction

Structural business statistics (SBS) aim at describing the structure, behaviour and performance of businesses across the European Union on a yearly basis,.

The Italian National Statistical Institute (Istat) has developed a new system for the production of estimates on economic accounts of businesses for the SBS regulation (Reg. EU 58/1997 e 295/2008). The system is based on the use of administrative and fiscal data as primary source of information (at firm level), integrated with direct survey data as complementary information on specific variables or businesses' sub-populations, for which administrative information is not available. Thus, a new system of a multidimensional set of estimates (FRAME in the following) at an extremely refined level of detail can be annually released (AA.VV. 2016). The FRAME system is made up of several components, the main one being a statistical register of a number of key SBS variables that are available at firm level for the overall target population (~4,4 million units), which is linked to the Business Register Asia (BR).

For the SBS variables that are not covered by the administrative sources, further quality gains are achieved by improving the direct sample surveys and/or by developing alternative estimation strategies exploiting as much as possible the increased amount of available auxiliary information.

This paper focuses on the new estimation of the variable of hours worked by employees that belongs to the set of variables required by the SBS Regulation. It represents the most suitable measure for quantifying the real use of labour in the income production process. The statistical measurement of this variable has always been challenging, because there is not a unique way of registering it into the business accounts system and neither there is a variable defined properly in the administrative data.

A new methodology has been introduced to deliver the estimates of the hours worked for the SBS regulation, starting from the reference year 2014. Many studies on the available information, both direct and correlated to the target variable, have been done to identify the eventual statistical structure underlining all the data and the proper model representing it.

Several Istat surveys, designed to satisfy different EU regulations, comprehend the hours worked among their target variables. Some of them are primarily designed to gather high quality information about this variable, but they can differ with respect to coverage and level of disaggregation. On the other side, some administrative sources contain information about the remunerated days, that properly elaborated within a statistical register can produce a proxy of the paid hours, that is a variable very correlated to the target one.

The variety of information at disposal has made it necessary to adopt a mixed-sources statistical process as a solution, merging them into a physically-consistent dataset and rendering them suitable for a model application across all the population enterprises.

The final integrated dataset would deliver all the available information, from both administrative and surveys sources: the target variable hours worked measure for the surveys observed units and the administrative hours paid data on all the population units that represents auxiliary information correlated to the target one.

In this way, all the available information about hours worked and about the correlated hours paid on the whole target population is organized.

The prediction theory addresses the problem to estimate the total of a finite population variable from a sample to be equivalent to predict the total of the non-sample values (Valliant et al., 2000). In these terms, the prediction approach means to model the relationship between the two variables on the observed units and to impute the same relationship on the unobserved units on the basis of the model based estimation results.

In this context, the basic assumption is that there is a strong relation between hours worked and hours paid, so building a consistent integrated dataset of all related variables would create the suitable dataset to perform the estimation model representing the relationship properly.

The whole estimation scheme has been designed in several steps, to guarantee at each stage the required coherence between the observed units and the underlying concepts among the sources at disposal. In particular, to define the proper classification in specific sub-populations of the so obtained set of observed units has been challenging. The usual variables of stratification have been taken into consideration; nevertheless many other aspects to profile the enterprises structure have been defined. Furthermore, also other peculiar issues suggested by subject matter experts have been deeply analysed. Many aspects proved to need to be constantly monitored, in order to define the proper model stratification and the best model specification.

This paper aims at describing the new methodology used for the SBS estimates of hours worked and focuses on the main aspects that have resulted to be important to profile the enterprises' structure, to better estimate the target parameter.

The results indicate important differences between the new estimates of hours worked with respect to the previous ones based on the direct surveys that are designed to satisfy the SBS population.

The description of the statistical process of every source and the several aspects of the final estimation scheme are useful to explain the reasons of the differences and to give a final assessment of the entire mixed-source estimation process.

2. Informative context

Statistics on working time are needed to construct economic indicators, such as the average hourly earnings, the average labour cost per unit of time and labour productivity, and to evaluate policies and programs, as well as to estimate time-related underemployment. For these purposes statistics on hours worked need to refer to the same reference period and cover the same group of workers as are covered in statistics of, e.g., earnings, labour cost, employment-related income and production.

Many different working time arrangements exist, due to different scheduling of the hours of work that can be combined on a daily, weekly or monthly basis. Working schedules other than regular full-time, such as night work, shift work, part-time work and flexible working time arrangements are very frequently connected to certain economic activities and are also intended to enable workers to achieve a balance of work and family life (ILO, 2005).

For these reasons, the accurate measurement of the hours worked involved in the production is complicated, because it can depend on many different aspects. What is usually measured from the enterprise side are the component of regular hours paid by the business, both for actual hours worked and for paid but not worked hours.

The usual components of paid hours are:

1. Normal hours worked
2. Overtime
3. Not worked but paid hours: vacations, permits, etc.

Several EU regulations on business statistics require the measurement of hours worked. Nevertheless, they can refer to different population coverages, timeliness and details of the variable components.

In order to identify information useful to perform the estimations of hours worked for the SBS regulation all the available sources have been studied, both the surveys and the administrative data.

Since the main goal is to model the relationship between the hours worked and hours paid, the definition of the two variables and their differences have been widely studied across the sources, in order to fix the rules to build a coherent set of units and variables, into an integrated dataset. The relation has resulted to be influenced not only by the usually considered variables, that is by size and economic activity of the enterprise, but also by many aspects related to the internal organisation, as the share of peculiar work contracts, the use of short-time working (STW)[7] etc. Indeed, the chance to describe in a very detailed way the events and the actual input of

7 Short time working occurs when employees are laid off for a number of contractual days each week, or for a number of hours during a working day.

work, through the information of the register on all the population units, has made it possible to test several hypotheses that resulted to be significant to the type of relationship.

In the following, all the sources are described and compared from the coverage and the process point of view, so as to have the full picture on both definitions and measurement aspects. Many features of the surveys' design and processes have been studied (the survey design, the editing and imputation procedures, the calibration methods and the specific operational arrangements), in order to allow an assessment of eventual source effects in the measurement of the final variables (see § 3.3). Furthermore, also specific issues that could influence the target relationship are considered.

2.1 Survey data

Four Istat business surveys collect data on hours worked. Two of them are annual and are aimed at producing the estimation of the number of hours worked by employees in the reference year for the SBS EU Regulation: PMI and SCI, respectively covering enterprises with less and more than 100 persons employed, classified in Nace Rev. 2 sections from B to S with the exclusion of K and O. The two other surveys are short-term and have the production of indicators on per capita hours worked among their main targets: GI, a monthly survey covering enterprises with at least 500 employees classified in Nace Rev. 2 sections from B to S with the exclusion of O, and VELA, a quarterly survey covering enterprises with 10 to 499 employees in the same economic activities. GI and VELA microdata are used jointly to produce quarterly indicators of hours worked for both national publication and the STS EU Regulation (Regulation EC n. 1165/1998 of the Council and its revisions and amendments).

The surveys statistical design and processes are different in several aspects. Starting from the sample design: PMI and VELA are sample surveys, while SCI and GI are censuses of their enterprises' target population.

About the measurement of the variables: in both PMI and SCI employees are defined as the annual average of the end of month stock for each month of the reference year. In both surveys, this measure includes managers.

PMI and SCI measure hours worked as the total number of hours worked by all the enterprise's employees (managers included) in the reference year. Hours worked include both normal time and overtime and the sum of the two components is measured as a unique variable. Furthermore, SCI also collects data on the total number of hours paid during the reference year by all the enterprise employees (managers included).

GI collects monthly data on employees, hours worked by employees (distinguishing between normal time and overtime), hours paid but not worked, wages and employers' social contributions. Data on employees include managers while those on hours and labour costs do not cover them. Employees are observed as the stock at the end of the reference and previous months. The number of employees who are not managers can be calculated by subtracting the number of managers from the total number of employees.

VELA collects quarterly data on employees, hours worked by employees (distinguishing between normal time and overtime) and hours paid but not worked. All data cover only employees who are not managers. Employees are observed as the stock at the end of the reference and previous quarters.

Based on GI and VELA data, total hours paid can be calculated as the sum of total hours worked (including normal time and overtime) and hours paid but not worked. This sum supplies the correct number of hours paid provided that no compensatory time or "time off in lieu" scheme is used in the enterprise[8].

The PMI and SCI surveys provide variables that describe the whole scheme of accounts and balance sheet of the enterprise. To this aim, the process of validation is standardised to achieve a coherent full set of information. Hence, there is not a specific editing and imputation process related to each specific variable, besides those that guarantee the given relation among them. Therefore the procedures to edit, impute and validate the microdata on hours worked aim mainly at obtaining a plausible average estimate with regards to the whole set of account balance rules. In the PMI survey, item non responses are imputed through the mean over respondent units, unit non-responses are imputed through the use of administrative data, that covers almost the full balance sheet, for which the hours worked are treated as item non responses. Being PMI a sample survey, a calibration to the known totals of employment and number of enterprises in the population is then carried out. In the SCI survey, the unit non-responses are treated as in PMI.

On the other hand, data on hours worked are treated with specific attention in both GI and VELA surveys, due to their relevance in the disseminated aggregate indicators. In particular, enterprise experts first check GI data of the responding units. In case of non-responses, normal time and overtime hours worked are imputed separately, using data of the longitudinal profile of the firm itself. Influential observations are then identified and checked by the experts (Rocci and Serbassi, 2008).

Also within the VELA production process normal time and overtime hours worked are checked and validated separately. The first checks are performed during

8 These schemes allow employees to compensate a longer than normal working time in a given period with less hours of work in another one. However, paid hours refer to the normal working time. Hence, in these cases, the sum of hours worked and hours paid but not worked is higher than the actual number of hours paid when the employees work longer than normal hours and lower when they work shorter than normal hours

data collection: the largest share of responses are obtained via CATI and are in this process subjected to an extensive set of plausibility controls on hours worked. Moreover, the validation procedures on collected data include both interactive checks of outliers and influential observations, carried out by subject matter experts, and automated editing and imputation procedures based on the per capita number of hours worked in the same enterprise in the same quarter of the previous year (taking into account working days changes), wherever this information is available, and hot deck nearest neighbour donations, in the remaining cases.

To be used in the estimation procedure of SBS hours worked, GI and VELA higher frequency data need to be annualized.

To this aim, the monthly data collected by GI are transformed into quarterly ones at the enterprise level, by summing monthly hours worked over the quarter and by measuring quarterly employees (managers excluded) as the average of the stocks at the end of the previous quarter and of the reference quarter. In this way, the definitions of average quarterly employees and quarterly hours worked are identical in the quarterly GI and VELA microdata.

Starting from quarterly GI and VELA microdata, annual data are calculated by summing quarterly hours worked and averaging quarterly employees across the four quarters of a year. This step requires the availability of microdata for each considered enterprise for all four quarters of the year. GI microdata satisfy this condition: all unit non-responses are imputed to produce the target monthly indicators of the survey (on jobs, hours worked, wages and labour costs). These imputed unit non-responses are used in the production of the quarterly indicators on hours worked. VELA microdata, on the other hand, are affected by wave non-responses for which a correction via calibration is carried out in the process aimed at the production of quarterly hours worked indicators. Therefore, for the estimation of SBS hours worked, an imputation procedure for per capita hours worked and employees, based on hot deck nearest neighbour donations, is used to compensate for VELA wave non responses.

In the following, the results of several assessment analyses are described. The importance of hours worked in the Short Term surveys plays a fundamental role in the choice of the rules to be followed to build the integrated dataset.

2.2 Register data

The administrative information on remunerated time, that has been used to build the integrated data set, is produced within the Statistical Register on Wages, Hours and Labour Cost at job level (hereafter RACLI).

The RACLI register is mainly based on social security data. It belongs to a system of registers, where it represents the extension of the Employment Register

[Istat, 2016] for variables related to wages, labour cost and labour input for all the employees of the enterprises in the private sector, agriculture excluded (with a sectorial coverage wider than that required by SBS Regulation). Both registers have a Linked Employer Employee Data (LEED) structure, based on the compulsory monthly information at job level that employers have to send to the National Social Security Institute. This implies the availability at enterprise level of details on the jobs and on the employees. The employment estimates obtained through the Employment Register are the source of BR Asia and they are the same used in FRAME. The RACLI variables at enterprise level are obtained by the summarization of the same variables' estimates at job level.

In this paper the focus is on the information related to labour input and to the employment characteristics available in RACLI, such as the type and the duration of the contract, the working time, etc. that makes the register a very rich and detailed source, very useful to delineate which enterprises' features mostly affect working time.

The evolution of the informative contents of the social security source on working time has led to the availability of new and more detailed data that have been used to improve the estimation method of hours paid.

At the moment, the estimates on labour input variables are produced in RACLI exploiting the administrative information available.

Information on labour input have to be declared to the social security system in different units of measure, depending on the type of employee contract. For full-time employees, in particular, enterprises have to declare the monthly paid days, while for the other categories the paid time is registered in terms of hours. This leads to derive a proxy variable of annual hours paid by each enterprise that is calculated using weekly and monthly information at job level.

For full-time employees information on the monthly paid days is declared, according to a standard social security calendar[9]. The number of hours paid have been derived multiplying the number of declared paid days by the contractual working time of the job established in the collective labour agreements. It is important to stress that for administrative aims, one remunerated hour in the day is sufficient to have an entire paid day declared, where paid means totally or partially at the expense of the employer. This has two main implications: i) a paid day may include hours not at all paid by the employer (i.e. hours of strike, etc.); ii) paid days can be both partially and totally remunerated by the employers (i.e. if the employee is sick and this day is partially paid by the employer and partially paid by the social security system it is declared as an entire paid day). The effect could be an overestimation of paid hours due to hours of absence not paid within paid days and days of absence partially remunerated by the employer.

9 The standard of the social security calendar is 26 days in each month, 312 days and 52 weeks in a year.

For part-time and job-on-call employees, employers have to declare the monthly numbers of hours paid in term of equivalent paid weeks of a full-time employee. Using the contractual time of the full-time employee, it is possible to derive the number of hours paid that should not have the measuring problems just described for the paid days of the full-time employee.

Nevertheless, it is evident that both the information on paid time and the proxy variable of hours paid derived on a contractual basis do not include overtime hours. This can cause a bias in the estimates of the levels of hours paid, that however does not prevent its use as auxiliary variable for hours worked.

The richness of all RACLI information can be very useful to distinguish different enterprises' structure and to study the effect on the relationship between hours worked and hours paid.

In this view, it is important to underline the possibility to identify clearly the jobs with contracts that have peculiar working time arrangements, like job-on-call, that characterize the labour input within enterprises that use them extensively. In particular, these types of jobs in general tend to reduce per capita hours worked and to affect the relationship with hours paid (Congia and Pacini, 2010).

Furthermore, very useful information is also available about the large use of short-time working schemes by some firms in many economic sectors during the economic crises (Congia and Pacini, 2012), that is expected to affect the relationship under study.

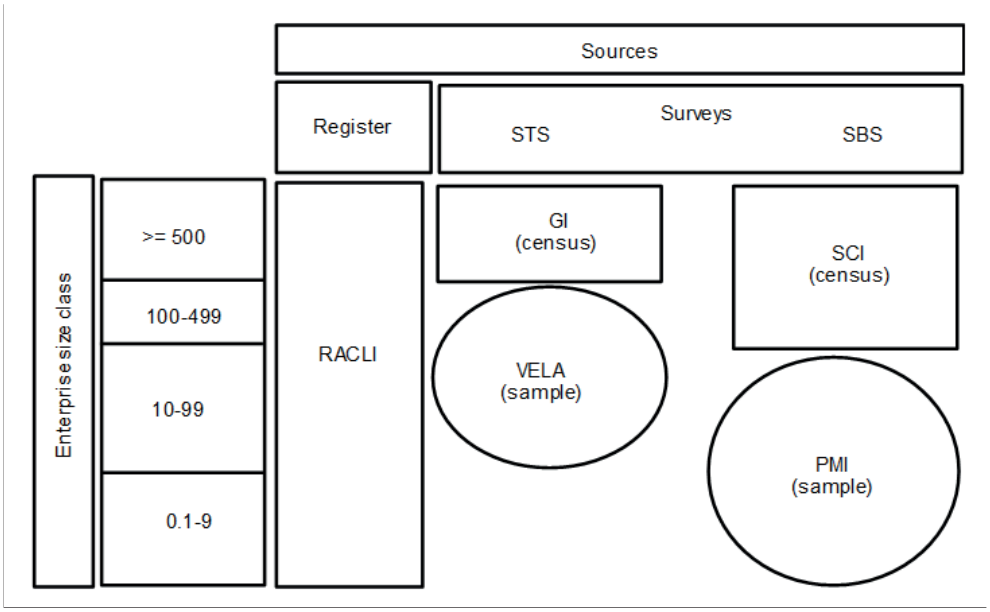
All the information on such employment features has been very useful to identify specific groups of enterprises according to the characteristics of their jobs and to specific events which may concern employees' working time.

In the following sections, the use of such a wide set of information and the way in which the proxy of hours paid has been employed as auxiliary variable to estimate hours worked are illustrated.

2.3 An outline of the data sources coverage

There are many different kinds of information related to the time of work, so all the sources described above have been combined to evaluate their main characteristics. Figure 2.1 represents the available data in terms of coverage and target population.

Figure 2.1 – Sources available on hours paid/worked and their coverage of enterprises with at least one employee in the private sector



Source: Authors' depiction

Furthermore, a summary of the coverage in terms of variables and their components is shown in Table 2.1.

From this point of view, the proxy variable on hours paid from RACLI is available for each unit of the target population. On the other side, the information on hours worked variables from the surveys covers only parts of the population. The big enterprises are almost fully covered, because censuses are run over enterprises with at least 100 employees (SCI, for units with at least 100 employees included, and GI, for those with at least 500 employees included), the response rate are very high and all the unit no-responses are imputed.

Moreover, several samples on the enterprises with less than 100 employees are available from PMI and VELA. Since not all units in the theoretical sample are respondents, the actual response rates have to be taken into account.

From the variable point of view, the definition of hours worked is the same across the surveys but the measurement is not completely homogenous, mainly due to the different coverage of employees, because of the managers which are not always included and when included the data referring to them cannot always be separated from that for all the other employees.

Table 2.1 – Variables on hours paid and worked in the surveys and register

Data source	Reference period	Employment coverage	Hours worked (w)			Hours paid not worked (p)	Hours paid (w+p)
			Total (w_o+ w_s)	Normal (w_o)	Overtime (w_s)		
PMI	Year	managers included	x				
SCI	Year	managers included	x				x
VELA	Quarter	managers excluded	x	x	x	x	x
GI	Month	managers excluded	x	x	x	x	x
RACLI	Year	breakdown for managers					X[(a)]

Source: Authors' depiction
(a) It does not include overtime hours.

It is important to underline that the surveys under analysis have different data collection and validation processes. Consequently, here all aspects have been compared in order to monitor any significant difference in the measurement of the same variable.

On the other side, the register proxy variable has to be compared with survey variables in order to assess the effect of the lack of the overtime component in the register data.

A very deep and careful analysis of the data and the eventual underlying relationship structure has followed, to assess the possibility to build a data set of statistical units with consistent variables coming from different sources.

The steps that have been followed to establish a coherent mixed-sources process are described in the following. First of all the analyses on the measurement of variables by different sources and the rules according to which the coherence between units and variables can be ascertained are described. Then the methods applied in order to identify suitable groups of enterprises, with specific worth considering characteristics, are presented.

3. Assessment analysis and integration of the data sources

In the following, the main results of the analysis of the difference in measurement in the several sources are shown. As first step, the analysis of the released aggregate estimates are presented to assess the coherence at macro-level. Because of some unexpected results, different hypotheses have been analyzed. Since the final aim is to build a micro-integrated dataset of statistical units with homogenous variables, a main aspect has been studied: whether given the same definition, there is any difference

in the measurement of the variable due to the effects of the validation process. Some evidences, in this regard, led to deeper investigation, assessment and comparison among the sources. Hence, the following studies have been done over the group of respondent statistical units common to the several surveys. Once they are linked, the assessment of the similarity or difference in the measurement of the same variable on the same units could help in identifying a possible survey measurement bias.

The final aim has been to identify clear rules to build the integrated data set, covering the entire list of units for which information from various sources would be at disposal in a chessboard way.

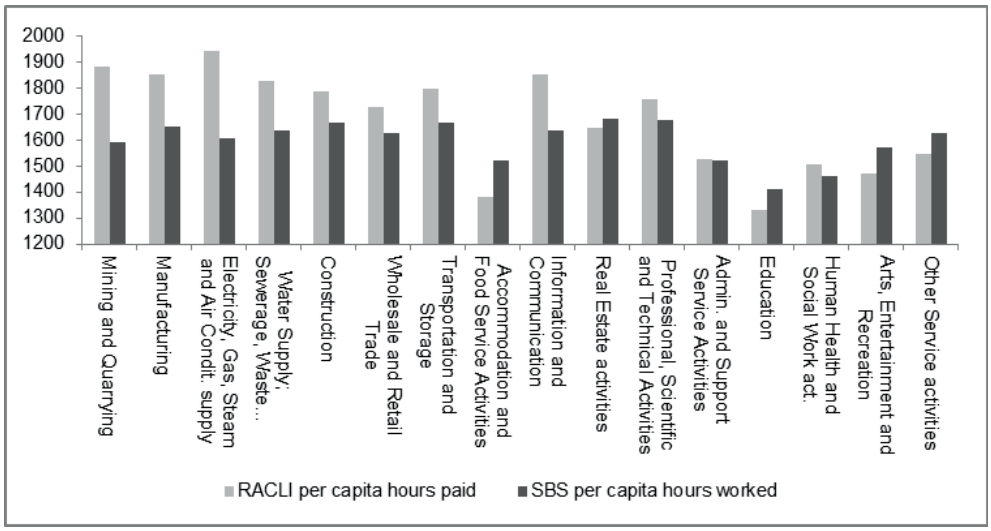
3.1 The comparison of hours worked estimates

In order to assess the coherence among the several sources, a comparison between register data and survey macro estimates officially released for SBS and STS regulations have been carried out. All results are shown in per capita terms.

The proxy variable of per capita hours paid has been compared to the per capita hours worked released by the surveys by economic activity, at the Nace Rev. 2 section level.

For the total population of enterprises with at least one employees, RACLI per capita data on hours paid may be compared with per capita hours worked from SBS estimates, see Figure 3.1.

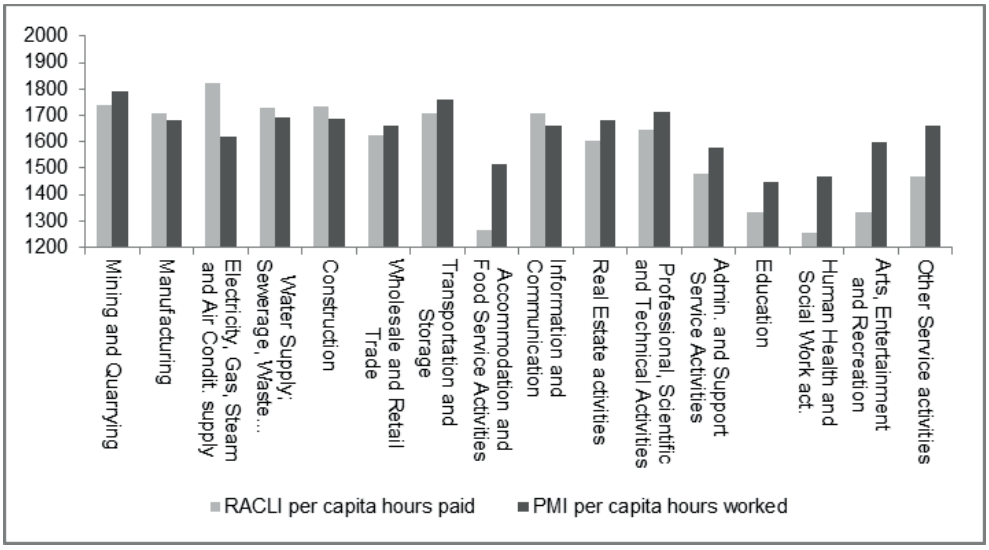
Figure 3.1 – SBS per capita hours worked and RACLI register hours paid in all enterprises with employees by economic activity. Year 2012



Source: Istat, RACLI Register, SBS disseminated data

Hours worked are expected to be lower or at most equal to hours paid, because the latter variable includes the first one, plus hours paid but not worked due to paid holidays, sickness, work permits, etc.[10]. Nevertheless, in some economic activities in the services sectors[11] the estimate of per capita hours worked calculated based on the structural PMI and SCI surveys is higher than the RACLI based estimate of per capita hours paid. To investigate this unexpected result, the population is divided in two groups of enterprises, with less and more than 10 employees, respectively covered only by PMI and covered by both PMI-SCI and GIVELA. As it is shown (Figure 3.2), the above-described result is mainly due to enterprises with less than 10 employees, where the estimate of hours worked based on PMI is higher than the RACLI based estimate of per capita hours paid for almost all of the services sections.

Figure 3.2 – PMI per capita hours worked and RACLI register hours paid in enterprises with less than 10 employees by economic activity. Year 2012



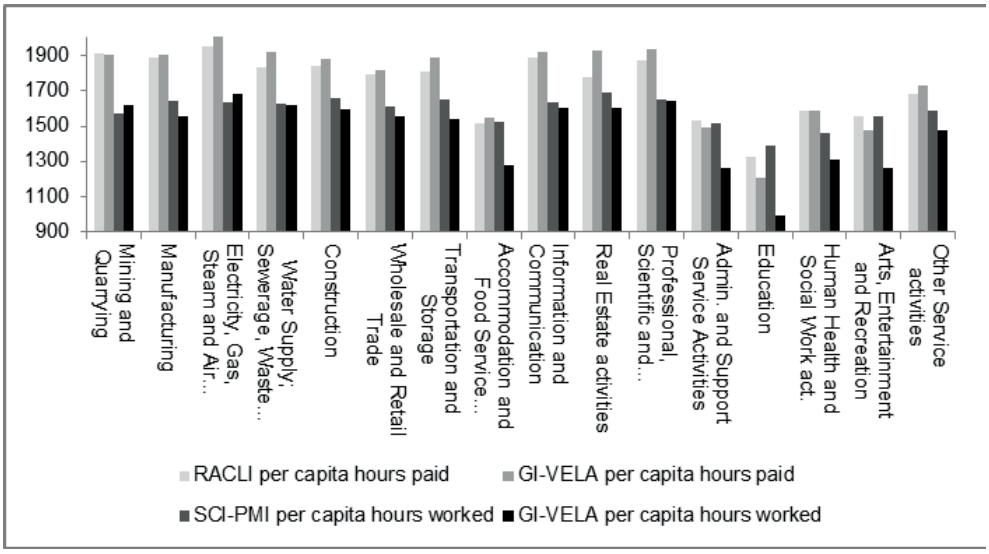
Source: Istat, RACLI Register and PMI Survey

On the population of enterprises with at least 10 employees, which is covered also by the short-term GI and VELA surveys, the estimate of per capita hours worked calculated on the basis of these two surveys is lower than the RACLI based estimate of per capita hours paid, as expected (see Figure 3.3). For this size class also SCI+PMI per capita hours worked are in almost all sections lower than the RACLI per capita hours paid even if they are always higher than the GI-VELA estimates.

10 The only case when hours worked can be higher than hours paid is in enterprises where there is a compensatory time or "time off in lieu" scheme and in the considered time period above average hours worked have not been compensated by an equal amount of time off.

11 Industry includes sections from B to F, while sections G to S are classified in services.

Figure 3.3 – SCI-PMI and GI-VELA (a) per capita hours worked and RACLI register hours paid in enterprises with 10+ employees by economic activity. Year 2012



Source: Istat, RACLI Register, SCI-PMI and GI-VELA Surveys
(a) GI-VELA hours paid includes overtime hours.

Many complex issues have been identified as underlying the unforeseen results in the services sector. This has driven further analyses aimed at investigating all the aspects of how different production processes can generate different estimates of the considered variables. Thus, an assessment of how the RACLI register proxy variable on hours paid is affected by errors of measurement (underestimation of the overtime hours and overestimation due to hours not paid in paid days) has been made. A more in depth exploration of hours paid has been carried out comparing RACLI and GI-VELA microdata at enterprise level (see § 3.2), that helps to focus on the specific component of overtime hours and which is the effect of not measuring it in the register data.

Beyond such issue, all the aspects described in the previous paragraphs have been taken into account in order to identify how the different production processes can explain the discrepancies in hours worked measurement and to delineate a quality assessment of the several sources, with respect to the construction of a consistent integrated dataset (see § 3.3).

3.2 Quality assessment of hours paid

The subset of linked respondent statistical units across all surveys has been built, to proceed with the necessary deeper analyses about the quality and characteristics of the relevant variables and their appropriateness for the model design.

To assess the suitability of the hours paid estimated in RACLI as auxiliary variable in a predictive model, this variable has been compared at micro level with

Table 3.1 – Difference between RACLI and GI-VELA hours paid, total and net of overtime hours, on GI-VELA respondents (net of managers) by economic activity. Year 2012

Economic activity	Num. of enterprises	Difference between RACLI and GI-VELA hours paid (%)				Difference between RACLI and GI-VELA hours paid net of overtime (%)			
		Mean	Q1	Median	Q3	Mean	Q1	Median	Q3
B - Mining and Quarrying	382	-0.9	-8.4	**-2.2**	3.5	1.9	-4.2	**1.0**	6.2
C - Manufacturing	4,625	0.2	-6.1	**-1.1**	4.0	2.5	-2.9	**1.5**	6.0
D - Electricity, Gas, Steam and Air Conditioning Supply	243	-2.7	-7.1	**-2.3**	1.4	0.6	-2.6	**0.9**	4.2
E - Water Supply; Sewerage, Waste Manag. and Remediation Activities	661	-3.4	-8.7	**-2.8**	2.0	0.3	-4.3	**1.2**	5.4
F - Construction	907	-0.4	-11.0	**-2.5**	4.2	2.7	-9.3	**0**	5.9
G - Wholesale and Retail Trade; Repair of Motor Vehicles and Motorcycles	1,235	-1.5	-7.6	**-1.5**	3.2	0.9	-4.7	**0.6**	5.3
H - Transportation and Storage	734	-3.0	-12.3	**-3.2**	3.6	-0.3	-9.5	**-0.2**	6.0
I - Accommodation and Food Service Activities	1,625	2.8	-12.2	**-1.9**	7.6	4.6	-9.9	**-0.3**	9.4
J - Information and Communication	542	8.9	-6.3	**-0.3**	5.5	10.5	-4.1	**1.4**	6.6
K – Finance and Insurance Activities	477	-0.2	-5.3	**0.3**	4.6	1.2	-3.4	**1.6**	6.2
L - Real Estate activities	158	-1.8	-6.8	**-1.5**	3.3	-0.2	-5.5	**-0.1**	4.7
M - Professional, Scientific and Technical Activities	552	1.0	-6.2	**-0.8**	4.2	2.7	-4.4	**1.1**	6.0
N - Administrative and Support Service Activities	580	-1.8	-10.1	**-1.8**	3.6	1.6	-7.2	**0.9**	7.6
P - Education	218	10.4	-7.1	**1.9**	13.4	11.1	-5.5	**3.2**	14.2
Q - Human Health and Social Work act.	406	-1.8	-8.5	**-2.4**	3.5	-0.5	-7.0	**-1.0**	4.9
R - Arts, Entertainment and Recreation	262	20.5	-13.6	**-2.4**	8.7	26.0	12.4	**-0.2**	10.7
S - Other Service activities	178	-3.4	-11.1	**-3.6**	5.1	-1.2	-9.1	**-0.1**	6.8
Industry net of Construction	5,911	-0.4	-6.7	**-1.4**	3.7	2.1	-3.1	**1.3**	5.8
Industry	6,818	-0.4	-7.1	**-1.5**	3.7	2.2	-3.5	**1.2**	5.8
Services	6,967	1.5	-8.9	**-1.4**	5.0	3.7	-6.8	**0.4**	7.0
Total	**13,785**	**0.6**	**-8.0**	**-1.5**	**4.3**	**2.9**	**-5.1**	**0.9**	**6.3**

Source: Authors' calculation on RACLI Register data and GI-VELA Surveys

the hours paid measured by the surveys. This comparison has been possible only with GI and VELA data, because the PMI survey does not measure hours paid and both structural surveys measure total hours worked without distinguishing between normal time and overtime.

For GI-VELA respondents both the values of the hours paid measured in the survey and those elaborated in the RACLI register are available. In Table 3.1 the main indicators of the distribution of the percentage difference between the two measures of the total amounts of hours paid are presented[12]. The median of the differences (1.5%) indicates a lower, albeit very near, level of the RACLI proxy of hours paid with respect to the GI-VELA measure. The size of the difference is almost equal in industry and services, although it shows more variation but nearly always the same sign across economic activity sections.

Furthermore, when the comparison is carried out on hours paid net of overtime, the GI-VELA measure and the RACLI proxy are even closer and the register measure is in median 0.9% higher than the survey one.

By enterprise size it seems that the overestimation in the RACLI hours paid data is negligible for the small enterprises (under 100 employees), while it is more relevant for largest ones (Table 3.2).

Table 3.2 – Difference between RACLI and GI-VELA hours paid, total and net of overtime hours paid by size, on GI-VELA respondents (net of managers). Year 2012

Employees size class	Enterprises	Difference between RACLI and GI-VELA total hours paid (%)				Difference between RACLI and GI-VELA hours paid net of overtime (%)			
		Mean	Q1	Median	Q3	Mean	Q1	Median	Q3
<=9	2,016	-1.1	-11.5	**-1.9**	4.5	0.3	-10.2	**-0.2**	5.7
10-20	4,741	0.7	-7.8	**-1.6**	3.6	2.9	-5.1	**0.4**	5.2
21-99	4,435	1.8	-7.6	**-1.7**	4.1	4.3	-4.5	**0.9**	6.2
100-499	1,249	4.8	-5.9	**0.0**	7.3	7.8	-2.6	**2.7**	9.9
500+	1,344	4.0	-4.5	**0.1**	5.3	7.7	-1.0	**3.2**	9.2
Total	**13,785**	**0.6**	**-8.0**	**-1.5**	**4.3**	**2.9**	**-5.1**	**0.9**	**6.3**

Source: Authors' calculation on RACLI Register data and GI-VELA Surveys

This evidence confirms that the RACLI proxy of hours paid is affected by both the underestimation due to the exclusion of not contracted overtime and the overestimation due to the inclusion of hours of unpaid absence in paid days and indicates that the size of each of these effects is not so relevant.

12 To this aim, managers are not included the hours paid in RACLI elaboration

Nevertheless, such effects are much characterized by the economic activity and by size, hence as far as it affects the relationship under study, it should be accounted for by the definition of a suitable stratification as part of the final model specification.

3.3 Issues to be considered on hours worked

Further analyses to understand the discrepancies occurring between the survey hours worked estimates has led to outline different issues that need to be tackled.

Firstly, the presence of a bias response effect in enterprises has been investigated considering that it has already emerged in earlier studies for the turnover variable (Oropallo, 2010; Casciano et al, 2011) and for hours worked (Baldi et al., 2013). To this aim, specifically designed analyses have been exploited, that compare main indicators of the correlated variable hours paid on the whole population, properly divided between respondents and non-respondents. For each survey the sample or the entire target population have been divided between the respondents and the not respondents to the survey[13]. In general, the survey respondents appear to have per capita hours paid higher than non-respondents average of hours paid based on the RACLI data. The set of respondents seems to be affected by a self-selection phenomenon negatively correlated to the enterprise size, as a consequence it is stronger for smaller enterprises. This phenomenon seems more relevant for structural surveys, in particular for PMI on the subpopulation of enterprises with less than 10 employees, characterized by lower response rate, the self-selection appears as more intense.

The calibration phase of the PMI survey shows that it cannot correct for the bias in per capita hours worked of the respondents. In fact, the grossing up to the reference population carried out by the PMI survey is based on a calibration on known totals on number of enterprises and jobs and it is not designed to solve any bias problem.

A similar consideration can be done for the editing and imputation procedures for SCI unit and item non-responses and PMI item non-responses: within these procedures, the imputation of hours worked is based on the respondents values (bias affected), observed on the previous year release. Hence, it can cause an overestimation of the final imputed data.

Hence, SBS estimates of hours worked, calculated based on the SCI and PMI surveys may be upwardly biased.

On the other hand, the short-term business surveys (GI and VELA) are much less affected by this self-selection phenomenon due to their much higher response rates. Furthermore, per capita hours worked enter in the sample design of the VELA survey. Its sample allocation is in fact carried out minimizing sample size under

13 This means that the entire target population has been divided into two sets: survey's respondents and all the other population units.

constraints on the maximum CVs on a set of variables which includes per capita hours worked.

Finally, for enterprises responding to both a structural and a short-term survey, the data on hours worked were compared to assess whether the questionnaire the sample unit was responding to and/or the data collection mode affected the measurement of the variable of interest. On average, in these cases hours worked were measured as lower by the short term surveys than by the structural ones. Several aspects of the production processes seems to cause differences in the variable measurement among surveys. We could assume that the timing of the survey could affect measurement accuracy, and result in more precise measures for monthly/quarterly data than for annual one. The discrepancies could also be a consequence of the different information system within the company used to calculate the requested information.

The much greater relevance of the hours worked data in GI and VELA with respect to PMI and SCI and the implications in terms of editing and imputation and validation procedures have led to give the priority to GI-VELA data in the building of an integrated data set of survey data. Therefore, when for a given enterprise and a given year a response is available both from a short term and a structural survey the GI-VELA response is selected. SCI and PMI respondents data (excluding the imputed ones) are used for enterprise non available in the STS surveys.

All these analyses on respondents and non-respondents stressed the importance of defining a proper stratification strategy, to consider all the aspects that influence the relation between hours worked and paid, to optimize the model estimation in every strata. The availability of detailed structural information in the RACLI register allowed to identify specific labour or enterprises structures and economic events to deal with, such as a high share of part-time or job-on-call workers or a large share of employees in short-time working. Moreover, the analyses have shown that also the subpopulation of enterprises with less than one employee has to be considered separately. These very small enterprises are difficult to be surveyed, showing a lower response rate also in the PMI survey, and a peculiar relationship between hours worked and hours paid.

To this purpose different sub-populations have been identified for a more correct estimation of the model parameter (see § 4.2.1).

3.4 The final integrated dataset

The first step of the estimation process is the construction of an integrated dataset of microdata containing all available information on the target population, defined by the active enterprises with employees of the Business Register (BR Asia) belonging to the Nace economic activities covered by the SBS regulation. The final dataset would gather the auxiliary variable on every units, while the target variable would

be available only on the observed units properly chosen from the surveys as to be representative of the remaining part of the population. This set of units would be used to perform the estimation model.

The starting point for the construction of the dataset is the definition of a criterion to identify clearly the same unit in all the considered sources (RACLI, SBS and STS surveys). This identification can be non-trivial because the different processes of the sources can generate problems in the identification of the 'same enterprise' from the point of view of the economic significance. A record linkage operation has therefore been carried out, keeping an acceptance level of mismatch errors close to zero with the aim to avoid any misalignment (Zhang, 2012).

The key linking variable used for matching units is the Statistical Business Register code, available in all the sources, as unique business identification number (BIN). Despite an accurate pre-matching process, some problems in using the BIN equality function as unique key still remain, causing 'not matched' and 'false matched' pairs. The first event occurs very rarely and is due to surveys' coding errors (formal errors) or refers to codes that identify enterprises that have ceased to exist, while 'false matches' occur when codes are linked correctly but data refer to substantially different units. The main reasons explaining this phenomenon are the different rules about registration and statistical treatment of business longitudinal changes together with the different timing of the data collection across the various surveys and RACLI (Baldi et al., 2011).

For surveys 'not matched' units a match with units in the register is attempted using the company name, but only for large firms or for specific economic activities with small populations.

The detection of 'false matches' is done through an indicator function, based on the difference in the annual average of the number of employees between RACLI and a specific survey. These differences are calculated separately for each unit for which RACLI and survey data have been linked via the BIN and are based on a measure of the annual average of jobs in RACLI harmonized with how this variable is obtained in a specific survey. The table below illustrates the differences across surveys in the calculation of this variable.

Table 3.3 – Annual average of employees in the SBS and STS surveys

Source	Annual average of employees
PMI - SCI	average of the end of month stock for each month of the year
VELA	average of quarterly employees across the four quarters of the year (managers excluded) (quarterly employees are averages of end of previous quarter and current quarter stocks)
GI	average of monthly employees across the twelve months of the year (monthly employees are averages of end of previous month and current month stocks)

Source: Authors' depiction

Thanks to the availability in RACLI of monthly information on employment for each job within each enterprise, it has been possible to calculate the employment variable according to the different survey definitions above described.

The match is assessed by comparing the value of the indicator function with a threshold and is accepted if the difference in the annual average of the number of employees between RACLI and a specific survey is below the threshold. For large firms and for specific sectors, characterized by a high turnover or seasonal workers that can imply wider differences in the number of employees measured by different sources, a higher threshold is considered.

Furthermore, some large units with differences above the threshold are also assessed by experts on the basis of the information in the specific data base on enterprises demography and changes of the BR and in the GI data base. For most of these units (about 113 in 2012) survey data are, however, used for the final estimate of hours worked but not in the model estimation. The table below shows the difference between matched units by BIN and validated ones. This set of units will be the set over which the model will be performed.

Table 3.4 – Enterprises by class of employment and result of the match. Year 2012 (number and percentage)

Size class of employment	Target population RACLI (num. of enterpr.)	Observed units in surveys	Not matched units Racli-surveys (%)	Matched units Racli- surveys by BIN only (%)	Validated matched units Racli- surveys (%)
1-9	1,444,468	11,345	1.4	98.6	84.0
10-19	123,096	10,071	0.5	99.5	89.5
20-49	51,277	7,422	0.3	99.7	88.2
50-249	21,288	6,601	0.6	99.4	92.3
250-499	2,087	1,366	0.0	100.0	96.6
500+	1,510	1,429	0.0	100.0	97.3
Total	**1,643,726**	**38,234**	**0.7**	**99.3**	**88.7**

Source: Authors' calculation on RACLI Register data, GI-VELA and SCI-PMI Surveys

As there are overlaps in the target populations of the four surveys (GI - SCI/VELA - SCI/VELA- PMI), for some units data from more than one survey are available, leading to more than one possible record matching with RACLI. The much greater relevance of the hours worked data in GI and VELA with respect to PMI and SCI and the implications in terms of editing, imputation and validation procedures described in paragraph 2.1 have guided the choice to prefer GI and VELA microdata over PMI and SCI when for a given enterprise and a given year a response is available both from a short term and a structural survey.

The final dataset contains for all the enterprises in RACLI the information related to employees and paid hours from the register itself and, for the linked enterprises, the information on hours worked and STW hours coming from the selected survey.

The Table 3.5 shows the breakdown in the final dataset of the selected units by source in terms of employees that are available for the estimation model. The composition of the validated set of units, by source, put the light on the issue that the data from the short term surveys play an important role for the estimation model while the PMI is fundamental for the very small enterprises stratum.

Table 3.5 – The final dataset in terms of employees: RACLI population and linked enterprises by survey and class of employees. Year 2012 (number and percentage)

Employees size class	RACLI population (number of employees)	Linked units (number of employees)	Employee coverage (percentage)	Linked units by survey (percentage)		
				Vela-GI	PMI	SCI
< 1 empl.	414.714	899	0.2	-	100.0	-
1-9 empl.	2.722.556	28.095	1.0	27.4	72.6	-
10-99 empl.	3.559.995	377.065	10.6	49.4	50.5	0.1
100-249 empl.	1.107.889	304.607	27.5	27.0	-	73.0
250-499 empl.	697.297	267.673	38.4	40.6	-	59.4
>=500 empl.	2.793.292	2.360.063	84.5	88.2	-	11.8
Total	**11.295.743**	**3.338.402**	**29.6**	**73.9**	**6.3**	**19.8**

Source: Authors' calculation on RACLI Register data, GI-VELA and SCI-PMI Surveys

4. Estimation model

The final aim is to estimate the total amount of hours worked for the domains required by the SBS regulation. Model-based sampling theory begins by recognizing that problems of estimation of finite population characteristics are naturally expressed as prediction problems (Valliant et al., 2000). To estimate a finite population total from a sample is equivalent to predict the total of the non-sample values. In this context, the values of the target variable are available on a subset of units, observed by several surveys. On the other side, the auxiliary variable, which is strongly correlated with the target one, is available for each enterprise in the target population from an administrative source.

In the following, the general scheme of the prediction approach is presented and the specifications tailored to this issue are described. The aspects that have been analysed in depth are about how the final target parameter is influenced by the description of the enterprise structure. Indeed, the definition of different sub-populations has resulted to be very significant, in order to tackle all the challenges

arisen during the preliminary studies on the estimation of hours worked in the presence of such a rich but complex informative context.

4.1 Prediction theory

Supposing that the number of units N in the finite population is known and that a number y_i is associated to each unit, the prediction approach treats the numbers $y_1,\ldots y_N$ as realized values of random variables $Y_1,\ldots, Y_N$. Once a sample $s<N$ is observed, the estimation of a function of the data h $(y_1, \ldots, y_N)$ entails predicting a function of the unobserved y_r (Valliant et al., 2000). Hence, the whole population can be divided into two set of units: the first set s made by the units observed by the surveys $\{y_i, i\in s\}$, the second set r including all the unobserved units $\{y_i, i\in r\}$. The final aim is to learn about the second set by studying the first.

A further assumption is that an auxiliary variable is available on the whole population, for which the following general linear model is valid:

$$M: Y= \beta X + \varepsilon$$

where

$$E(Y_i) = \beta x_i$$
$$\text{var}(Yi) = \sigma^2 \gamma_i,$$
$$\text{cov}(Y_i, Y_j) = 0,\ i\neq j.$$

Under these assumptions, the model based method starts from the estimation of the relationship between the interest and auxiliary variables on the observed units. The variable of interest values on the non-observed units are then imputed applying the estimated relationship to those of the auxiliary variable. The best linear unbiased predictor (BLUP) $\hat{\beta}$ of β under model M is:

$$\hat{\beta} = \sum_s (x_i\ y_i/\gamma_i) / \sum_s (x_i^2/\gamma_i)$$

whose final expression depends on that for γ_i, that defines the variance shape.

The problem to estimate the population total for Y is solved as follows:

$$\hat{T} = \sum_s y_i + \sum_r \hat{y_i} = \sum_s y_i + \sum_r \hat{\beta}\ x_i$$

Therefore the estimate of the target variable in each of the study domains is the result of summing up the target variable values, both observed and imputed through the model, on the units belonging it.

In the case of the estimation of hours worked, it is possible to model the problem according to this scheme considering the proxy variable of hours paid as the auxiliary one. Indeed, the integrated dataset includes both the set s of observed units, in which both the target and auxiliary variable values are available, and the set r for which only the auxiliary variable values are available.

An extensive set of analyses on the two variables and the linking function have been carried out in order to obtain a robust formalization of the problem. Indeed, the form of the expression defining γ_i, that describes the variance of the target variable and the definition of a stratum design strategy allowing the estimation on homogenous groups of units resulted to be very important to tackle all the issues that arose.

The classification variables on which strata are defined need to be linked to the target variables. Usually, economic activity and size class are used, for their relevance and for their relatively easy availability on all population units. In this case, the rich set of information available for each enterprise in RACLI has been used to profile units according also to working time and the relation between hours worked and paid. In particular, this has allowed considering in the stratum design strategy variables related not only to STW hours but also to surveys' response bias.

In the following, the main results about which factors have resulted to affect the variable of hours worked per each enterprise are presented. Indeed, other variables besides economic activity and class size have been found to be very significant in properly estimating the target parameter.

4.2 Model specification

In this case, the target variable is the number of hours worked and the proxy variable of hours paid is the auxiliary one. The whole population is defined by the list of active enterprises belonging to the Business Register (BR Asia), the set *s* is given by the data observed by the surveys, suitably chosen to form the integrated dataset, and the remaining units of the population define the set *r*.

Therefore:

Y ≡ HW number of hours worked

X ≡ HP number of hours paid

Based on preliminary analyses, the following model specification has been found to better fit the data:

1. as target variable, the total amount of hours worked for each enterprise is considered (instead of a per capita value which was originally suggested because more easily interpretable);
2. the parameter β is estimated through a heteroscedastic model. More precisely, the parameter γ_i is a linear function of the auxiliary variable, describing the increase in the variance of hours worked with enterprise size. Thus, the expression for γ_i is the following:

$$\gamma_i = hp_i$$

This means that the BLUP estimator for β in model M can be written as:

$$\hat{\beta} = \sum_{s} hw_i \,/\, \sum_{s} hp_i$$

3. the stratification on the basis of economic activity and enterprise size allows to estimate hours worked including the overtime component on the basis of the proxy of hours paid, even if the auxiliary variable does not include this component. This happens because the incidence of overtime hours over total hours worked is strongly associated with the two stratification variables;
4. to identify sub-populations of enterprises, defined by specific characteristics relating to several events and the type of work remuneration, is important. Indeed, finding the proper strategy for defining the suitable strata is necessary to better represent the non-respondent units on the basis of what is observed on the respondent ones. To this aim, beyond the usual classification variables of economic activity and size class, many other aspects have been taken into account that can be analysed through the variety of information from administrative data.

Hence, the whole scheme outline is as follows: at first, the whole target population is divided into the identified four sub-populations, defined as described below (see § 4.2.1). Afterwards, each sub-population is stratified according to economic activity and size. In this way, almost 250 different strata have been defined. They do not coincide with the study domains, for which the estimates of the total number of hours worked are obtained as sum of the observed and estimated values of hours worked on all enterprises in all the strata in the domain itself.

The hypothesized relationship is for each stratum C:

$\forall$ stratum C:

$$HW_C = \beta_C \cdot HP_C + \varepsilon$$

According to the classification made to build the integrated dataset, each stratum C can be partitioned into three sets:

s of observed units, for which data on hours worked are available from the surveys, used to estimate the model

z of observed units, accepted with a bigger threshold, so that are judged to be self-representative, not used to estimate the model

r of unobserved units, for which the values of this variable need to be estimated so that is $C \equiv (s \cup r \cup z)$.

The BLUP estimate of β_C is obtained as follows:

$$\hat{\beta}_C = \sum_{i \in S} hw_i \,/\, \sum_{i \in S} hp_i$$

Hours worked in the *j*-th not observed enterprise are then calculated as follows, based on the above indicated parameter estimate $\widehat{\beta}$ and of the number of hours paid available in RACLI, HP_j:

$\forall$ stratum *C* and $\forall$ unit $j \in r \subset C$:

$$\widehat{hw_j} = \hat{\beta}_C \, hp_j$$

Finally, the aggregated estimates of hours worked for each study domain *D* are calculated in the following way:

$\forall$ domain *D*:

$$\widehat{HW}_D = \sum_{i \in S} hw_i + \sum_{i \in Z} hw_i + \sum_{i \in R} \widehat{hw}_i$$

the sum of hours worked on all the units, observed or estimated, belonging to domain *D*. For each domain, the value of the total depends in different regards on the percentage of observed units or on the percentage of the estimated values.

4.2.1 Description of the sub-populations

Once the dataset has been built, according to standard quality requirements, a sample of observed data is available. To apply the predictive approach it has been important to assess whether the sample is representative of the remaining population, to be imputed, and which is the right classification in order to tackle all the issues that have arisen during the preliminary analyses.

A regression tree method has been used to test for the factors affecting more significantly the relation between hours worked and hours paid, represented by the parameter β. On every eventual classification so suggested, further assessment has been done studying the comparison between the hours paid on the two group of units used for the estimation for the model and the remaining units in the same set, to be imputed with the same model. These pervasive analyses have pointed out the necessity to go beyond the usual classification of size and economic activity, to weaken the effect of the bias response, that have always to be taken into account.

In this way, the relevant variables, their thresholds and the hierarchy with which the variables have to be considered, that is all the elements needed to identify the sub-populations, have been recognised. These sub-populations constitute a partition of the total population, due to the hierarchy established through the application of the regression tree method.

Four types of sub-populations have been identified as those to be considered before defining the strata on the basis of economic activity and size. They are defined as follows:

- enterprises with at most 1 employee: for such population the hypothesis under which the relation between the two variables is exactly equal to 1 has been

tested. For a consistent part of those enterprises such hypothesis has been accepted, for the remaining part a proper model has been estimated. The relevance of this size threshold has been shown by the regression tree method to dominate those of the other variables mentioned below (incidence of STW and job-on-call employees). Therefore, for enterprises with at most 1 employee these additional classification variables do not need to be considered;

- enterprises with STW incidence above a pre-defined threshold: the very detailed information on the phenomenon in the RACLI register has allowed to delineate four different profiles of STW use;
- enterprises with incidence of job-on-call employees above a pre-defined threshold: this kind of enterprises have been always under study, because for the workers on this type of contract the information about the day paid can be directly elaborated in terms of hours paid. Furthermore, they resulted to be affected by the response bias. For units with incidence of job-on-call employees above the threshold, hours worked are estimated as equal to hours paid. While for those below the threshold, hours worked are estimated through the usual model together with other units.
- the remaining enterprises (generally called no event).

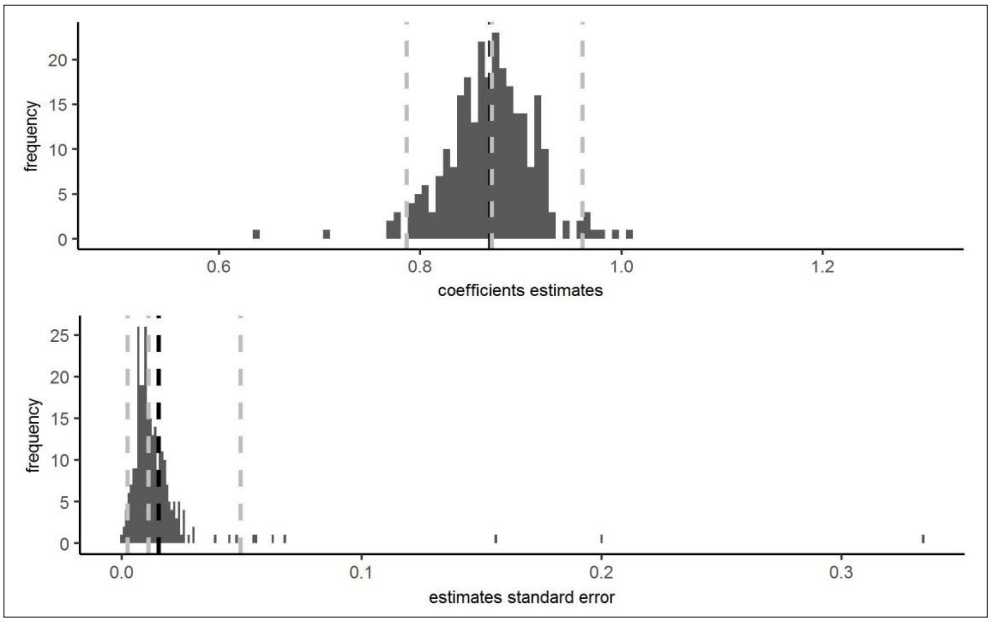
The stratification on each sub-population is carried out maximising the number of strata under the constraint that each stratum needs to include at least a minimum number of units. When this constraint requires to aggregate neighbour strata, the priority is given to keeping separate enterprises with different sizes, rather than with different economic activities as it is more common. In fact, all the analyses show that hours worked are much more influenced by the enterprise size rather than by its economic activity. This holds especially for the very small enterprises, where the sensitivity to size is very high. The level of disaggregation for the strata definition resulted to vary across sub-populations, the finest being based on 2-digit NACE and 6 size classes.

4.3 Model estimates' assessment

For all the strata, the null hypothesis of the estimate of the coefficient β_C being equal to zero is rejected. In the following, some graphs are presented (Figure 4.1) to give an idea of the overall behaviour of such model across all the strata (more than 200) used for the final estimation.

Both distributions are concentrated, especially the one of the estimates' standard error. The strata for which the coefficients' estimates are very small are those for very specific group of enterprises as the ones that use short-time working (STW).

Figure 4.1 – Histogram of coefficient estimates and their standard errors (quantiles in light grey, mean in black)

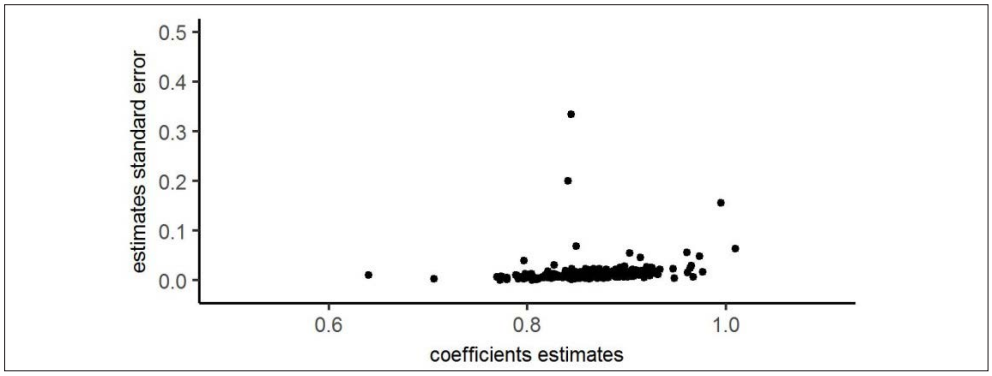


Source: Authors' estimates based on RACLI Register data, GI-VELA and SCI-PMI Surveys

The distribution of the standard errors of the estimates has also a concentration over a specific range: all of them are less than 0.07, only three of them reach values close to 0.2.

Finally, a scatter plot between the coefficients' estimates and the estimates' standard errors is analysed, to assess whether there is a relationship between the two.

Figure 4.2 – Scatter plot of coefficients' estimates and their standard errors



Source: Authors' estimates based on RACLI Register data, GI-VELA and SCI-PMI Surveys

A test on the correlation coefficient between the coefficients' estimates and their standard errors across all strata has confirmed that there is no significant linear relationship between the two of them.

5. Results and concluding remarks

5.1 Results and comparisons with previous hours worked estimates

The methodology to estimate hours worked described above has been tested on two years, 2012 and 2013. This has allowed validating the estimation's results both in level and in dynamics. In this way, the robustness of the criteria used for the identification of sub-populations and strata and the related model's parameters could be assessed.

As described above, the use of an integrated dataset as result of a mixed-source approach offers the opportunity to take into account many different kinds of information related to all the aspects of the working time structure. Indeed, a large amount of information is provided by the administrative source, not only with respect to hours paid but also to all the features characterizing different patterns of working time.

Table 5.1 – Annual hours worked per employee by economic activity and employment size class: comparison with SBS official data. Year 2013 (percentage differences between new and previous SBS data)

Employment size class	Economic activity[b]			
	Industry net of construction	Construction	Services (a)	Total Economy (a)
0-9	-9.2	-6.0	-15.9	-13.3
10-19	-4.4	-3.4	-8.6	-6.5
20-49	-3.7	-4.5	-7.7	-5.7
50-249	-2.6	-4.2	-7.8	-5.3
250 +	-3.3	-4.3	-6.1	-5.2
Total	**-4.2**	**-4.9**	**-9.9**	**-7.6**

Source: Authors' estimates based on Istat SBS data
(a) Excluding financial and insurance activities and Public Sector.

The comparison of the results of the new methodology and the old one is shown in Table 5.1 for the year 2013 (the results confirmed what was registered on the previous year 2012). The new estimates of hours worked are consistently lower than those previously disseminated for the SBS regulation and based only on the two SBS surveys. More specifically, the new estimates indicate that for the entire economy

hours worked are lower than the previously calculated figure by 7.6 per cent , respectively -4.2 per cent in industry net of construction , -4.9 per cent in construction and -9.9 per cent in services. It can be noted that the differences are strongly related to size and economic activity: the largest ones are recorded for smaller enterprises and services. It is worth pointing out that, for every sector, enterprises with less than 10 employees present a far bigger difference with regard the difference registered for the other size classes.

To explain such differences, it is useful to recall all the aspects regarding the statistical processes that have been considered and the composition of the final sample, used both for the model estimation and for the final vector on which the estimates for the domain total are built. As highlighted in Table 3.5, the sample coverage of the target population is increasing with size, together with the use of the short term data. This means that for the bigger size class, the differences can be explained mostly as a substitution effect between the SBS surveys and the STS ones.

On the other side, for the smaller size class enterprises, the coverage is consistently due to the SBS surveys, so the differences can be ascribed to the model estimation scheme. In this regards, it is important to underline that the imputation is done on units whose values can be measured only through the register data. In this regard, the knowledge of the factors influencing actual working hours in Italian enterprises has proven very relevant and it has been enhanced significantly. In particular, relatively small sub-populations of enterprises with specific characteristics implying peculiar patterns of hours worked have been identified. The measurement of hours worked in these sub-populations can present additional difficulties, but even when this is not the case the small sizes of the sub-populations make it difficult to represent them adequately through a sample survey. This concerns in particular small and micro-enterprises, firms with a significant share of low labour input employment contracts (e.g. jobs-on-call) or of absence events, or units operating in specific economic activities such as arts, entertainment, recreation and other service activities (sections R and S of the Nace rev.2 classification).

Hence, as final consideration, among the many reasons behind the relevant differences between the previously released SBS estimates and the newly produced ones, it can be recalled that the sample data used for the new estimates include where possible GI or VELA data rather than SCI or PMI ones and that in the aggregate estimates of per employee hours worked produced by the STS sources are lower than those produced by the SBS ones (see § 3.1 Figure 3.3). Moreover, SBS estimates of per employees hours worked are high also when compared with the RACLI measure of hours paid used as independent variable in the models (see § 3.1 Figures 3.1-3.3).

Finally, the prediction approach provide the chance to take into account every kind of units, also the ones that in any regards are more difficult to be reached from

the direct surveys. In these terms, the identification of specific sub-populations at first helps in avoiding the distortion effect of the bias-response on the parameter estimation. Furthermore, the final imputation on the remaining part of the population, identified through the register data, entails the final estimation to represent also the type of units considered to be very elusive.

5.2 Concluding remarks

In general, the new methodology has produced lower estimates of hours worked in comparison with those based on the SCI-PMI surveys. Furthermore, the most relevant result is that the estimates show a far greater variability across economic activity and size class of enterprises. In particular, these differences increase as the enterprise size decreases.

The break in the SBS series is not negligible, it has been deeply studied, in order to understand whether the reasons are structural or not. These differences are mostly due to the fact that the additional sources used with respect to those of the previous SBS estimates measure hours worked quite differently from SCI and PMI surveys. Furthermore, an extensive use has been made of the detailed information available for the entire population of enterprises in RACLI. The register coverage with regards to the target population allows to appropriately represent specific sub-populations that are more difficult to measure via sample surveys. Therefore, the use of administrative data as auxiliary information has helped in shedding light on phenomena that tend to be very elusive.

The results, tested on two years, have been considered statistically reliable in terms of the basic assumptions, choice of the models and coherence with the labour cost variable of FRAME. In particular, the production of the new estimates on consecutive years has allowed to test all the issues raised by the consulted experts.

Hence, starting from the reference year 2014, the estimates produced with the method described here have been disseminated officially at national level and transmitted to Eurostat to fulfil the SBS EU Regulation for the variable “hours worked”.

Despite the stability of the model across the considered period, it is suggested that in the next years the definition of the sub-populations and their threshold values are tested as a preliminary step before carrying out the estimates. Moreover the future evolution of the informative contents of the social security source on working time, the enlargement of the target population of VELA, to cover enterprises with less than 10 employees starting from 2016 and the inclusion, from the same year, of managers among the employees whose hours worked are measured by GI and VELA will provide new opportunities for improving the estimation.

References

AA.VV. 2016. *Rivista di Statistica Ufficiale*. N. 1/2016. Roma: Istat.

Baldi, C., C. De Gregorio, A. Giordano, S. Pacini, F. Solari, and M.Sorrentino. 2013. Joint use of survey and administrative sources to estimate the hours actually worked. *1st Southern European Conference on Survey Methodology (SESM) and VI Congreso de Metodología de Encuestas*, Barcelona, 12-14 December.

Baldi, C., D. Bellisai, F. Ceccato, S. Pacini, L. Serbassi, M. Sorrentino, and D. Tuzi. 2011. The system of short term business statistics on labour in Italy. The challenges of data integration. *ESSnet Data Integration Workshop*, Madrid, 24-25 November. http://www.ine.es/e/essnetdi_ws2011/ppts/Baldi_et_al.pdf.

Casciano, M.C., V. De Giorgi, F. Oropallo, and G. Siesto. 2011. Estimation of Structural Business Statistics for Small Firms by Using Administrative Data. *Rivista di statistica ufficiale*. 2-3: 55-74. Roma: Istat.

Congia, M.C., and S. Pacini. 2012. La stima da fonti amministrative di indicatori retributivi congiunturali al netto della cassa integrazione guadagni. *Rivista di Statistica Ufficiale*. 2-3: 19-40. Roma: Istat.

Congia, M.C, and S. Pacini. 2010. L'utilizzo del lavoro a chiamata da parte delle imprese italiane. *Approfondimenti Istat*. Roma: Istat.

Ilo. 2005. *General Survey of the reports concerning the Hours of Work (Industry) Convention, 1919 (No. 1), and the Hours of Work (Commerce and Offices) Convention, 1930 (No.30)*. Genève: International Labour Organization. http://www.ilo.org/public/english/standards/relm/ilc/ilc93/pdf/rep-iii-1b.pdf.

Istat. 2016. Il censimento delle imprese. *Atti del 9° Censimento Generale dell'Industria e dei Servizi e Censimento delle Istituzioni Non Profit*. Roma: Istat.

Oropallo, F. 2010. Analisi delle differenze strutturali nella performance economica tra unità rispondenti e unità non rispondenti nella rilevazione dei risultati economici delle piccole e medie imprese (PMI). *Contributi Istat*. N.7/2010. Roma: Istat.

Rocci, F., and L. Serbassi. 2008. The process of Editing and Imputation on Large Firms survey: between experience on field and computational standardization. *Proceedings of 2008 European Conference on Quality in Official Statistic*, Roma, 8-11 July.

Valliant, R., A. H. Dorfman, and R. M. Royall. 2000. *Finite Population Sampling and Inference, a prediction approach.* Hoboken, New Jersey, U.S.: Wiley, Series in Probability and statistics.

Zhang, L.-C. 2012. Topics of statistical theory for register-based statistics and data integration. *Statistica Neerlandica*. 66 (1): 41-63.

Exploiting the integration of businesses micro-data sources

Giovanni Seri, Daniela Ichim, Valeria Mastrostefano, Alessandra Nurra [1]

Abstract[2]

The new Statistical information system for estimating structural economic variables on business accounts (Turnover, Value Added, ...) based on the primary use of integrated administrative/fiscal data, "complemented" with survey data, FRAME SBS, has been released in 2012-2013. FRAME-SBS is by now the pillar of the new system of economic statistics in Italy. As further development, the exploitation of new opportunities of economic analysis by the integration of Frame SBS with other sources of data stemming from sample surveys has been particularly promoted. The paper reports on the work done to define a methodological approach for the production of economic indicators involving variables stemming both from Frame SBS and from two sample surveys: Community Innovation Survey (CIS) and Information and Communication Technologies Survey (ICT).

Keywords: SBS, Information and Communication Technologies Survey (ICT), Community Innovation Survey (CIS), data integration, balancing, calibration.

1 Researcher (Istat), e-mail: {seri,ichim,mastrost,nurra}@istat.it.

2 A version of this paper was presented at the 4th European Establishment Statistics Workshop held in Poznan (07-09 September 2015) and is available on the workshop website (http://enbes.wikispaces.com). The authors are grateful to Orietta Luzi who contributed with useful suggestions. Although the article is the result of a joint work, paragraphs 1, 2.1, 2.3, 3.2, 5 has been drafted by Giovanni Seri, paragraph 2.2 by Alessandra Nurra and Valeria Matrostefano, paragraph 3.1 by Daniela Ichim and paragraph 4 by Giovanni Seri and Daniela Ichim. The views expressed in this paper are solely those of the authors and do not involve the responsibility of their Insitutions.

1. Introduction

Recently the Italian National Statistical Institute (Istat) is evolving towards an integrated production system of SBS statistics. In this model, the core of the information content is represented by administrative sources while sample surveys are conducted in order to estimate only not directly available specific sub-populations information. For the majority of enterprises, the core of SBS variables, such as Turnover, Purchases of goods and services or Personnel costs, are often registered in different administrative sources, such as Financial statements, Tax Authority data or social security data. Consequently, the core of SBS variables may be estimated at an extremely refined resolution level. The SBS variables obtained through administrative data collection or statistical estimation procedures are registered in an exhaustive archive, called Frame, covering the whole population of enterprises as defined by the SBS Regulation.

The paper describes the analyses performed in order to integrate the Frame main information with data stemming from two structural businesses sample surveys. The main objective has been the production of economic indicators exploiting the interaction between two data sources, a register and a sample survey.

The considered sample surveys are the Community Innovation Survey (CIS) and the Information and Communication Technologies (ICT) survey. It is worth noting that the core variables of each survey are not registered in the Frame. Examples of such variables are binary indicators regarding innovation status, type of innovation activities, use of mobile devices or involvement in e-commerce activities.

Two different statistical approaches were examined, a macro and a micro one. The first one applies to aggregated (tabular) data stemming from the linking of the Frame and the chosen survey. The aggregates are then conveniently modified and subjected to the constraint that the marginal distributions independently derived from the two sources are maintained. The micro approach applies to a linked microdata file obtained by merging a survey dataset and the Frame archive. Practically, a 'new' set of weights is calibrated in order to preserve the consistency with the surveys' disseminated statistics. Simultaneously, the totals derived from the Frame variables are accounted for. Several calibration strategies are compared in this study.

In the paper we will illustrate the results obtained by implementing the two considered approaches. In Section 2 a brief description of the data used is given. In Section 3 the implementation of the different methods is described. Section 4 is devoted to comment on the results obtained. Some conclusions are drawn in Section 5.

2. Data description

2.1 Frame SBS

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 increasing stability, timeliness, coverage and accuracy of firm-level information available in some administrative sources on businesses' economic accounts has made it possible to develop a new SBS estimation system mainly based on the direct use of administrative data as primary source of information, integrated with the SME and LE survey data. Firm-level data for the main economic aggregates are directly obtained from the integrated sources, as they cover about 95% of the whole target population. As a consequence, aggregates of the most important SBS variables can be determined at an extremely high precision level, while for other SBS variables, more complex statistical modeling strategies might be required. The resulting statistical data warehouse covering the whole SBS target population and variables is named Frame.

It should be pointed out that each combined source actually covers different, possibly partially overlapping, subpopulations of enterprises, and that some sources provide data on a, possibly partially overlapping, set of variables. The overlapping information has been used for assessing the quality of input data and harmonizing classifications and definitions with SBS concepts described by the SBS regulation. Specific analyses have been devoted to manage inconsistencies among data from different sources. The registration of a key identifier facilitates the linkage of the data sources. (in each administrative archive enterprises are uniquely identified and classified based on a complex procedure performed at the Business Register construction stage).

The reference year for the data used in this work is 2012. The main objective of the project is to develop a strategy for deriving economic performance indicators by combinations of register and survey information. The economic indicators considered in this work are all registered in Frame, e.g. Value added per person employed, ratio between Value added and Turnover, etc. The spanning variables are defined as pairs of structural information registered in Frame (principal economic activity; size class)

and survey indicators. Possible inconsistencies between the data sources were solved by assuming that the true/real information was registered in Frame.

2.2 ICT and CIS surveys: purposes and indicators

The used survey data concern the ICT survey conducted in the year 2013[3], and the most recent edition of the CIS survey (the 7th Europe-wide CIS) referring to enterprises innovation activities between 2010 and 2012.

As for the ICT survey[4] a set of six indicators, characterizing the enterprises with at least ten persons employed operating in industry and non financial services were defined. Different topics belonging to the ICT questionnaire were included in the analysis through the derivation of composite indicators. Two main criteria were used for the indicators selection. The first criterion is intrinsically represented by the aim of the current project. Indeed, the trivial replication of already disseminated information is out of scope. Secondly, as Frame information is yearly registered and archived, 'core' questions and areas which are observed each year were identified. Consequently, biennial indicators or those belonging to the one-off sections characterizing the dynamic nature of observed ICT phenomena were avoided.

Based on national or international experiences (European Commission, OECD composite indicators related to the following areas of interest were included in this work: downloading speed of Internet connection declared by businesses (e_speed), intensity of use of the network in terms of persons employed using Pc connected to the Internet for work reasons, dematerialization and integration of organizational processes, levels of maturity reached by the company in e-commerce (from those only buying on line to those firms selling and buying on line or having also their own website offering opportunities to place on line orders for goods and services). The indicators choice leave open the possibility to update and/or extend their definition in order to better monitor the ICT improvement. Indeed, the classification of maturity levels may be easily changed, the speed or ICT usage classes may be updated, as well as the surveyed technologies (for example from Pc to mobile devices intensity usage).

The Community Innovation Survey is one of the major sources of innovation data. Based on a 'subject' approach aimed at identifying the innovative behavior of firms, its main goal is to overcome some drawbacks of the traditional long-established indicators based on the science-push model of innovation (R&D and

3 The reference year should be 2013, but since the survey was conducted in the first half of 2013 it is possible to consider the required qualitative information as referring to the end of 2012.

4 Since 2004, data collection on ICT is based on a European Regulation which ensures that the data are harmonized among Member Countries and in line with strategic European framework for the information society. ICT survey produces indicators for Digital Agenda Scoreboard (one of the seven pillars of the Europe 2020 Strategy) and it is annually implemented to better respond to evolving needs by users and decision makers.

patents indicators). CIS provides data on a diverse range of ways of innovating and captures forms of 'dark innovation' that don't rely on formal in-house creative activities such as R&D and which are seldom patented. CIS explores as well small-scale innovation or technology adoption of the "off-the-shelf innovators".

In particular, the CIS survey covers innovation activities of the Italian enterprises with at least ten persons employed operative in industry and services and focuses on four macro-typologies of innovation: product, process, organizational and marketing innovation, even if just for the first two categories it collects more detailed information on the expenditures, outcomes, linkages, sources for knowledge and technology transfers, factors hampering and objectives of innovation.

The survey is part of the Eu Innovation Survey (CIS), carried out on a two-year basis (from 2004 onwards) by all the Eu Member States and candidate countries. In order to ensure a sound comparability across countries, the CIS is carried out on the basis of a standard core questionnaire and a harmonized survey methodology developed by Eurostat, in close cooperation with the participating countries. Since 2000, the CIS has become one of the major sources of data for the European Innovation Scoreboard, and it has been confirmed by the European Commission as one of the flagship initiatives for measuring the performances of the Innovation Union within the Eu2020 strategy.

In this preliminary phase, in the selection of the most suitable indicators we have privileged some complex indicators based on the responses to different nominal level questions, more revealing of firms strategies than simple indicators and best capturing the propensity of the Italian firms to innovate, here defined as the attitude to carry out any kind of innovation activity (product, process organizational and marketing innovation, R&D driven or not) and regardless of whether the activity resulted in the implementation of a commercially successful innovation.

2.3 ICT and CIS surveys: methodological framework

Both ICT and CIS are surveys ruled by specific European Regulations requiring estimates for given domains of the target population, i.e. enterprises employing at least 10 persons and belonging to given NACE codes[5]. The sampling design of both ICT and CIS surveys is one-stage stratified random sampling. The strata are defined by combining the economic activity (Nace classification), size class (Number of persons employed) and region (Nuts classification) according to the domains of interest. Equal selection probabilities are assigned to enterprises belonging to the same stratum. The sample size in each stratum is mainly defined according to the

5 Hereafter for Frame we intend the dataset including enterprises belonging to the theoretical population of the considered survey (196186 units for the ICT survey and 160909 units for the CIS survey).

Bethel procedure (Bethel, 1989) as the minimum sample size ensuring that the coefficient of variation of estimates in predefined domains does not exceed a given threshold. Estimates are then derived through calibration methodology (Deville, Särndal, 1992; Casciano *et al.*, 2006) to compensate nonresponse and to match known population totals (benchmarks) of selected auxiliary variables (Number of persons employed, Number of enterprises). The population totals are computed using the Italian Statistical Business Register (ASIA). According to the time schedule of the surveys, the reference year of the ASIA register is 2011 and 2012 respectively for ICT and CIS. When linking Frame and ICT datasets, due to the different reference years, around 1400 units of the 19114 units cannot be linked. The main reason is given by the changes in the number of persons employed. Consequently, the majority of the non-linked enterprises did not satisfy the criteria defining the survey target population (at least 10 persons employed). Additionally, several NACE misclassifications and demographic events caused around 100 ousting of units. As regards the integration of Frame and the CIS survey these kinds of problems have a very low impact, as the Frame and the survey sampling frame (the most updated version of the official statistical business register Asia) both refer to the same reference year (2012). Anyway, the linking does not cover the whole CIS target population that includes the Financial Services sectors which are considered in Frame.

3. Methods

3.1 Macro-integration

Following a macro integration approach we considered estimates as two way tables involving both Frame and survey variables. Particularly, the spanning variables are structural information registered in Frame as NACE or size class combined with a survey indicator. The cells contain aggregations of an economic variable/indicator registered in Frame as Value added or the Number person employed. The tables were then modified by a multiplicative algorithm and by imposing constraints on the marginal row and column totals.

Following the macro approach we first considered the method known as Balancing (Nicolardi, 1998; AAVV, 2012) where a set of estimates in the form of tabular data stemming from different sources and having some common marginal have to be reconciled in order to achieve consistency on these margins. The method is usually used in the compilation of the National Accounts and it is implemented as a constrained optimization problem[6]. For our purpose, marginal totals were

6 The method is implemented in an R routine developed at Istat.

determined from the two different sources while the initial cell values were computed on the linked dataset: the dataset including statistical units belonging to both the sources (the Frame reduced to the theoretical target population of the survey and the sample data set of the survey). Unfortunately, the implementation does not impose non-negativity constraints for the cell values. In our application, the solution diverges to unacceptable solutions (negative frequency counts). Therefore we tested the Iterative Proportional Fitting procedure (IPF[7]). IPF requires as input the two given marginal distributions and an initial set of cell values. IPF iteratively adjusts the cell values to achieve consistency with the marginal row and column totals. The method may be easily implemented. Since it uses a multiplicative algorithm to achieve the consistency with a given marginal distribution, there is no risk to obtain inadmissible solutions. IPF may be applied independently in each table. On one side, this feature increases its applicability. On the other side, without further control or constraints, inconsistencies in linked tables are possible. It is worth noting that the marginal distribution of Frame quantitative variables with respect to survey categorical variables cannot be known; therefore it was estimated by means of the corresponding distribution derived from the linked dataset. We report IPF as method A when presenting the results.

3.2 Micro-integration

In the micro-integration approach, through calibration, the sampling weights (or a set of initial weights) were modified in order to achieve numerical consistency between estimates and 'known' population totals.

Different calibration strategies were tested[8] (Deville, Särndal, 1992; AAVV, 2012; Leadership Group SAM, 2003). First we applied the calibration strategy used by the survey. Indeed, the population totals of the variables Number of persons employed and Number of enterprises for given domains were derived from the Frame. Then, these totals were used as known population totals when calibrating the weights corresponding to the linked dataset.

In order to achieve consistency on the productivity indicator Value added per person employed, the second strategy, named method B to present the results, consists in enriching the set of auxiliary variables by Value added. In order to guarantee the convergence of the calibration algorithm, the geographical information was removed from the list of variables defining the estimation domains. Moreover, the calibration process is generally set to achieve a-priori defined lower and upper bounds for weights.

7 Implemented in the R package Teaching Sampling available in R.

8 The generalized software ReGenesees was used. The software has been developed at Istat (http://www.istat.it/it/strumenti/metodi-e-strumenti-it/strumenti-di-elaborazione/regenesees).

When combining the Frame and ICT information, an additional method C was implemented. Instead of using the linked dataset, the ICT survey dataset may be directly used through its original set of calibrated weights assigned to all the units in the survey sample. In this case, the Frame plays twice the role of secondary source: firstly, the Frame balance sheet information may be integrated for common units; secondly, Frame may be used for computing the known population totals. Although we achieved numerical consistency with the known totals, the main drawback of this strategy is that the theoretical target population selected by the Frame does not include the whole sample and this can represent a sort of incoherency. For the CIS survey the linked data set and the full sample survey dataset are extremely similar (no significant differences between the two sets of weights resulting from the two strategies were observed). That's why this strategy was applied only to the ICT survey.

Finally, we tested a strategy, called D in the results, where the known totals computed using the Frame auxiliary variables were combined with the ICT estimates derived for a categorical variable (e.g. the number of enterprises performing ICT sector[9] activities or not). The idea behind the strategy was to simultaneously calibrate on known population totals for the variables involved in the computation of the productivity indicator and to be consistent with the selected (and maybe published) estimates of the ICT indicator. Subsequently, by means of Consistent Repeated Weighting – CRW (AAVV, 2012), different ICT selected indicators were added. In general, in our tests, when the ICT estimates used as known population totals were zero or very small, the algorithm did not converge.

4. Results

The integration strategies illustrated in the previous section were applied for different economic indicators and for different combinations of spanning variables. A selection of the results obtained is reported in Tables 4.1 to 4.6.

In Table 4.1 the Value added per person employed (VA/PE) is reported for the subpopulations of enterprises defined by cross-classification of the NACE categories and the downloading speed of the broadband Internet connection; the latter variable is called e_speed. The NACE categories were grouped in "inside" and "outside" the ICT sector while the categories of the binary ICT indicator "e_speed" were defined using a threshold equal to 10 Mbit/sec. The shown results allow for the comparison of the four strategies: (A) IPF; (B) calibration of the 'linked dataset' using known

9 ICT sector in NACE Rev. 2 (based on the 2006 OECD definition) is defined by the following economic activities: 261, 262, 263, 264, 268, 465, 582, 61, 62, 631, 951 (https://ec.europa.eu/eurostat/cache/metadata/en/isoc_se_esms.htm). It is possible to distinguish ICT Manufacturing activities (261, 262, 263, 264, 268) and ICT services activities (465, 582, 61, 62, 631, 951).

totals derived by the Frame; (C) calibration of the 'survey dataset' using known totals derived by the Frame and (D) calibration of the linked dataset using known totals derived from Frame and ICT estimates, respectively. It is worth noting that each third column is constant, proving the convergence to the known values of the marginal distribution (differences reported for method D are within the admissible error range).

In Table 4.2, through the relative differences of the values reported in Table 4.1, the IPF method is compared with calibration approaches (B, C and D), while the method C is compared with the method B. Similar conclusions may be drawn for other comparisons that were performed for different cross-tabulations involving more detailed NACE levels and other ICT indicators.

In Table 4.3 the percentage of Value added out Turnover is reported. The cross-classifying variables of the Table 4.1 are used. In this case the economic indicator involves the Turnover information which was not considered as auxiliary data during the stratification and calibration processes.

Table 4.1 – Value added per person employed for ICT and non-ICT economic activities and e_speed values: comparison of methods A, B, C and D

VA/PE	IPF (method A) e_speed			Linked datasets (method B) e_speed			Survey' dataset (method C) e_speed			Table (method D) e_speed		
NACE	0	1	Tot_ICT	0	1	Tot_ICT	0	1	Tot_ICT	0	1	Tot_ICT
Outside ICT sector	49,082	62,457	55,065	48,905	62,976	55,065	48,529	63,435	55,065	50,520	61,363	55,056
Inside ICT sector	52,313	123,658	104,070	51,801	122,924	104,070	50,770	124,932	104,070	54,442	125,040	104,265
Tot_e_ speed	49,168	67,433	57,600	48,977	68,006	57,600	48,588	68,483	57,600	50,625	66,725	57,600

Source: Authors' calculation on ICT and FRAME data - reference year 2012

Table 4.2 – Relative differences (%) Value added per persons employed for ICT and non-ICT sectors and values of e_speed: calibration methods B, C and D compared to method IPF (A) and of method C with respect to B

Rel Diff VA/PE	(A-B)/A e_speed			(A-C)/C e_speed			(A-D)/D e_speed			(B-C)/B e_speed		
NACE	0	1	Tot_ICT	0	1	Tot_ICT	0	1	Tot_ICT	0	1	Tot_ICT
Outside ICT sector	0.4	-0.8	0.0	1.1	-1.6	0.0	-2.9	1.8	0.0	0.8	-0.7	0.0
Inside ICT sector	1.0	0.6	0.0	3.0	-1.0	0.0	-4.1	-1.1	-0.2	2.0	-1.6	0.0
Tot_e_ speed	0.4	-0.9	0.0	1.2	-1.6	0.0	-3.0	1.0	0.0	0.8	-0.7	0.0

Source: Authors' calculation on ICT and FRAME data - reference year 2012

Table 4.3 – Value added out Turnover (%) for ICT and non-ICT sectors and values of e_speed: comparison of method A, B, C and D

VA/TURNOVER	IPF (method A) e_speed			Linked datasets (method B) e_speed			Survey' dataset (method C) e_speed			Table (method D) e_speed		
NACE	0	1	Tot_ICT	0	1	Tot_ICT	0	1	Tot_ICT	0	1	Tot_ICT
Outside ICT sector	20.9	19.6	20.2	21.4	19.9	20.6	21.2	20.4	20.8	21.4	19.4	20.4
Inside ICT sector	30.9	43.5	41.1	31.1	43.6	41.4	30.5	41.2	39.3	28.9	42.9	39.9
Tot_e_speed	21.1	21.4	21.2	21.6	21.7	21.6	21.4	22.1	21.8	21.5	21.2	21.4

Source: Authors' calculation on ICT and FRAME data - reference year 2012

Similarly, we present in Table 4.4 the values of Value added per Person Employed (VA/PE) computed on the CIS data. We consider a different classification of the NACE code in six categories defined by the Eurostat/OECD technological intensity classification[10] combined with the CIS binary indicator "PPI" identifying the 'enterprises carrying out product or process innovation'. As stated before, only two methods are considered for the CIS survey: (A) IPF and (B) calibration of the 'linked dataset' on known totals derived from the Frame. Table 4.5 reports the comparison of these methods through relative differences of the values given in Table 4.4. In Table 4.6 the percentage of Value added over Turnover is reported for the same cross-classification of Table 4.4.

Table 4.4 – Value added per person employed for technological intensity categories and values of PPI: comparison of methods A and B

VA/PE	IPF (method A) PPI			Linked datasets (method B) PPI		
PAVITT	0	1	Tot_PPI	0	1	Tot_PPI
Not elsewhere classified	66,945	110,452	81,831	67,515	112,120	55,065
High-technology	89,509	88,624	88,837	90,627	88,231	
Medium-high-technology	54,347	71,570	67,341	56,533	70,933	
Medium-low-technology	50,065	61,042	56,180	50,603	60,703	
Low-technology	41,953	61,195	52,800	43,984	59,747	
Knowledge-intensive services	64,292	114,504	95,853	65,103	115,239	
Lessknowledge-intensive services	47,237	58,403	51,877	47,879	58,302	104,070
Tot_PPI	52,489	73,223	63,332	53,423	73,000	57,600

Source: Authors' calculation on CIS and FRAME data - reference year 2012

10 https://www.oecd.org/sti/ind/48350231.pdf
https://ec.europa.eu/eurostat/statistics-explained/index.php/Glossary:High-tech_classification_of_manufacturing_industries

Table 4.5 – Relative differences (%) Value added per persons employed for technological intensity categories and values of PPI: calibration methods (B) compared to method IPF (A)

VA/TURNOVER	(A-B)/A PPI		
PAVITT	0	1	Tot_PPI
Not elsewhere classified	-0.9	-1.5	0.0
High-technology	-1.2	0.4	0.0
Medium-high-technology	-4.0	0.9	0.0
Medium-low-technology	-1.1	0.6	0.0
Low-technology	-4.8	2.4	0.0
Knowledge-intensive services	-1.3	-0.6	0.0
Lessknowledge-intensive services	-1.4	0.2	0.0
Tot_PPI	-0.9	-1.5	0.0

Source: Authors' calculation on CIS and FRAME data - reference year 2012

Table 4.6 – Value added over Turnover (%) for technological intensity categories and values of PPI: comparison of methods A and B

Rel Diff VA/PE	(A-B)/A PPI			Linked datasets (method B) PPI		
PAVITT	0	1	Tot_PPI	0	1	Tot_PPI
Not elsewhere classified	2.9	13.5	16.7	19.2	12.5	15.6
High-technology	25.4	33.3	31.0	23.1	29.7	27.7
Medium-high-technology	21.3	23.2	22.8	21.0	22.0	21.8
Medium-low-technology	16.9	19.7	18.5	15.7	19.0	17.5
Low-technology	20.5	21.3	21.0	21.7	21.6	21.6
Knowledge-intensive services	34.9	39.4	38.2	32.6	41.3	38.6
Lessknowledge-intensive services	14.2	16.4	15.2	15.4	17.1	16.1
Tot_PPI	18.0	20.6	19.5	18.1	20.3	19.3

Source: Authors' calculation on CIS and FRAME data - reference year 2012

As expected, the analysis of Table 4.1 shows greater value added per person employed values for companies with higher Internet connection speed than companies connecting at speeds below 10 Mbit/sec confirming a positive correlation between potential for greater use of the technologies and higher economic efficiency.

Similarly on the base of the CIS-Frame data we can convey that there is a positive correlation between higher values of the economic performance indicators (value added per person employed and ratio between value added and turnover) and the implementation of innovation activities.

In both cases, the results presented here are partial. Other aspects related to the compliance of data obtained with the expertise of the phenomena or efficiency of the

estimators detected require further study. However, the proposed methods lead to sensible conclusions both from the mathematical and subject-matter points of view.

5. Conclusions and future work

In this work we dealt with methodologies suitable to exploit the potential of data integration of two sources of business data. The datasets considered in this study are represented by an exhaustive archive, called Frame, supplying the main balance sheet variables for the whole population of enterprises as defined by the SBS Regulation and a sample survey dataset adding thematic variables not registered in the Frame. A macrointegration and a microintegration approach were tested. A general comparison of the two strategies is a difficult task as it should depend on the available data and on the aim of the integration project. In any case, subject-matter experts should always be involved in the quality analysis of each integration project. As for the macrointegration approach methods referring to reconciliation of tabular data were tested: Balancing that can simultaneously deal with a set of tables, and IPF that deals with a single table, independently on any other information. Balancing generated inadmissible solutions. On the other side, IPF was not deemed flexible enough to be applied on a large set of tables. Consequently, they were not further investigated in our case study. Eliminating these drawbacks, microintegration was preferred. Microintegration was implemented through calibration taking into account the detail of domains of estimates that allow for convergence. In particular, the calibration of the linked dataset, i.e. method B, may be preferred as the direct calibration of the survey dataset reduces the importance of the Frame. Moreover, the calibration on known totals stemming from disseminated estimates did not always achieve convergence. Further studies on calibration methodology will be done considering different sets of auxiliary variables to produce alternative indicators. Moreover we could test also the possibility to define different sets of weights for different target indicators.

Another way to exploit the information supplied by the Frame in favor of sample survey is to consider the Frame as the business register to draw samples using economic variables not elsewhere available.

Finally we should mention the possibility of simultaneously integrating the two sample surveys and the Frame. This objective will be pursued by statistical matching techniques (D'Orazio et al., 2006). The 'common universe' allowing for statistical matching analysis on CIS and ICT data is only the Industry (NACE divisions 10 to 33), thus excluding the service sector. When applying this method, Frame would represent the overlapping information linking the two surveys. The methods to be used to best exploit the results of a statistical matching procedure will be studied.

References

AA.VV. 2012. *Essnet on Data Integration Final Reports.* https://ec.europa.eu/eurostat/cros/content/data-integration_en.

Bethel, J. 1989 Sample allocation in multivariate surveys. *Survey methodology*, 15. 1989: 47-57.

Casciano, C., P.D. Falorsi, S. Filiberti, A. Pavone, and G. Siesto. 2006. Principi e metodi per il calcolo delle stime finali e la presentazione sintetica degli errori di campionamento nell'ambito delle rilevazioni strutturali sulle imprese. *Rivista di Statistica Ufficiale*, N. 1. 2006: 67-102. Roma: Istat.

Deville, J.C., and C.E. Särndal. 1992. Calibration Estimators in Survey Sampling. *Journal of the American Statistical Association*, N. 87. 1992: 376-382.

D'Orazio M., M. Di Zio, and M. Scanu. 2006. *Statistical Matching, Theory and Practice*. Hoboken, New Jersey, U.S.: Wiley.

Eurostat. 2012. *Final Report ESSnet on Linking of Microdata on ICT Usage*, (link: ec.europa.eu/eurostat/documents/341889/725524/2010-2012-ICT-IMPACT-2012-Final-report.pdf).

Leadership Group SAM. 2003. *Handbook on Social Accounting Matrices and Labour Accounts, Population and Social Conditions* 3/2003/E/N23.

Nicolardi, V. 1998. *Un sistema di bilanciamento per matrici contabili di grandi dimensioni*, (*A balancing method for big accounting matrices*). Quaderni di ricerca, N. 4, 1998. Roma: Istat.

Spiezia, V. 2011. Are ICT Users More Innovative? An analysis of ICT-enabled Innovation in OECD Firms. OECD *Journal: Economic Studies*, Vol. 2011/1, http://www.oecd.org/economy/growth/are%20ict%20users%20more%20innovative.pdf.

Information for the authors

The Rivista di statistica ufficiale is the international journal published by the Italian National Institute of Statistics - Istat. It strongly encourages and welcomes works related both to the development of official statistics and to the functioning of statistical systems.

The *Rivista di statistica ufficiale* represents therefore an area of discussion open to researchers and experts from statistical institutions and other scientific organisations both at national and international level.

The scientific contributions selected deal with cross-cutting topics ranging from statistical methods and indicators to economics, public finance, demography, society and environment.

All the proposals are double-blind reviewed by experienced referees in the different areas of interest who evaluate their originality, the quality of the papers, the validity of the conclusions, the importance and the impact of the researches and analyses illustrated.

To contact the editorial board and to send articles, please write to: rivista@istat.it.

La Rivista di Statistica Ufficiale accoglie lavori che hanno come oggetto la misurazione e la comprensione dei fenomeni sociali, demografici, economici ed ambientali, la costruzione di sistemi informativi e di indicatori come supporto per le decisioni pubbliche e private, nonché le questioni di natura metodologica, tecnologica e istituzionale connesse ai processi di produzione delle informazioni statistiche e rilevanti ai fini del perseguimento dei fini della statistica ufficiale.

La Rivista di Statistica Ufficiale si propone di promuovere la collaborazione tra il mondo della ricerca scientifica, gli utilizzatori dell'informazione statistica e la statistica ufficiale, al fine di migliorare la qualità e l'analisi dei dati.

La pubblicazione nasce nel 1992 come collana di monografie "Quaderni di Ricerca ISTAT". Nel 1999 la collana viene affidata ad un editore esterno e diviene quadrimestrale con la denominazione "Quaderni di Ricerca - Rivista di Statistica Ufficiale". L'attuale denominazione, "Rivista di Statistica Ufficiale", viene assunta a partire dal n. 1/2006 e l'Istat torna ad essere editore in proprio della pubblicazione.

1B012016003000000

ISSN 1828-1982

€ 15,00