Proceedings of Statistics Canada Symposium 2001
Achieving Data Quality in a Statistical Agency: A Methodological Perspective

A METHOD FOR SHORT-TERM ESTIMATION OF LABOUR INPUT USING CURRENT PRELIMINARY DATA FROM ADMINISTRATIVE SOURCES HAVING COVERAGE ERRORS

Alessandro Pallara, Ciro Baldi, Piero Demetrio Falorsi, Raffaella Succi[1]

Aldo Russo[2]

ABSTRACT

This paper proposes a method for short-term estimation of labour input indicators using administrative data from Social Security Database (SSD). The rationale for developing this methodology originated from the need for national statistical offices to meet the standard quality criteria in the *Regulation No. 1165/98* of the European Community concerning short-term business statistics. Information requested in the Regulation involves such a detailed disaggregation that it would impossible to meet all the requirements through direct data collection. Administrative data, because of their timeliness and detailed coverage, represent a valuable source for obtaining estimates of business population aggregates that meet such quality requirements.

KEY WORDS: Administrative data; Short-term indicators; Coverage error; Superpopulation model; Selection bias

1. INTRODUCTION

The use of administrative sources for statistical purposes is receiving increasing attention in National Statistical Institutes (NSI) as a means to keep costs and response burden of surveys at a reasonable level and it is also suggested in official regulations of international statistical organizations as a recommended practice for supplementing or substituting direct data collection on business, farms and institutions in order to improve quality of survey data.

Successful use of administrative data for statistical purposes obviously depends on the availability of effective source. In Italy social security (*Istituto Nazionale della Previdenza Sociale*, *INPS*) database (SSD) represents the most important administrative source for data on employment and wages in private and public sector and it has been used as the primary source of information for the design of a Quarterly Survey of Employment, Payrolls and Labour Cost for all enterprises in the industry and service sector, combining data from direct business surveys and administrative data.

This survey is primarily aimed at satisfying the requirements of the European *COUNCIL REGULATION No. 1165/98 concerning short-term business statistics* (STS), regarding provision in each Member state of quarterly estimates of: *i*) number of persons employed, *ii*) wages and salaries and *iii*) hours worked.

The survey design project consists of two parts. First part is aimed at setting up a method for current estimation of two variables, namely: *i*) number of persons employed, *ii*) gross wages and salaries, for enterprises with less than 500 employees, using only administrative data. The second part will focus on the problems related to combining estimates for the enterprises with less than 500 employees, yielded using only the administrative source, with the information obtained for the largest enterprises, which are directly surveyed each month by the Italian NSI.

[1] Italian National Institute of Statistics, Via Depretis 74/b, 00184, Rome (Italy).
[2] Department of Political Institutions and Social Sciences, University Roma Tre; Rome (Italy).

The relevance of this project may be thoroughly evaluated in view of the following issues: (*i*) information requested in STS Regulation involves such a detailed sectoral disaggregation (2-digit level of Nace Rev.1 classification) that it would impossible to meet all the requirements through direct data collection; (*ii*) short term statistics on employment and wages have severe coverage problems in Italy, in that, at the moment, only one direct sample survey is currently executed, covering very large businesses (>500 employees) and specific groups of economic activities; (*iii*) using administrative data has in this case the advantage that definitions are in many cases coherent with the requirements of STS Regulation: it is therefore possible to obtain most information without additional loading on enterprises.

This paper will focus on the first part of the survey design mentioned above, illustrating the methodology developed for yielding quarterly estimates of employees and wages and salaries using administrative data as the only source of information. The proposed estimation method extends some results obtained in previous studies (cf. Falorsi *et al.*, 2000). Estimates are obtained through a prediction model based on current data on a subset of units from SSD, which is extended to unobserved *y* values for all the units in the register of Social Security. Indeed, using SSD data involve some problems:

i. the subset of units with current data represents a *non-random* sample of the population. To the extent that the data generating mechanism is informative on the estimation process, estimates may be subject to selection bias (Royall, 1988). One tool to balance the estimates is constructing the model within homogeneous subgroups (Hidiroglou et al., 1995);

ii. SSD may have *coverage errors*, because of problems with keeping-up-to-date the register (with entries to, removals from, transformation of units) or late fulfilment for some units of their administrative duties. Hence SSD may differ systematically from the target population. Inclusion in the administrative register is modelled as the outcome of a Bernoulli process, with "success" probability fixed within subgroups of the population.

The paper is organised as follows: in section 2 the parameters of interest are introduced and then it is defined the statistical model used for the estimation; in section 3 an explicit form for the estimates is presented, discussing some practical aspects involved in the computation of the parameters of interest as well as of their Mean Square Errors; finally, some empirical results are presented in section 4, concerning quarterly estimates of the level of employment and wages and salaries as well as their relative MSE.

2. PARAMETERS OF INTEREST AND STATISTICAL MODELS

Let P_t be the (finite) population of active enterprises for current time interval *t* (e.g., month or quarter). The parameter of interest are the totals of the variables *employment* and *wages and salaries* for the target population P_t, namely:

$$Y_t = \sum_{i \in P_t} y_{ti} \qquad (1)$$

where y_{ti} denotes the value of the variable of interest *y* (e.g. number of persons employed) for the enterprise *i* at time *t*. We are interested in estimation of Y_t using auxiliary information from the administrative source.

Each enterprise registered in the SSD when sending each month its remittance has to fill a form containing information on: *i*) number of employees, *ii*) wages and salaries, *iii*) social security costs.

The basic unit forming the register does not correspond to any of the standard definitions, like those listed, e.g., in the *Council Regulation* (EEC) *No. 696/93 on the statistical units for the observation and analysis of the production system in the Community*. Each form pertains to only one enterprise; however an enterprise may fill more than one form and it is not easy to map all forms belonging to the same enterprise. Moreover, the list of all units registered, for each reference period, in SSD suffers of some overcoverage problem, because typically it takes some months before a dead unit cancels out from the social security register.

Because of differences among units in the transmission procedures, in each reference period, information for current time interval *t* are available only for a subset of units (currently, some 300,000 out of about 1,100,000 units in the population of enterprises with less than 500 employees in the industry and service sector in Italy). Although quite large, this subset of units is not a random sample, selected according to a specific design. Indeed, non randomness of the observed units may involve bias of the standard estimators of (1).

Let A_t be the set of *active* enterprises in the SSD. A_t can be regarded as the available representation of the target population P_t according to the administrative register. Because of the presence of overcoverage errors in the administrative source, it is $P_t \subseteq A_t$. As mentioned previously, this is due mainly to the circumstance that it may take some months before a dead unit cancels out from the social security register. Therefore, A_t may include recent deaths (among non sample units) that we have to account for in order to yield unbiased estimates of the variable of interest; therefore the total of interest may be represented as

$$Y_t = \sum_{i \in A_t} y_{ti} \, \lambda_{ti} \qquad (2)$$

where λ_{ti} is an indicator variable, which equals 1 if enterprise $i \in P_t$ (i.e. it is an *active* unit and therefore *included correctly* in administrative register) and equals 0 otherwise .

In order to introduce the estimation method, let partition A_t as follows:

$$A_t = s_t \cup \bar{s}_t \qquad (3)$$

where

- s_t is the subset of units on which we observe the value of the variable of interest for the current time interval *t*; in what follows we will refer to s_t as the *sample.* Note that, $\forall\, i \in s_t$, we have λ_{ti} =1, since s_t includes the enterprises which, having sent their payroll forms for the current period to social security using a special transmission procedure, are deemed as *active* with *certainty*;
- $\bar{s}_t$ is the set of units on which we do not observe the variable of interest at time *t*, hereafter referred to as the *nonsample*.

The proposed estimation method is based on a predictive model, assuming that a relation exists between the variable of interest and some auxiliary variables. The parameters of the models are then estimated on the subset of units containing data for the reference period. When modelling the relationship between the *sample* and the *nonsample* proper attention has to be given to two different potential sources of bias:

- non *randomness* of s_t. As mentioned above, s_t includes units which choose by their own to send the payroll forms by a special computer transmission procedure, while the *normal* transmission procedure to social security is by mail. Among the proposed methods to deal with *selection bias* (cf. Royall, 1988) the solution adopted in this paper has been to fit the model within subgroups of the population (*regression* groups), assuming that the model coefficients are the same for the units in each regression group. A good criterion for defining the regression groups is proposed in Hidiroglou *et al*. (1995);
- *coverage errors* in $\bar{s}_t$. List errors due to unregistered deaths in the administrative register may involve that some subsets of A_t differ systematically from the corresponding target population. It may be assumed that a partition of A_t exists such that in each of the above subsets, the probability of correct inclusion in the administrative source is nearly equal for all the units of the subset; in what follows these subsets will be denoted as *register error groups.*

Hence, let

– $\{A_{tw}\}_{w=1,..,W}$ be a partition of A_t into *W regression groups*. We will denote with $A_{tw} = s_{tw} \cup \bar{s}_{tw}$ the *w*-th *regression* group, namely the subset of units belonging to A_t for which the *w*-th statistical models is defined;

– $\{A_{tr}\}_{r=1,....,R}$ be a partition of A_t, into *R register error groups*. It is assumed that within each *register error* group A_{tr} the probability of coverage errors is fixed and roughly constant over the time span of the estimation.

Denoting with E (·) and C (·), the model expectation and the model variance, respectively, it is assumed that for $w = 1,\ldots,W$ the following superpopulation model holds:

$$\begin{aligned} &\mathrm{E}\left(y_{ti} \middle| \lambda_{ti}=1\right)=\boldsymbol{\beta}'_{tw}\,\mathbf{x}_{ti} && \forall i \in A_{tw} \\ &\mathrm{C}\left(y_{ti}, y_{t'i'} \mid \lambda_{ti}, \lambda_{t'i'}\right)=\begin{cases} c_{ti}\sigma^2_{tw} & \text{for } (t=t')\cap(i=i')\cap(\lambda_{ti}=1) \\ 0 & \text{otherwise} \end{cases} && \forall (i,i') \in A_{tw} \end{aligned} \tag{4}$$

where $\boldsymbol{x}_{ti}$ is a vector of M_{tw} auxiliary variables for the enterprise *i* at time *t*, $\boldsymbol{\beta}_{tw}$ is a column vector of M_{tw} regression coefficients and c_{ti} is a known constant related to the dimension of the enterprises reported in the administrative source.

Furthermore, we assume that *correct inclusion* in SSD is defined by a sequence of independent Bernoulli trials, independently of the value of *y*. These trials are assumed to be identically distributed within each register error group A_{tr} ($r = 1,\ldots,R$). Therefore, it is postulated that for $r = 1,\ldots,R$ the following model holds:

$$\begin{aligned} &\mathrm{E}(\lambda_{ti}) = p_{ti} = \theta_{tr} && \forall i \in A_{tr} \\ &\mathrm{C}\left(\lambda_{ti}, \lambda_{t'i'}\right)=\begin{cases} \theta_{tr}\,(1-\theta_{tr}) & \text{for } (t=t')\cap(i=i') \\ 0 & \text{otherwise} \end{cases} && \forall (i,i') \in A_{tr} \end{aligned} \tag{5}$$

With this set-up, the following model is derived for $w = 1,\ldots,W$ and $r = 1,\ldots,R$

$$\begin{aligned} &\mathrm{E}(y_{ti}\ \lambda_{ti}) = \boldsymbol{\beta}'_{tw}\,\boldsymbol{x}_{ti}\,\theta_{tr} && \forall i \in A_{t(w,r)} \\ &\mathrm{C}(y_{ti}\lambda_{ti}, y_{t'i'}\ \lambda_{t'i'})=\begin{cases} c_{ti}\ \sigma^2_{t\,w}\,\theta_{tr} + (\boldsymbol{\beta}'_{tw}\,\mathbf{x}_{ti})^2\,\theta_{tr}(1-\theta_{tr}) & \text{for } (t=t')\cap(i=i') \\ 0 & \text{otherwise} \end{cases} && \forall (i',i) \in A_{t(w,r)} \end{aligned} \tag{6}$$

where $A_{t(w,r)} = A_{tw} \cap A_{tr}$.

Using a predictive approach, having obtained the estimates $\tilde{\boldsymbol{\beta}}_{tw}$ ($w = 1,\ldots,W$) and $\tilde{\theta}_{tr}$ ($r = 1,\ldots,R$), of the parameters involved in models (4) - (6), the estimate of the total Y_t is given by

$$\tilde{Y}_t = \sum_{w=1}^{W} \sum_{r=1}^{R} \left[\sum_{i \in s_{t,(w,r)}} y_{ti} + \sum_{i \in \bar{s}_{t,(w,r)}} \tilde{\boldsymbol{\beta}}'_{tw}\,\boldsymbol{x}_{ti}\,\tilde{\theta}_{tr} \right] \tag{7}$$

where $s_{t(w,r)} = s_t \cap A_{t(w,r)}$ and $\bar{s}_{t(w,r)} = \bar{s}_t \cap A_{t(w,r)}$.

That is, conditionally on the partition $A_{t(w,r)}$ the estimation of population total of variable *y* is similar to a *model based inference* for *domains totals* (Chambers, 1997).

In next section it is discussed in detail an explicit form for the estimates $\tilde{\boldsymbol{\beta}}_{tw}$ ($w = 1,\ldots,W$) and $\tilde{\theta}_{tr}$ ($r = 1,\ldots,R$) as well as some practical aspects involved in the definition of the partitions $\{A_{tw}\}_{w=1,\ldots,W}$ and $\{A_{tr}\}_{r=1,\ldots,R}$.

3. ESTIMATION METHOD

There are two main issues related to the estimation method proposed:

(i) estimates of $\boldsymbol{\beta}_{tw}$ and θ_{tr} have been obtained separately for each set of parameters;

(ii) the most important problem to solve in order to yield (roughly) unbiased estimates has been the definition of the partitions $\{A_{tw}\}_{w=1,\ldots,W}$ and $\{A_{tr}\}_{r=1,\ldots,R}$.

3.1 Estimation of $\boldsymbol{\beta}_{tw}$

In order to solve the problem of selection bias, the estimates of $\boldsymbol{\beta}_{tw}$ in (7) have been obtained separately in each regression group, trying to define a partition $\{A_{tw}\}_{w=1,\ldots,W}$ such that in each group there was as much homogeneity as possible of the $\boldsymbol{\beta}$ parameters between the *sample* and the *nonsample* part of the group. The availability at time *t* of data on the variables of interest for all the enterprises in SSD referred to a year earlier the current reference period allows to verify the homogeneity of $\boldsymbol{\beta}$'s using actual data for the previous period, assuming stability over time of the models.

The regression groups have been defined for three subsets of A_t having a different amount of auxiliary information: (*a*) units established from more than one year on which a complete set of auxiliary variables is available. The auxiliary variables used are number of employees, wages and salaries and other components of labour costs in the previous year and the total number of enterprises registered (about 310 groups have been defined for these units); (*b*) units established from more than one year on which the *only* auxiliary variable available is the number of registered enterprises (100 groups, approximately); (*b*) units established from less than one year. For these units the auxiliary variables available are the number of employees at the time in which the enterprise first registered in SSD and the total number of enterprises (60 groups approximately). Moreover, for each unit included in A_t information on geographical location and economic activity (NACE REV.1) is also available.

The definition of the partition has been done for each of the three above subsets (a), (b) and (c) using natural classes resulting from cross-classification of *economic activity* (2-digits of NACE rev. 1), *geographical area* (regions) and *size class* (for subsets (a) and (c) only) defined in terms of number of employees [observed either in the previous year (subset (a)), or at the time in which the enterprise first registered in SSD (subset (c))]. Using the above criteria for regression groups definition aims to obtaining subsets as homogeneous as possible which result in (nearly) unbiased estimates at low levels of aggregation and regression groups which do not cut across the *domains of interest* for publishing final estimates.
In some particular cases, it has been necessary to collapse groups not having a sufficient sample size, in order to provide reliable estimates of regression coefficients; conversely, with some groups a further *subdivision* of the partition obtained has been done (using, e.g., 3-digits of NACE rev.1 classification) in order to ensure a greater homogeneity of the estimated regression coefficients between the sample and the nonsample part of the resulting groups.

Estimates of the coefficients for the generic group w ($w = 1,\ldots,W$) have then been obtained as

$$\tilde{\boldsymbol{\beta}}_{tw} = \left[\sum_{i \in s_{tw}} \boldsymbol{x}_{ti} \boldsymbol{x}_{ti}^{'} / c_{ti} \right]^{-1} \sum_{i \in s_{tw}} \boldsymbol{x}_{ti}\, y_{ti} / c_{ti} \qquad (w = 1,\ldots,W). \tag{8}$$

3.2 Estimation of θ_{tr}

In order to deal with the potential source of bias due to *coverage* errors in SSD, the parameters θ have been introduced in (5) for defining the probability for a unit in the register to belong to the target population. This parameters have been assumed to be constant within each subset A_{tr} of the partition $\{A_{tr}\}_{r=1,\ldots,R}$ introduced in section 2. The procedure for defining the partition aimed at obtaining subgroups which are maximally internally homogenous with respect to the probability of correct inclusion in the register while maximising *between groups* heterogeneity. Furthermore the partition has been defined in order to ensure stability of the probability over time; this was done trying to reduce the potential bias related to the information set-up of SSD, such that data on coverage errors of the administrative register are available only for time periods preceding one year the *current* reference period. The partition $\{A_{tr}\}_{r=1,\ldots,R}$ has been obtained modelling the relationship between the dichotomous variable λ introduced in (2), (and available at t with reference to one year before the current period), and a set of covariates in SSD. The partition is based on classes resulting from cross-classification of the values of the most influential covariates (age of the enterprise, geographical region, size class, economic activity) selected through non parametric regression (Breiman et al., 1984) .

For the r-th ($r = 1,\ldots,R$) register error group, the estimate of the probability of correct inclusion at time t has been obtained as the *proportion* of units belonging to the target population observed one year earlier the current time, namely

$$\tilde{\theta}_{tr} = \hat{\theta}_{t-12r} = \sum_{i=1}^{N_{t-12r}} \lambda_{t-12\,i} \Big/ N_{t-12r} \qquad (r = 1,\ldots,R) \tag{9}$$

where N_{t-12r} denotes the number of units belonging to the r-th register error group a year earlier the current reference period, where it is $\mathrm{E}\left(\hat{\theta}_{t-12r}\right) = \theta_{t-12r}$.

Substituting expressions (8) and (9) in (7), and after some algebra, the estimator of population total may be represented in linear form as

$$\tilde{Y}_t = \sum_{i \in s_t} y_{ti}\, w_{ti} \tag{10}$$

being

$$w_{ti} = 1 + \left[\sum_{r=1}^{R} \sum_{l \in s_{t,(w,r)}} \boldsymbol{x}_{tl}\, \hat{\theta}_{t-12r} \right]^{'} \left[\sum_{r=1}^{R} \sum_{l \in s_{t,(w,r)}} \boldsymbol{x}_{tl} \boldsymbol{x}_{tl}^{'} / c_{tl} \right]^{-1} \boldsymbol{x}_{ti} / c_{ti} \qquad \text{for } i \in A_{tw} \tag{11}$$

where l denotes the l-th enterprise in $s_{t(w,r)}$.

3.3 Mean Square Error of $\tilde{\tilde{Y}}_t$

The mean square error (MSE) of estimator (7) is the sum of model variance and of the square of model bias:

$$\mathrm{MSE}(\tilde{\tilde{Y}}_t) = \mathrm{E}\left(\tilde{\tilde{Y}}_t - Y_t\right)^2 = \mathrm{V}(\tilde{\tilde{Y}}_t - Y_t) + \left[\mathrm{E}(\tilde{\tilde{Y}}_t - Y_t)\right]^2 = \mathrm{V}(\tilde{\tilde{Y}}_t - Y_t) + \left[\mathrm{Bias}(\tilde{\tilde{Y}}_t)\right]^2 . \tag{12}$$

There are two main sources of model bias:

1. a poor definition of the regression groups, such that the vector of coefficients $\boldsymbol{\beta}$ may differ systematically between the sample and the nonsample part. It may then be assumed that in each regression group the vector of coefficients $\boldsymbol{\beta}_{\bar{s}_{tw}}$ ($w = 1,\ldots,W$) of the nonsample part may be represented as the sum of $\boldsymbol{\beta}_{tw}$ and a vector $\boldsymbol{\alpha}_{tw}$ of fixed effects; namely $\boldsymbol{\beta}_{\bar{s}_{tw}} = \boldsymbol{\beta}_{tw} + \boldsymbol{\alpha}_{tw}$;
2. an invalid assumption of a constant probability of correct inclusion in the register between two consecutive years.

Considering the above points, an expression for first order approximation of the square of bias of $\tilde{\bar{Y}}_t$ may be given by:

$$\left[\mathrm{Bias}(\tilde{\bar{Y}}_t)\right]^2 \cong \left[\sum_{w=1}^{W}\sum_{r=1}^{R}\left[\boldsymbol{\beta}_{tw}\,\theta_{t-12r} - (\boldsymbol{\beta}_{tw}+\boldsymbol{\alpha}_{tw})\theta_{tr}\,\right]' \sum_{i\in\,\bar{s}_{t(w,r)}} \boldsymbol{x}_{ti}\right]^2 .$$

The variance of $\tilde{\bar{Y}}_t$ in (12) can be represented as

$$\mathrm{V}(\tilde{\bar{Y}}_t - \bar{Y}_t) = \mathrm{V}_1 + \mathrm{V}_2 = \mathrm{V}\left[\sum_{w=1}^{W}\sum_{r=1}^{R}\sum_{i\in\bar{s}_t(r,w)} y_{ti}\,\lambda_{ti}\right] + \mathrm{V}\left[\sum_{w=1}^{W}\sum_{r=1}^{R}\tilde{\theta}_{tr}\,\tilde{\boldsymbol{\beta}}'_{tw}\sum_{i\in\bar{s}_{t(r,w)}}\boldsymbol{x}_{ti}\right]. \quad (13)$$

With the assumptions of model (6) V_1 is given by

$$\mathrm{V}_1 = \sum_{w=1}^{W}\sum_{r=1}^{R}\sum_{i\in\bar{s}_t(w,r)} c_{ti}\sigma^2_{tw}\,\theta_{tr} + \theta_{tr}(1-\theta_{tr})\,\boldsymbol{\beta}'_{tw}\,\boldsymbol{x}_{ti}\,\boldsymbol{x}'_{ti}\,\boldsymbol{\beta}_{tw} \quad . \quad (14)$$

First order approximation of V_2 may be expressed as

$$\mathrm{V}_2 \cong \mathrm{V}\left[\sum_{w=1}^{W}\sum_{r=1}^{R}\sum_{i\in\bar{s}_{t(r,w)}}\left[\theta_{t-12r}\,\tilde{\boldsymbol{\beta}}'_{tw}\,\boldsymbol{x}_{ti} + \boldsymbol{\beta}'_{tw}\,\boldsymbol{x}_{ti}\tilde{\theta}_{t-12r}\right]\right] = \quad (15)$$

$$= \sum_{w=1}^{W}\sum_{r=1}^{R}\sum_{i\in\bar{s}_{t(r,w)}}\left[\theta^2_{t-12r}\,\sigma^2_{tw}\,\boldsymbol{x}'_{ti}\left(\sum_{i\in s_{tw}}\boldsymbol{x}_{ti}\,\boldsymbol{x}'_{ti}/c_{ti}\right)^{-1}\boldsymbol{x}_{ti} + (\theta_{t-12r}(1-\theta_{t-12r}))\,\boldsymbol{x}'_{ti}\boldsymbol{\beta}_{tw}\boldsymbol{\beta}'_{tw}\,\boldsymbol{x}_{ti}\right].$$

Using actual data on the entire population which are available, as mentioned, with a reference period preceding one year current reference period, it is possible to obtain an estimate of the MSE as

$$\widetilde{\mathrm{MSE}}(\tilde{\bar{Y}}_t) = \left(\tilde{\bar{Y}}_t - \bar{Y}_t\right)^2 \quad (16)$$

where, using expression (12), it is $\mathrm{E}\left[\widetilde{\mathrm{MSE}}(\tilde{\bar{Y}}_t)\right] = \mathrm{E}\left(\tilde{\bar{Y}}_t - \bar{Y}_t\right)^2 = \mathrm{MSE}(\tilde{\bar{Y}}_t)$.

Sample estimate of the variance at current time may be found plugging sample estimates of the unknown parameters σ_{tw}^2, $\boldsymbol{\beta}_{tw}$, θ_{t-12r}^2, θ_{t-12r} in expressions (14) and (15), thus yielding:

$$\tilde{\mathrm{V}}(\tilde{Y}_t - Y_t) = \sum_{w=1}^{W}\sum_{r=1}^{R}\sum_{i\in\bar{s}_{t(w,r)}} c_{ti}\tilde{\sigma}_{tw}^2\,\tilde{\theta}_{tr} + \theta_{tr}\left(1-\tilde{\theta}_{tr}\right)\tilde{\boldsymbol{\beta}}_{tw}'\,\boldsymbol{x}_{ti}\,\boldsymbol{x}_{ti}'\,\tilde{\boldsymbol{\beta}}_{tw} + \tag{17}$$

$$+\sum_{w=1}^{W}\sum_{r=1}^{R}\sum_{i\in\bar{s}_{t(w,r)}}\left[\left(2\hat{\theta}_{t-12r}^2-\hat{\theta}_{t-12r}\right)\tilde{\sigma}_{tw}^2\,\boldsymbol{x}_{ti}'\left(\sum_{i\in s_{tw}}\boldsymbol{x}_{ti}\,\boldsymbol{x}_{ti}/c_{ti}\right)^{-1}\boldsymbol{x}_{ti} + \left(\hat{\theta}_{t-12r}\left(1-\hat{\theta}_{t-12r}\right)\right)\boldsymbol{x}_{ti}'\tilde{\boldsymbol{\beta}}_{tw}\,\tilde{\boldsymbol{\beta}}_{tw}'\,\boldsymbol{x}_{ti}\right]$$

with

$$\tilde{\sigma}_{tw}^2 = \frac{1}{n_{tw}-M_{tw}}\sum_{i=1}^{n_{tw}}\left(y_{ti}-\tilde{\boldsymbol{\beta}}'_{tw}\,\boldsymbol{x}_{ti}\right)^2/c_{ti}\quad; \tag{18}$$

where n_{tw} is the number of units in the sample s_{tw}, while ($2\hat{\theta}_{t-12r}^2-\hat{\theta}_{t-12r}$) represents an unbiased estimate of the square of the probability θ_{t-12r}^2.

Therefore one year after the current time, an unbiased estimate of the square bias may be obtained as

$$\left[\tilde{\mathrm{Bias}}(\tilde{Y}_t)\right]^2 = \left(\tilde{Y}_t - Y_t\right)^2 - \tilde{\mathrm{V}}(\tilde{Y}_t - Y_t)\,. \tag{19}$$

In order to obtain a *current* estimate of the bias it may be useful to use an approach similar to the one typically adopted with *generalised variance functions* (Wolter, 1985), assuming that the relative squared bias is a decreasing function of the *sampling fraction* (*n/N*). The relative square bias may then be modelled as

$$\ln\left[\left(\tilde{\mathrm{Bias}}(\tilde{Y}_t)\right)^2/\tilde{Y}_t^2\right] = a_t + b_t\ln\left(n_t/N_t\right)+\varepsilon_t \tag{20}$$

where a_t and b_t represent fixed coefficients and ε_t is a random residual. Assuming that the coefficients a_t and b_t are nearly constant over time, an estimate of the *current* squared bias may be obtained as:

$$\left(\tilde{\tilde{\mathrm{Bias}}}(\tilde{Y}_t)\right)^2 = \tilde{Y}_t^2\,exp\left(\tilde{a}_{t-1}+\tilde{b}_{t-1}\ln\left(n_t/N_t\right)\right) \tag{21}$$

where $\tilde{a}_{t-1}$ and $\tilde{b}_{t-1}$ are the estimates of the parameters *a* and *b* obtained one year before the current time.

4. SOME EMPIRICAL EVIDENCES

The methodology presented in sec. 2 and 3 is now applied to estimating total employment and gross wages and salaries in Italy in the first quarter of 2000 for the population of enterprises with less than 500 employees in the industry and service sector.

For each variable of interest, estimates of the variance and the squared bias have been computed using expressions (17) and (19), respectively. In order to assess performance of the estimates and the relative influence of the bias component of the mean square error with respect to the variance, the following measures have been calculated: (*i*) percent Relative Root Mean Square Error: $\mathrm{RRMSE}(\tilde{Y}) = 100\sqrt{\tilde{\mathrm{MSE}}(\tilde{Y})}/\tilde{Y}$, which represents the amount of the root of the MSE with respect to the estimates; (*ii*) the percent Relative Bias: $\mathrm{RB}(\tilde{Y}) = 100\,\tilde{\mathrm{Bias}}(\tilde{Y})/\tilde{Y}$; the percent Relative Component Bias: $\mathrm{RCB}(\tilde{Y}) = 100\left[\tilde{\mathrm{Bias}}(\tilde{Y})\right]^2/\tilde{\mathrm{MSE}}(\tilde{Y})$.

In table 1, appearing in next page, the above indexes are shown for employment and gross wages and salaries, for total economy and for each industry sector.

Observation of the tables suggests a good precision of the estimates in terms of RRMSE: estimates of the number of employees and gross wages and salaries for the total of the population have a RRMSE of 0,45% and 0,03% respectively. As expected, variability of the estimates is mainly due to the bias components.

5. DISCUSSION AND CONCLUDING REMARKS

In this paper we have illustrated a methodology for obtaining estimates of labour input indicators using administrative data. A growing attention is being devoted in many NSI to data obtained from administrative sources as a means to complement or substitute direct surveys, in order to meet an increasing demand for timely and detailed statistical data, especially for survey of businesses, farms and institutions.
In particular, the social security system allows a rich set of variables on the labour market and poses no additional burden on enterprises. However, using SSD for constructing short-term statistics on employment and wages has some shortcomings that we have tried to address in this paper. Specifically:

1. basic unit in the SSD do not correspond to a definition of the statistical unit suitable for the analysis of an economic system;
2. the subset of data with information for the current time interval used to estimate model coefficients is not a random sample;
3. data gathering procedures and problems with keeping-up-to-date SSD (with entries to, removals from, transformation of units) involve coverage problems in the administrative source such that the list of active units defined through SSD may differ systematically from the target population.

Delineation of units has involved a preliminary processing of SSD to map all payroll forms deemed to belong to the same economic entity in a single record, using different key variables to match different records in the database.

Selection bias due to non randomness of the subset of observed units for the current time interval has been dealt with fitting the model within homogeneous subgroups of the population (*regression* groups).

Overcoverage errors largely due to unregistered deaths in the SSD have been accounted for modelling the probability for a unit in the administrative register to belong to the target population within subsets of SSD (*register error groups*) and assuming that in each subset the probability is nearly equal.

There are some open problems with the estimation method herein proposed, concerning primarily an explicit modelling of coverage errors using, e.g. an approach base on state-space models (cf. Tam; 1987). This modelling would allow to account for variability over time of probability of correct inclusion in the administrative register depending from variation in the economic cycle. A similar approach may be adopted for modelling $\boldsymbol{\beta}$ parameters using a recursive approach for updating current estimation through an updating equation starting from the observed values of the previous period.

TABLE 1. – MSE and its components for the variables *Employment* and *Gross wages and salaries* by industry division for the population of enterprises with less than 500 employment in Italy - I quarter 2000

Industry division	Employment				Gross wages and salaries		
	Value Observed	RRMSE	RB	RCB	RRMSE	RB	RCB
Manufacturing	3,236,249	0.87	0.85	95.95	0.03	0.028	87.12
Construction	746,553	1.03	1.00	93.95	1.88	1.86	98.85
Wholesale and Retail Trade	1,157,101	1.28	1.27	98.10	0.85	0.82	94.66
Accomodation, Food and Beverages	326,975	2.31	2.28	97.49	4.83	4.82	99.38
Transportation and Communication	396,060	1.20	1.07	80.39	0.34	0.30	78.24
Finance, Real Estate, Insurance	148,052	1.33	1.00	56.61	4.35	3.19	53.83
Business Services	850,568	0.47	0.33	49.77	0.24	0.20	70.18
Total	6,861,556	0.45	0.44	94.29	0.03	0.029	93.65

REFERENCES

Breiman, L., Friedman, J.H., Olshen, R.A. and Stone, C.J. (1984), *Classification and Regression Trees*, Wadsworth International, Belmont, CA.

Chambers, R. L. (1997), "Weighting and calibration in sample survey estimation", in: Malaguerra, C., Morgenthaler, S., Ronchetti, E. (eds), *Proceedings of the Conference on Statistical Science honouring the Bicentennial of Stefano Franscini's Birth*, Basel: Birkhäuser Verlag.

Falorsi P. D., Pallara A., Succi R., and Russo A. (2000), "Estimating indicators of labour input from administrative records having coverage and measurement errors", *ICES II – Proceedings of the Second International Conference on Establishment Surveys*, (available at web site *http://www.eia.doe.gov/ices2/#errata*).

Hidiroglou, M.A., Latouche, M., Armstrong, B., and Gossen, M. (1995), "Improving Survey Information Using Administrative Records: the Case of the Canadian Employment Survey," *Proceedings of the Annual Research Conference, U.S Bureau of the Census*, pp. 171-197.

Royall, R.M. (1988), "The Prediction Approach to Sampling Theory", in: Krishnaiah, P.R., Rao, C.R. (eds.), *Handbook of Statistics, vol. 6*, Elsevier Science Publishers, pp. 399-413.

Tam, S.M. (1987), "Analysis of Repeated Surveys Using a Dynamic Linear Model," *International Statistical Review*, 55, pp. 63-73.

Wolter, K.M. (1985), *Introduction to variance estimation*, New-York; Springer-Verlag.