European Conference on Quality in Official Statistics (Q2016)
Madrid, 31 May-3 June 2016

A latent class model to estimate labour cost from Multi-Source data

U. Guarnera[1], S. Pacini[1], R. Varriale[1]

[1] *National Statistical Institute, Italy.*

Abstract
In recent years, statistical analysis based on different data-sources has become an active area of research in both theoretical and applied statistics. In particular, due to the increasing availability of administrative data, problems concerning the use of multiple sources for estimation purposes have been receiving an increasing attention in Official Statistics. National Statistical Institutes (NSIs) frequently try to combine data from available sources in order to build "statistical" archives to be used in different phases of the statistical production process.
In this work we describe a procedure to predict the labour cost of enterprises when measurements are available from two different administrative data sources. Contrary to the situations where the information coming from all the sources can be considered a (possibly erroneous) measure of the amount of interest (see, for example, Guarnera and Varriale (in press); Pavlopoulos and Vermunt, 2015; Scholtus and Bakker, 2013; Bakker and Daas, 2012), in this context only one source is considered to be a direct measure of the response variable, while the information coming from the other source is treated as auxiliary information. In particular, we propose to use a latent class model, where latent classes correspond to different error patterns. The proposed model produces individual predictions (estimates) of the response variable, obtained taking expectations of the true data distribution conditional on the observed data. These predictions can be used for different purposes. First, they can be directly used as "smoothed" estimates of the response variable in presence of errors. Alternatively, they can be used for editing activities. This approach can also be used to assess the quality of the data sources in terms of the model parameters.

Keywords: Multi-source statistics, Administrative data, Latent class models.

1. The use of administrative data in NSIs

In recent years, statistical analysis based on different data sources has become an active area of research in both theoretical and applied statistics. Massive use of "external" data is considered by NSIs as an important alternative to the traditional approaches based only on survey data. In fact, this approach allows NSIs to move resources previously allocated in

conducting surveys to other activities, reducing at the same time the response burden on respondents. Moreover, statistical analysis based on large datasets may result in more accurate estimates than the ones that can be obtained through sample surveys.

On the other hand, combining data to build a statistical information system is a complex task. In fact, administrative data are typically collected by different institutions for specific purposes (for instance, data on enterprises provided by the tax agency have "fiscal nature") and may not be usable in their original form for statistical purposes. Thus, a lot of "pre-processing" work has to be done in activities, such as harmonization of definitions, variable standardization, etc., aiming at providing users with data that satisfy their informative requirements. Another important issue is related to the possibility of partial (or total) overlapping among informative contents from different sources.

Contrary to the situation where the information coming from all sources can be considered a (possibly erroneous) measure of the amount of interest (Guarnera and Varriale, 2016); Pavlopoulos and Vermunt, 2015; Scholtus and Bakker, 2013; Bakker and Daas, 2012), in this work we focus on situations where there is only one source directly measuring the response variable and the information coming from other sources is associated to the response variable through a linear relationship, but it never coincides with it. In this context, the aim of the analysis is to edit the response variable and obtain a microdata file containing individual adjusted values. The methodological approach we propose is a latent class model dealing with continuous data variables, where latent classes correspond to different error patterns. The proposed approach produces individual predictions (estimates) of the response variable, obtained taking expectations of the true data distribution conditional on the observed data.

In this work we describe a procedure to predict the labour cost of enterprises with at least 100 employees, when measurements are available from two different administrative data sources. The paper is organized as follows. Section 2 presents the research problem and the available information, Section 3 describes the model and Section 4 contains the description of the application and some conclusions.

2. Labour cost, the available administrative data

In 2012 the Italian National Statistical Institute (Istat) developed the archive FRAME SBS, that is a new framework for the production of structural business statistics (SBS) based on

the massive use of administrative and fiscal data integrated with direct survey (Luzi *et al.*, 2014). At the beginning the archive was limited to small and medium enterprises (i.e., with less than 100 employees) and was extended in 2014 to large enterprises (i.e., with at least 100 employees).

In this work, we focus on the variable *total labour cost* of large enterprises, which has been introduced in FRAME SBS in 2013. The massive use of administrative data for the production of business statistics implied the need to carefully analyze definitions and concepts contained in the available sources and to harmonize them with the statistical requirements defined by the European Regulations on SBS. In case of the variable labour cost, the main sources of information are the Financial Statements from the Chamber of Commerce (hereinafter BIL) and the Istat register on wages, hours and total labour cost at employee-employer level (whose Italian acronym is RACLI), largely based on social security information. As a result of a detailed analysis of these sources focused on the adequacy with respect to the statistical requirements from SBS regulation (Arnaldi *et al.*, 2015), BIL has been chosen as the reference source, while data from RACLI has been used as an auxiliary source to check BIL data. In fact, while RACLI definitions and concepts are not completely compliant with the statistical requirements, the SBS regulation explicitly refers to the financial statements for all requested variables, so that the choice of BIL as primary administrative source for all variables, included labour cost, should guarantee internal coherence.

Nevertheless, in Italy there are specific situations where the values of labour cost reported in BIL are not coherent with the statistical definition. This is the case of costs for workers like *agency* workers and *external workers* (for example project workers) that should be excluded from personnel costs according to SBS (they should be included in the *intermediate costs*), but that are *usually* not distinguished from the costs for employees in the company accounts (and consequently in BIL). By contrast, data from RACLI register, essentially based on social security information, do not contain any information on agency or external workers among labour cost, and can be used to correct BIL data.

Table 2 shows the quartiles of the distribution of the relative wage differences between RACLI and BIL separately for firms with only employees and firms with at least one agency or external worker. The information on agency and external workers for each

enterprise, together with their costs for the enterprises, is available from the new Istat Business Register. Out of the 8,866 enterprises with at least 100 employees that are present in both BIL and RACLI, only 6% has zero agency or external workers. The comparison between the two distributions shows that larger positive differences between BIL and RACLI are observed for enterprises with agency or external workers. This is probably due to the inclusion of costs for these types of workers in the BIL source.

Tab. 2 – Distribution of relative differences (%) of total labour cost between BIL and RACLI in enterprises with at least 100 employees matched in the two sources. Year 2012.

Economic Activity Sector	*External or agency worker*										*Total*				
	None					*At least one*									
	N.	*Mean*	*Q1*	*Q50*	*Q3*	*N.*	*Mean*	*Q1*	*Q50*	*Q3*	*N.*	*Mean*	*Q1*	*Q50*	*Q3*
Industry	134	9.8	-1.3	1.2	5.2	4,207	6.4	0.7	3.0	7.1	4,341	6.5	0.7	3.0	7.1
Services	405	3.9	-0.8	2.1	6.7	4,120	7.2	0.7	2.9	7.5	4,525	6.9	0.6	2.8	7.5
Total	539	5.4	-1.0	1.9	6.4	8,327	6.8	0.7	3.0	7.3	8,866	6.7	0.6	2.9	7.3

Summarizing, the BIL data could be a reference source if one would be able to adjust data for the presence of items not to be included in the labour cost. Adjusting data for this kind of error may have important consequences for estimation of economic aggregates, because moving amounts from labour costs to intermediate costs causes a decrease in the Value Added.

In the next paragraph, we describe a latent class model used to correct the variable *total labour cost* from BIL for the possible presence of costs that are not related to employees, using data on total labour cost from RACLI register as auxiliary information.

3. The latent class model

Let Y_i^* be the variable associated with the "true value" of the *per-capita labour cost* on the i-*th* unit (enterprise) in the dataset of analysis, and Y_i the corresponding variable available from BIL. We consider Y_i as an *imperfect measure* of the target variable, assuming that the measurement error is intermittent, i.e., the probability of the event $\{Y_i = Y_i^*\}$ is strictly greater than zero (Guarnera and Varriale, 2015). By contrast, we consider the information from RACLI (variable X_i) as merely auxiliary, assuming the linear regression model:

1) $Y_i^* = \beta X_i + U_i$,

where U_i ($i=1,..,n$) are independent zero-mean Gaussian variables with variance σ^2. Hereafter the model 1) will be referred to as the *true data model*.

As illustrated in Section 1, the value of the labour cost reported in BIL can be affected by a random error and by two different and independent types of systematic error due to the erroneous inclusion of costs for agency and external workers, respectively.

In order to adjust for these types of errors we have to specify the *measurement model*. Let d_i be the number of employees working in the *i*-th enterprise (*i=1,…,n*) and $C_i^B = Y_i d_i$ the corresponding total labour cost from BIL. Analogously, let $C_i^R = X_i d_i$ be the total value of the labour cost reported in RACLI for the *i*-th enterprise. Let us denote the total number of people working as agency and external workers as n_{age} and n_{ext}, respectively, and the corresponding total labour cost amounts as C_i^{age} and C_i^{ext}. We introduce the per-capita values $V_i^{age} = C_i^{age} / d_i$ and $V_i^{ext} = C_i^{ext} / d_i$, representing the error components in Y_i due to the inclusion in the labour cost of C_i^{age} and C_i^{ext}, respectively. In this context, n_{age}, n_{ext}, C_i^{age} and C_i^{ext} are considered as known quantities (actually, they are available in the new Istat Business Register).

The intermittent error mechanism is modeled through the equation:

2) $$Y_i = Y_i^* + Z_{i,age} V_i^{age} + Z_{i,ext} V_i^{ext} + Z_{i,\varepsilon} \varepsilon_i$$

where $Z_{i,age} \sim Be(\pi_{age})$ and $Z_{i,ext} \sim Be(\pi_{ext})$ are two Bernoullian random variables representing the indicators for the erroneous inclusion of agency and external workers in the computation of the labour cost, and $Z_{i\varepsilon} \sim Be(\pi_\varepsilon)$ is a third Bernoullian random variable which is one or zero depending on whether a random measurement error is present or not in BIL per-capita labour cost measure Y_i. The error components ε_i (*i=1,…,n*), are supposed to be independent Gaussian random variables with zero mean and variance $\sigma_\varepsilon^2 = \alpha\sigma^2$. The equation 2) can be interpreted as follows: the value Y_i reported in BIL, when not including costs for agency workers ($Z_{i,age}$=0 or n_{age}=*0*) and external workers ($Z_{i,ext}$=0 or n_{ext}=*0*), and in absence of random errors ($Z_{i\varepsilon}$=0), is a correct measure of the per-capita labour cost Y_i^*. Thus, if all the parameters π_{age}, π_{ext}, π_ε are strictly smaller than 1, there is a chance to observe in BIL the *true* value of the labour cost. Note that in equation 2), the second and third terms are products of an observed variable (V_i^{age}, V_i^{ext}) and a latent variable ($Z_{i,age}$, $Z_{i,ext}$), while in the third term both the involved variables ($Z_{i\varepsilon}$ and ε_i) are unknown.

From equations 1) and 2), it follows that the BIL per-capita Y_i can be expressed in terms of both X_i and the error terms as:

3) $Y_i = \beta X_i + Z_{i,age} V_i^{age} + Z_{i,ext} V_i^{ext} + Z_{i,\varepsilon} \varepsilon_i + U_i$.

Equation 3) expresses a mixture of linear regressions having a common regressor (X) and different patterns of covariates (V^{age}_i, V^{ext}_i) depending on the occurrences of the systematic errors associated with the Bernoullian variables $Z_{i,age}$ and $Z_{i,ext}$. The residual variance of each regression model is σ^2 or $(1+\alpha)\,\sigma^2$ depending on whether $Z_{i\varepsilon}$ equals 1 or 0 (presence of "random" error). Thus, the membership of each unit i is associated with an "error state vector" $\mathbf{Z}_i = (Z_{i,age}, Z_{i,ext}, Z_{i\varepsilon})$ with corresponding *a priori* probabilities (mixing proportions)

4) $p_{z_i} = P(\mathbf{Z}_i = \mathbf{z}_i) = P(Z_{i,age} = z_{i,age}) \times P(Z_{i,ext} = z_{i,ext}) \times P(Z_{i,\varepsilon} = z_{i,\varepsilon})$,

and posterior probabilities

5) $\tau_{z_i} = P(\mathbf{Z}_i = \mathbf{z}_i \mid X_i, Y_i) = \dfrac{p_{z_i} f_{z_i}}{\sum_{\{\mathbf{z}_i\}} p_{z_i'} f_{z_i'}}$.

The sum in the denominator in the expression 5) is over all the possible state vectors $\mathbf{z}_i$ and $f_{z_i} = N(y_i; \mu_{z_i}, \sigma^2_{z_i})$ is the Gaussian density with mean and variance $\mu_{z_i} = \beta x_i + Z_{i,age} V_i^{age} + Z_{i,ext} V_i^{ext}$ and $\sigma^2_{z_i} = (1 + z_{i\varepsilon}\alpha)\sigma^2$, respectively.

We have implemented an appropriate Expectation Maximization (EM) algorithm for the maximum likelihood estimation (MLE) of the model parameters ($\beta, \sigma^2, \alpha, \pi_{age}, \pi_{ext}, \pi_{\varepsilon}$), based on the available sample of enterprises.

The main applications of the model illustrated above are based on the expectations $E(Y_i^* \mid X_i, Y_i)$ of the true value of the per-capita labour cost Y_i^* conditional on the available information (X_i, Y_i; $i=1,..n$). Starting from Equations 1)-5) and using the Bayes formula, it is easily seen that the desired expected values are:

6)
$$E(Y_i^* \mid X_i = x_i, Y = y_i) = \sum_{\{z_i\}} \tau_{z_i} E(Y_i^* \mid X_i = x_i, Y = y_i, \mathbf{Z}_i = \mathbf{z}_i)$$
$$= \sum_{\{z_i\}} \tau_{z_i} \frac{y_i + \alpha z_{i\varepsilon} \beta x_i - z_{i,age} V_i^{age} - z_{i,ext} V_i^{ext}}{1 + \alpha z_{i\varepsilon}} \quad .$$

Note that in case of $Z_{i\varepsilon} = 0$ (no random error), Y_i^* is a deterministic function of Y_i given $Z_{i,age}$ and $Z_{i,ext}$. In fact, it can be obtained by simply subtracting the systematic error component $z_{i,age}\ V^{age}{}_i + z_{i,ext}\ V^{ext}{}_i$ from Y_i. In the opposite situation, as α becomes large (large measurement error), the observed value of Y_i provides little additional information with respect to the regression predictor βX_i.

If the estimates of the model parameters are plugged in formulas 6) we obtain predictions $\hat{y}_i$ that can be used as "robust" estimates of the per-capita value of the labour cost. Robustness refers to both systematic errors (erroneous inclusion of some amounts in the labour cost) and to the possible presence of outliers originated by "random" errors. Analogously, parameter estimates can be used to derive estimates $\hat{\tau}_{z_i}$ of the posterior probabilities τ_{z_i}. Besides directly using the predictions $\hat{y}_i$ as "smoothed" estimates of the per-capita labour cost in presence of both systematic and random errors (*predictive approach*), there are other *probabilistic approaches* that can be used to correct the per-capita labour cost based on the quantities $\hat{\tau}_{z_i}$ (and the probabilities that can be defined in terms of them):

- *systematic approach*: the per-capita labour cost is corrected by subtracting from the observed labour cost the expected value c_i of the systematic component of the error:

$$c_i = \hat{P}(Z_{i,age} = 1)V_i^{age} + \hat{P}(Z_{i,ext} = 1)V_i^{ext};$$

- *classification approach*: the error components are subtracted from the observed per-capita labour cost only when the posterior error probabilities are higher than a certain value, e.g., 0.5.

Other strategies are possible based on combining different approaches. For example, one could manually check all the cases where discrepancies between predictions and observed values are large (i.e., exceed a prefixed threshold) because of the presence of random errors, and treat the other cases automatically by using the predictive approach or classification approach (*selective editing* perspective). In the next section an application of the different methods to a real dataset is shown.

4. The estimation of labour cost

In this section, we present the analysis of the total labour cost of enterprises with at least 100 employees using the latent variable model presented in the previous section. As explained above, the aim is to obtain reliable values of the per-capita labour cost for each enterprise and, based on these microdata, the estimate of the total labour cost for the entire population.

The application has been conducted on 8,866 enterprises with at least 100 employees with information from BIL in 2012. Out of them, 4,755 (53.6 %) have agency workers and 7,165 (80.8 %) external workers. In Figure 1 the scatter plot of RACLI vs BIL per-capita labour cost is represented: while in most situations the two measures are quite similar, there are some units where very large discrepancies are observed.

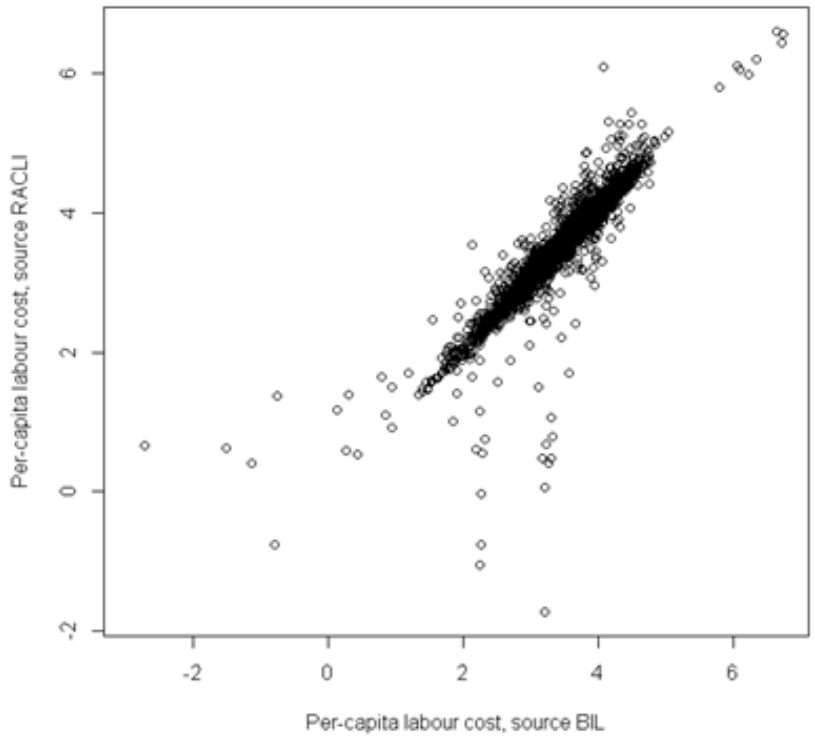

Figure 1. Total labour cost from sources BIL and RACLI, logarithm scale

The latent model described in Section 3 has been fitted to the data of analysis. The parameter estimates are: $\hat{\beta} = 1.02$, $\hat{\sigma}^2 = 1.37$, $\hat{\alpha} = 274$, $\hat{\pi}_{age} = 0.62$, $\hat{\pi}_{ext} = 0.22$, $\hat{\pi}_{\varepsilon} = 0.12$. According to the parameter estimates, it seems that the erroneous inclusion in the total labour cost of costs for atypical workers is more frequent for agency workers ($\hat{\pi}_{age} = 0.62$)

than for external workers ($\hat{\pi}_{ext} = 0.22$); moreover, the random component of the measurement error seems to be important both in terms of frequency and magnitude. The estimate of the coefficient β (1.02) seems to indicate that RACLI provides on average a value of the labour cost about 2% lower than the true one.

We have analyzed the impact of the proposed methodology on the estimate of the total labour cost by comparing the estimate based on raw data from BIL with that one obtained with the different approaches illustrated in the previous section. Using the predictive approach (that is replacing raw BIL data with predictions from the model) results in a decrease of 4.13% of the total labour cost. However, only a small fraction of the difference is due to the systematic component of the error. Precisely, subtracting from each BIL value y_i the expected value c_i of the systematic error (systematic approach, see Section 3) results in an estimate of the total labour cost only 1.13% lower than the raw estimate. This confirms a strong impact of the error random component.

As explained in Section 3, another possible use of the model is to adjust data for the (possible) presence of systematic error using a classification approach, i.e. to correct data according to the error posterior probabilities. In the present application we have corrected for the *agency worker error* or *external worker error* whenever the corresponding posterior probabilities are greater than 0.5. The (probabilistic) classification approach can be directly compared with the deterministic approach currently used in the production process. According to the latter approach, the cost of agency and/or external workers is deducted from BIL total labour cost if both costs or at least one is smaller than the difference between BIL and RACLI total labour cost. In the latter case, costs for agency workers, for external workers, or both, are subtracted from the BIL value depending on which operation results in the closest value to RACLI.

The comparison between the current deterministic approach and the correction of systematic error based on probabilistic classification of the enterprises is shown in Table 2.

Out of the 8,866 enterprises in the dataset, 2,814 (31.7%) observations are classified as "correct" by both methods, and 15.1% are classified as having the same error pattern. The remaining 4,711 (53.1 %) enterprises are differently classified by the two approaches.

Table 2 Number (and percentage) of enterprises corrected using deterministic and classification approach

		Classification approach				
		No correction	Only external	Only *agency*	Both	Total
Deterministic approach	No correction	2,814 (31.7%)	9 (0.1%)	744 (8.4%)	0 (0.0%)	3,567 (40.2%)
	Only external	1,842 (20.8%)	131 (1.5%)	407 (4.6%)	5 (0.1%)	2,385 (26.9%)
	Only *agency*	2 (0.0%)	2 (0.0%)	1,072 (12.1%)	1 (0.0%)	1,077 (12.1%)
	Both	0 (0.0)	0 (0.0%)	1,699 (19.2%)	138 (1.6%)	1,837 (20.7%)
	Total	4,658 (52.5%)	142 (1.6%)	3,922 (44.2%)	144 (1.6%)	8,866 (100.0%)

Generally, the deterministic approach tends to correct the total labour cost more often than the classification approach. In fact, the probabilistic method, differently from the deterministic one, is based on statistical modeling the relation between the two sources of information so that discrepancies are not explained only in terms of systematic errors.

Moreover, the effects of the two approaches are different for the two error types: while the deterministic approach corrects more enterprises for the external worker error, the classification approach has stronger effect on the agency worker error. This is probably due to the explicit modeling of the random error in the probabilistic approach. When the discrepancy between the BIL and RACLI values is very high and cannot be justified only in terms of systematic error, the BIL value does not provide strong information on the presence of the systematic error, and the corresponding posterior probabilities tend to the *a priori* probabilities. Thus, the high percentage of units that are classified as affected by an agency worker error is due to the fact that $\hat{\pi}_{age}$ (0.62) is higher than the classification threshold (0.5). On the other hand, for the enterprises with external workers, the threshold results to be much higher that the a priori probability $\hat{\pi}_{ext}$ (0.22).

From the application it results that the impact of the correction of the systematic error component is very similar in deterministic and probabilistic approaches. Nevertheless, modeling the error through a statistical model provides more information on the error structure, allowing different strategies for data analysis. Indeed, as described in Section 3, the model output can be used to obtain "smoothed" estimates or to identify possible influential errors to be treated through interactive editing.

References

Arnaldi S., Baldi C., Filippello R., Mastrantonio L., Pacini S., Sassaroli P. and Tartamella F. (in press), The labour cost variables in the building of the frame, Rivista di Statistica Ufficiale.

Bakker B.F.M. and Daas P.J.H. (2012), Methodological Challenges of Register-Based Research, Statistica Neerlandica, 66(1), pp. 8-17.

Guarnera U. and Varriale R. (2016), Estimation from Contaminated Multi-Source Data Based on Latent Class Models, Statistical Journal of the IAOS. In press.

Guarnera U. and Varriale R. (2015), Estimation and editing for data from different sources. An approach based on latent class models**,** UNECE Work Session on Statistical Data Editing, Budapest, 14-16 Settembre 2015.

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

Pavlopoulos D. and Vermunt J.K. (2015), Measuring temporary employment. Do survey or register data tell the truth?, Survey Methodology, 41(1), pp. 197-214.

Scholtus S. and Bakker B.F.M. (2013), Estimating the validity of administrative and survey variables through structural equation modeling. A simulation study on robustness. Discussion paper (201302), Statistics Netherlands, The Hague/Heerlen [Internet].