WP. 24
ENGLISH ONLY

UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Oslo, Norway, 24-26 September 2012)

Topic (iii): Editing and Imputation in the context of data integration from multiple sources and mixed modes

IMPROVEMENT OF THE TIMELINESS OF THE ITALIAN BUSINESS REGISTER VIA IMPUTATION OF MISSING DATA

Prepared by Davide Di Cecco[1], Danila Filipponi[2] , Italian National Institute of Statistics (Istat).

I. INTRODUCTION

1. The Italian Business Register (BR) records all the active enterprises of industry and services and their structural attributes, by integrating information coming from both administrative sources, managed by public agencies or private companies, and statistical sources owned by ISTAT. Up to now, the construction process of the BR relative to the year t, has been carried out as follows: the set–up process starts in the last quarter of the year $t+1$, when the yearly data supplies from the main sources become available. After a process of normalization and standardization, the data are integrated and the main structural and identification variables are estimated for each integrated unit (see Garofalo [1998]). Finally, the BR for the reference year t is available within the first quarter of the year $t+2$.

2. The BR has been conceived as a base defining the reference universe of the business population, to be used as a sampling frame for the preparation and co-ordination of the economic surveys and for grossing up the sampling results. Currently, the BR is also considered the official source for statistical information on the companies structure and demography. The BR is more and more used as a dissemination tool, i.e., a source for the statistical analysis of the economic system, especially with regards to the territorial aspects of the economic dynamics (see Consalvi [2005]). Because of this new and important role conferred to the BR, it is necessary to get a timely representation of the actual evolution of the economic structure.

3. In recent years, in particular since 2010, new administrative sources have been introduced in the realization of the BR, changing radically the set–up process, the information available and the timeliness. The source that primarily changed the way the process is carried out, is the Social Security Database (EMENS). The EMENS is the database of the declarations that Italian companies are due to submit for each employee to Social Security. Thus, the structure of the database is that of a Linked Employer–Employee Database (LEED). That is, we are switching from a structure where the information was available at an enterprise–level only, to an individual level Database. Moreover, the new source allows an improvement in the timeliness of the BR dissemination.

[1] dicecco@istat.it
[2] dafilipp@istat.it

4. The aim of the project we present, is to improve the BR timeliness, providing information concerning the structure of the business population at six months from the reference year, exploiting only the administrative and the statistical sources available in the first quarter of the year $t+1$ and the longitudinal information. In this paper we restrict our attention to the enterprises with dependent employees. Clearly, an improvement of the timeliness will reduce the accuracy of the information. Indeed, the administrative information available within the first quarter of the year may be incomplete because of delays in the Social Security declarations. In this work we present a methodology to detect and impute missing jobs declarations data in the EMENS, in order to set up an earlier release of the BR. Section II describes administrative and statistical sources available within the first quarter of the year $t+1$, with attention to their completeness. Section III describes the statistical models used to impute the missing data. Some concluding remarks are made in Section IV.

II. The input sources

5. The statistical sources available within the first quarter of the year $t+1$, are all the short-term surveys on the enterprises carried out by ISTAT, and a specific survey on the enterprises of large dimensions having multiple establishment (IULGI). In particular, IULGI is a survey carried out in December on a yearly basis, having the specific purpose of correct/update some structural variables of the BR, for enterprises having more than 50 employees.

6. As regards to the administrative sources, those that are available within the first quarter of the year $t+1$, can be grouped as follows:

- Fiscal data (VAT), managed by the Ministry of Economy and Finances, that records all natural and legal persons operating over the national territory, who are required to comply with fiscal legislation;
- Social Security data (EMENS), managed by the Social Security Authority that records monthly employer declarations on jobs position for persons subject to the payment of social security contributions;
- Chambers of Commerce data (CCIAA) including compulsory declarations to be submitted by anyone who wants to start a new enterprise.

All these archives are continuously updated, as new submissions constantly arrive in a flow. This involves the addition of new units and/or the updating of the values of existing records. ISTAT does not have a continuous access to the data, indeed it acquires the entire datasets in two distinct occasions during the year. The continuous updating of the administrative archives may cause misalignments in the reference population when the archive is downloaded in different periods of the same reference year. This means that using an earlier supply of the administrative data with respect to the timing regularly adopted for the realization of the BR, might lead to problems of completeness, both in terms of coverage (units) and in terms of quality (variables).

7. The VAT data and the CCIAA data are used exclusively for the attribution of the main characteristics of the enterprises like economic activity code and legal status. Our methodological contribution will focus on the EMENS database. The EMENS is the main administrative source to identify the universe of enterprises with employees and the number of employees per company. The EMENS data contain monthly employer declarations on job positions, and it covers all persons with a dependent employment contract, i.e., self-employed, business owners or pensioners are not included. As we have said, it has a LEED structure: each record refers to a single job contract and reports various information pertaining it. We have various variables describing the type of employment contract of the worker (Fixed-term or Permanent, salary, particular tax regimes...), and many others describing

the continuity of the employment (the number of worked days per month, the eventual periods of sick leave, maternity leave...). The structure of the data create a direct relationship between employees and enterprises, that is, all the employees present in the EMENS data are assigned to an employer (enterprise). This means that together with the basic enterprise attributes, detailed information on employment and employment composition at enterprise level (e.g., gender and age composition) are achievable by summing over individual information.

8. As we have said, we want to use an earlier supply of the EMENS dataset for the BR construction. In order to evaluate the possibility of using the EMENS supply available in the month of March of the year $t+1$, we compare it with the supply of October of the same year, which is the one regularly adopted for the realization of the BR. The first two columns of Table 2 show the total number of job positions per month in the two supplies. The differences in the number of job positions is not constant throughout the year, but is markedly higher in latest months, especially in December where it reaches a maximum of 1.4%. This fact suggests that the differences are mainly due to delays in the delivery of job declarations, and that an imputation is necessary to reduce the misalignments of the monthly information on employment and employment composition in the reference population.

III. THE IMPUTATION PROCESS

9. Our main interest is in the presence of the working relationship of the employee for the company in each month. So, formally, the data we will focus on consist of a set of binary 12–ples $(X_{i,1}, \ldots, X_{i,12})$ where $X_{i,j} = 1$ if the i–th working relationship is taking place in the j–th month, i.e., if a declaration of the company identifying the individual as working in the j–th month is present in the database, $X_{i,j} = 0$ otherwise.

10. As we have said in Section II, by comparing the two EMENS supplies, we can find a certain rate of missing monthly job positions, especially in latest months, and we want to compensate for this difference through an imputation process. Anyway, the problem we are facing cannot be afforded by the usual imputation techniques. In fact, for the very nature of the data under study, we cannot single out the missing data, even if their presence is beyond dispute. That is, we cannot distinguish between a "genuine" 0 (identifying the employee as not working for that company in that month) and a partial missing information due to errors or delayed declarations arrival. For this reason, we will adopt a probabilistic approach and follow a two steps process:

Step 1. we construct a model defining the probability of each zero value of being missing;
Step 2. we construct an imputation model defining the conditional probability of each zero value of being one, given that it is missing.

Then, the imputation will consist of changing some zero values into ones; precisely, each zero value will be changed with probability equal to the product of the two probabilities resulting from the two models above.

A. Step 1 – The probabilities of being missing

11. As regards to the first Step, before we define the probabilities of being missing, we further refine our strategy by adopting some preliminary solutions, that will severely limit the set of records which are eligible for imputation:

- on a "macro–level", we utilize aggregate employment data at a company–level to identify the anomalous fluctuations in the monthly patterns in the number of workers, and, consequently,

identify a (relatively small) subset of companies which are more likely to have missing data, and limit the imputation process to the records of those companies workers;
- on a "micro–level", we utilize the available auxiliary information to exclude some 12–ples from the imputation process. That is, we asses a zero probability of having missing values to those 12–ples which are consistent with particular information. We will say that some zero values are identified as "sure zeros", the remaining as "uncertain zeros", and just the second can be missing values.

A.1. *Detecting the anomalous companies.*

1. We want to analyze the sequences

$$Y_e = \left(\sum_{i:\, i \in e} X_{i,1}, \ldots, \sum_{i:\, i \in e} X_{i,12} \right)$$

of the total number of working employees per month in each company e, in order to identify the anomalous monthly patterns, and, hence, the companies whose single worker records are likely to have missing values. To define an "anomalous" pattern, we have to choose a suitable model for longitudinal data and then find the cases which significantly depart from the expected values.

2. One common method to analyze longitudinal data is based on the specification of a two-levels random effects model (Laird and Ware [1982]). Several variations on this basic theme are presented in Crowder and Hand [1990], Lindstrom and Bates [1990], and Lindsey [1993]. The model equation suggested by Laird and Ware is a two level model: level 1 introduces population parameters and within-person variation and level 2 between-person variation. To outline the general structure of those models for longitudinal data, we first introduce some notation:
Let Y_e denotes a vector of n repeated measurements on subject e. Let α denotes a p–dimensional vector of unknown population parameters, and W_e be a known $n \times p$ design matrix linking α and Y_e. Let b_e be a k–vector of unknown individual effects and Z_e be a known $n \times k$ design matrix linking b_e and Y_e. Then the model is the following:
Level 1. For each subject e:

$$Y_e = W_e\alpha + Z_e b_e + \epsilon_e \tag{1}$$

where each ϵ_e is an n–vector normally distributed as $N(0, R)$, and the ϵ_e are assumed to be independent. At this stage, α and b_e are considered fixed.
Level 2. The b_e are independent random effect distributed as $N(0, D)$, and the b_e are independent of the ϵ_e.
Marginally, the Y_e are normally distributed with mean $W_e\alpha$ and variance and covariance matrix $R + Z_e D Z_e^T$. Maximum likelihood estimation for this model is described in Laird and Ware [1982].

3. In our case, the units e are the about 2 millions enterprises, and the repeated measurements per unit are the monthly number of employees (12 measurements). We based our analysis on a simple model in the family of (1). In particular, we used a random intercept model (i.e., Z_e is the unit vector $\mathbf{1}$), and, as fixed effects, we used the monthly number of employees per company of the previous year, and the yearly average number of employees as resulting from IULGI (when available). To get a better fit to a Normal distribution, we use a log transformation of the Y_e. So, our model is specified as:

$$Y_{e(t)} = \mathbf{1}\alpha_1 + Y_{e(t-1)}\alpha_2 + u_e\mathbf{1}\alpha_3 + \mathbf{1}b_e + \epsilon_e,$$

where $Y_{e(t)}$ denotes the vector of the monthly employees of company e for the reference year t on a logarithm scale $\left(\log \sum_{i\in e} X_{i,1}, \ldots, \log \sum_{i\in e} X_{i,12}\right)$, $Y_{e(t-1)}$ refers to the previous year, and u_e denotes the average number of employees of company e as resulting from IULGI. We assume that $b_e \sim N(0, \sigma)$ and $\epsilon_e \sim N(0, R)$, where R is a 12×12 matrix of variance and covariance following a first order AR

structure.

The huge number of companies in the present application required some compromises. First, since we are dealing with repeated count data, probably a Generalized Mixed effect model based on a Poisson distribution could be a better choice. But the available algorithms for the estimation procedure of that model are far less efficient and stable. So, we were forced to use a normal Mixed effect model on a log transformation of the counts. Moreover, we restricted our analysis to the enterprises with more than 5 employees – about 400,000 units. Despite this, our task was still too computational onerous, so that we had to split the dataset into subpopulations based upon some structural variables of the enterprises like the economic activity.

4. In a random effect model each subject's data is referred to as a curve. The goal of our diagnostics is to determine, for each curve, wether there is an evidence of model violations at any of the hierarchical level. In particular, we want to identify, for each enterprise, if there exists an anomalous change in the number of employees in some of the months of the reference year. In order to identify what the model regards as an outlier, we provide an indicators for each curve based on the analysis of the conditional residuals $r_e = (r_{e,1}, \ldots, r_{e,12})$ indicating which components (months) of the e-th curve exhibits an anomaly. Once we get the vector r_e of conditional residual

$$r_e = Y_{e(t)} - \widehat{Y}_{e(t)} = Y_{e(t)} - \mathbf{1}\widehat{\alpha}_1 + Y_{e(t-1)}\widehat{\alpha}_2 + u_e\mathbf{1}\widehat{\alpha}_3 + \mathbf{1}\widehat{b}_e$$

for each subject e, we define the outliers indicator $out_e = (out_{e,1}, \ldots, out_{e,12})$ as

$$out_{e,j} = \begin{cases} 0 & \text{if } |r_{e,j}| \leq 2\widehat{\sigma}_{e,j}; \\ 1 & \text{otherwise.} \end{cases}$$

where $\widehat{\sigma}$ is the estimated variance. For more details on the variance estimation see Harville [1990].

5. Table 1 shows the results of the methodology used to identify the anomalous enterprises. We can evaluate the accuracy of the procedure by comparing the classification of the enterprises as being identified as anomalous or not by the model, and as having an actual updating or not in the second EMENS supply of Social Security data. Table 1 shows that, for the month of November, 78.4% of enterprises with an updated number of employees in the second supply are not identified as anomalous. Similarly for the month of December. However, if the same error is measured in terms of job positions to be updated, we can see that we detect 73% of the job positions updated in November and 80% for the month of December. This result shows how difficult is to identify the enterprises with small variations in terms of employees between the two data supply.

A.2. *Defining the probability of being missing.*

1. The r.v.s $X_{i,j}$ only assume the values 0 or 1, so the missing values are unobservable. However, once we compare the two EMENS supplies, we can see that in the latest, in some records, some information have been updated, both in the data arrays of monthly presences $(X_{i,1}, \ldots, X_{i,12})$ and in the auxiliary variables. At this point, we apply to both supplies a common set of deterministic criteria that exclude some zero values from the probability of being missing, and hence, from the possibility of being imputed. That is, as we have said at the beginning of this Section, we identify the sure zeros and the uncertain zeros on the basis of the consistency of each 12–ple with some auxiliary variables pertaining the record. For example, a criterion we utilize is the following: some (but not all) of the records report a starting and/or an ending date of the job contract, and a consistent number of worked days: those cases have been considered as unlikely to have missing information. As a matter of fact, we observe that some of the values classified as uncertain zeros in the first supply became sure zeros or ones in the second, revealing their nature of missing values. So, what we are going to model is a

TABLE 1. Distribution of the companies classified by anomalous detection results and actual updating.

	Number of companies			
	Identified as		Identified as	
	Non Anomalous	Anomalous	Non Anomalous	Anomalous
	Absolute values		**Row Percentage**	
	November			
Having no update	363,467	21,630	94.4%	5.6%
Having at least an update	5,392	1,482	78.4%	21.6%
Total	368,859	23,112	94.1%	5.9%
Number of updated job positions	23,065	63,443	26.7%	73.3%
	December			
Having no update	363,075	20,379	94.7%	5.3%
Having at least an update	5,784	2,733	67.9%	32.1%
Total	368,859	23,112	94.1%	5.9%
Number of updated job positions	30,169	122,931	19.7%	80.3%

binary r.v. $U_{i,j}$ defined as:

$$U_{i,j} = \begin{cases} 1 & \text{if } X_{i,j} \text{ is updated;} \\ 0 & \text{otherwise.} \end{cases}$$

2. Then, an estimate of the probability of being missing $Pr(U_{i,j} = 1)$ is given by the proportion of uncertain zeros updated as sure zeros or ones. That proportion varies considerably during the year and is markedly higher in latest months (in fact, the difference in the number of employees in December was the evidence that motivated the present work). So, in creating a simple model for the construction of the missing probabilities, we take into account the following factors:

- The reference month; i.e. we define the missing probabilities as a function of the month;
- The type of employment contract of the worker: Fixed-term or Permanent (we will keep this distinction in the imputation model too);
- Finally, we deem it appropriate to include some kind of conditioning to the structure of the 12–ple the uncertain zero value belong to (for example, as is intuitive, we observed that a zero value in a 12–ple with other 11 ones and one in a 12–ple with a convoluted pattern have different probability of being missing). To avoid an excessive anchoring to the contingencies of the current year, we condition the missing probabilities to the number of sure zeros, uncertain zero and ones in the 12–ple, avoiding to consider more in detail the structure.

3. As an example, consider a record of a Fixed-term contract having the following presence sequence:

$$(x_1 = 0, x_2 = 0, x_3 = 1, x_4 = 1, x_5 = 1, x_6 = 1, x_7 = 1, x_8 = 1, x_9 = 1, x_{10} = 0, x_{11} = 0, x_{12} = 0) \quad (2)$$

where x_{10} and x_{11} are the uncertain zeros. Then, the estimate of $Pr(U_{i,10} = 1)$ is the proportion of uncertain zeros in October, among records of Fixed term contracts, of sequences consisting of 3 zeros, 7 ones and 2 uncertain zeros that have been updated as sure zeros or ones in the second supply. Note that, in this way, we are hypothesizing the $U_{i,j}$ as independent. That hypothesis may be unsatisfactory, but can be overcame by including a common empirically estimated correlation between the $U_{i,j}$.

4. Note that we want to define a strategy that depends as little as possible from the contingencies of the year under study. For this reason, we use the comparison between the two supplies just for the

construction of the probability of missing, and not for the probability of imputation, in the hypothesis that the "distance", defined as the number of updates between the two supplies, remains approximately constant in the subsequent years.

B. Step 2 – The imputation model

5. Defining an imputation model means to define a distribution modelling the joint probability of the binary sequence $(X_1, \ldots, X_{12})$. In choosing such a distribution, we have to consider some factors:

- We are dealing with longitudinal data, so we want a model which somehow take into account the time dependency;
- Our data shows some clustering aspects: you can identify a few kind of patterns among the set of binary sequences that show up more frequently, and characterize distinct groups, and we want to represent this heterogeneity;
- Finally, we want the model to be simple enough to be reliable even in future years.

As regards to the heterogeneity, there are major evident differences in the patterns of binary vectors $(x_{i,1}, \ldots, x_{i,12})$ among employees with different kind of job contracts. Due to this, we estimate two distinct models: one for the employees with a Fixed term contract, one for those with a Permanent contract.

6. The model chosen for the imputation is a finite mixture of Markov chains. To define this model we introduce some notation and refer the reader to literature (first of all [Heckerman et al. 2003]) for details. A first order, time-homogeneous Markov chain for binary data is defined by an initial probability $q = Pr(X_1 = 1)$ and a 2×2 matrix of transition probabilities $P = \left(\begin{smallmatrix} p_{0,0} & p_{0,1} \\ p_{1,0} & p_{1,1} \end{smallmatrix}\right)$ where $p_{i,j}$ is the transition probability from i to j, $Pr(X_t = j \mid X_{t-1} = i)$. Denote as $f_h(x_1, \ldots, x_{12})$ the probability $Pr(X_1 = x_1, \ldots, X_{12} = x_{12})$ of observing that binary array $(x_{i,1}, \ldots, x_{i,12}) \in \{0, 1\}^{12}$ under a Markov chain of parameters q_h and P_h where $P_h = \left(\begin{smallmatrix} {}_hp_{0,0} & {}_hp_{0,1} \\ {}_hp_{1,0} & {}_hp_{1,1} \end{smallmatrix}\right)$, that is:

$$f_h(x_1, \ldots, x_{12}) = q_h^{x_1} \prod_{t=1}^{11} {}_hp_{x_t, x_{t+1}}.$$

A finite mixture of Markov chains with k components is the distribution defined as:

$$Pr(X_1 = x_1, \ldots, X_{12} = x_{12}) = \sum_{h=1}^{k} \pi_h \, f_h(x_1, \ldots, x_{12}),$$

where f_h, $h = 1, \ldots, k$, are the distributions of the k components Markov chains, and the π_h, $h = 1, \ldots, k$, are the weights of the mixture $\left(\sum_{h=1}^{k} \pi_h = 1\right)$. We can say that the choice of this model answers quite well to the questions raised above, in fact:

- The time series we are analyzing are rather short (12 months) . Our model is suitable for the analysis of longitudinal data even for short series like those of our study, whereas more complex models for longitudinal data are unusable;
- The model is parsimonious in the number of parameters. The total number of possible distinct outcomes for a binary 12–ple is $2^{12} = 4096$, which would require, in a saturated multinomial model, an equal number of parameters. As opposed to this, a mixture of Markov Chains model of k components has $4k - 1$ independent parameters;
- Despite the small number of parameters, the model is flexible as we can vary the number of components to adjust the fitting to the data, and it has an immediate interpretability in terms of cluster analysis. That is, we can capture the heterogeneity of the data, by modeling different homogeneous subgroups of the population by the different components of the mixture.

7. Actually, we observed from our results that the model reproduces well the nature of the data. Indeed, whatever the number of components, (at least) one of the component Markov chains resulting from the estimated model was found to have a transition probability $p_{0,0}$ very close to one (0 is an absorbing state), and at least a component was found to have the transition probability $p_{1,1}$ very close to one (1 absorbing state). That is, the mixture model handles well the patterns that we observe more often in the data: sequences like $(0,\ldots,0,1,\ldots,1)$ starting with all zeros and ending with all ones, representing the employees who, once started working for a company, do not stop (at least in the current year) and are modelled by the Markov chain component with the absorbing state 1 and, conversely, sequences like $(1,\ldots,1,0,\ldots,0)$ starting with all ones and ending with all zeros, relative to those employees who stopped working and do not restart (represented by the component Markov chain with the absorbing state 0). Obviously, the observed 12–ples having multiple transitions between 0 and 1 are rare, and decrease in frequency as the number of transitions 0-1 and 1-0 increase.

8. So, we estimate the model parameters on the basis of all available 12–ples not containing any uncertain zero (about 80% of the 12.5 millions available records). We use the EM algorithm to find the maximum likelihood estimate of the parameters (see Heckerman et al. [2003] for details). For the model selection (that is, to choose the number k of components of the mixture), we follow the common practice, particularly recommended in case of finite mixtures, of choosing the best model on the basis of the AIC (Akaike Information Criterion) to take into account both the goodness of fit and the parsimony of the model. Once the mixture model is estimated, we associate each 12–ple containing uncertain zeros to a component according to the following scheme. We evaluate its posterior probabilities of "belonging to" each component of the mixture. That probability is defined as a likelihood ratio, that is, for the h-th component it is the weighted likelihood of the 12–ple under the h-th Markov Chain divided by the likelihood under the estimated mixture model:

$$\frac{\pi_h\, f_h(x_1,\ldots,x_{12})}{\sum_{h=1}^{k} \pi_h\, f_h(x_1,\ldots,x_{12})}\,. \tag{3}$$

In calculating those probabilities, the uncertain zeros have to be considered as missing values, so, we have to calculate the multiple-step transition probabilities: $p_{i,j}(n) = Pr(X_t = j \mid X_{t-n} = i)$. For example, the sequence of example (2), where x_{10} and x_{11} are missing, has the likelihood

$$f_h(x_1,\ldots,x_{12}) = (1-q_h)\,{}_hp_{0,0}\,{}_hp_{0,1}\,{}_hp_{1,1}^{6}\,{}_hp_{1,0}(3)$$

To calculate the n-step transition probabilities, $n = 2,\ldots,12$, we have to calculate the n-th power matrix of ${}_hP$ for each n and h. Note that we could use the data containing uncertain zeros in the estimation procedure too, in which case each record would contribute to the likelihood with the "non-missing part" as above, but the power matrices should be calculated for each iteration of the EM algorithm, which would be computationally onerous, and the procedure would slow down too much.
Once we have calculated the posterior probabilities (3), we associate the 12–ple to the component of the mixture having the highest value, and define the imputation probabilities for each uncertain zero of the sequence on the basis of the parameters of that component. For example, say the sequence (2) is associated to component d, then x_{10} and x_{11} will be changed into ones with probability $\frac{{}_dp_{1,1}^2}{{}_dp_{1,0}(3)}$. Given the structure of the 12–ple, the sequence will be likely associated with the component with the highest $p_{1,1}$ probability, and, consequently, the probability of x_{10} and x_{11} of being imputed will be high.

Some computational aspects.

9. The huge number of records in our hands required some ad hoc solutions to overcome the related computational difficulties: As regards to the estimation procedure, we note that the total number of possible outcomes of the sufficient statistic for our model (which is the initial state and the

number of transitions between states in the sequence), is not large; in fact, for a sequence of n elements it is $2\binom{n}{2}+2$ (see Di Cecco [2009]), i.e., in our case, $2\binom{12}{2}+2=134$. Since each record contributes to the likelihood through the sufficient statistic only, we have just 134 distinct values to calculate. Then we group the data according to the value of the sufficient statistic and the log-likelihood can be written as:

$$\sum_{s\in S} n_s \lg f_s$$

for s, the sufficient statistic, ranging over the 134 values of set S of possible outcomes for binary 12–ples, n_s being the number of observed sequences consistent with s, ($\sum_{s\in S} n_s$ =12.5 millions), and f_s being the relative likelihood. Note that, in case of sequences of unequal lengths, or if we want to use the sequences with missing values in the estimate, things become more complicated.
As regards to the calculation of the n–th power of P_h for each component h and each $n=2,\ldots,12$, we can make use of a specific result reported in McLaughlin [2004] for the special case of 2×2 matrices, which express the power matrix as a simple formula. This result is particularly useful in case one want (or is forced) to use all available data in the estimation procedure, that is, to include the sequences with missing values in the EM algorithm.
Lastly, we mention the fact that, in programming (with R) the imputation procedure, we keep storage of the binary 12–ples by the relative decimal number (e.g., 000000111110 = 62). That trick allow us to effectively handle millions of sequences, leading an incredible acceleration to the process.

IV. CONCLUDING REMARKS

10. Up to now, the Italian BR has been published with a two years delay w.r.t. the reference year. We improved the BR timeliness exploiting the quality of the administrative sources available in the first quarter of the year following the reference year. We focused on the estimate of the monthly number of employees for each company. The number of employees is estimated by summing over individual information (employee-level data) coming from the Social Security Authority. Each month, companies are due to submit a declaration for each employee. That is, the data under study consist of each employee's array of presence/absence per month in a given company, i.e., 12–ples of binary data. Declarations continuously arrive in a flow, and new submissions constantly update the database. However, at our disposal we just have two supplies of the Social Security data, one in March, the other in October. To reduce the delay of the register, we utilized the first supply and impute some missing information. We made an in–depth comparison of the two supplies to estimate the amount of missing values and their location (at an individual–level). As a matter of fact, we used auxiliary variables, and a mixed-effect model to single out possible candidate (zero values) to imputation, and the relative probability of being missing. Finally, data classified as missing are changed in presence data according to a finite mixture of binary Markov Chains model.

In Table 2 we show a summary of the results of the imputation process by giving the total number of job positions per month in the two supplies and in the dataset resulting from the imputation. By comparing the result of the imputation with the second supply, we can see that our method is able to cover about 40% of missing values, but we want to underline the good results relative to the latest months, which have the highest number of missing information, with about 70% of the missing job positions imputed for the month of November and about 80% for December.

TABLE 2. Results of the imputation process on the total number of employees per month

Month	Total number of employees			Difference		% of Imputed Missing values
	Supply 1 (A)	Supply 2 (B)	Supply 1 + imputation (C)	(B)-(A)	(B)-(C)	
1	12,164,371	12,192,056	12,164,959	27,685	27,097	2.1%
2	12,167,883	12,197,322	12,171,731	29,439	25,591	13.1%
3	12,311,622	12,340,123	12,314,453	28,501	25,670	9.9%
4	12,442,147	12,473,120	12,443,552	30,973	29,568	4.5%
5	12,550,630	12,587,977	12,553,053	37,347	34,924	6.5%
6	12,711,253	12,753,026	12,718,615	41,773	34,411	17.6%
7	12,666,379	12,708,363	12,667,959	41,984	40,404	3.8%
8	12,457,136	12,502,072	12,459,575	44,936	42,497	5.4%
9	12,603,734	12,668,348	12,623,254	64,614	45,094	30.2%
10	12,440,801	12,514,929	12,472,751	74,128	42,178	43.1%
11	12,373,326	12,478,341	12,443,435	105,015	34,906	66.8%
12	12,292,116	12,470,754	12,429,478	178,638	41,276	76.9%

References

M. Consalvi. The role of the business register as an informative source for business analysis. 19th International Roundtable on Business Survey Frames, Cardiff, 2005.

M.J. Crowder and D.J. Hand. *Analysis of repeated measures*. Monographs on Statistics and Applied Probability. 41. London:Chapman and Hall, 1990.

D. Di Cecco. A class of models for multiple binary sequences under the hypothesis of Markov exchangeability. *Electron. J. Stat.*, 3:1113–1132, 2009. URL http://dx.doi.org/10.1214/09-EJS478.

G. Garofalo. The asia project. setting-up of the italian business register. synthesis of the methodological manual. 12th International Roundtable on Business Survey Frames, Helsinki, 1998. URL http://www.voorburggroup.org/Documents/1998%20rome/papers/1998-045.pdf.

D. A. Harville. Blup (best linear unbiased prediction), and beyond. pages 239–276. Advances in Statistical Methods for Genetic Improvement of Livestock, Springer-Verlag, 1990.

D. Heckerman, C. Meek, P. Smyth, and S. White. Model-based clustering and visualization of navigation patterns on a web site. *Data Min. Knowl. Discov.*, 7(4):399–424, 2003.

N. M. Laird and J. H. Ware. Random-effects models for longitudinal data. *Biometrics*, 38:963–974, 1982.

J.K. Lindsey. *Models for repeated measurements*. Clarendon Press, Oxford, 1993.

M. J. Lindstrom and D. M. Bates. Nonlinear mixed effects models for repeated measures data. *Biometrics*, 46:673–687, 1990.

J. McLaughlin. Combinatorial identities deriving from the n–th power of a $2\times$ 2 matrix. *Integers: The Electronic Journal of Combinatorial Number Theory*, 4:A19, 2004.