Integration of administrative sources and survey data through Hidden Markov Models for the production of labour statistics

Danila Filipponi, Ugo Guarnera

1. Introduction

The increased availability of large amount of administrative information at the Italian Institute of Statistics (Istat) makes it necessary to investigate new methodological approaches for the production of estimates, based on combining administrative data with statistical survey data.
Traditionally, administrative data have been used as auxiliary sources of information in different phases of the production process such as sampling, calibration, imputation. Basically, the classical approach, that could be defined *supervised*, relies on the assumption that, at least after some data editing procedures to remove occasional measurement errors, the survey data provide correct measures of the target variables, so that the use of external sources is essentially limited to the reduction of the sampling error. This is because the measures provided by administrative sources usually do not correspond to the target variables. On the other hand, although surveys are designed to meet the statistical requirements, also statistical data could be affected by measurement errors that may seriously compromise the accuracy of the target estimates.
In order to take into account deficiencies in the measurement process of both survey and administrative sources, a more symmetric approach with respect to the available sources can be adopted. A natural strategy, according to this approach, (*unsupervised* approach), is to consider the target variables as latent (unobserved) variables, and to model the measurement processes through the distributions of the observed variables conditional on the latent variables.

In this latent modeling approach it is useful to classify the variables in three groups:

1. variables Y^* representing the "true" target phenomenon. These are the variables that we would observe if data were error free. In general, Y^* are considered latent variables because they are not directly observed.
2. variables Y^g ($g=1,..G$) representing imperfect measures of the target phenomenon. These variables are the ones actually observed from G different data sources.
3. covariates X^L and X^M associated respectively to the latent variables Y^* and to the measures Y^g through statistical models.

The statistical model is composed of two components specified via the conditional probability distributions:

$$P(Y^* \mid X^L) \qquad (\textit{latent model}), \tag{1}$$

$$P(Y^1,..,Y^G \mid Y^*, X^M) \qquad (\textit{measurement model}) \tag{2}$$

From the conditional distributions (1) and (2) one can derive the marginal distribution $P(Y^1,..,Y^G \mid X^L, X^M)$ of the imperfect measures.:

$$P(Y^1,..,Y^G \mid X^L, X^M) = \int P(Y^1,..,Y^G \mid Y^*, X^M) P(Y^* \mid X^L) dY^*$$

Then, model parameters can be estimated using a likelihood approach, based on the data observed from the *G* different sources. Once the model parameters have been estimated, we can derive the marginal distributions $P(Y^g \mid Y^*, X^M), \quad g = 1,..,G$ from (2). These distributions can be used to assess the accuracy of each source and the sources can be ranked accordingly. Using Bayes theorem one can derive the distribution of the latent variables conditional on the available information (*posterior distribution*):

$$P(Y^* \mid Y^1,..,Y^G, X^M, X^L)$$

and use the expectations from this distribution to obtain predictions of the true values for each unit.

2. Use of administrative and statistical data for labour statistics

The main sources available for the production of labour statistics are the Italian Labour Force Survey (Lfs) and administrative sources mainly providing social security and fiscal data.

The Italian Lfs is a continuous survey carried out during every week of the year. Each quarter, the Lfs collects information on almost 70,000 households in 1,246 Italian municipalities for a total of 175,000 individuals (representing 1.2% of the overall Italian population). The Lfs provides quarterly estimates of the main aggregates of labour market (employment status, type of work, work experience, job search, etc.), disaggregated by gender, age and territory (up to regional detail).

Administrative data relevant for the labour statistics come mainly from social security, Chambers of Commerce and fiscal authority. Data are organized in an information system having a linked employer-employees (LEED) structure. From this data structure it is possible to obtain information on the statistical unit of interest, i.e., the worker. The main goal of this analysis is twofold: 1) to produce statistics on the employment status by small geographical domains in order to fulfill the population census requirements; 2) to improve the accuracy of the labour force estimates.

Within the general framework described above appropriate models are Hidden Markov Models (HMM). In fact the methodological choices have to take into account that the variable of interest is categorical and the data are longitudinal.

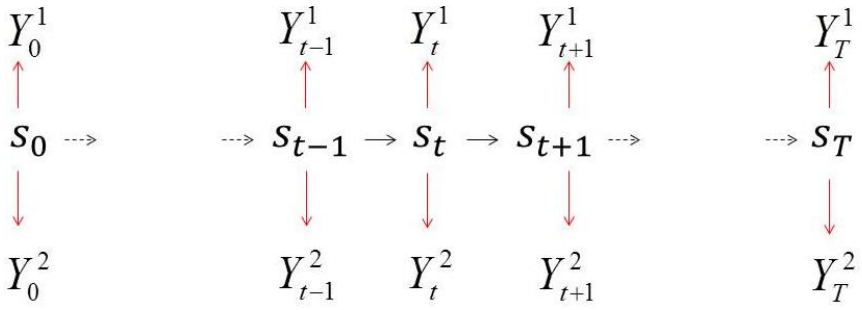
According to the HMM modeling, the latent variable at time *t*, S_t takes values on a finite set of size *r* that we can identify, without loss of generality, with the set (*1,2,..,r*). For a given final time *T*, the values (s_0, s_1, s_2,…, s_T) represent the realization of an unobserved random process *S* at discrete times 0,*1,..,T*. We assume that the stochastic process *S* is a first order Markov process, that is $P(S_{t+1} \mid S_1, S_2, .., S_t) = P(S_{t+1} \mid S_t)$. The law of this process is specified through the *initial probabilities* $p^0_j = P(S_0=j)$ $(j=1,..,r)$, and the *transition probabilities* $p^t_{j|i} = P(S_{t+1}=j \mid S_t=i)$ $(i,j=1,..,r;\ t=1,..,T)$.
Furthermore, we assume that, at each time *t*, a set of *G* imperfect measures Y_t^g $(g=1,..G)$ is also available. If we consider the manifest variables Y_t^g as measures with error of the target variable S_t, it is natural to assume that they take values on the same set (*1,2,..,r*) associated with the categories of S_t. However, in some circumstances it is useful to allow for more general situations where the latent process and the manifest variables take values on different domains. For instance, this is the case if S_t takes values 1=*employed*, 2=*unemployed*, 3= *economically inactive*, while the categories of Y_t^g are only *employed* (1) and *not employed* (2+3).

In the basic version, the measurement process is modeled by assuming *local independence* among the manifest vartiables:

$$P(Y_t^1,...,Y_t^G \mid S_t = s_t) = \prod_{g=1}^{G} P(Y_t^g \mid S_t = s_t). \qquad (3)$$

The meaning of the equation (3) is that the *G* measures Y_t^g are conditionally independent, given the true value of the target variable S_t (see Figure 1 for a graphical representation of the conditional independence structure of the HMM in the case *G*=2).

Figure 1. Hidden Markov Model with G=2 data sources.



Estimates of the model parameters can be obtained via likelihood methods provided that the model is identifiable. The latter condition is trivially not valid if the number of parameters to be estimated is higher than the numbers of distinct combinations of values of the observed variables.
In cases where there exists a one-to-one relation between the categories of the variable Y_t^g and the states of the latent process *S*, the (estimates of the) probabilities $\psi_{j|i}^g \equiv P(Y_t^g = j \mid S_t = i)$ can be used to evaluate the accuracy of the measurement process associated to the source *g*.

The methodology can be easily extended by introducing covariates in the latent process as well as in the measurement model. This is usually done by relating the involved probabilities to the covariates through multinomial-logit models. Moreover, mixtures of HMMs can be used in order to account for possible heterogeneity among the units of the population.

If the latent model is not only used to assess the quality of the available sources, but also to directly provide estimates of some finite population quantities, one can use the Bayes formula to derive the posterior probabilities of the true target variable conditional on the available information (manifest variable and covariates). Specifically, given *G* sequences $Y_{k,1:T}^g \equiv (Y_{k1}^g,...,Y_{kT}^g)$ of values of the manifest variables and values of the covariates $X = (X^L, X^M)$ for each unit *k* of the population, the relevant probability distribution is:

$$P(s_{k1},..s_{kT} \mid Y_{k,1:T}^1,...,Y_{k,1:T}^G; X^L, X^M). \qquad (4)$$

Different usages of distribution (4) are possible. For instance, estimates of linear aggregates referring to time *t* can be obtained by taking expectations from the conditional distributions $P(s_{it} \mid Y_{i,1:T}^1,...,Y_{i,1:T}^G; X^L, X^M)$, resulting by marginalization of (4). Furthermore in a general purpose estimation context, one can build a synthetic micro-data file by random drawing from distribution (4).

3. Experimental study

In this section, we illustrate a simulation study where the methodology described above is applied in different scenarios. The main goal of the study is to assess the robustness of the methodology with respect to departure from the model assumptions. In particular, we are interested in evaluating the robustness of the measurement error parameter estimates and of the aggregate estimates based on prediction of the latent variable given the observed measurements.

In all scenarios, N arrays of T binary values are drawn from a discrete time process which is assumed to be the latent process. The t component of the ith array S_{it} ($i=1,..N;\ t=1,..T$) represents the true employment status of the ith individual at time t in a population of size N. Since according to the international regulation, the reference time for the employment status is the week, we set T=52 which is the number of weeks in a year.

For each individual and each time, two different imperfect measures (Y_{it}^A , Y_{it}^L) of the latent process are also simulated by independently drawing two binary values at each time, conditionally on the realized values of the latent process. In other words, given a realization of the latent variable S_{it}, we draw the two imperfect measures Y_{it}^A, Y_{it}^L from the conditional distributions $P(Y_{it}^A, Y_{it}^L \mid S_{it} = s_{it}) = P(Y_{it}^A \mid S_{it} = s_{it})P(Y_{it}^L \mid S_{it} = s_{it})$.

In order to mimic the real scenario where one of the sources (labour force survey) is available only on a sample of size n, we assume that the measure Y^L is observed only on n units. For the sake of simplicity, we do not take into account the labor force sampling design and we draw the n sample units according to a simple random sampling.

Moreover, in order to reproduce the missing pattern of the labour force survey implied by the sample design, we drop values in the Y^L source so that for each individual i, the corresponding measure is available not more than twice in the year and not more than once in a quarter.

In the following, several simulation scenarios are described differing for the distribution generating both the "true" data and the manifest measures. For each scenario we try to fit data through some latent models and we obtain the model parameters estimates via maximum likelihood estimation. We split data in two datasets E and P: E is used to estimate the model while P is used as test set. Specifically, given the parameter estimates obtained from E, these estimates are used to predict values of the latent variables at different times conditionally on the available information. Evaluation is performed by comparing the estimates of the annual averages of "employed" based on predictions with the corresponding number of true "employed". Moreover, in order to evaluate the capability of the method to correctly assess the quality of the different sources of information, the estimates of the parameters associated with the measurement processes (classification errors) are also compared with the corresponding true values.

For the conditional distributions associated with the measurement processes we will use the following notation:

$$\psi_{k|j}{}^{L} \equiv P(Y_{it}^L = k \mid S_{it} = j),\ \psi_{k|j}{}^{A} \equiv P(Y_{it}^A = k \mid S_{it} = j);\ \ i = 1,..,n;\ \ j,k \in (0,1);\ \ t = 1,..,T$$

while initial probabilities and transition probabilities from state i, to state j will be denoted by $p^0{}_i$ and $p_{j|i} = P(S_{t+1} = j \mid S_t = i)$ respectively. Note that, since the latent processes are supposed to be time homogeneous in all scenarios, dependence on time has been removed from the notation.

A Monte Carlo simulation study is carried on considering three different scenarios. In all scenarios, R=100 replications have been simulated. For each replication N=1000 binary arrays with *T=52* time occasions and two imperfect measures are drawn. In the following, the different simulation scenarios are described.

S1) In this scenario we simulate the latent process as a two state Markov chain with 52 time occasions and the two measurement processes through the specification of the corresponding conditional distributions.

Two experiments $S1_a$ and $S1_b$ are conducted differing for the set of parameters of the measurement process:

$$S1_a: \ \psi_{1|0}{}^{L} = 0.05, \psi_{0|1}{}^{L} = 0.1; \ \psi_{1|0}{}^{A} = 0.2, \psi_{0|1}{}^{A} = 0.1$$

$$S1_b: \ \psi_{1|0}{}^{L} = 0.4, \psi_{0|1}{}^{L} = 0; \ \psi_{1|0}{}^{A} = 0.2, \psi_{0|1}{}^{A} = 0.1$$

The second set of parameters corresponds to situations where one of the two sources measures the target variable correctly when the true value is equal to one, while the probability of misclassification is high when the true value is zero.

The probability at *t=0* and the independent parameters of the transition matrices are in both cases:

p01 = 0.4, p1|0 = 0.07, p0|1= 0.05

For each set of parameters we estimate two models corresponding to different choices of the reference time for the dynamic of the employment status. Specifically, in the first case, we suppose, according to the simulation model, that the reference time for the Markov chain is the week (52 times). In the second model, we synthetize the weekly available information at month level by considering only one value per month of the manifest variables (12 times). In detail, for each month of the year we take for Y^L (representing the labour force survey) the unique available value (when present) as representative of the month. The week representing Y^A in the month is the same as Y^L when it is present, and is randomly selected otherwise. Collapsing information from week level to month level could be an option for dimensionality reduction. Thus, we performed this experiment in order to analyze the impact of the approximation on the accuracy of the estimates.

The main goal of the other experiments is to investigate robustness of the methodology with respect to misspecification of the underlying model.

S2) In this scenario we simulate the latent process as a mixture of two Markov chains *C1* and *C2* with probabilities at *t=0* and transition parameters $p^1_{j|i}$ and $p^2_{j|i}$ given by:

$p^{01}{}_1 = 0.6$ and $p^{01}{}_i = 0.5$; $P^1{}_{1|0} = 0.07$, $p^1{}_{0|1} = 0.05$; $p^2{}_{1|0} = 0.3$, $p^2{}_{0|1} = 0.4$.

The mixing weight of the mixture is $\pi = 0.7$

The measurement processes are simulated according to the following values of the probabilities for the classification errors:

$$\psi_{1|0}{}^{L} = 0.05, \psi_{0|1}{}^{L} = 0.1; \ \psi_{1|0}{}^{A} = 0.2, \psi_{0|1}{}^{A} = 0.1$$

The scenario represents situations where individuals can be classified in two groups with different characteristics in terms of employment dynamics. Total employment and classification errors are

estimated by modeling data both with a simple HMM (basic model) and with the appropriate mixture of HMMs (mix model).

S3) In the last group of experiments we simulate heterogeneity among the units by allowing that the initial probabilities and transition matrix to depend on a set of four binary covariates $X_1,..,X_4$. Dependence is modeled through logit functions. The measurement processes are simulated according to the following values of the probabilities for the classification errors:

$$\psi_{1|0}^{\ L} = 0.05,\ \psi_{0|1}^{\ L} = 0.1;\ \psi_{1|0}^{\ A} = 0.2,\ \psi_{0|1}^{\ A} = 0.1$$

We obtain predictions and estimates of the classification errors using 4 models:

1) A simple HMM (basic)
2) A two component mixture of HMMs (mix2)
3) A three component mixture of HMMs (mix3)
4) The appropriate HMM where covariates for the latent process are correctly specified (cov).

Bias and RMSE for the estimation of the parameters of the measurement processes are reported in Table 1. Figures 1-3 show the distributions of the estimation errors, respectively for scenarios *S1-S3*.

Table1. Bias and RMSE for the estimates of the parameters of the measurement processes

		BIAS				RMSE			
Simulated Model	**Estimated model**	$\psi_{0\|1}^{\ A}$	$\psi_{1\|0}^{\ A}$	$\psi_{0\|1}^{\ L}$	$\psi_{1\|0}^{\ L}$	$\psi_{0\|1}^{\ A}$	$\psi_{1\|0}^{\ A}$	$\psi_{0\|1}^{\ L}$	$\psi_{1\|0}^{\ L}$
S1a scenario	*Basic 52 weeks*	-0.0007	0.0003	-0.0010	-0.0007	0.0050	0.0034	0.0174	0.0114
	Basic 12 months	0.0081	-0.0014	-0.0049	-0.0039	0.0677	0.0145	0.0302	0.0192
S1b scenario	*Basic 52 weeks*	0.0001	-0.0003	0.0009	-0.0028	0.0048	0.0035	0.0025	0.0235
	Basic 12 months	-0.0022	-0.0063	0.0044	-0.0039	0.0216	0.0136	0.0084	0.0261
S2 scenario	*Basic*	0.0581	0.0102	0.0354	-0.0041	0.0588	0.0112	0.0414	0.0148
	Mixture 2 comp	0.0220	0.0242	0.0242	0.0170	0.0927	0.0874	0.1318	0.0794
S3 scenario	*Basic*	0.0272	0.0197	0.0007	0.0020	0.0544	0.0231	0.0371	0.0532
	Mixture 2 comp	0.0087	0.0051	0.0028	0.0006	0.0502	0.0159	0.0364	0.0546
	Mixture 3 comp	0.0029	0.0015	0.0047	0.0010	0.0501	0.0142	0.0374	0.0544
	Covariates	-0.0006	-0.0013	0.0085	0.0057	0.0220	0.0149	0.0376	0.0351

Table 1 shows that the estimates of the parameters $\psi_{j|i}^{\ A}$ and $\psi_{j|i}^{\ L}$ are not biased in all scenarios. The accuracy level seems quite high in all cases except for the scenario *S2* where in some cases the RMSE is around 10%. In particular, the results for scenarios *S1a* and *S1b* show that, as expected, the accuracy level decreases as we move from the weekly reference period to the monthly reference period. Moreover, different sets of parameters in the simulated model do not seem to imply significant change of the accuracy level. All these findings are confirmed in Figure 1. It is worthwhile noting that when the true error parameter $\psi_{0|1}^{\ L}$ is equal to zero the corresponding estimation error vanishes.

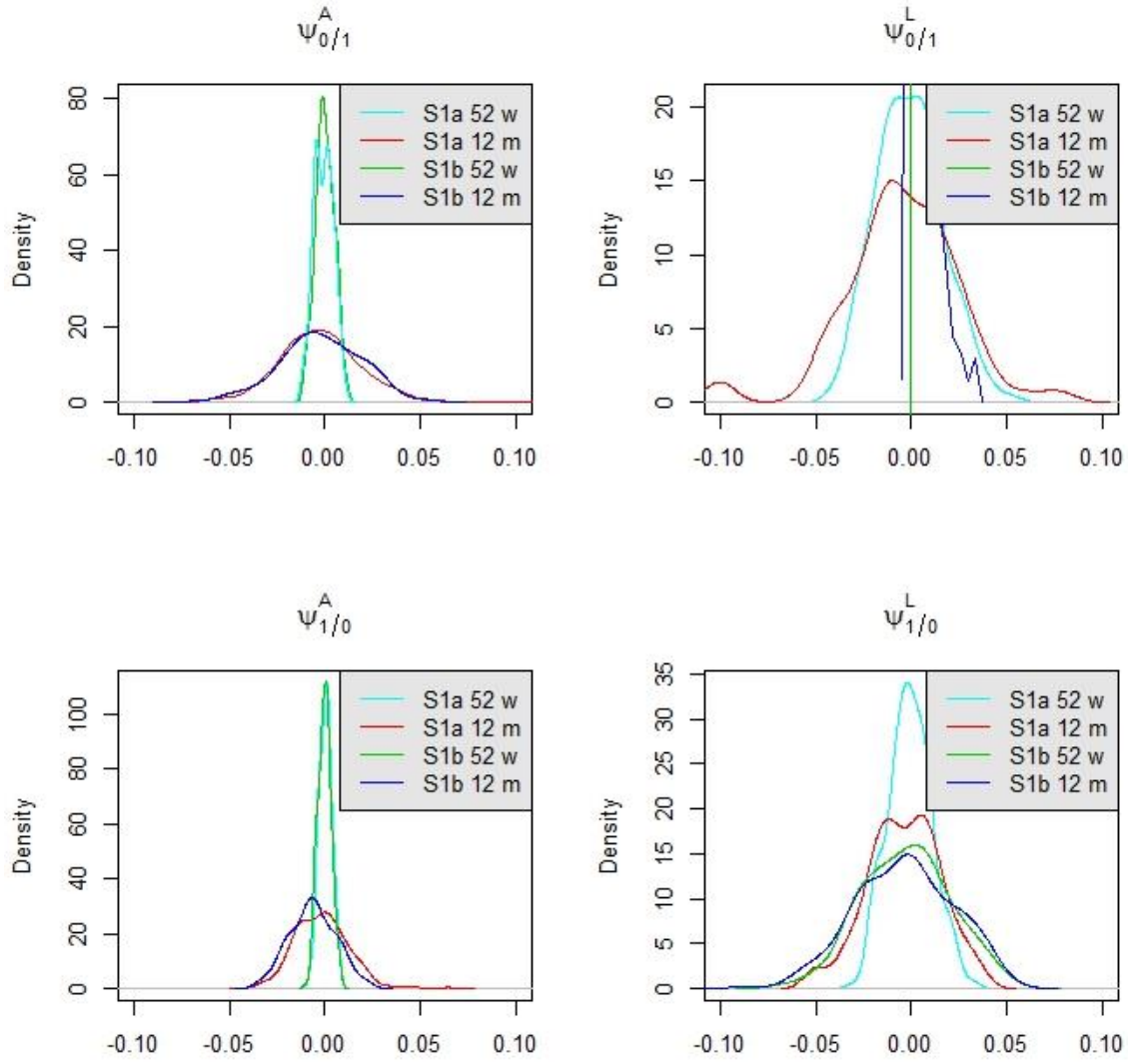


Figure 1. Distributions of the estimation errors for the parameters of the measurement processes - scenarios S1a and S1b

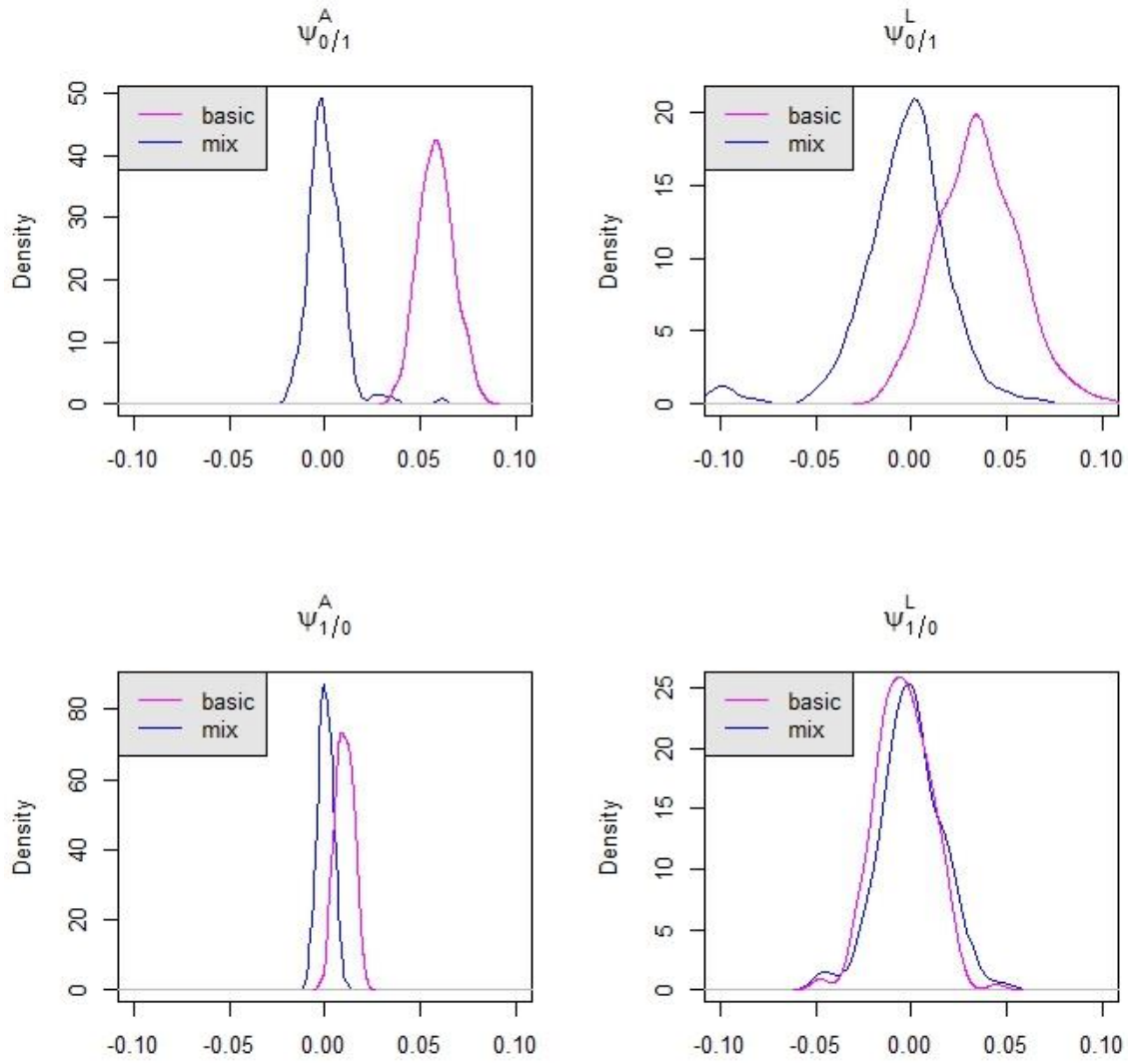


Figure 2 Distributions of the estimation errors for the parameters of the measurement processes - scenario S2

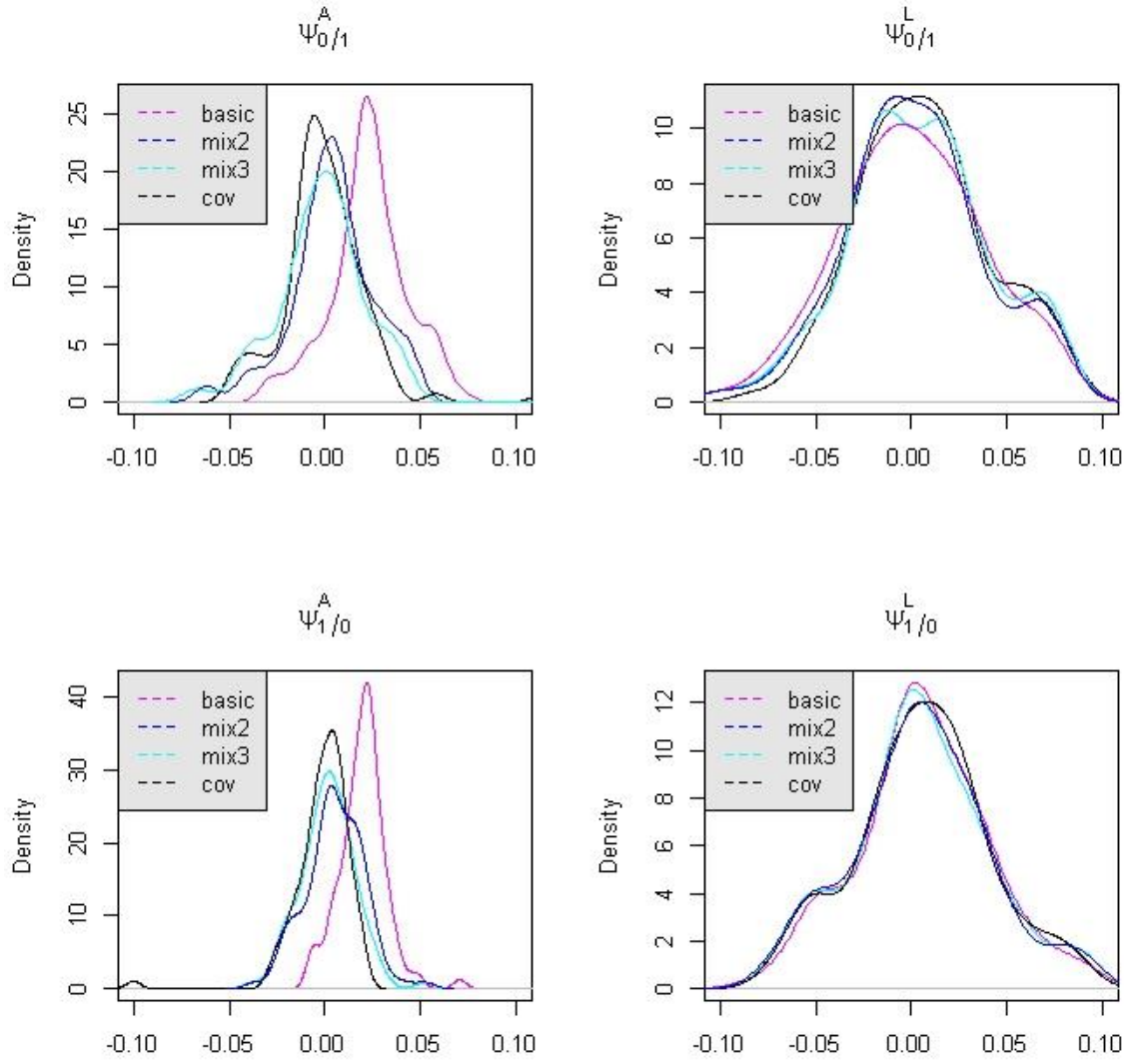


Figure 3. Distributions of the estimation errors for the parameters of the measurement processes - scenario S3

As far as the robustness of the estimation method is concerned, we notice that when we simulate true data from mixture of HMMs, or HMM with covariates (scenarios *S2* and *S3*), the estimates obtained via the basic HMM in most cases are biased. It is also interesting to note that the accuracy level of the error parameters is lower for the measure with an high rate of missing values (ψ^L).

Table2. BIAS and RMSE for the prediction errors

		Prediction errors	
Simulated Model	**Estimated model**	**BIAS**	**RMSE**
S1a scenario	*Basic 52 weeks*	-0.0628	1.6978
	Basic 12 months	0.3714	5.2928
S1b scenario	*Basic 52 weeks*	-0.0621	4.0159
	Basic 12 months	0.2215	9.4780
S2 scenario	*Basic*	7.5007	7.8658
	Mixture 2 comp	0.7219	6.2633
S3 scenario	*Basic*	-1.3861	4.7556
	Mixture 2 comp	-0.4677	4.7199
	Mixture 3 comp	-0.5599	4.6703
	Covariates	-0.2059	4.5738

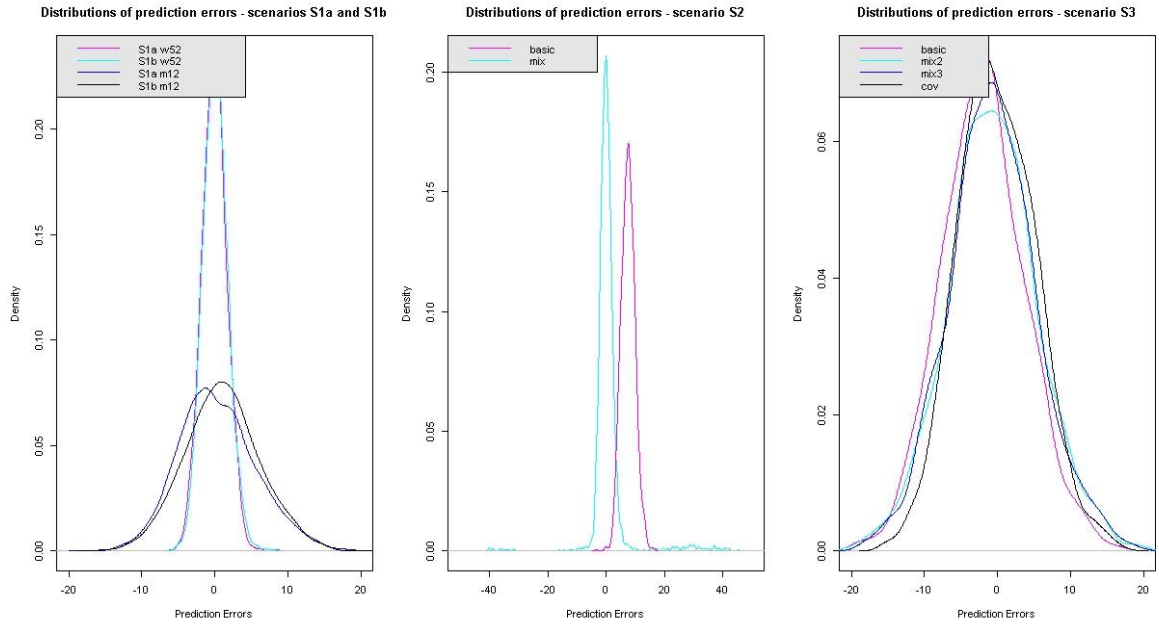


Figure 4 Distributions of the prediction errors in scenarios S1-S3

The results in Table 2 and Figure 4 concerning the distributions of prediction errors agree with the previous findings. In particular, the accuracy level of the predictions is definitely lower when moving from week to month for the reference time. Furthermore, in presence of heterogeneity among individuals (scenarios S2 and S3), the basic model provides strongly biased estimates whereas the mixture of HMMs seems to approximate quite well the true data distribution.

4. Issues for discussion

1. In this paper inference based on latent models is presented as a possible useful approach in presence of different imperfect measures of the variables of interest. One of the possible usages of latent models is to assess the quality of the available sources. Is it appropriate to directly produce estimates based on latent variables in the context of Official Statistics?

2. Instead of directly using the posterior probabilities for estimation, the estimates of misclassification errors could be used to correct estimates based only on the survey data. Can be this approach considered as a valid alternative?

3. For longitudinal categorical data, a natural approach is to model true variables through Hidden Markov Chains. However, in some situations (e.g., when the target variable is the employment status), Markov property does not seem a realistic assumption. Can the approach be extended in order to account for possible departure from the Markov assumption?

4. What valid inferential alternatives can be considered when none of the available sources (survey included) are assumed to be error free?

References

Bartolucci F., Farcomeni F., Pennoni F. (2012). Latent Markov Models for Longitudinal Data. Chapman and Hall/CRC.

Biemer, P.P., and Bushery, J.M. (2000). On the validity of markov latent class analysis for estimating classification error in labor force data. *Survey Methodology*, 26, 2, 139-152.

Biemer, P.P. (2011). *Latent Class Analysis of Survey Error*. New Jersey: John Wiley & Sons, Inc.

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