ST1_1 Suitability Test of Employment Rate for Employees (Wage Labour Force) (ERWLF)

A certain amount of businesses are classified wrongly with regards to activity codes in the business register (BR). Wrong classifications affect the published figures from the ERWLF.

We will perform a suitability test, in order to examine the quality of the ERWLF as a function of the quality of the BR. The test is intended to reveal the margin of error of number of wage laborers within activity groups due to wrong NACE classifications in the Business Register (BR). We will make simulations, where we will simulate distributions of activity codes and examine the effect on published figures from the ERWLF.

Background

The Employment rate for Wage Labor Force (ERWLF) statistics is an indicative statistics reflecting the activity in the Danish labor market. The purpose is creating a fast and efficient indicator rather than creating precise numbers of the Danish employment (labor force). All income taxes from wage earners in Denmark are collected through the E-income register.

The results from the ERWLF are presented on aggregated levels of activity groups. The monthly statistics is published on level 0 with 10 groups. The quarterly statistics is published on level 1 with 20 activity sections.

- Level 0: 10 sections identified by numbers 1 to 10.
- Level 1: 19 sections identified by alphabetical letters A to S. There are 20 sections if X, unknown activity, is included.

Activity is linked to ERWLF through the BR. In order to secure possible reproduction of the statistic, it is based on a frozen version of the BR. The early publication of ERWLF forces the statistic to use the preliminary version of the BR.

Output: Number of Danish wage earners, in total, by NACE activity groups and by regional breakdowns.

Sources: E-income register, population register, business register (BR).

- The E-income register is the backbone of the ERWLF statistics.
- The BR is used to connect business activity with businesses, so it is possible to calculate number of wage earners on business activity groups.
- The first preliminary frozen version of BR is used.
- Numbers of wage earners are also presented on a geographical level. The population register is used to place wage earners geographically in cases where they cannot be placed through their working place.
- It is planned for the ERWLF to also present wage earners on other demographical break downs, that will require an intensified use of the population register.

Undercoverage:

- Self-employed persons who never get their wage through e-income.
- Other wage earners who do not get their wage through e-income. Danish wage earners working in e.g. Sweden or Norway do not get their wage through e-income and are not covered in the ERWLF statistics.
- Unreported work. Due to high taxing level unreported work is somewhat common in Denmark, but only a very small fraction of the working force is believed to work fully unreported and therefore not covered by the e-income register at all.

Overcoverage:

Non-Danish employees payed by Danish employers through E-income register. These wage earners are by definition counted with in the ERWLF figures, even though they strictly taken not are Danish wage earners.

Misclassifications:

When businesses are founded, the businesses themselves complete NACE classifications in the BR register. There is limited check of new registrations in the BR, but often when businesses participate in surveys errors are observed. Such errors are reported back to the BR, where they are corrected.

Small businesses with less than one full time employee (FTE) can be wrongly classified for a long period or even permanently. The quality of 'medium size businesses' (2-10 employees), is better than that of small businesses, but does still contain a proportion of misclassifications. For 'larger businesses' (More than 10 employees), it is believed that misclassifications always will be corrected at some point in the BR. All larger businesses participate in a series of surveys conducted by Statistics Denmark and NACE misclassifications will eventually be discovered when business are selected for a survey, it not is meant to be a part of. Misclassifications for larger businesses, is primarily due to delays in updatings, when businesses merge, split or change activity. Updating of NACE codes in the BR is sometimes, but not always aligned with events of this kind.

Table 1. Number of enterprises and employees by size of enterprise (Enterprises with employees in march 2016).

	Number of enterprises	Number of employees	Share of enterprises	Share of employees
1 or fewer employees	51,961	44,274	27.4%	1.7%
2 - 10 employees	91,189	406,424	48.2%	15.6%
More than 10 employees	46,156	2,160,508	24.4%	82.7%
Total	189,306	2,611,206	100%	100%

Table 1 shows that by far the most wage earners are employed at enterprises with more than 10 employees. 'Small businesses' (0-1 employees) only represent a small fraction of the total and only 1.7% of the employees work at enterprises with 1 or fewer fulltime employees. Hence the relatively large misclassification of small business does only have a limited influence on totals.

Statistics Denmark has previously, conducted surveys for measuring the quality of NACE codes in the Danish BR. These surveys are not up to date, but might still give a good impression of the correctness of the NACE codes. Feedback from both continuous and ad-hoc surveys is used to correct information in the BR. Larger enterprises participate in more surveys and the information on larger enterprises is in general of better quality than the one on smaller enterprises.

Delays in updates in the BR

In some cases businesses are split into smaller businesses or merged with other businesses. In other cases businesses simply change main branch. Updates in BR are not always performed at exactly the same moment as they occur. Misclassifications caused by timeliness issues are included in misclassifications mentioned earlier and will not be treated separately.

By using the old surveys and rates of corrections from newer surveys Statistics Denmark has estimates of the level of misclassifications over time (t) of the NACE codes split on size of businesses.

Frozen versions of the BR are saved, so it is possible to see the effect of corrections over time. This is sometimes referred to as the progressiveness of the register. It is assumed that the Business Register never reaches perfect classifications, but larger enterprises participating in national surveys are expected at some time the achieve correct classification in the BR. We will examine impact on results from the ERWLF using older frozen versions and compare with results when using newer and more updated versions for the same reference period, always assuming that the most updated versions are the most valid ones.

Handling of quality issues in ERWLF

Clearly there are quality issues regarding ERWLF that can be addressed. Since ERWLF primarily is meant to be a fast indicator of the activity in Denmark, many of these issues are dealt with in the definition of the ERWLF. The E-income register cover all "normal" payments of wage laborers in Denmark. The construction of the E-income register is solid and there is no reason to believe that the E-income register will cover more or less payments in near future. Instead of trying to make ERWLF cover unreported work or other undercovered labor, the ERWLF is simply defined as 'The number of wage laborers receiving payment through the E-income register'. Hence by definition ERWLF does not suffer from neither undercoverage nor overcoverage. Solving the coverage issue by definition of the statistics can be considered as a too simplistic solution. On the other hand quantification of coverage problems on ERWLF is very difficult if not impossible to estimate. Due to taxes all (legal) wage payments in Denmark are payed through the E-income register, so it does make perfectly good sense, to define the number of persons in ERWLF as the total number of persons receiving wages through the E-income register.

Accuracy simulations

The ERWLF statistics results in total number of employees and number of employees by NACE sections. When all data are gathered the total number of employees is fixed. We will perform simulations that can assess the impact from NACE misclassifications on output estimates in the ERWLF. We will also perform sensitivity analysis investigating the effect of misclassification rates being higher or lower than the rates estimated by Statistics Denmark.

Table 2. Number of wage earner by activity section in Denmark in March 2016.

Activity section	Description	Number of wage earners
A	Agriculture, forestry and fishery	38,268
B	Mining and oil industry	4,394
C	Industry	291,214
D	Energy	10,231
E	Water and renovation	11,139
F	Construction	144,467
G	Trade	408,681
H	Transport	136,453
I	Hotels and restaurants	98,619
J	Information and communication	100,471
K	Finance and insurance	78,691
L	Real estate and rent	36,811
M	Knowledge based services	140,754
N	Travelling, cleaning and other operations services	141,088
O	Public administration, defense and police	138,963
P	Education	230,469
Q	Healt and social security	491,261
R	Culture and leisure	49,482
S	Other services	59,430
X	Unknow activity	319
Total		2,611,205

Proportion of enterprises with wrong activity codes

Once all data are available the total number of wage earners is fixed since by definition the total number of wage earners equals the total number of persons in the E-income register. Wrong NACE classifications on the other hand will lead to wrong figures on number of wage earners by activity section. With known fractions of wrongly coded enterprises in the BR it is possible to simulate 'likely distributions' of wage earners by activity section.

Accuracy simulations

Data from March 2016 are used to simulate distributions of wage earners in Denmark by activity section. There were 189,306 enterprises paying salaries to 2,611,206 wage earners. The enterprises are grouped at a 20 category level with number of wage earners ranging from 4,394 to 491,261 in each category (Table 2).

Simulation model:

The input needed for the simulations is data on enterprise level with number of employees and NACE classifications and expected number of misclassifications by size of enterprise.

Statistics Denmark has conducted three so called recoding projects where the aim was to correct NACE codes. Two surveys with 3,000 enterprises in 2006 and 2009, and one survey with 50,000 enterprises in 2007. Even though no surveys have conducted with the same purpose since 2009, the impression from experienced employees at the BR department at Statistics Denmark is that the proportions of misclassifications are roughly the same today as when the above mentioned surveys were conducted. This impression is primarily based on questions regarding activity, which are a part of any business survey conducted at Statistics Denmark. The estimated proportions of misclassifications within size group of business and NACE aggregation level can be found in Table 3.

Table 3. Proportion of missclassified businesses by size of business in Full Time Employees (FTE) and aggregated level of business activities.

	Aggregation level of business activities	
Size of business (FTE)	Level 0 (10 cat.)	Level 1 (20 cat.)
Enterprises with 1 FTE or less	8%	10%
Enterprises with 2 to 10 FTE	6%	8%
Enterprises with more than 10 FTE	3%	4%

In order to simulate accuracy on activity sections it is not enough to know the proportion of wrongly classified enterprises. It is also necessary to know which activity sections wrongly coded enterprises are likely to belong to. Hence a confusion matrix, with the expected distribution of wrongly coded business is required. Each row in a confusion matrix will add to 1 and the values in the diagonal reflect the probability of correct coding within each activity group, derived as complement of the corresponding figure in Table 3. Table 4 shows the confusion matrix for smaller enterprises and level 1 NACE classification.

Table 4. Confusion matrix used for simulation between business sections.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	X
1	0.90	0.02	0.02	0.02	0.00	0.01	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0
2	0.01	0.90	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0
3	0.01	0.01	0.90	0.01	0.01	0.01	0.01	0.01	0.01	0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0
4	0.01	0.01	0.01	0.90	0.01	0.01	0.01	0.01	0.01	0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0
5	0.01	0.01	0.01	0.01	0.90	0.01	0.01	0.01	0.01	0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0
6	0.01	0.01	0.01	0.01	0.01	0.90	0.01	0.01	0.01	0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0
7	0.00	0.00	0.00	0.00	0.00	0.00	0.90	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.00	0.00	0.01	0.01	0.01	0
8	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.90	0.01	0.01	0.01	0.01	0.01	0.01	0.00	0.00	0.01	0.01	0.01	0
9	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.90	0.01	0.01	0.01	0.01	0.01	0.00	0.00	0.01	0.01	0.01	0
10	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.01	0.90	0.02	0.01	0.01	0.01	0.00	0.00	0.00	0.01	0.01	0
11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.90	0.05	0.01	0.01	0.01	0.00	0.00	0.00	0.00	0
12	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.05	0.90	0.02	0.02	0.01	0.00	0.00	0.00	0.00	0
13	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.01	0.90	0.02	0.01	0.01	0.01	0.01	0.01	0
14	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.01	0.02	0.90	0.01	0.01	0.01	0.01	0.01	0
15	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.90	0.03	0.03	0.02	0.02	0
16	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.90	0.03	0.02	0.02	0
17	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.03	0.90	0.02	0.02	0
18	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.01	0.01	0.01	0.90	0.05	0
19	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.01	0.01	0.01	0.05	0.90	0
20	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.10	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0

The confusion matrix in Table 4 is constructed by letting the diagonal be the proportion of correct classified enterprises. The sum of each row equals 1 and corresponds to the probabilities for correct business sections. This transition matrix is used for simulating small enterprises on a 20 group level. Note the 0.9 in the diagonal corresponds to estimated proportion of smaller business that are classified correctly (Table 3) Hence the remaining probability is distributed in other sections, taking similarities between sections and size of sections into consideration. Another way to construct the confusion matrix would be to observe movements between sections over time and thereby construct an evidence based confusion matrix rather than a confusion matrix based on subjective relationships between business sections. Note that the last column (X = unknown activity) is zero. No enterprise correctly belongs to “unknown”, hence the zeros.

A simulation study has been performed where the activity section is simulated by using the observed activity and confusion matrices as the one in Table 4. When all enterprises have a simulated activity code, simulated number of wage earners can be calculated in the same way as in the public figures. In the case where 0.9 is in the diagonal for enterprises with 1 or fewer FTE’s, confusion matrices with 0.92 and 0.96 are used for enterprises with 2-10 FTE’s and more than 10 FTE’s respectively.

Fout! Verwijzingsbron niet gevonden. shows results from the simulations. The coefficient of variation varies between 0.9% and 13.2% with a relationship towards higher coefficient of variation for businesses with few wage earners (Figure 1). In the simulations all enterprises with unknown activity have been placed in other activity sections and hence the simulated number of wage earners with unknown activity always becomes zero.

Table 5. Results from simulation study of estimated number of wage earners by business section with basic proportion of correct classified enterprises equals 90% for small enterprises.

Activity section	Official figures	Std dev	Coeffcient of variation	5% quantile	95% quantile
A	38,268	1,246	3.3%	37,307	41,974
B	4,394	579	13.2%	2,488	4,490
C	291,214	3,562	1.2%	244,662	258,059
D	10,231	542	5.3%	8,043	10,050
E	11,139	467	4.2%	9,461	11,133
F	144,467	1,731	1.2%	128,678	135,190
G	408,681	3,484	0.9%	385,552	400,527
H	136,453	3,388	2.5%	120,790	133,705
I	98,619	2,076	2.1%	98,629	106,300
J	100,471	2,508	2.5%	93,355	103,324
K	78,691	2,376	3.0%	77,041	86,227
L	36,811	2,379	6.5%	54,884	64,748
M	140,754	2,786	2.0%	143,532	154,776
N	141,088	2,750	2.0%	132,085	142,471
O	138,963	3,479	2.5%	123,074	136,463
P	230,469	4,653	2.0%	219,074	238,078
Q	491,261	6,904	1.4%	467,298	495,908
R	49,482	3,254	6.6%	65,945	78,654
S	59,430	4,422	7.4%	98,701	116,425
X	319	0	0.0%	0	0

In order to study the sensitivity of the simulations confusion matrices with 0.85 and 0.95 correct classification proportions for small enterprises are also performed. Proportions of correct classified enterprises in other size groups are adjusted proportionally in the same directions. Results from the simulations can be seen in

Table 6. When percentage of wrong classified is doubled (correct changed from 90 to 95% for small enterprises), the standard deviation is reduced by just short of factor of square root of two in average. Seemingly the condition of a constant total of number of wage earners reduces the variation, compared with a the variation in a simple random sample.

Table 6. Standard deviations for simulated number of employees within NACE 19 group classification with three different base probabilities of correct classification.

NACE 19 classification group	85% correct classified	90% correct classified	95% correct classified
A	1,457	1,246	884
B	666	579	407
C	4,108	3,562	2,479
D	684	542	433
E	598	467	363
F	2,042	1,731	1,280
G	4,492	3,484	2,652
H	4,013	3,388	2,639
I	2,384	2,076	1,364
J	3,072	2,508	1,904
K	2,910	2,376	1,775
L	3,028	2,379	1,751
M	3,417	2,786	2,046
N	3,215	2,750	1,929
O	4,084	3,479	2,427
P	5,737	4,653	3,343
Q	8,585	6,904	4,989
R	3,902	3,254	2,284
S	5,577	4,422	3,222
X	0	0	0

Comparison of frozen versions

The Business Register (BR) is a living register, where information continually is updated. That means that the same query in the BR today can give another result tomorrow, e.g. because of business closure or reclassification of NACE group. Published statistics are always based on frozen versions of the BR. This makes it possible to reconstruct published statistics, by using the same frozen version as the one used originally. A change in the preliminary frozen version of the same period is triggered once information about enterprises is corrected back in time in the BR.

Table 7. Number of enterprises that have changed activity compared to the preliminary version of the BR and number of employees in these enterprises. The first two lines are changes in frozen versions of the same period, while the next two lines are changes from the preliminary frozen version in the given quarter to the preliminary frozen version in the following quarter.

Year	2014				2015				2016	
Quarter	1	2	3	4	1	2	3	4	1	2
Changes from preliminary version to final frozen version										
Enterprises	3,479	7,656	3,984	2,001	3,645	3,652	2,478	3,757	4,789	3,029
Employees	5,443	9,421	5,930	5,018	7,045	8,417	6,575	7,061	15,507	3,745
Changes compared to next final frozen version										
Enterprises	18,387	23,648	17,297	20,862	15,888	16,620	18,143	24,656	19,591	19,318
Employees	22,950	18,818	14,654	19,954	17,737	17,737	17,601	22,189	27,099	13,748
Ratio between corrections from preliminary to final frozen and changes to next final frozen version										
Enterprises	18.9%	32.4%	23.0%	9.6%	22.9%	22.0%	13.7%	15.2%	24.4%	15.7%
Employees	23.7%	50.1%	40.5%	25.1%	39.7%	47.5%	37.4%	31.8%	57.2%	27.2%

The ERWLF uses the preliminary frozen version of the BR. In Table 7 the numbers of enterprises that have changed activity from the preliminary version to the last frozen version are shown. Below is the number of enterprises changing activity from one period to another activity the following period.

The BR is updated more than three months back, so an update in the final version of a frozen version compared to the preliminary version, does not always compare to a change from one quarter to another. The frozen versions of the BR are corrected approximately one year back, compared to the preliminary version.

On the other hand if an update from one period to another is not registered within three months, but within a year it will always result in a change from the preliminary version to the final version. Hence the large amount of changes from one period to another reflects that by far the most changes of the BR are made within a time period, so that the first preliminary version is correct. The ratios between corrections of frozen versions and changes between preliminary versions are between 10 and 32% of the enterprises and 24 and 57% of the employees. These figures reflect that a fairly large part of the changes in the BR are corrected in time for the quarterly preliminary version to be updated with changed activity codes.

Conclusion

The total number of wage earners is not affected by wrong business classifications, but on activity section level official figures are affected by wrong NACE classifications. If there are estimates of the fractions of enterprises with wrong classifications we have demonstrated a method to calculated standard deviations and variation of coefficients of official figures as a function of wrong NACE classifications. Previous surveys conducted at Statistics Denmark have provided figures that can be used as input for these simulations.

Comparisons of frozen versions show that by far the most changes of activity the BR are made quickly after the change. By comparing changes of frozen versions over the same period with changes between different periods it is possible to conclude that most changes are made in time to be corrected already in the first preliminary version on a quarterly level.

Appendix. Other tables with changes in frozen versions.

Table 8: Number of enterprises in BRWLF split on NACE branches in first and last frozen version of BR.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	X
A	9477	0	3	1	0	5	5	4	1	0	2	6	1	3	0	0	0	1	1	0
B	0	209	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0
C	1	0	11016	1	1	6	31	1	3	3	15	4	9	5	0	0	0	1	1	0
D	1	0	1	557	0	0	2	0	0	1	0	0	1	1	0	0	0	0	0	0
E	0	0	0	0	1187	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0
F	4	0	15	0	1	16778	8	1	1	4	22	10	12	11	1	2	0	0	1	0
G	7	0	43	1	0	10	38526	16	10	9	45	28	28	28	0	3	2	1	5	0
H	3	8	4	0	0	9	8	7305	0	1	6	0	6	14	0	3	1	1	3	0
I	1	0	1	0	0	4	16	3	11090	1	11	6	4	4	1	0	0	1	4	0
J	0	1	3	0	0	1	10	2	0	7929	8	0	22	9	0	2	1	2	0	0
K	1	0	3	0	0	3	6	0	1	7	5400	6	13	3	0	0	0	0	1	0
L	7	0	10	0	0	10	5	3	3	0	18	9863	6	5	0	0	0	3	7	0
M	0	2	10	2	0	10	18	2	2	23	39	14	15008	23	0	3	0	0	3	0
N	2	0	4	0	0	19	11	6	6	2	11	9	15	8686	0	0	6	3	1	0
O	0	0	0	0	0	0	0	0	0	1	0	0	1	1	1791	1	5	1	0	0
P	0	0	0	0	0	0	1	0	1	0	4	0	8	2	1	5708	6	2	0	0
Q	0	0	1	0	0	1	2	0	4	1	12	7	3	2	6	5	20169	3	3	0
R	0	0	2	0	0	0	3	0	3	2	2	1	0	2	0	1	1	4772	3	0
S	0	0	2	0	0	2	5	0	2	0	1	3	1	4	0	2	3	0	9573	0
X	0	0	0	0	0	0	0	0	1	1	0	2	0	0	0	0	0	0	0	25

Table 9. Comparisons of preliminary version of BR and final frozen version for 1st quarter 2016. Only enterprises with wage earners are counted.

	A	B	C	D	E	F	G	H	I	J
Wage earners first frozen version	38268	4394	291223	10231	11139	144479	408717	136453	98626	100471
Wage earners last frozen version	38292	4535	291173	10214	11140	144491	408922	135585	98659	100239
Enterprises first frozen version	9687	214	11252	571	1197	17049	39528	7536	11527	8132
Enterprises last frozen version	9681	223	11272	569	1197	17037	39423	7507	11508	8127
Change in number of wage earners	0.1%	3.2%	0.0%	-0.2%	0.0%	0.0%	0.1%	-0.6%	0.0%	-0.2%
Change in number of enterprises	-0.1%	4.2%	0.2%	-0.4%	0.0%	-0.1%	-0.3%	-0.4%	-0.2%	-0.1%
	K	L	M	N	O	P	Q	R	S	X
Wage earners first frozen version	78693	36811	140756	141089	138963	230469	491287	49482	59430	319
Wage earners last frozen version	79298	37080	140564	141366	139068	230395	491291	49534	59145	310
Enterprises first frozen version	5538	10029	15314	8926	1878	5830	20533	4833	9698	49
Enterprises last frozen version	5691	10050	15293	8948	1877	5827	20508	4832	9706	45
Change in number of wage earners	0.8%	0.7%	-0.1%	0.2%	0.1%	0.0%	0.0%	0.1%	-0.5%	-2.8%
Change in number of enterprises	2.8%	0.2%	-0.1%	0.2%	-0.1%	-0.1%	-0.1%	0.0%	0.1%	-8.2%

ST2_1 Overlapping numerical variables without a benchmark: 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:

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

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

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. 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.

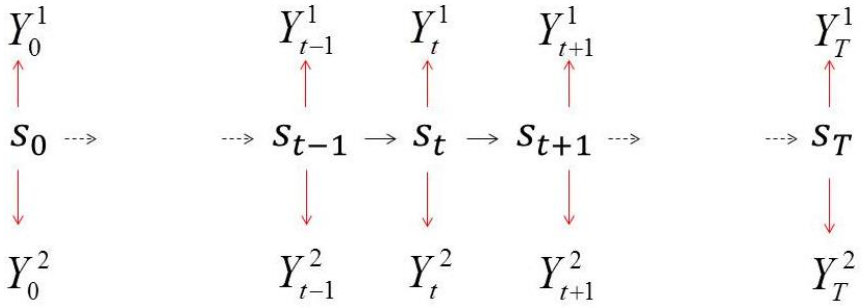
According to the HMM modeling, the latent variable at time *t*, S_t takes values on a finite set of size *r* that we can identify, without loss of generality, with the set (*1,2,..,r*). For a given final time *T*, the values (s_0, s_1, s_2,…, s_T) represent the realization of an unobserved random process *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 take 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 state 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 distribution $P(s_{it} \mid Y_{i,1:T}^1,..., Y_{i,1:T}^G; X^L, X^M)$, 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 distribution $P\,(Y_{it}^{A}, Y_{it}^{L} \mid S_{it} = s_{it}) = P(Y_{it}^{A} \mid S_{it} = s_{it})P(Y_{it}^{L} \mid S_{it} = s_{it})$.

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 (S1-S3) 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 (Root Mean Squared Error) 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.

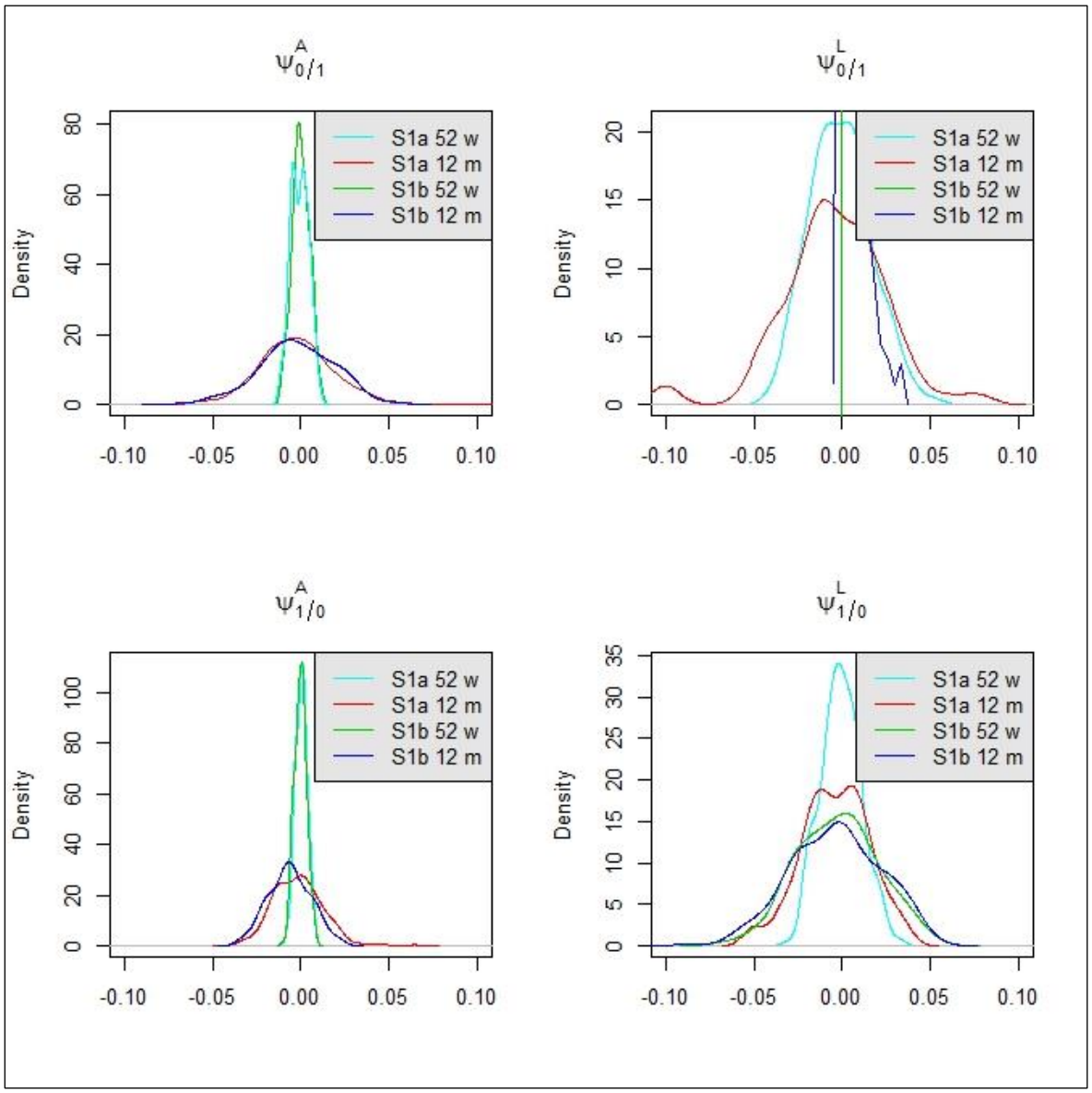


Figure 1. Distributions of the estimation errors for the parameters of the measurement processes - scenarios S1a and S1b

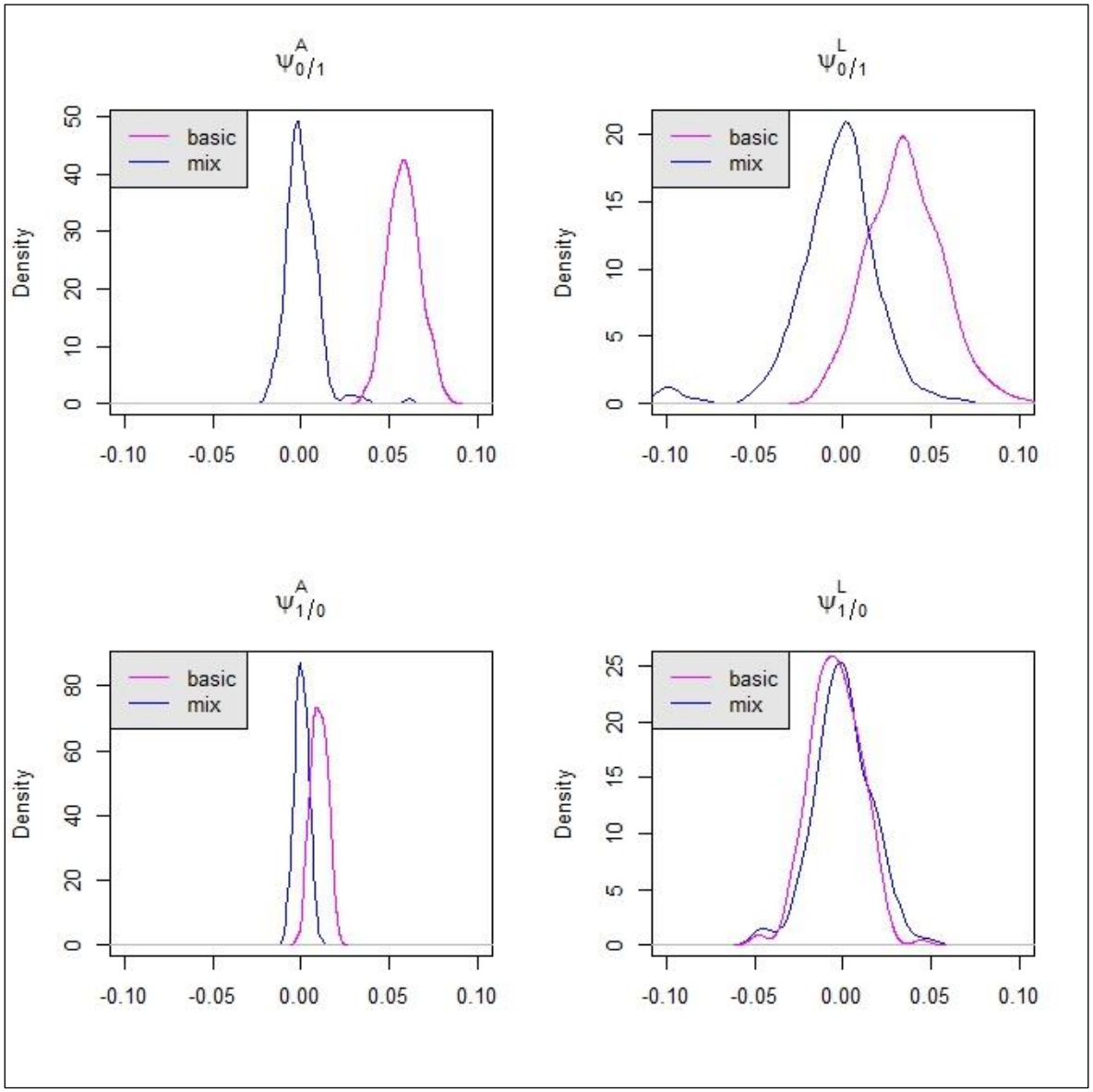


Figure 2 Distributions of the estimation errors for the parameters of the measurement processes - scenario S2

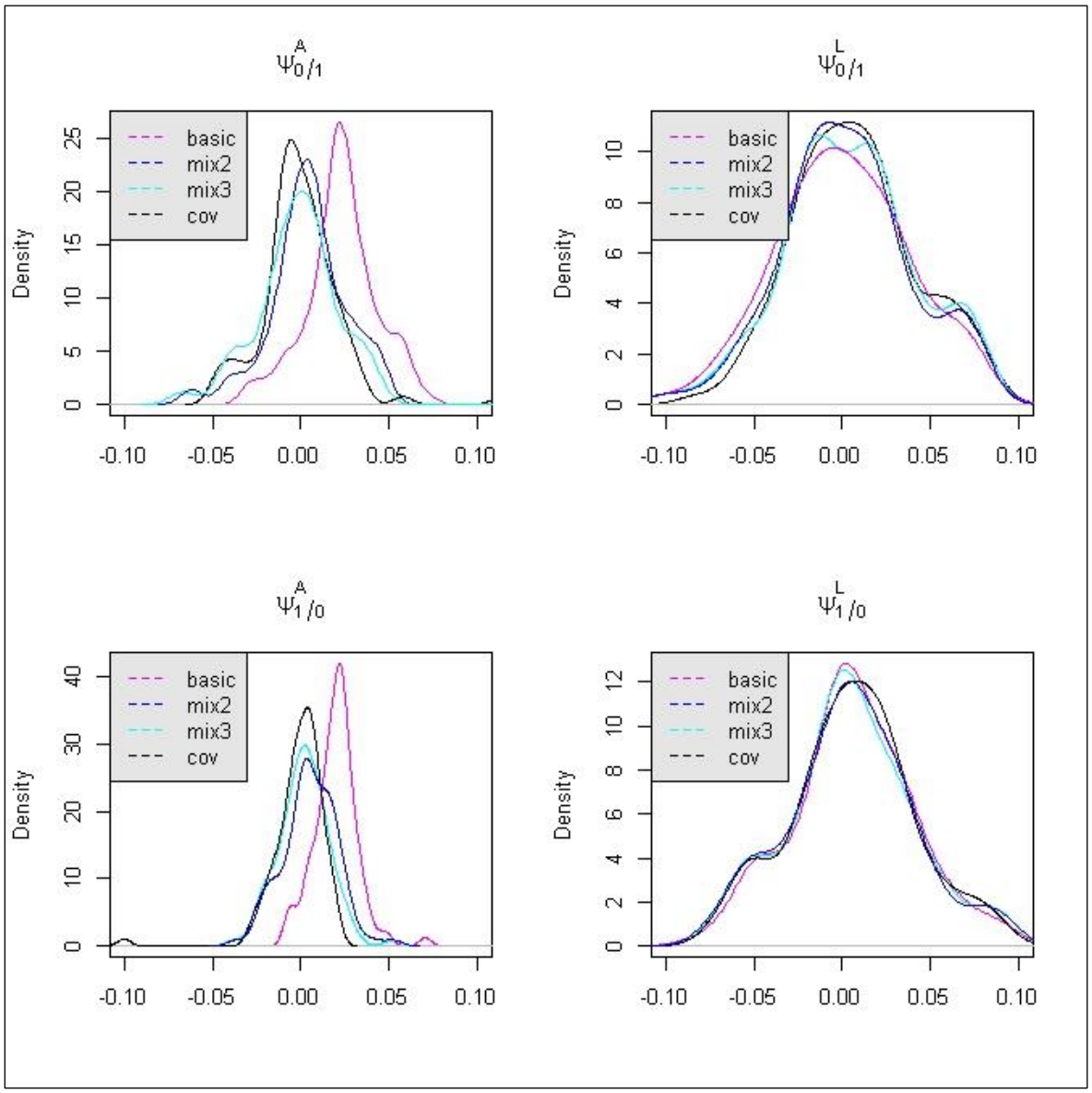


Figure 3. Distributions of the estimation errors for the parameters of the measurement processes - scenario S3

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 a 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

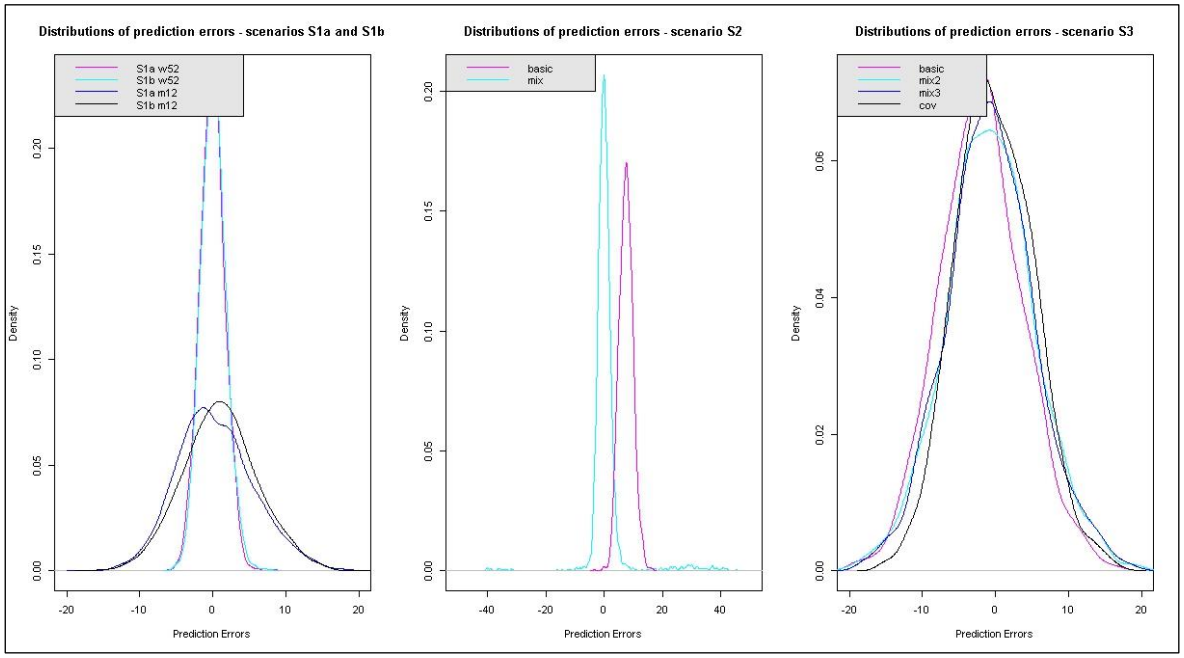


Figure 4 Distributions of the prediction errors in scenarios S1-S3

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.

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.

ST2_2 Overlapping numerical variables with a benchmark

Andrea Fasulo, Fabrizio Solari

Data configuration ISTAT has also examined basic data configuration 2 for the situation where there are overlapping units as well as overlapping variables in the data sources to be combined and one of the data sources (typically a survey measure) is considered as error free and the administrative measures are merely used as auxiliary information.

Thus, differently from the situation in Istat Suitability test n.1, the error free variable (Y) is modelled as a *response* variable and all the other measures (X) are considered *covariates*. This supervised approach can be adopted in a model based inference as well as in a design based inference. In the latter case, the covariates can be used to specify a *working* model (model assisted approach). The choice of the methodological approach depends both on the informative content and the quality of the available data sources.

In this situation, where administrative data play the role of auxiliary variables, evaluation of accuracy is primary based on the estimation of sampling error.

If we assume that administrative and survey data have an asymmetric role, then adequate statistical models can be applied only to the units belonging to survey samples and the corresponding estimates can then be used to predict the phenomena over the unsampled population units, either using a model-based (not considering sample weights) or a design-based approach (considering information deriving from the survey sampling scheme).

The choice of a projection unit level estimator is connected with (i) the purpose of producing a microdata archive which can be used to spread information at the required territorial levels, such as municipal level and (ii) and the need to produce coherent estimates with those provided by different ISTAT surveys dealing with the same thematic areas of interest (employment area in this case).

Also in this setting one can use the employment status provided by the Labour Force Survey (LFS), which provides the target variable that need to be estimated, and the administrative data as auxiliary determinants in estimating the phenomena of interest.

Different options for the estimator will be tested. The first is the model assisted linear estimator proposed by Kim and Rao (2012). Other estimators are the composite estimator deriving from the linear mixed model where the random effects are the Labour Market Areas (LMAs).

Data used Again the goal of the study is to combine administrative and survey data in order to build a "labour register" to be used for producing estimates on the employment status at fine level. To this aim, data from the Italian Labour Force Survey (LFS) and administrative data strongly associated with the target phenomenon are used. The experimental study is conducted on data coming from the Italian regions Trentino-Alto Adige and Marche.

Results In order to analyse the feasibility of the application of a projection estimator to employment LFS data considering as auxiliary variables administrative information already mentioned before, with the aim to produce different territorial levels estimates (provinces, macro LMAs, LMAs and municipalities) a simulation scheme has been implemented. This simulation scheme consists of drawing different samples,

estimating the model over the samples and projecting the results over the entire population. The performance is assessed in terms of bias and mean squared errors by means of:

$$\text{MARE} = \frac{1}{D}\sum_{d=1}^{D}\left|\frac{1}{R}\sum_{r=1}^{R}\hat{y}_{rd} - Y_d\right|, \; ARRMSE = \frac{1}{D}\sum_{d=1}^{D}\frac{1}{R}\frac{\sqrt{\sum_{r=1}^{R}(\hat{y}_{rd}-Y_d)^2}}{Y_d}$$

The following model specifications have been used in the experimental study.

Projection on pooled sample

Full model: Marital status, educational level, citizenship, not in labour force, cross classification gender-age, register variable by region

Reduced model: Marital status, educational level, citizenship, cross classification gender-age, register variable by region

Projection on register

Minimal model: Marital status, citizenship, cross classification gender-age, register variable by region

The main results in terms of bias and mean squared errors for the target variables employment status and unemployment status are given in Tables 1 and 2 below.

Table 1. MARE and ARRMSE for the variable employment status

	GREG LFS	**Projection LFS on Pooled Reduced model**	**Projection LFS on Pooled Full model**	**Projection LFS on Register Minimal model**	**Projection Pooled on Register Minimal model**	**Pooled**	**EBLUP**
Mean Absolute Relative Error - Employed							
R^2	-	0.89	0.95	0.89	0.89	-	-
Province (7)	0.33	0.55	0.56	0.12	0.08	0.60	-
Macro LMA (14)	1.97	0.47	0.49	0.14	0.09	0.44	2.42
LMA (54)	232.12	73.32	74.04	1.15	1.11	72.32	2.74
Municipality (527)	1779.27	550.48	550.09	1.97	1.96	551.12	3.48
Average Relative Root Mean Squared Error - Employed							
Province (7)	4.126	5.52	5.49	1.51	0.95	5.4	-
Macro LMA (14)	21.36	3.15	2.96	2.07	1.31	2.73	3.74
LMA (54)	264.32	109.25	110.96	2.60	1.91	108.25	4.08
Municipality (527)	1791.43	608.87	608.87	3.10	2.52	610.36	4.62

Table 2. MARE and ARRMSE for the variable unemployment status

	GREG LFS	**Projection LFS-Pooled Reduced model**	**Projection LFS-Pooled Full model**	**Projection LSF-Register Minimal model**	**Projection Pooled-Register Minimal model**	**Pooled**	**EBLUP**

Mean Absolute Relative Error - Unemployed							
R^2	-	0.15	0.33	0.15	0.14	-	-
Province (7)	0.88	1.04	2.12	0.96	0.46	0.61	-
Macro LMA (14)	2.53	1.25	1.40	1.36	1.05	0.98	34.19
LMA (54)	243.45	68.22	44.52	12.32	12.78	82.3	49.28
LMA in-sample (26)	8.06	7.23	6.28	5.25	5.22	2.78	36.55
Average Relative Root Mean Squared Error - Unemployed							
Province (7)	16.12	15.56	15.01	14.55	9.33	12.78	-
Macro LMA (14)	30.65	22.52	22.21	22.12	14.37	15.85	42.82
LMA (54)	312.93	111.71	99.37	29.08	21.09	136.72	57.45
LMA in-sample (26)	54.58	34.63	34.54	22.22	15.83	33.11	44.37

Analysis of results

The results for the variable *employed* are shown in Table 1 in which the R^2 shows very high values for all the models. The MARE and the ARRMSE indicators are computed for the four type of domains described above. At province level the best results are obtained by the Projection estimator using the register, but also good performances are obtained for direct GREG estimator. At Macro-LMA level the GREG estimator loose its good properties showed at provinces level, presenting a huge increase of variability (ARRMSE 21.36%). The Pooled estimators presents still good results on this level, very closed to the Projection estimator using the register. At LMA and at municipality level the estimators both based on the LF data or on the pooled sample show very poor results with respect to those referred to macro-LMA. This is due to the fact that on 54 LMA areas included in the regions only 26 are always present in the 200 simulations, while for the municipalities on 572 areas only 27 are always included in the simulations. For this reason, the synthetic estimators (projection on register and SAE estimator) show similar results in terms of bias and variability as well.

The results for the variable un*employed* are shown in Table 2. For this variable the R^2 value is similar using the *reduced* and the *minimal* model (14-15%) while goes up to the 33% using the *full* model. As well as for the employed, at provinces level and at Macro-LMA good results are obtained from the GREG estimator and from the Pooled estimator, especially in terms of bias. At LMA level only the projection on register estimator show good performance both with the MARE and the ARRMSE indicators below the threshold of the 13% and the 30%. The table 3 shows that considering only the 26 LMAs always sampled in the 200 simulations, the bias estimates goes down up to the 5%.

Gap analysis: A very useful aspect to be accounted for in the employment status estimation at a certain time *t*, is the one relying to the individual previous times employment status information. Considering the longitudinal aspect is not always trivial. An interesting research topic we propose to develop in the future is inherent to the utilization of longitudinal models in case of not balanced samples and in a projection model-assisted approach for micro data. This would make the developed approach more generally applicable.

References

Kim J.K., Rao J.N.K. (2012) Combining data from two independent surveys: a model-assisted approach, *Biometrika*, Vol. 99(1), pp. 85-100.