Statistics Finland

Accuracy of imputation:

a Case Study on the Finnish Labour Force Survey

Kari Djerf, Atte Lintilä, Riku Salonen, Ari Veijanen

BaNoCoSS 2015

Contents

- Characteristics of the Finnish LFS
- Missing data in January 2015
- Missing data treatment
- Evaluation of imputation

LFS design

- A monthly survey

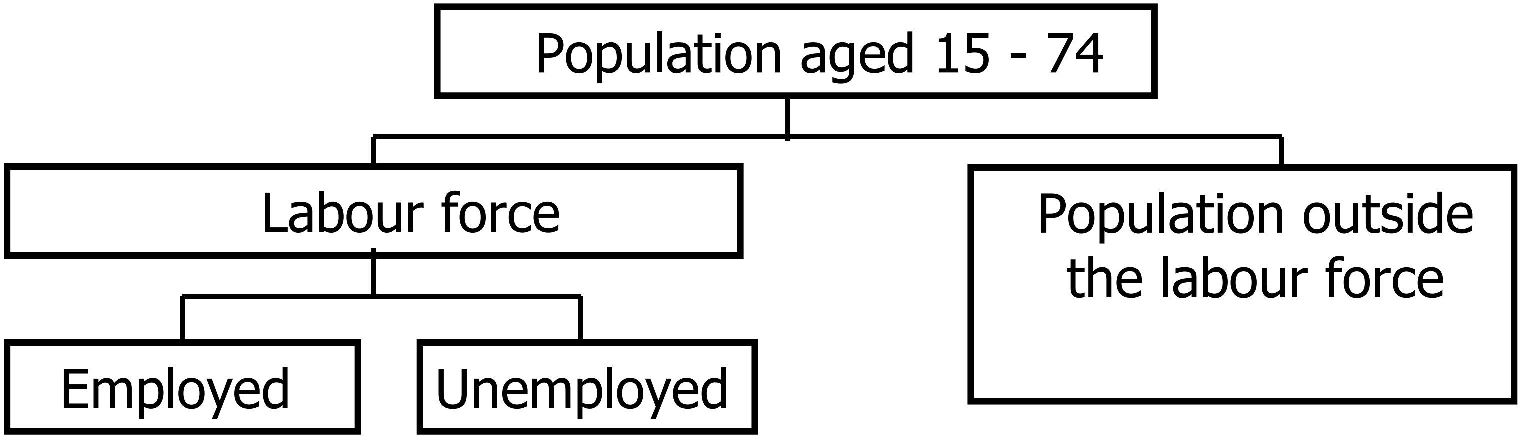
- Target population individual persons, aged 15-74

- Rotating panel design: 5 waves with 3/month interval
- Sample size 12 500 a month with 2 500,
  Quarterly 37 500

- Telephone interview

- Unit nonresponse rate now 25-30%
- Item nonresponse very small, e.g. working time information sometimes missing

Labour force status definition



Population aged 15 - 74

Labour force

Population outside the labour force

Employed

Unemployed

Labour force status definition – 2

- An **employed** person is a person who:
  - did at least one hour of paid work during the survey week, OR
  - has been temporarily absent from work, e.g. on vacation or ill.
- If the person has been absent from work during the survey week, he or she is classified as employed if:
  - the reason for the absence is the person’s own illness or maternity or paternity leave, **or**
  - the absence has lasted for less than three months, **or**
  - the person is still being paid wages, salary or other income-related compensation corresponding to at least half of what he or she receives while normally employed.

Labour force status definition – 3

- An **unemployed** person is a person who:
  - is without work, **and**
  - has taken specific steps during the last four weeks to seek employment, **and**
  - is available to start work within two weeks,
OR
  - is waiting for an agreed job to begin within three months, **and**
  - would still be available to start work within two weeks.

- All others are **outside labour force**.

Missing data in January 2015

- There is a number of questions to determine whether the respondent is unemployed or not
- In January one crucial question was, however, missed due to very last-minute change in the CAPI program
- Question EE13 determines whether the unemployed person is ready take a new job in two weeks time.
- 840 respondents should have replied to the question until the problem was found
- After the problem was exposed the CAPI program was corrected and all cases with missingness were sent back to field-work for re-interview

Missing data treatment - strategy

- Because of very short fieldwork time all missing cases were imputed:
  - Those who were already interviewed in earlier panel waves and who had been unemployed were treated with cold-deck imputation, i.e. reply from previous interview provided that many questions leading to EE13 were replied in the same way
  - Those of the first wave or those whose labour force status had changed since the previous interview were treated with model-imputation: weighted sequential hot deck (SUDAAN: Proc Impute, single imputation)
  - Pool of donors was taken from all respondents in 2014 modeled with logit model: each donor was taken only once.

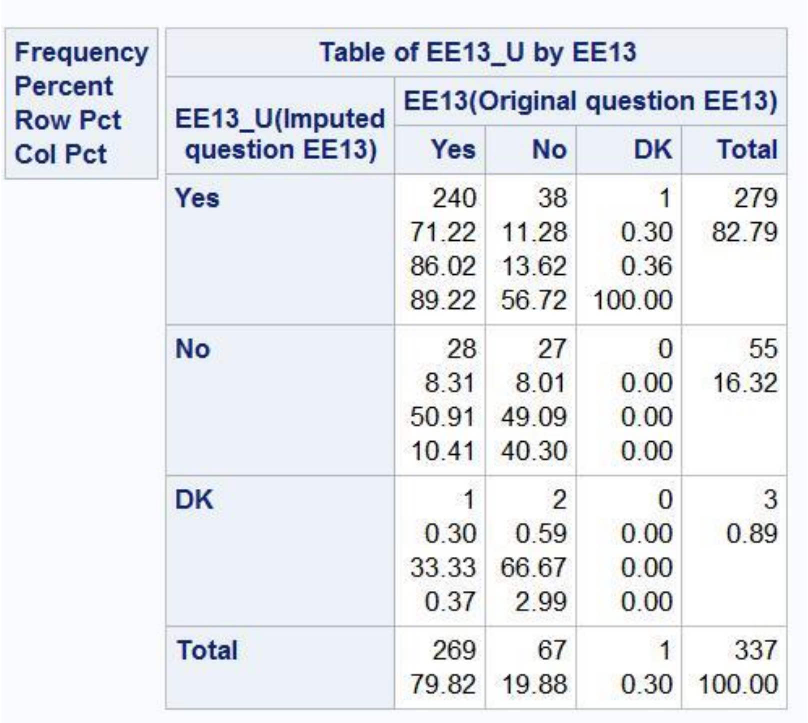
Accuracy of imputation methods

- Out of 840 cases 636 replies were received, i.e. 76%

- We can check the accuracy of imputations on those cases:
  - 337 cases were treated with cold deck,
  - 299 with hot deck

- We expected that cold deck imputation is more accurate than hot deck model since the information is from the same respondent and the leading questions were conditioned to be replied exactly the same manner

Accuracy of Cold deck imputation – EE13

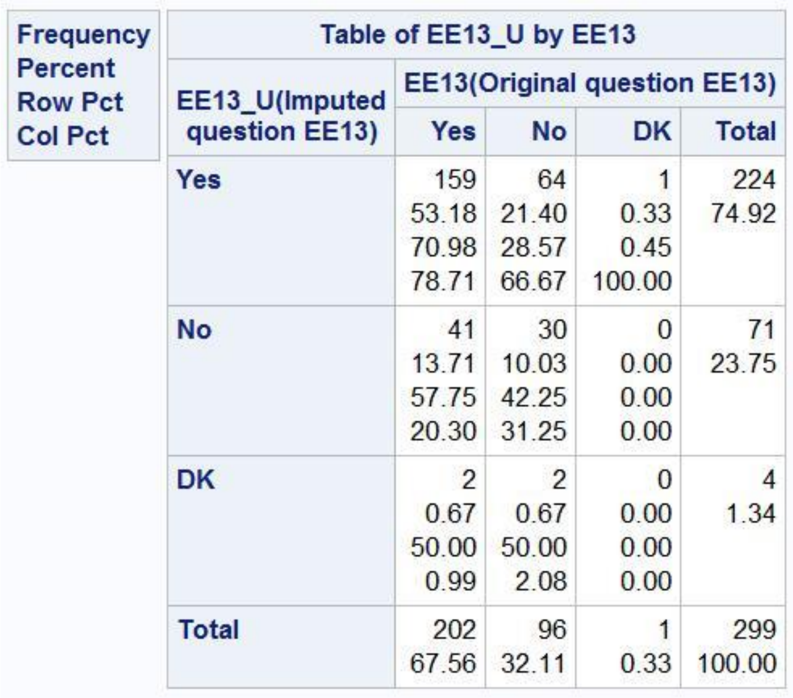


Frequency Percent Row Pct Col Pct	Table of EE13_U by EE13				
	EE13_U(Imputed question EE13)	EE13(Original question EE13)			
		Yes	No	DK	Total
	Yes	240 71.22 86.02 89.22	38 11.28 13.62 56.72	1 0.30 0.36 100.00	279 82.79
	No	28 8.31 50.91 10.41	27 8.01 49.09 40.30	0 0.00 0.00 0.00	55 16.32
	DK	1 0.30 33.33 0.37	2 0.59 66.67 2.99	0 0.00 0.00 0.00	3 0.89
	Total	269 79.82	67 19.88	1 0.30	337 100.00

Exactly correct 79%,

Type I error rate 12%
Type II error rate 9%

Accuracy of Hot deck imputation – EE13

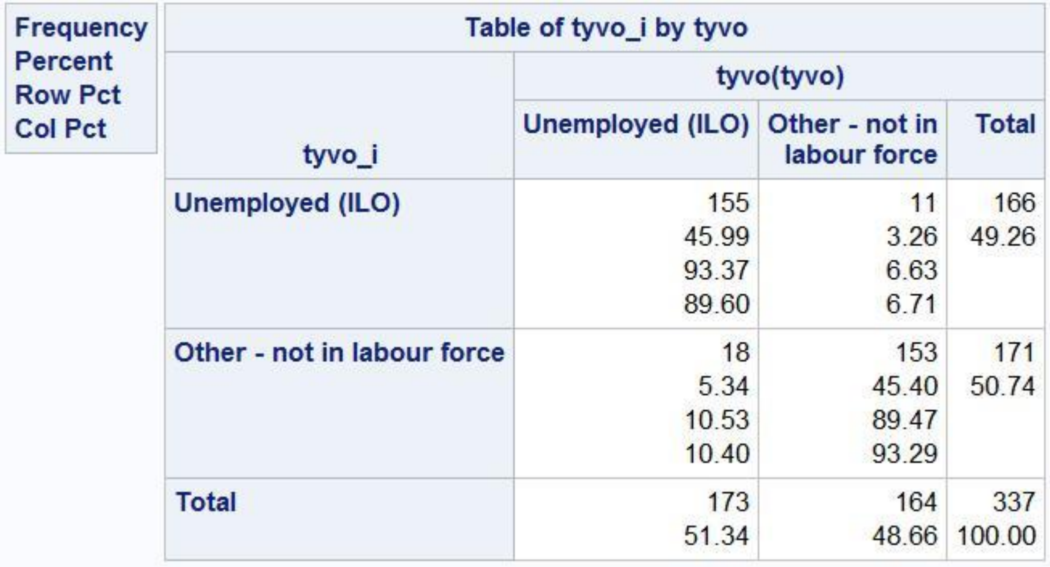


Frequency Percent Row Pct Col Pct	Table of EE13_U by EE13				
	EE13_U(Imputed question EE13)	EE13(Original question EE13)			
		Yes	No	DK	Total
	Yes	159 53.18 70.98 78.71	64 21.40 28.57 66.67	1 0.33 0.45 100.00	224 74.92
	No	41 13.71 57.75 20.30	30 10.03 42.25 31.25	0 0.00 0.00 0.00	71 23.75
	DK	2 0.67 50.00 0.99	2 0.67 50.00 2.08	0 0.00 0.00 0.00	4 1.34
	Total	202 67.56	96 32.11	1 0.33	299 100.00

Exactly correct 63%,

Type I error rate 22%
Type II error rate 15%

Accuracy of Cold deck imputation – labour force status

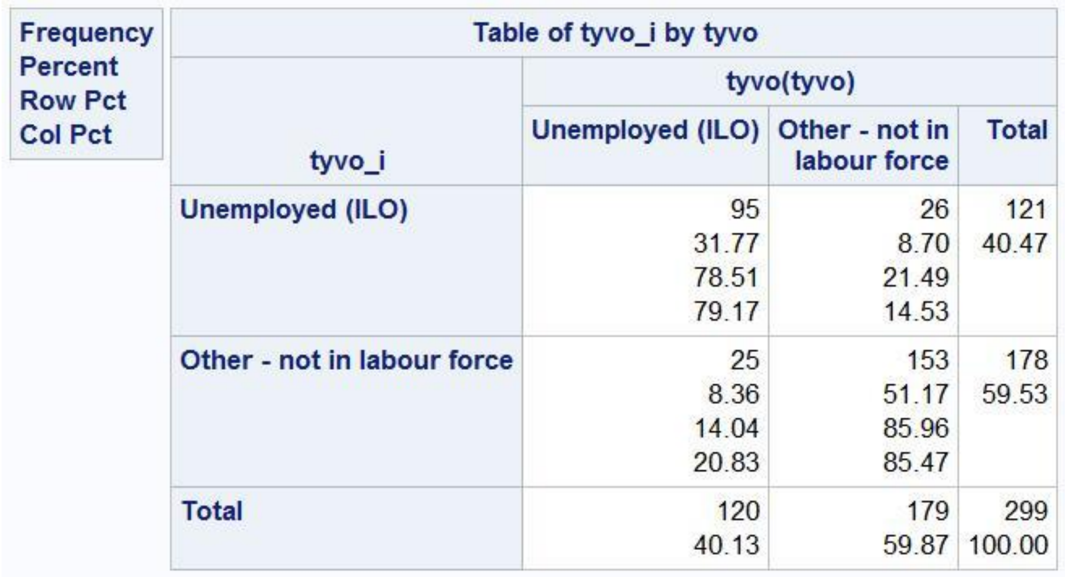


Frequency Percent Row Pct Col Pct	Table of tyvo_i by tyvo		
	tyvo(tyvo)		
tyvo_i	Unemployed (ILO)	Other - not in labour force	Total
Unemployed (ILO)	155 45.99 93.37 89.60	11 3.26 6.63 6.71	166 49.26
Other - not in labour force	18 5.34 10.53 10.40	153 45.40 89.47 93.29	171 50.74
Total	173 51.34	164 48.66	337 100.00

Correct 91%

Net error for unemployed: -7 persons

Accuracy of Hot deck imputation – labour force status

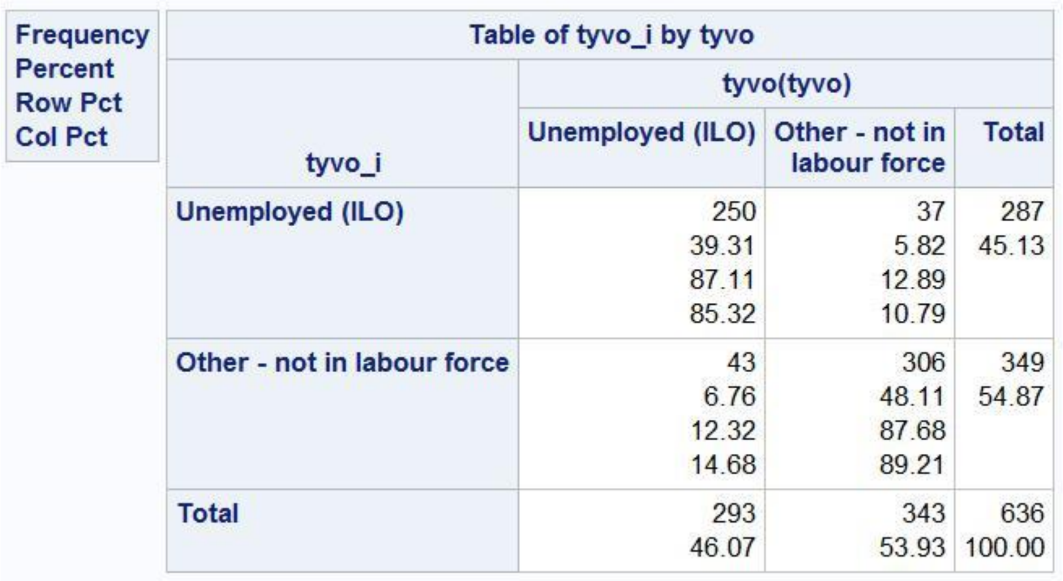


Frequency Percent Row Pct Col Pct	Table of tyvo_i by tyvo		
	tyvo(tyvo)		
tyvo_i	Unemployed (ILO)	Other - not in labour force	Total
Unemployed (ILO)	95 31.77 78.51 79.17	26 8.70 21.49 14.53	121 40.47
Other - not in labour force	25 8.36 14.04 20.83	153 51.17 85.96 85.47	178 59.53
Total	120 40.13	179 59.87	299 100.00

Correct 83%

Net error for unemployed: +1 person

Accuracy of imputation – labour force status



Frequency
Percent
Row Pct
Col Pct

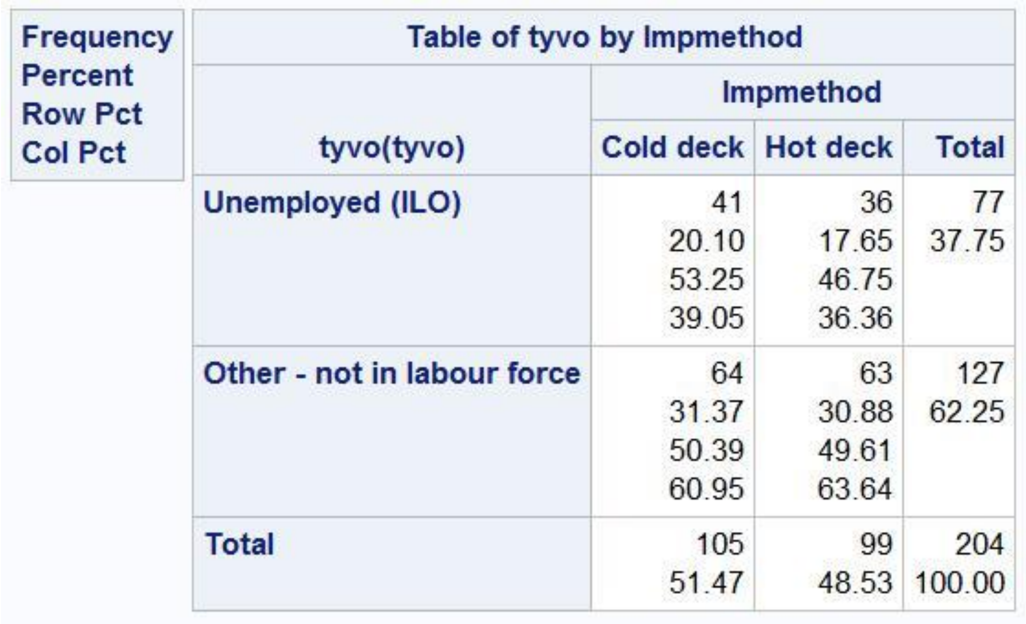
Table of tyvo_i by tyvo

tyvo_i	tyvo(tyvo)		
	Unemployed (ILO)	Other - not in labour force	Total
Unemployed (ILO)	250 39.31 87.11 85.32	37 5.82 12.89 10.79	287 45.13
Other - not in labour force	43 6.76 12.32 14.68	306 48.11 87.68 89.21	349 54.87
Total	293 46.07	343 53.93	636 100.00

Correct 87%

Net error for unemployed: -6 persons

Really imputed cases – comparison of labour force statys by imputation method



Frequency
Percent
Row Pct
Col Pct

Table of tyvo by Impmethod			
	Impmethod		
tyvo(tyvo)	Cold deck	Hot deck	Total
Unemployed (ILO)	41 20.10 53.25 39.05	36 17.65 46.75 36.36	77 37.75
Other - not in labour force	64 31.37 50.39 60.95	63 30.88 49.61 63.64	127 62.25
Total	105 51.47	99 48.53	204 100.00

No difference between methods

Really imputed cases – 2

- It is almost impossible to evaluate the accuracy of those imputed cases. After some ”worst-case scenarios” we assumed the effect to be ± 0.2 per cent in the umployment rate, i.e. about ± 5 380 persons.

- If the net error share of about one per cent (-6/636) from observed basic data analysis holds we can assume that there was an underestimate of 2 persons for unemployed, weighted about 900 persons which would have much smaller effect in unemployment rate than expected : - 0.03 per cent.

- Multiple imputation was applied to the hot deck part and based on that the error rate was evaluated ± 7 persons which is very close to empirical findings.

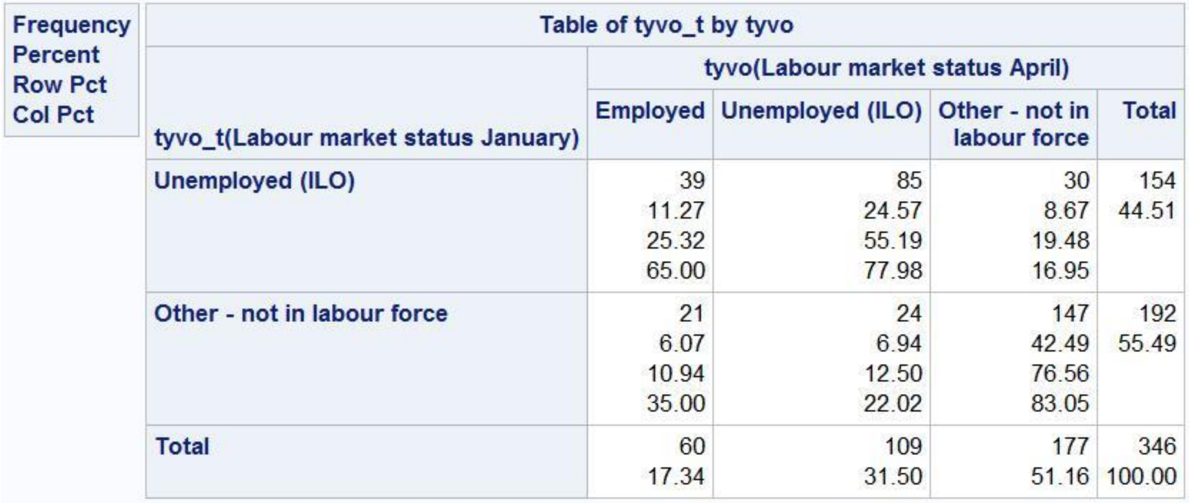
Really imputed cases – 3

- The next wave to about 60% of cases took place in April:
  - Total: 488 originally missing cases in the field, 454 replied (93%)
  - Imputed: 118 cases in the field, 108 replied (92%)

- Changes in labour market status occur:
  - Some people become employed
  - Some retire
  - Some start education etc.
  - Some stay the same

- Those changes correlate strongly with age

Comparison of labour market status: January-April



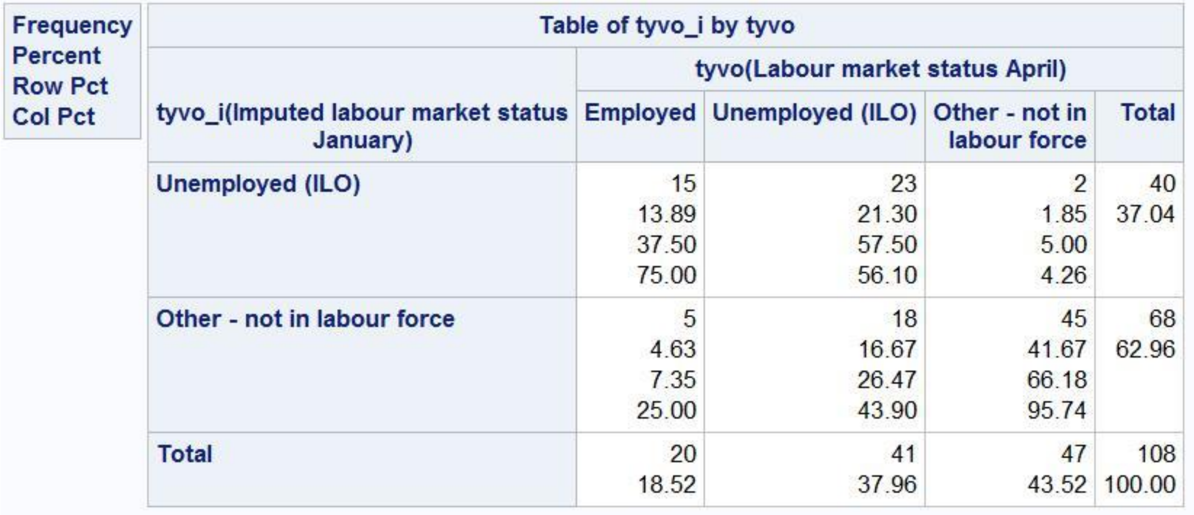
Frequency
Percent
Row Pct
Col Pct

Table of tyvo_t by tyvo

	tyvo(Labour market status April)			
tyvo_t(Labour market status January)	Employed	Unemployed (ILO)	Other - not in labour force	Total
Unemployed (ILO)	39 11.27 25.32 65.00	85 24.57 55.19 77.98	30 8.67 19.48 16.95	154 44.51
Other - not in labour force	21 6.07 10.94 35.00	24 6.94 12.50 22.02	147 42.49 76.56 83.05	192 55.49
Total	60 17.34	109 31.50	177 51.16	346 100.00

Re-interviewed cases

Same status: 67 %



Frequency
Percent
Row Pct
Col Pct

Table of tyvo_i by tyvo

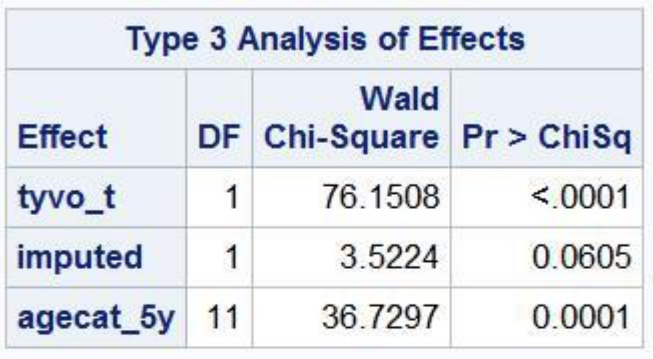
	tyvo(Labour market status April)			
tyvo_i(Imputed labour market status January)	Employed	Unemployed (ILO)	Other - not in labour force	Total
Unemployed (ILO)	15 13.89 37.50 75.00	23 21.30 57.50 56.10	2 1.85 5.00 4.26	40 37.04
Other - not in labour force	5 4.63 7.35 25.00	18 16.67 26.47 43.90	45 41.67 66.18 95.74	68 62.96
Total	20 18.52	41 37.96	47 43.52	108 100.00

Imputed cases

Same status: 63 %

Comparison of labour market status: January-April - 2

- A simple logistic regression analysis of the pooled data did not show significant effect from imputation:



Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
tyvo_t	1	76.1508	<.0001
imputed	1	3.5224	0.0605
agecat_5y	11	36.7297	0.0001

- We could not get significant difference between the two imputation methods, either.

Conclusions

- Imputation was deemed necessary to obtain information for the labour force status
- Cold deck imputation very accurate
- Model-based hot deck imputation almost as good as cold deck with respect to labour force status
- Imputation error was finally evaluated small; underestimate about 1,000 unemployed persons

Happy to hear your questions and comments!

THANK YOU!

Statistics on Finland 150 years:

Trust data. Grab statistics.