Università Ca' Foscari Venezia

Manual versus semi-automatic coding of occupation: Recode SHARE data for NL using CASCOT

Michele Belloni

Agar Brugiavini

Elena Meschi

Occupational Coding in Multi-national Surveys: CASCOT Training Workshop

University of Warwick - Venice
10-11 April 2014

Università Ca' Foscari Venezia

Recoding exercise: CASCOT performance on SHARE data

- The exercise:
  - recode open-ended questions in SHARE wave 1 using CASCOT-NL
  - we do not aim at assessing which method is "better", but rather…
- Our aims:
  - Highlight the complexity of occupational coding: this issue is often neglected in economics and the variable 'occupation' is often taken as free of error (see 'Further Steps')
  - Describe what are the occupations that are more frequently coded differently (differently coded=miscoded in the rest of the talk)
  - Understand which characteristics (such as gender, education, or industry) are associated to the probability of miscoding

Università Ca' Foscari Venezia

SHARE data

- SHARE: Survey of Health, Ageing and Retirement in Europe
  - Cross-national panel database
  - micro data on health, socio-economic status and social and family networks of more than 85.000 individuals aged 50 or over
  - Started in 2004 (now 4 waves available)

Università Ca' Foscari Venezia

SHARE data - occupation

- SHARE w1 (interview year 2004) collected open-ended questions concerning current job & last job (for retired individuals)
- Question:

EP016_ NAME OR TITLE OF JOB
What is your [main/secondary] job called? Please give the exact name or title.

(interviewers ask to be specific, examples are given – similar question for last job)

Università Ca' Foscari Venezia

SHARE data - occupation

- These strings were manually coded ex-post into ISCO 88.
  - separate management by each country team in original language
- SHARE coders made use of ancillary information on sector (NACE) and the unpublished variable "What training or qualifications are needed for this job?". Coders were suggested to code vague responses by means of trailing zeros.
- We focus on the Dutch SHARE sample. Dutch coders coded SHARE data into ISCO-88 3-digit.
  - Choice of NL due to availability of CASCOT in Dutch

Università Ca' Foscari Venezia

Recoding with CASCOT

- To conduct our recoding exercise, Dutch persons at CenterData run the software on Share original text strings
  - In the CASCOT mode "process one record at time": manual changes were few for high score levels

- CASCOT-Netherlands codes into ISCO-08 4-digit.
- We converted output of CASCOT from ISCO-08 into ISCO-88 using official conversion rules. We then consider only 3-digits
- **Comparison between SHARE and CASCOT is performed in terms of ISCO-88 3-digit metric.**

Università Ca' Foscari Venezia

Conversion ISCO-08 to ISCO-88

- No one-to-one correspondence between ISCO-08 to ISCO-88
- In some cases multiple ISCO-88 codes associated to the same 4 digit ISCO-08
- In these cases we attach to one person multiple ISCO-88 codes
  - About 220 individuals (about 1/5 of the sample) have multiple codes
- We define an occupation as **miscoded** if the ISCO code in Share **is not equal to any of** the ISCO-88 resulting from the conversion of CASCOT output

Università Ca' Foscari Venezia

Vague responses

- In CASCOT-NL there is a separate category for vague responses: “99..” (“afleidingscodes”: teacher, entrepreneur…).
- Vague and inadequate responses were excluded from the comparison

	Last job		**Current job**	
	Freq.	*Percent*	*Freq.*	*Percent*
Comparable	1,083	62.1	607	60.82
Not comparable	661	37.9	391	39.18
Total	1,744	100	998	100

Università Ca' Foscari Venezia

Results - Incidence of miscoding

Last job

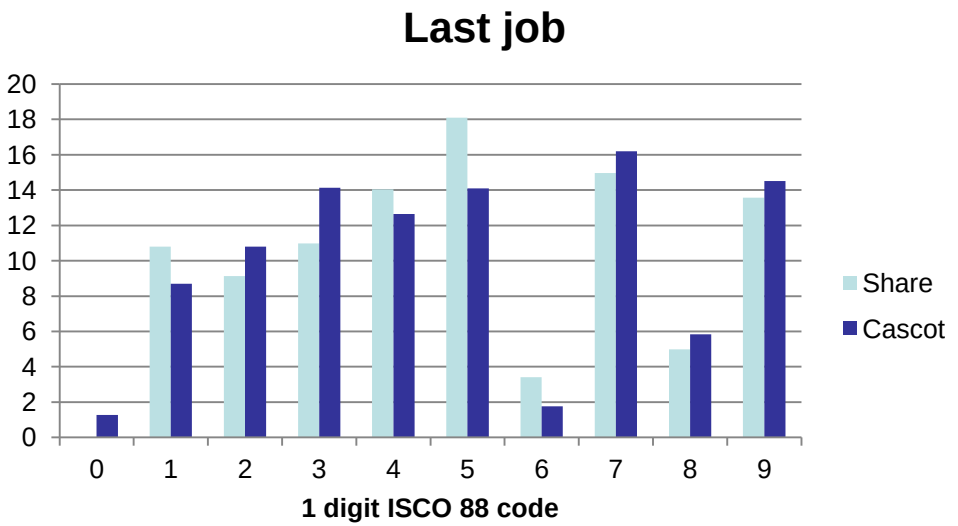
	1 digit		2 digit		3 digit	
	Freq.	*%*	*Freq.*	*%*	*Freq.*	*%*
same coding	750	69.25	683	63.07	572	52.82
different coding	333	30.75	400	36.93	511	47.18
Total	1,083	100	1,083	100	1,083	100

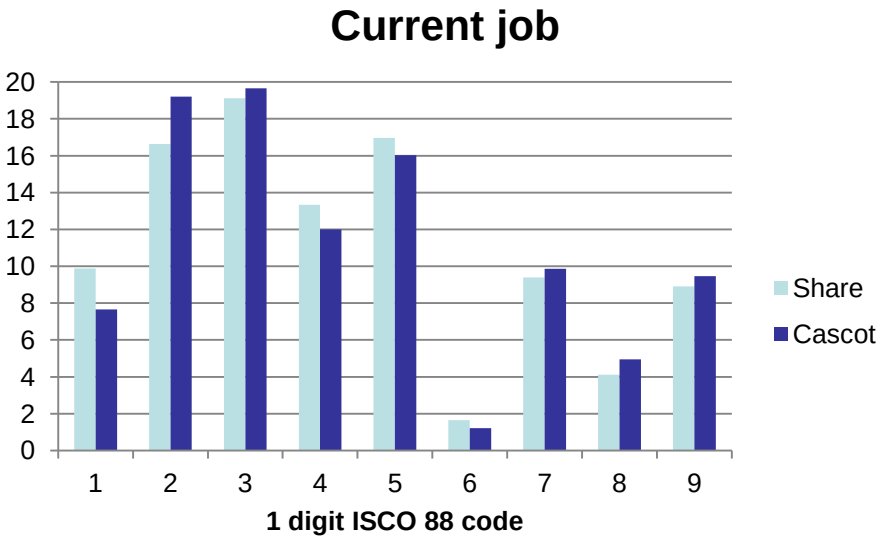
Current job

	1 digit		2 digit		3 digit	
	Freq.	*%*	*Freq.*	*%*	*Freq.*	*%*
same coding	401	66.06	347	57.17	272	44.81
different coding	206	33.94	260	42.83	335	55.19
Total	607	100	607	100	607	100

Università Ca' Foscari Venezia

Distribution of occupation 1 digit – Cascot and Share coding





Università Ca' Foscari Venezia

How *distant* are the two distributions?

Transition matrix – LAST JOB

Cascot → / Share ↓	1	2	3	4	5	6	7	8	9	Total
1	**41.56**	13.31	22.72	7	3.5	0	9.81	0.7	1.4	100
2	1.17	**63.05**	27.02	3.5	3.5	0	0.58	1.17	0	100
3	6.98	29.62	**44.5**	5.11	7.15	0.51	3.06	0	3.06	100
4	1.05	4.18	18.82	**69.69**	1.39	0	1.39	0	3.48	100
5	11.29	1.65	4.96	0.55	**70.52**	0	3.31	3.31	4.41	100
6	55.56	0	0	0	0	**8.89**	4.44	0	31.11	100
7	0	0	1.29	0	0.65	0	**90.32**	3.87	3.87	100
8	0	2.11	4.93	4.23	0	0	16.9	**63.38**	8.45	100
9	0.35	1.38	1.38	6.92	2.42	0.35	4.15	1.38	**81.66**	100
Total	7.64	11.01	13.2	13.27	15.49	0.32	18.28	4.79	16.01	100

Università Ca' Foscari Venezia

Miscoding by occupation ISCO 1-dgt

	Last job				Current job			
		% of miscoded				% of miscoded		
ISCO 1-dgt	*N*	*3 digit*	*2 digit*	*1 digit*	*N*	*3 digit*	*2 digit*	*1 digit*
1	**117**	**82**	**65**	**59**	**60**	**80**	**53**	**47**
2	99	44	37	34	101	38	31	28
3	**119**	**64**	**52**	**50**	**116**	**70**	**59**	**53**
4	152	52	33	31	81	48	36	32
5	196	40	39	31	103	38	36	26
6	37	24	24	22	**10**	**80**	**70**	**60**
7	162	30	20	09	57	61	35	16
8	54	44	39	28	25	32	24	16
9	147	39	25	17	54	72	56	31

Università Ca' Foscari Venezia

Miscoding by education

	Last job				Current job			
		% of miscoded				*% of miscoded*		
	N	*3 digit*	*2 digit*	*1 digit*	*N*	*3 digit*	*2 digit*	*1 digit*
ISCED 0-1	237	35	27	20	42	60	52	38
ISCED 2	465	44	34	27	208	55	42	29
ISCED 3-4	227	53	42	37	155	54	43	34
ISCED 5-6	137	67	54	49	197	55	41	37
Total	1066	47	37	31	602	55	43	34

There is a clear positive education-miscoding gradient for last job. However, this gradient is not present for current job.

Università Ca' Foscari Venezia

Miscoding by gender

	Last job				Current job			
		% of miscoded				% of miscoded		
	N	3 digit	2 digit	1 digit	N	3 digit	2 digit	1 digit
Male	536	59	46	38	332	61	45	35
Females	547	36	28	24	275	48	40	32
Total	1083	47	37	31	607	55	43	34

Males more likely to be miscoded: is it because they are more concentrated in ISCO categories that are more likely to be miscoded?

Università Ca' Foscari Venezia

Gender composition and educational attainment across ISCO1 categories

ISCO 1-dgt	% primary	% lower secondary	% upper secondary	% tertiary	Mean years of education	% of female
1	5.6	30.4	29.9	34.1	14.0	20.3
2	0.8	14.2	21.2	63.7	16.1	54.6
3	3.2	22.8	35.1	38.9	14.0	41.5
4	7.8	50.4	32.6	9.2	12.6	72.4
5	18.9	54.7	21.6	4.8	11.6	81.9
6	20.0	61.4	12.9	5.7	11.2	42.3
7	31.5	48.2	17.5	2.8	9.8	20.6
8	29.8	49.7	17.1	3.3	10.9	20.0
9	35.3	50.5	10.7	3.6	9.9	70.6
Total	*15.1*	*40.2*	*23.7*	*21.0*	*12.5*	*51.2*

Note: Share coding; pool current and last job

Università Ca' Foscari Venezia

Multivariate analysis

Last job

	Miscoding at 3 digit		Miscoding at 2 digit		Miscoding at 1 digit	
Females	-0.168***	-0.108***	-0.155***	-0.107***	-0.108***	-0.056
	(0.030)	(0.038)	(0.029)	(0.037)	(0.028)	(0.036)
ISCED 2	0.103***	0.068	0.100***	0.077*	0.090**	0.073*
	(0.038)	(0.043)	(0.038)	(0.042)	(0.037)	(0.041)
ISCED 3-4	0.195***	0.135***	0.166***	0.145***	0.192***	0.161***
	(0.045)	(0.050)	(0.044)	(0.049)	(0.043)	(0.048)
ISCED 5-6	0.342***	0.293***	0.350***	0.348***	0.339***	0.335***
	(0.051)	(0.060)	(0.051)	(0.059)	(0.049)	(0.057)
Industry dummy	*no*	*yes*	*No*	*yes*	*no*	*yes*
P-value joint signif		*0.0108***		*0.0001****		*0.0003****
Observations	1,081	933	1,081	933	1,081	933
R-squared	0.081	0.114	0.077	0.120	0.068	0.105

Università Ca' Foscari Venezia

Multivariate analysis

Current job

	Miscoding at 3 digit		Miscoding at 2 digit		Miscoding at 1 digit	
Females	-0.140***	-0.083*	-0.061	-0.033	-0.033	0.001
	(0.041)	(0.048)	(0.041)	(0.049)	(0.039)	(0.047)
ISCED 2	-0.035	-0.017	-0.098	-0.087	-0.086	-0.067
	(0.084)	(0.087)	(0.084)	(0.089)	(0.080)	(0.085)
ISCED 3-4	-0.055	-0.021	-0.096	-0.083	-0.038	-0.038
	(0.086)	(0.092)	(0.086)	(0.093)	(0.082)	(0.089)
ISCED 5-6	-0.035	0.028	-0.109	-0.049	-0.009	0.027
	(0.084)	(0.093)	(0.084)	(0.095)	(0.080)	(0.091)
Industry dummy	*no*	*yes*	*No*	*yes*	*no*	*yes*
P-value joint signif		*0.0068****		*0.0808**		*0.1715*
Observations	602	531	602	531	602	531
R-squared	0.020	0.113	0.007	0.079	0.006	0.071

Università Ca' Foscari Venezia

Summary of the results

- **Last job**
  - *Ceteris paribus,* males more likely to be miscoded
  - More educated individuals more likely to be miscoded, even after controlling for gender, industry and ISCO dummy (not shown here)
    - When controlling for ISCO dummy, only ISCED 5-6 remains significant
- **Current job**
  - No effect of education
  - Males more likely to be miscoded, only when looking at 3 digit level

Università Ca' Foscari Venezia

Further steps

- Interpretation of the results: just sorting across occupation or other factors (i.e. literacy)?
- Investigate the impact of coding on occupation-health relationship
- Measure the *distance* between distributions using job characteristics*? (i.e risk exposure?)*
- Use other datasets to increase sample size and/or avoid conversion issues…?