SURVEY FUTURES

SURVEY DATA COLLECTION METHODS COLLABORATION

How consistent is industry and occupation coding across data collection modes? Findings from the Census Non-Response Link Study (CNRLS)

Cristian Domarchi[1] • Olga Maslovskaya[1] • Lisa Calderwood[2] • Matt Brown[2]

[1]University of Southampton, UK; [2]Centre for Longitudinal Studies, University College London, UK

Survey Futures Workshop: Industry and Occupation Coding • University College London • 4 June 2026

University of Essex

University of Southampton

UKRI Economic and Social Research Council

CITY UNIVERSITY OF LONDON EST 1894 · Ipsos · LSE The London School of Economics and Political Science · Manchester 1824 The University of Manchester · National Centre for Social Research · NCRM National Centre for Research Methods · Office for National Statistics · UCL · Ulster University · University of Leeds · Unil Université de Lausanne · verian · Warwick The University of Warwick

Disclaimer

This analysis was conducted using data made available on the Integrated Data Platform hosted by the Office for National Statistics (ONS). Conclusions and views expressed in the analysis are the researchers' own, and do not necessarily reflect those of the ONS or any other organisations whose data were used in the analysis.

Introduction

SURVEY FUTURES
SURVEY DATA COLLECTION METHODS COLLABORATION

- **Industry and occupation** are key measures in social surveys:
  - They are an indicator of socioeconomic status
  - They are strongly linked to income, health, and lifestyle
- Collecting industry and occupation data in self-administered surveys can be challenging, as interviewers are not present to ensure respondents provide the necessary information
- There is limited evidence on how industry and occupation coding data differ due to the survey's mode of administration (and particularly, due to the presence of interviewers)

Objectives

SURVEY FUTURES
SURVEY DATA COLLECTION METHODS COLLABORATION

- This research aims to investigate differences in industry and occupation coding between data collected via the primarily online 2021 Census and interviewer-administered surveys
- We use data from the Census Non-Response Link Study (CNRLS), which allows for a direct comparison of industry and occupation variables for the same individuals collected through different modes
- We aim to evaluate:
  - How industry and occupation codes differ across modes
  - The types of mismatches between codes
  - The factors influencing the level of agreement between the two approaches

CNRLS Dataset

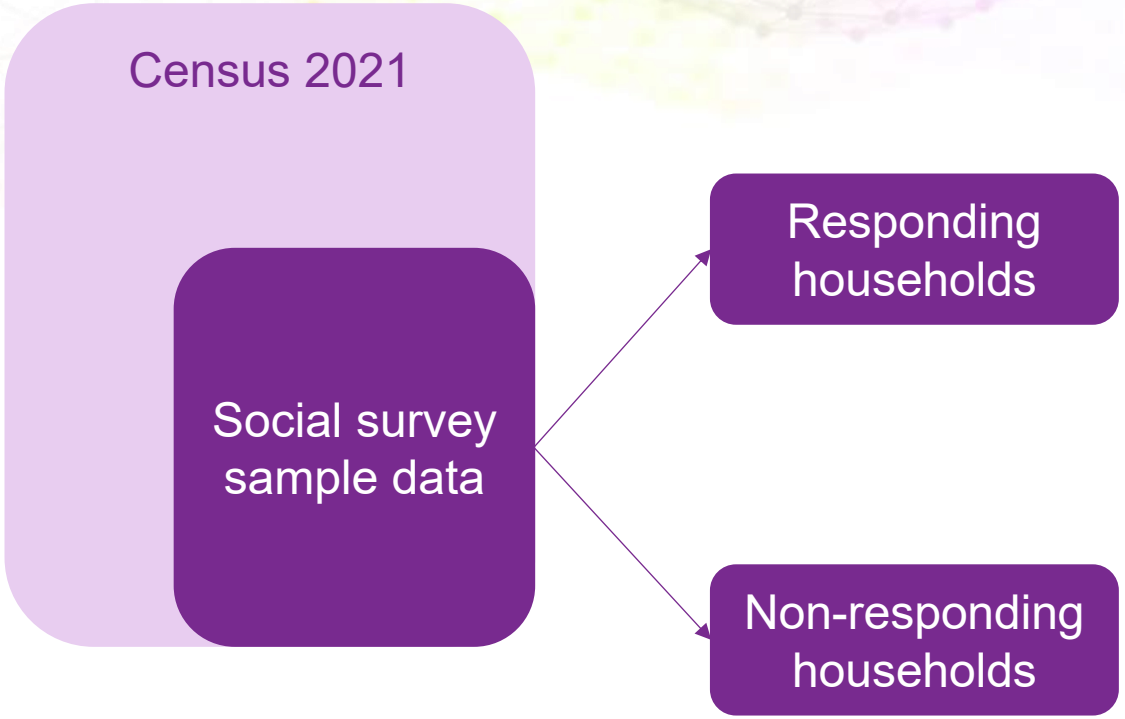
SURVEY FUTURES
SURVEY DATA COLLECTION METHODS COLLABORATION

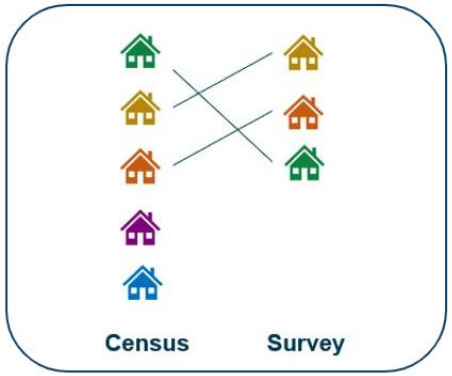
- The **Census Non-Response Link Study (CNRLS)** matches households and the people within them who completed an ONS survey taken around the time of the 2021 Census for England and Wales, with their corresponding census returns (O'Farrell et al., 2021)
- The CNRLS used data collected between January and June 2021 from the following surveys:
  - Survey on Living Conditions (SLC)
  - Living Costs and Food Survey (LCF)
  - Labour Market Survey (LMS)
  - **Labour Force Survey (LFS)**

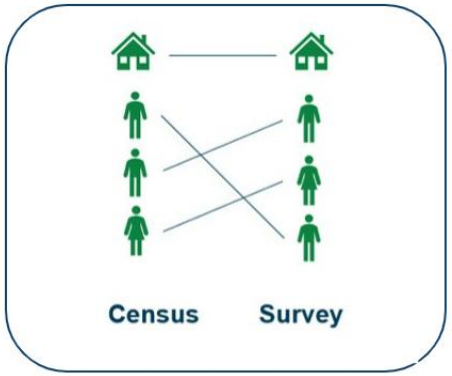
CNRLS Dataset

SURVEY FUTURES
SURVEY DATA COLLECTION METHODS COLLABORATION

- Schematic view of CNRLS linkage:







University of Essex | University of Southampton | UKRI Economic and Social Research Council

Source: Adapted from O'Farrell et al. (2021)

CNRLS – Census

SURVEY FUTURES
SURVEY DATA COLLECTION METHODS COLLABORATION

- The census aimed to count **all usual residents** and households in England at Wales as at Census Day (**21 March 2021**), using an address register and targeted follow-up of non-responding households and harder-to-reach population sub-groups [Source]
- The first **predominately online** census in England and Wales:
  - Around 89% of household responses were submitted online
  - The remaining 11% of responses were completed on paper (either sent upfront or requested by households), alongside additional assisted completion routes (e.g. telephone support) [Source]
- Household response rates were about 97%

CNRLS – Census

33 In the last seven days, were you doing any of the following?

Tick all that apply

Include casual or temporary work, even if only for one hour

- ☐ Working as an employee → GO TO 39
- ☐ Self-employed or freelance → GO TO 39
- ☐ Temporarily away from work ill, on holiday or temporarily laid off → GO TO 39
- ☐ On maternity or paternity leave → GO TO 39
- ☐ Doing any other kind of paid work → GO TO 39
- ☐ **OR** none of the above

39 Answer the remaining questions for your main job or, if not working, your last main job.

Your main job is the job in which you usually work (worked) the most hours

34 Which of the following describes what you were doing in the last seven days?

Tick all that apply

- ☐ Retired (whether receiving a pension or not)
- ☐ Studying
- ☐ Looking after home or family
- ☐ Long-term sick or disabled
- ☐ Other

35 In the last four weeks, were you actively looking for any kind of paid work?

- ☐ Yes
- ☐ No

36 If a job became available now, could you start it within two weeks?

- ☐ Yes
- ☐ No

37 In the last seven days, were you waiting to start a job already accepted?

- ☐ Yes
- ☐ No

38 Have you ever done any paid work?

- ☐ Yes, in the last 12 months
- ☐ Yes, but not in the last 12 months
- ☐ No, have never worked → GO TO 51

SURVEY FUTURES
SURVEY DATA COLLECTION METHODS COLLABORATION

University of Essex · University of Southampton · UKRI Economic and Social Research Council

CNRLS – Census

33 In the last seven days, were you doing any of the following?

Tick all that apply

Include casual or temporary work, even if only for one hour

- ☐ Working as an employee ➡ GO TO 39
- ☐ Self-employed or freelance ➡ GO TO 39
- ☐ Temporarily away from work ill, on holiday or temporarily laid off ➡ GO TO 39
- ☐ On maternity or paternity leave ➡ GO TO 39
- ☐ Doing any other kind of paid work ➡ GO TO 39
- ☐ **OR** none of the above

39 Answer the remaining questions for your main job or, if not working, your last main job.

Your main job is the job in which you usually work (worked) the most hours

Occupation questions

Industry question

Individual questions – Person 1 continued

40 In your main job, what is (was) your employment status?

- ☐ Employee
- ☐ Self-employed or freelance without employees
- ☐ Self-employed with employees

41 What is (was) the name of the organisation or business you work (worked) for?

If you are (were) self-employed in your own business, write in your business name

☐ **OR** no organisation or work (worked) for a private individual

42 What is (was) your full job title?

For example, RETAIL ASSISTANT, OFFICE CLEANER, DISTRICT NURSE, PRIMARY SCHOOL TEACHER

Do not state your grade or pay band

43 Briefly describe what you do (did) in your main job.

44 What is (was) the main activity of your organisation, business or freelance work?

For example, CLOTHING RETAIL, GENERAL HOSPITAL, PRIMARY EDUCATION, FOOD WHOLESALE

If you are (were) a civil servant, write CIVIL SERVICE

If you are (were) a local government officer, write LOCAL GOVERNMENT and give the name of your department within the local authority

45 Do (did) you supervise or oversee the work of other employees on a day-to-day basis?

☐ Yes ☐ No

46 If you had a job last week ➡ GO TO 47
If you were temporarily away from work last week ➡ GO TO 47
If you did not have a job last week ➡ GO TO 51

47 In your main job, how many hours a week do you usually work?

Include paid and unpaid overtime

0 to 15 ☐ 16 to 30 ☐ 31 to 48 ☐ 49 or more ☐

48 How do you usually travel to work?

Tick one box only for the longest part, **by distance**, of your usual journey to work

- ☐ Work mainly at or from home
- ☐ Underground, metro, light rail, tram
- ☐ Train
- ☐ Bus, minibus or coach
- ☐ Taxi
- ☐ Motorcycle, scooter or moped
- ☐ Driving a car or van
- ☐ Passenger in a car or van
- ☐ Bicycle
- ☐ On foot
- ☐ Other

49 Where do you mainly work?

- ☐ At a workplace or report to a depot
- ☐ At or from home ➡ GO TO 51
- ☐ An offshore installation ➡ GO TO 51
- ☐ No fixed place ➡ GO TO 51

50 What is the address of your workplace or depot?

Postcode

51 There are no more questions for **Person 1**.

➡ GO TO questions for **Person 2**
OR if there are no more people in this household
➡ GO TO the **Visitor questions** on the back page
OR if there are no visitors staying here overnight
➡ GO TO the **Declaration** on the front page

University of Essex · University of Southampton · UKRI Economic and Social Research Council

CNRLS – Census

SURVEY FUTURES
SURVEY DATA COLLECTION METHODS COLLABORATION

- ONS developed an in-house coding tool for write-in text responses (including industry and occupation questions) in Census 2021
- The coding tool included the following stages (Office for National Statistics, 2023):
  1. **Parsing** or text cleaning
  2. **Exact matching** between the parsed data and the parsed index (SIC/SOC). When an exact match is found, the tool applies the relevant code.
  3. **Fuzzy matching** happens if an exact match is not found in the previous stage. It consists of two stages: word matching and phrase matching

CNRLS – Census • Data processing

SURVEY FUTURES
SURVEY DATA COLLECTION METHODS COLLABORATION

Initial dataset			1,159,830
Missing data from Census records	-72,130	=	1,087,700
Aged less than 16	-207,080	=	880,620
People with missing info on “Activity last week”	-10,870	=	869,750
People aged 16 or over who have never worked	-88,500	=	**781,250**

University of Essex
University of Southampton
UKRI Economic and Social Research Council

CNRLS data – Census

SURVEY FUTURES
SURVEY DATA COLLECTION METHODS COLLABORATION

- **All these records were successfully coded:**
  - For **industry**, the SIC 2007 framework **(**Standard Industrial Classification of Economic Activities) was used to encode open-text descriptions of industry at the 5-digit level: Division [2], Group [1], Class [1], Sub-class [1]
  - For **occupation**, the SOC 2020 framework (Standard Occupational Classification) was used to encode open-text descriptions of occupation at the 4-digit level: Major group; Sub-major group; Minor group; Unit group

CNRLS data – Labour Force Survey

SURVEY FUTURES
SURVEY DATA COLLECTION METHODS COLLABORATION

- The largest household study in the UK
- It provides the official measures of employment and unemployment
- In 2021, it was a telephone survey
- Basic indicators for 2021:
  - Issued sample: 378,832
  - Full responses: 98,237
  - Partial responses: 8,197
  - Response rate: 28.1%

CNRLS data – Labour Force Survey

- **Industry and occupation questions**:
  - **Job title**: *What was your (main) job in the week ending Sunday the (date)?*
  - **Job description**: *What did you mainly do in your job?*
  - **Industry**: *What did the firm/organisation you worked for mainly make or do (at the place where you worked)?*
- These questions apply to respondents currently in work or who have worked in the last eight years
- These open-text responses were coded interactively by the interviewer, as usual in the LFS. Selected occupations (SOC) were amended using the in-house coding tool that was used for Census

CNRLS data – Labour Force Survey • Data processing

SURVEY FUTURES
SURVEY DATA COLLECTION METHODS COLLABORATION

Initial dataset (Filtered Census data)			781,250
Non-respondents from surveys	-480,000	=	301,250
Non-LFS respondents	-218,725	=	82,525
Missing info on employment and economic activity	-22,925	=	59,600
Respondents not in employment	-23,935	=	34,125
Missing key sociodemographic variables	-65	=	**34,060**

- LFC: 4,603
- **LFS: 82,525**
- LMS: 210,118
- SLC: 4,004

University of Essex | University of Southampton | UKRI Economic and Social Research Council

Results

SURVEY FUTURES
SURVEY DATA COLLECTION METHODS COLLABORATION

- **Overall agreement rates:**

Type	Code length	Agreement rates (%)
SIC	1-digit	69.0
SOC	1-digit	70.3
	2-digit	64.3
	3-digit	58.5
	4-digit	50.2

- **SIC/SOC agreement:**

SIC (%)	SOC (1-digit) (%)	
	Agree	Disagree
Agree	**51.4**	18.4
Disagree	18.9	**11.3**

Industry (SIC – 1 digit)

Section	Description	LFS (%)	Census (%)	Agreement rate (%)	N (LFS)
A	Agriculture, forestry and fishing	1.0	1.0	68.6	330
B	Mining and quarrying	0.2	0.3	58.2	55
C	Manufacturing	7.5	9.2	65.5	2,565
D	Electricity, gas, steam & air conditioning	0.7	0.6	49.0	250
E	Water supply, sewerage and waste management	0.8	0.7	62.0	260
F	Construction	7.1	6.8	66.0	2,410
G	Trade (retail sales and wholesales) and vehicle repair	12.8	11.0	64.7	4,370
H	Transportation and storage	4.3	4.6	72.3	1,475
I	Accommodation and food service activities	3.3	3.3	71.2	1,135
J	Information and communication	5.3	4.8	62.9	1,800
K	Finance and insurance	4.5	4.3	73.2	1,540
L	Real estate	1.7	1.4	59.0	575
M	Professional, scientific and technical	7.8	8.9	58.8	2,665
N	Administrative and support services	4.3	4.2	53.9	1,455
O	Public administration and defence	7.6	7.8	65.8	2,600
P	Education	12.1	11.9	84.1	4,140
Q	Human health and social work	14.3	14.2	80.5	4,890
R	Arts, entertainment and recreation	2.4	2.3	56.4	835
S	Other service activities	2.2	2.7	64.1	735
T	Employed by a household as domestic staff	*	0.1	*	<15
U	International organisations and bodies	*	0.0	*	<15

- Distribution of industries differs between LFS and Census ($p < 0.001$)
- Dissimilarity index: **3.8%**

Industry (SIC – 1 digit)

- **Logistic regression model of agreement between Census and LFS SIC 1-digit codes:**
  - Higher likelihood of agreement was associated with:
    - Workers in a supervisor/managerial role
    - Respondents with White ethnicity
    - Fully matched addresses/individuals (as opposed to partially matched)
  - Lower likelihood of agreement was associated with:
    - Industry divisions D, J, L, M, N, P, Q, R (compared to A)
    - Female respondents (compared to male)
    - People aged 26-35 and 66+ (compared to <25)
    - People working 35 hours or more
  - No effects of household size, tenure, UK region, type of survey response (partial/full), or survey month

Occupation (SOC – 1 digit)

SURVEY FUTURES
SURVEY DATA COLLECTION METHODS COLLABORATION

Major group	Description	LFS (%)	Census (%)	Agreement rate (%)	N (LFS)
1	Managers, directors, and senior officials	14.3	12.0	58.7	4,100
2	Professional occupations	25.5	26.4	78.3	9,005
3	Associate professional and technical occupations	14.6	14.9	60.1	5,085
4	Administrative and secretarial occupations	10.9	11.8	71.0	4,015
5	Skilled trade occupations	8.3	9.0	78.0	3,060
6	Caring, leisure and other service occupations	7.8	7.5	74.0	2,545
7	Sales and customer service occupations	6.2	5.9	66.5	2,000
8	Process, plant and machine operatives	5.4	5.3	72.2	1,810
9	Elementary occupations	7.0	7.3	74.0	2,475

- Distribution of occupations differs across coding methods ($p < 0.001$)
- Mean dissimilarity index: **3.1%**

Occupation (SOC – 1 digit)

SURVEY FUTURES
SURVEY DATA COLLECTION METHODS COLLABORATION

- Confusion matrix comparing 1-digit level SOC codes between Census and LFS:

Major group (Census)	Description	Major group (LFS) (%)								
		1	2	3	4	5	6	7	8	9
1	Managers, directors, and senior officials	8.4	2.2	1.4	0.8	0.7	0.2	0.3	0.1	0.2
2	Professional occupations	1.1	19.9	1.9	1.1	0.5	0.6	0.1	0.1	0.1
3	Associate professional and technical occupations	0.9	2.3	8.8	1.0	0.3	0.5	0.5	0.1	0.3
4	Administrative and secretarial occupations	0.7	0.8	1.0	7.8	0.1	0.2	0.3	0.0	0.2
5	Skilled trade occupations	0.3	0.4	0.2	0.1	6.5	0.0	0.1	0.4	0.3
6	Caring, leisure and other service occupations	0.1	0.4	0.9	0.2	0.0	5.8	0.1	0.0	0.2
7	Sales and customer service occupations	0.3	0.2	0.4	0.6	0.1	0.1	4.2	0.1	0.4
8	Process, plant and machine operatives	0.1	0.1	0.2	0.1	0.4	-	0.1	3.9	0.4
9	Elementary occupations	0.1	0.1	0.2	0.2	0.3	0.2	0.3	0.4	5.2

- Indication of inter-group errors

Occupation (SOC – 1 digit)

- Coding accuracy indicators:

4-digit SOC ends in 9 in Census	4-digit SOC code ends in 9 in survey (%)	
	YES	NO
YES	**6.6**	5.2
NO	7.2	**80.9**

- Ending in 9: “not elsewhere classified”
- Highest proportions in major subgroups 4 (Administrative and secretarial occupations) and 8 (Process, plant and machine operatives).
- Lowest proportions in 6 (Caring, leisure and other service occupations)

Occupation (SOC – 1 digit)

SURVEY FUTURES
SURVEY DATA COLLECTION METHODS COLLABORATION

- **Logistic regression model of agreement rates between Census and LFS SOC 1-digit codes**
  - Higher likelihood of agreement was associated with:
    - Major sub-groups 2, 4, 5, 6, 7, 8, 9 (compared to 1)
    - Male respondents (compared to female)
    - People aged 46-55 and 56-65 (compared to <25)
  - Lower likelihood of agreement was associated with:
    - People working 35 hours or more
    - Workers in a supervisor/manager role
  - No effects of ethnicity, household size, tenure, UK region, type of survey response (partial/full), or survey month

Conclusions

- Agreement rates between self-administered (Census) and interviewer-administered (LFS) industry and occupation codes are around 70% at the 1-digit level.
- For 4-digit occupations, agreement rates are around 50%
- Mismatches in SIC are associated with:
  - Administrative and professional industries, but also with electricity, real estate, mining and the arts
  - Female and younger respondents and part-time workers
- Mismatches in SOC are mainly associated with:
  - Supervisory and managerial roles
  - Female and younger respondents and full-time workers

Limitations/Work in progress

SURVEY FUTURES
SURVEY DATA COLLECTION METHODS COLLABORATION

Limitations:

- We lack a “true value” to which we could compare responses
- We cannot be 100% sure that participants are referring to the same job in their Census and survey responses, although we used all the available variables to mitigate this risk
- We are unable to isolate the effect of data collection mode from coding approach

Work in progress:

- Additional variables that could help us understand coding mismatches: mode of response, character length for open-text responses
- In-depth look at coding mismatches, e.g. confusion matrices
- Crossed effects, e.g. are there any industries in which occupation mismatches are more likely?

References

- Office for National Statistics (2023) ONS website, methodology, Automated text coding: Census 2021
- O'Farrell, K., Higham-Lloyd, S., Kilner, H., Nelson, D. & Pike, O. (2022) *CNRLS 2021 Quality Report*. Office for National Statistics.

SURVEY FUTURES

SURVEY DATA COLLECTION METHODS COLLABORATION

How consistent is occupational coding across data collection modes? Findings from the Census Non-Response Link Study (CNRLS)

Cristian Domarchi[1] • Olga Maslovskaya[1] • Lisa Calderwood[2] • Matt Brown[2]

[1]University of Southampton, UK; [2]Centre for Longitudinal Studies, University College London, UK

Survey Futures Workshop: Industry and Occupation Coding • University College London • 4 June 2026

University of Essex

University of Southampton

UKRI Economic and Social Research Council

CITY UNIVERSITY OF LONDON EST 1894

Ipsos

LSE THE LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE

MANCHESTER 1824 The University of Manchester

National Centre for Social Research

NCRM NATIONAL CENTRE FOR RESEARCH METHODS

Office for National Statistics

UCL

Ulster University

UNIVERSITY OF LEEDS

UNIL | Université de Lausanne

verian

WARWICK THE UNIVERSITY OF WARWICK