Using administrative data vs survey data

Danilo Dolenc

GCC-STAT Regional Workshop

Muscat, Oman, 22-24 September 2019

Overview

- Three main points to discuss
  - Integration of administrative data with survey data
  - Harmonization of outputs using administrative data over different surveys
  - Validation of administrative based outputs with survey outputs

Important

- Administrative data could be used directly
    - In most cases statistical data have been derived and reuse in statistical sample surveys

Advantages of use of administrative data in surveys

- Not asking questions already available from data sources
  - Less time needed to collect data
    - Shorter questionnaires
- Reduction of field operation costs
- Decrease response burden
  - Increasing non-response huge problem in household surveys
- Improvement of the quality of outputs
- Harmonization of outputs

Disadvantages of use of administrative data in surveys

- Not synchronized field data collection with availability of administrative data
  - More time needed for final outputs
- More demanding data processing due to expect inconsistency between survey and administrative data
- Same systematic errors could appear in all surveys

Case: Survey on Income and Living Conditions (SILC)

- The most advanced EU register-based countries have problem with timely delivery of data to Eurostat
  - Use of taxation data for income of household
- Next problem – to which year sample survey data refers
  - Year of field collection (T)
    - In Slovenia first half of year
  - Year of income taxation data(T-1)

Income data – some observations

- Data on income are under-estimated by respondents
  - Memory effect (previous year)
  - Not include all income
    - Almost no income from interests from field survey
  - Psychological profile of respondents
    - Tension to cover income but not expenditure

Income data – editing strategy

- Priority rules in case of inconsistency between income and labour force status from survey
  - Priority to administrative data on income
    - New labour force status derived according to the type of income
      - Example: persons retired at the end of the year

Case: Business surveys

- Joining previous numerous surveys into one survey
- Cancelation of surveys
- Exclusion of less important business subjects from sample surveys
- Imputations based on aggregated administrative data

Core social variables project (1)

- Standardization of variables for all European statistical social sample surveys (SILC, LFS, HBS, AES, EHIS, TUS and ICT HH)
  - Harmonized definitions
  - Categories for the variable determined
  - Reference questions suggested
    - In case of field data collection
  - 28 common variables foreseen by now
- Data from administrative data are allowed
  - 17 variables available

Core social variables project (2)

- Data originated from four sources with full coverage but different periodicity and timeliness
  - Monthly stock data on employment (T + 2)
    - Industry, occupation, status in employment, full/part time job, permanency of job
  - Quarterly data on population (T + 4)
    - Sex, age, region of residence, citizenship

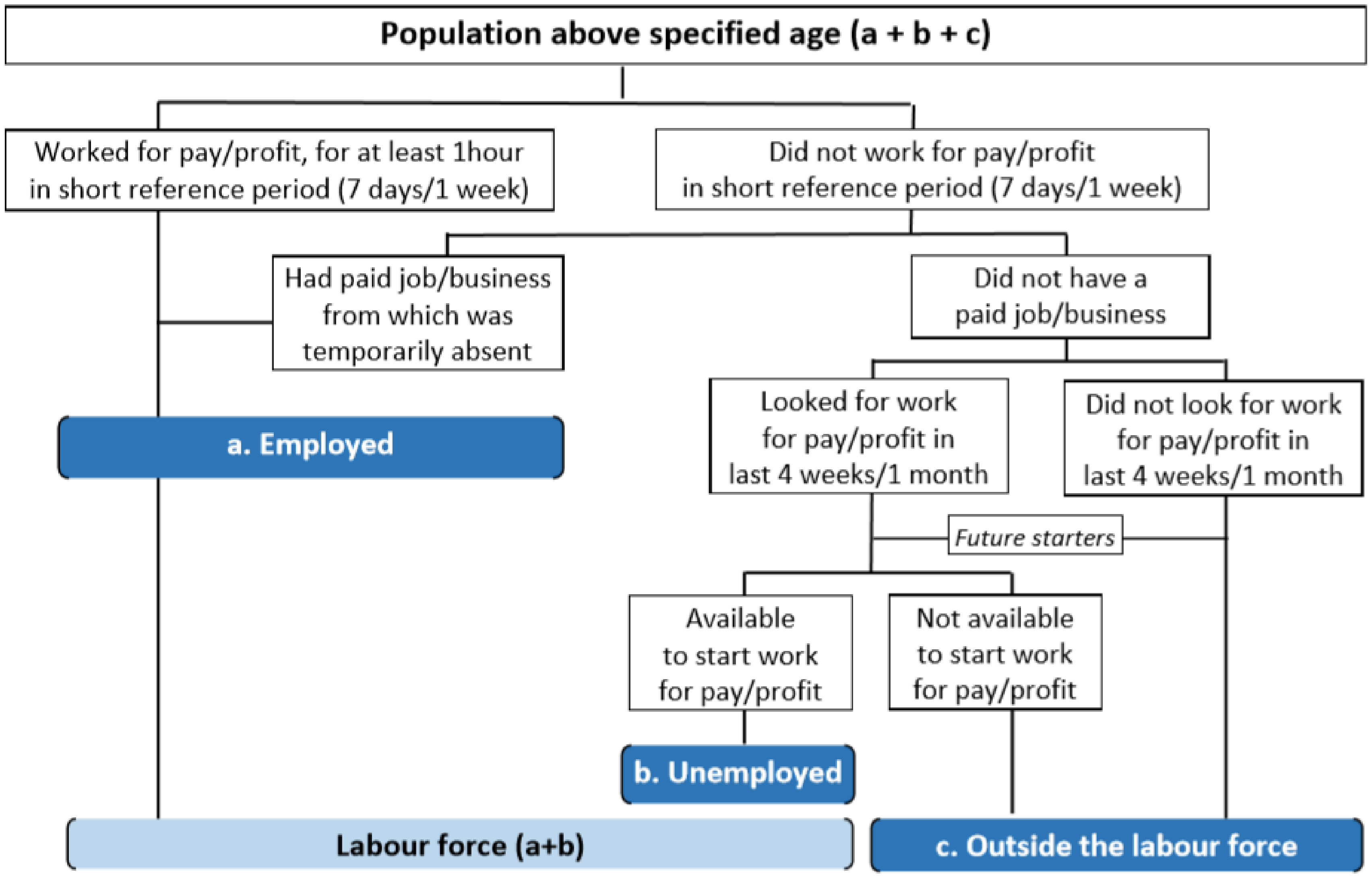
Core social variables project (3)

- Data originated from four sources with full coverage but different periodicity and timeliness
  - Annual data on population as of 1 January (T + 10)
    - Labour force status, educational attainment level, country of birth, country of birth of mother, country of birth of father, year of last immigration to the country
  - Annual data on formal educational enrolment from primary to tertiary level in current school year as of 1 October (T + 6)
    - Enrolment in formal education, level of current enrolment

Case: Labour force status from survey and administrative sources (1)

- 2 concepts from Labour Force Survey available
  - ILO definition based on one hour criterion of work / last week
    - Three questions needed
      - Working or not
      - Looking for work or not
      - Available to start work or not
  - Self-declared labour force status
    - Core social variables concept (Eurostat)

Chart 4 Classification of working age population by labour force status



Population above specified age (a + b + c)

Worked for pay/profit, for at least 1hour in short reference period (7 days/1 week)

Did not work for pay/profit in short reference period (7 days/1 week)

Had paid job/business from which was temporarily absent

Did not have a paid job/business

Looked for work for pay/profit in last 4 weeks/1 month

Did not look for work for pay/profit in last 4 weeks/1 month

Future starters

Available to start work for pay/profit

Not available to start work for pay/profit

a. Employed

b. Unemployed

Labour force (a+b)

c. Outside the labour force

Source: UNECE Recommendations for the 2020 Censuses, paragraph 500

Case: Labour force status from survey and administrative sources (2)

- Register-based (RB) labour force status
  - Very similar to the Eurostat concept of self-declaration
  - Methodology of priority / hierarchy of sources used

Case: Labour force status from survey and administrative sources (3)

- High Quality sources

  1 - Statistical Register of Employment (last week before reference day)

  2 - Registered unemployment (1 January)

  3 - Enrolment in formal education (current school year – as of 1 October)

  4 - Scholarships (1 January)

  5 - Pension recipients(1 January)

Case: Labour force status from survey and administrative sources (4)

- Lower Quality sources

  6 - Health insured persons under specific schemes (1 January)

  7 - Family members of health insured persons (1 January)

Case: Labour force status from survey and administrative sources (5)

- Outdated sources
  - High quality

  8 - Income taxation (previous calendar year)
  - Lower quality

  9 - Recipients of social transfers (previous calendar year)

Linking RB data and LFS data (1)

- Individual records linked using PIN‘s
  - RB data - 1 January 2014 + 1 January 2013
    - Persons that belong to stock at both reference dates
  - LFS - Q4/2013 + Q1/2014
    - Two consecutive databases joined together
    - Duplicate LFS records excluded
      - Due to panel nature of the survey
        - Data for Q1 obtained in case of duplication

Linking RB data and LFS data (2)

- Total number of records from LFS - 31,379
- Preparation of analytics database (exclusion criteria)
  - Younger than 15 as of 1 January 2014 - 2,851
  - Duplicate LFS records - 8,736
  - Unlinked to population database – 177
    - Not usual residents (short-term immigrants)
  - PIN’s with low probability – 410
    - Errors at field data entry

Outcomes - coverage

- Database consists of 19,205 records (1.1% of working age population)
  - Over-estimation of retired persons in survey
    - Lower refuse rate
  - Under-estimation of students in survey
    - Excluded from sample if live in student dormitory

Outcomes - comparing concepts

- RB vs. LFS self-declared status
  - 90% exact match using census classification
    - 95.4% for HQ sources (88% of records)
    - 54.5% for LQ sources (12% of records)
- Register-based vs. LFS ILO status
  - 87% exact match using census classification
- Surprisingly not significant difference between both concepts related to RB data

Final outcomes

- The main contributors to employed are not unemployed persons

Structure of working age population by labour force status

	Employed	Unemployed	Schooling	Retired	Other non-active
RB	45.5	7.2	9.8	30.5	7.1
ILO	50.9	6.2	8.4	28.4	6.1
Diff.	+5.4	-1.0	-1.4	-2.1	-1.0

- Very good quality of sources for producing RB labour force status
- Differences between RB concept and both LFS concepts much lower than expected in advance

Case: Usual residence from administrative sources and surveys (1)

- Residence status of the selected respondent in sample survey
  - Standardized data collection in all social sample surveys to measure
    - Internal redistribution (de facto : de iure)
    - Over-registration
    - Quality of field work of interviewers

Survey residence status - results

Survey	Total	Died	Unknown	Living elsewhere			
				Total	Slovenia	Abroad	No answer
HBS 2012	5.3	0.2	0.5	4.6	3.6	1.0	0.0
HBS 2015	9.4	0.2	1.0	8.2	5.5	1.8	0.9
LFS 2014	8.8	0.1	1.4	7.3	5.8	1.5	0.0
SILC 2014	6.2	0.7	0.5	5.0	2.6	1.6	0.8
SILC 2015	8.4	0.9	0.5	7.0	4.2	1.9	0.9
ICT-HH 2014	7.9	0.3	0.8	6.8	4.9	1.3	0.6

Interviewer non-response

Time delay

Internal redistribution

Over-registration

Case: Usual residence from administrative sources and surveys (2)

- Opposite approach - the residence status of interviewed household members
  - Based on linkage address from survey and address from administrative source to measure
    - Internal redistribution (de facto : de iure)
    - Under-registration

Administrative residence status - results

Type of administrative residence	*SILC 2014*		*SILC 2015*		*HBS 2015*	
	Number	*Share (%)*	*Number*	*Share (%)*	*Number*	*Share (%)*
Household members - total	**28,176**	**100**	**26,571**	**100**	**8,525**	**100**
Registered residence in the household	27,287	96.8	25,773	97.0	8,350	97.9
Residence registered in Slovenia	889	3.2	798	3.0	175	2.1
Belong to statistical population	877		780		171	
Outside statistical population	12		11		4	
Residence not registered in Slovenia	0		0		0	

Internal redistribution

Under-registration

Case: Target survey on over-registration

- Criteria for sample frame
  - Usual resident population (statistical)
    - No data on RB labour force status for 3 consecutive years from any source
    - Foreign citizens without RB labour force status data last year
    - Slovenian citizens with temporary residence only and without LFS data last year
- Presumption – people do not live in Slovenia

Target survey – methods

- Two methods applied using the same very short questionnaire – 2 pages (9 questions)
  - Postal method – letters sent to the official (registered) address
    - Prepaid envelope enclosed
  - Field inquiry (non-response follow-up)
    - Face to face interview using PAPI method
    - Selected regions only

Target survey – results (1)

- Total number of respondents – 11,678
  - Low response rate in postal survey expected in advance
    - 14% of letters returned by Post Office
      - Unknown recipient
    - 16% of letters returned (most filled-in)
    - Non-response – 70%
  - Final real response rate – 25.5%
    - Including not identifiable returns – 42.9%

Target survey – results (2)

- Three categories of responses could be recognized excluding non-identifiable returns
  - Over-registration
    - Persons living abroad (69%)
    - Administrative survivors (4%)
  - Correctness
    - Persons belong to usual population (27%)

Quality evaluation of administrative data (CPR)

- Side effect of the survey
  - 10% respondents deregister from CPR in less than six months after survey
    - 83% of them non-response

Conclusion

- There is still room for improvement CPR data by administrative authorities
  - But quality is better year by year
- Population data based on register are more than satisfactory quality
- Under-coverage is not statistically important phenomena