EU-LFS and EU-SILC:
Official microdata for research purposes

Heike Wirth and Christof Wolf
German Microdata Lab

1st European User Conference for EU-LFS and EU-SILC
Mannheim, March 5-6, 2009
Introduction

• Focus of conference is on microdata provided by Eurostat
  
  – These data are not produced for research purposes
  
  – Primary purpose to provide statistical information for policy making and administration at the European level

=>Consequently the data collection and dissemination are bound by political interests, legal constraints and regulations
Overview

• Why do researchers need European microdata?

• EU-LFS and EU-SILC: Official microdata for research purposes

• Room for Improvements

• Conclusions
Why do researchers need European microdata?

• Within countries
  – considerable research efforts have been made on poverty, labour market, income, education etc.

• But
  – little systematic knowledge on the social, economic and demographic structures and their changes over time from a European comparative perspective

• Still an open question whether the social, economic, demographic processes in Europe do evolve
  – towards an convergence or
  – towards a perpetuation of national differences?
Why do researchers need European microdata?

• Need to understand how the integration of social and economic structures work across Europe

  – descriptive, e.g.
    • How do unemployment rates, population skill levels, poverty rates vary among countries and how do they change over time?

  – analytical, e.g.
    • Why does parenthood in some countries cause an increase, in others countries a decrease in the gender wage gap?

  – methodological, e.g.
    • Regarding poverty and social exclusion: Is there a bias towards income based indicators, do we need more material deprivation indicators?
Why do researchers need European microdata?

• In order to investigate causality social sciences typically rely on variations across groups and through time

  – Assumption: aims, decisions and actions of individual actors are determined basically by the same mechanisms in all countries

  – However the contextual conditions (institutional and structural settings) under which individual decisions are made can vary considerably between countries
Why do researchers need European microdata?

- To understand similarities or differences in behaviour, contextual conditions have to be taken into account, e.g.
  - What are the core characteristics of institutional and structural settings (e.g. education and training system, social security system, labour market conditions)?
  - What is the impact of contextual conditions on cross-national behavioural differences?

=> Need for harmonised microdata at the European level namely for households, individuals, establishments etc.
EU-LFS and EU-SILC: Microdata for research purposes

- Typical problems of secondary data analysis:
  - Data are used for research topics for which they were not designed
  - Consequently data often do not match the requirements of a specific research interest, e.g.
    - important information is missing
    - information is included but not as detailed as needed
  - These problems are even more pronounced when using official microdata, e.g.
    - Definitions and concepts of variables are determined by policy and administrative needs and legal frameworks. These often do not match research needs!

=> Researchers have to settle for second best solutions
EU-LFS and EU-SILC: Microdata for research purposes

• Nevertheless EU-LFS and EU-SILC are important data source for research
  
  – Large sample sizes
    • EU-LFS:
      – 7,600 to 320,000 households; 12,000 to 540,000 persons
    • EU-SILC:
      – 3,000 to 22,000 households; 7,000 to 47,000 persons
  
  – For most countries individuals can be matched to households

  – High response rates
    • in a lot of countries participation in the LFS is compulsory
EU-LFS and EU-SILC: Microdata for research purposes

- Rich time series
  - EU-LFS microdata are available from 1983 onwards

- Longitudinal character of EU-SILC

- Harmonized data across time and countries

- International standard classifications
  - NACE for economic activity
  - ISCO 88(Com) for occupation
  - ISCED for education
EU-LFS and EU-SILC: Microdata for research purposes

- Timely release of scientific use files
  - Latest release 2007: EU-LFS; 2006: EU-SILC
  - Germany
    - latest release scientific use file: 2005: Mikrozensus (German LFS);
      German-SILC not available as scientific use file
Room for Improvements

- Suggestions for improvements might vary depending on research interests and data but there are some issues which might be of a more general interest

  - Data documentation
  - Data anonymisation
  - Costs
Room for Improvements – Data documentation

• Documentation for both datasets is available
  – user guides including descriptions of basic concepts and definitions, variables and the database
  – further information is provided on the internet, e.g.
    • description of national sample designs
    • quality reports
    • EU-LFS: national questionnaires 2005, 2006, 2008 (core questionnaire, ad-hoc modules as far as provided by the NSIs, however not all are available in English)
Room for Improvements – Data documentation

- However when it comes to detail the documentation is not always helpful, especially with regard to country specific data characteristics or inconsistencies over time

- By way of illustration: intergenerational module of EU-SILC 2005
  - Why do Germany and Portugal provide no information regarding ‘financial problems when the respondent was a teenager’?
  - Why is there such a high percentage of missing values for the occupation of fathers for Sweden and the UK?
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>AT</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>BE</td>
<td>7</td>
<td>8</td>
<td>7</td>
<td>8</td>
</tr>
<tr>
<td>CY</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>DE</td>
<td>5</td>
<td>2</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>HU</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>IE</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>IS</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>IT</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>LT</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>LU</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>LV</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>NL</td>
<td>7</td>
<td>0</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>NO</td>
<td>1</td>
<td>1</td>
<td>40</td>
<td>45</td>
</tr>
<tr>
<td>PL</td>
<td>5</td>
<td>1</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>SE</td>
<td>71</td>
<td>93</td>
<td>72</td>
<td>93</td>
</tr>
<tr>
<td>SI</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>14</td>
</tr>
<tr>
<td>UK</td>
<td>51</td>
<td>57</td>
<td>22</td>
<td>25</td>
</tr>
</tbody>
</table>
Room for Improvements – Data documentation

=> To be sure not every detail can be documented for every country, but it might be a first step to

  – name a person responsible who can be contacted in case of country specific questions

  – provide more country specific metadata

  – in particular provide questionnaires for all countries and all years for EU-LFS as well as for EU-SILC
Examples of country specific questions asking for supervisor-responsibilities in EU-LFS

<table>
<thead>
<tr>
<th>Country</th>
<th>Supervisor - Question</th>
</tr>
</thead>
<tbody>
<tr>
<td>Norway</td>
<td>Is it <strong>part of your responsibility</strong> to supervise or manage other employees?</td>
</tr>
<tr>
<td>Irland</td>
<td>Do you supervise the work of other people <strong>on a regular basis</strong>?</td>
</tr>
<tr>
<td>Austria</td>
<td>Do you work in a <strong>leading position</strong>? (also applying to less qualified jobs!)</td>
</tr>
<tr>
<td>Sweden</td>
<td>Do your <strong>tasks include managing and supervising</strong> the work of other employees?</td>
</tr>
<tr>
<td>United Kingdom</td>
<td>In your job, do you have <strong>formal responsibility</strong> for supervising the work of other employees?</td>
</tr>
<tr>
<td>Czech Republic</td>
<td>Does s/he have <strong>subordinate(s)</strong> in the job?</td>
</tr>
</tbody>
</table>

Room for Improvements - Anonymisation

- Considerable differences in the way EU-LFS and EU-SILC are anonymised

- EU-LFS is much more affected by aggregation than EU-SILC is:
  - Age:
    - 5-year age bands (year of birth is suppressed)
  - Marital status:
    - widowed, divorced or separated are combined in a single category
  - Nationality and country of birth are aggregated in 3 groups (same in EU-SILC):
    - Nationals
    - Non-nationals but nationals of an EU country
    - Non-nationals and non-nationals of an EU country
Room for Improvements - Anonymisation

- Occupation (ISCO):
  - aggregated to the 2-digits level
  - EU-SILC provides ISCO also only at the 2-digits level (not for reasons of data confidentiality but because the countries are only required to deliver ISCO-2digits)

- Professional status:
  - self employed with and without employees are combined in a single category

- Information on income:
  - till 2007 it was up to the NSIs whether they pass on income variables or not
  - suppressed for all countries and years for the time being
Room for Improvements - Anonymisation

- each of these aggregations means a loss of information which diminishing the value of these data, e.g.
  
  - Age is a core variable controlled for in most analyses
  
  - Nationality or citizenship is a core variable in migration and integration research
  
  - Occupation is not only a core variable for labour market research it is also used for generating occupational prestige scales (e.g. SIOPS; ISEI) and socioeconomic classifications (e.g. EGP; ESeC)

=> availability of ISCO-3digit = precondition
Room for Improvements - Anonymisation

- To sum up: EU-LFS is the largest and most important database regarding information on labor participation and related issues in Europe

  - However: the sum of aggregation applied to the core variables do reduce the benefits of these data for research purposes substantially

- Data confidentiality is an important issue - not only for official statistics but also for empirical research -

=> Nevertheless data confidentiality measures in the LFS seem to be overdone and ought to be revised
Room for Improvements – Costs

- **EU-LFS**
  - all years from 1983-2007, covering available data in 26 countries
    - 8,000 €
  - current year, covering available data in 26 countries
    - 2,000 €: new users
    - 1,000 €: users who bought the full data set for the previous year

- **EU-SILC**
  - one year cross-sectional **or** longitudinal data
    - 500 €; any following purchase: 250 €
  - annual subscription cross-sectional **and** longitudinal data
    - 1,000 €: first purchase; 750 €: annual subscription
Room for Improvements – Costs

• Although costs for the full EU-LFS can be justified with respect to its extent, 8,000 € often will be prohibitive for the typical researcher

• A further problem arises in multi-national research teams:
  – Does each national team/researcher have to buy their own EU-LFS/EU-SILC data, or
  – are there any possibilities for teams/researchers to share the data and split the costs?
Conclusions

• Important improvements in access to European official microdata
  – Data access has considerably been simplified over the last decade
  – Timeliness of data release is excellent
  – Data documentation is good but can be improved

=> Cross-national research activity has both broadened and deepened
Conclusions

• Preferable improvements

  – Researchers have to articulate their interests and needs more actively, systematically and coordinated

  – The dialog between researchers and Eurostat ought to be institutionalised

  – Anonymisation of EU-LFS data and the cost model should be reconsidered
Thank you for your attention!