Firm-level data pooling

Experiences from Poland

Michał Gradzewicz and Kamil Wierus

Narodowy Bank Polski (NBP)

11 Oct 2019, Bratislava
InGRID Data Forum - Data Pooling: Opportunities and Challenges

The views expressed herein belong to the author and have not been endorsed by Narodowy Bank Polski.

Motivation

- Narodowy Bank Polski has an access to two sets of firm-level data on enterprises in Poland
  - Financial statements and balance sheets (financial data survey - FDS)
  - An underlying data for Eurostat's Structure of earnings survey (SES)
- Problem: there is no common identifier across two data sources
- Vast prospects for the research agenda on combined data-sets:
  - relation between firm efficiency / distance to technological frontier and wage dispersion
  - relation between firm efficiency / distance to technological frontier and employment structure
  - entering the export markets and wage dispersion / changes in the structure of employment
  - firm-level minimum wage range and firm's profitability/efficiency
  - and many others

Financial data survey (FDS)

- Three levels of data frequency with different coverage
  - quarterly - census for 50+, mainly P&L accounts
  - half-year - census for 50+, sample for 10-49, mainly P&L accounts
  - annual - census for 10+, balances and elements of P&L accounts
- Some firms exist only in one of these sources
- There is basic information about the firm (detailed sector, ownership, region)
- Firms from the non-financial enterprises form the enterprise sector (limited number fo firms in the non-market services)
- FDS allows to:
  - measure efficiency - firm-level labor productivity
  - estimate production function, TFP, monopolistic markups - see e.g. Gradzewicz and Mućk (2019)
  - measure firm-level exports and imports
- FDS contains only information on
  - total employment (in full-time equivalent) and
  - total labor costs

Structure of earnings survey (SES)

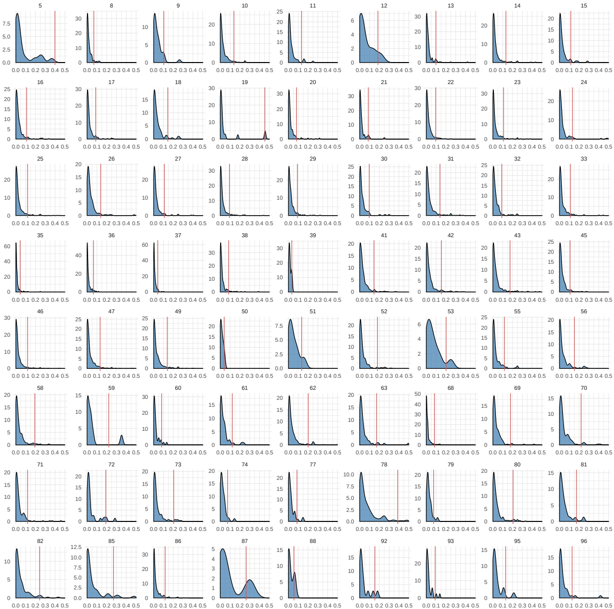
- SES is a bi-annual survey
- A sample of firms with employment 10+
- Includes also firms form the financial and non-market services sectors
- It is possible to generalize observations to relevant part of the enterprise sector
- Within a firm there is a sample of employees
  - employee's characteristics (occupation, gender, education, work-experience, age)
  - employee's wage structure (base wage, additional bonuses, overtime payments, etc)
  - employee's working time arrangements
- It is possible to generalize a sample of employees within firm to a firm's employment
- Very limited information on firm:
  - region (16 regions)
  - 3-digit NACE sector
  - detailed ownership status
  - total employment

Setting the thresholds

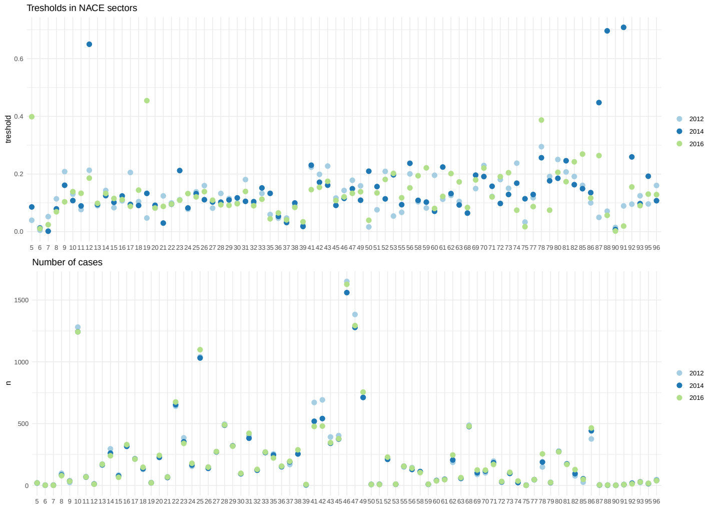
- The time of measurement of employment is different across data-sets
  - SES - information on number of employees in the end of October
  - FDS - number of employees either at the end of 3rd quarter (end of Sep) or at the end of the year (end Dec)

- Solution
  - We calculated (for each firm) the absolute percent difference of employment between Q4 and Q3
  - Within each date-industry cell we computed the 95th qunatile of these percent differences
  - and set them as thresholds for the differences of employment measures between two data-sets

Tresholds in NACE sectors for 2016



Tresholds and sector sizes



- tobacoo (12), oil (19), employment agencies (78) and some non-market services

Matching technique

- We use the R's package reclin (see van der Laan, 2018)
- We supplement it to use simple percentage absolute difference as a distance function (the package uses mainly text-oriented functions)
- We create date-region-ownership-industry cells
- Within cells we compare measures of employment in both datasets using our distance function
- We use probabilistic record matching method (see e.g. Winkler, 2006) and calculate a match-score for all possible pairs within a cell (see Sayers, Ben-Shlomo, Blom, and Steele, 2016)
- We choose pairs to optimize the total score of the selected records under the restriction that each record can be selected only once

Discrepancies of employment in quarterly FDS and SES

Figure: Discrepancies before applying thresholds

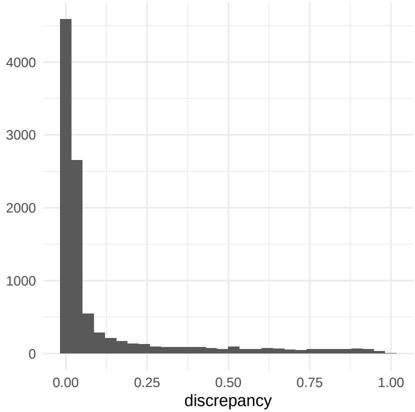
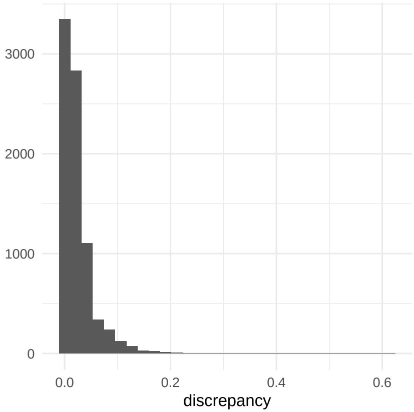


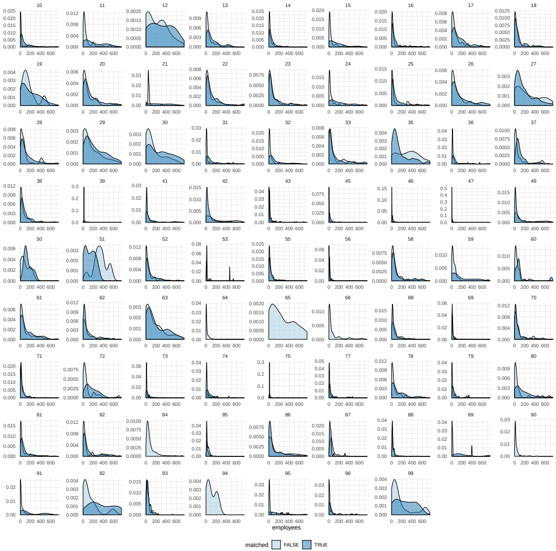
Figure: Discrepancies after applying tresholds



Sequnce of matching

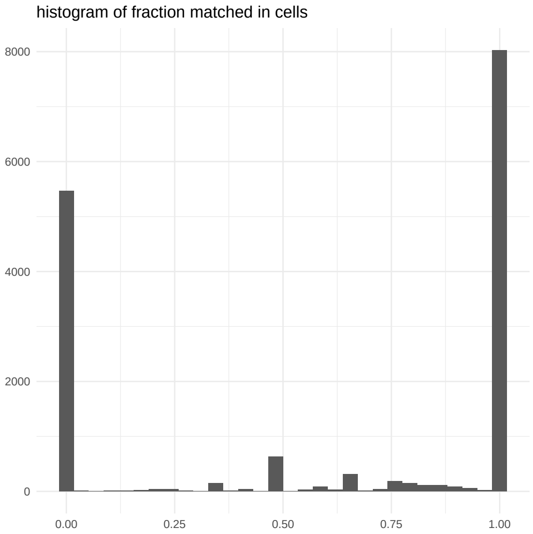
- Total number of firms to match is SES: 56k for years 2012, 2014, 2016
- **1st pass** - matching with quarterly data (51.5k firm-date observations)
- After applying thresholds - 8k successful matches
- **2nd pass** - matching with annual data (167.8k firm-date observations), after dropping matched cases in previous step from both datasets
- After applying thresholds - cumulative 21.3k successful matches
- **3rd pass** - matching with half-year data (151.9k firm-year observations)
- After applying thresholds - 22.2k successful matches
- **4th pass** - matching again with annual data, ownership classification is collapsed form detailed to three-levels (public, private domestic and foreign) and NACE 3-digit codes to NACE 2-digit codes
- After applying thresholds - 27.3k successful matches
- 0.484 matched cases

Distribution of size across industries



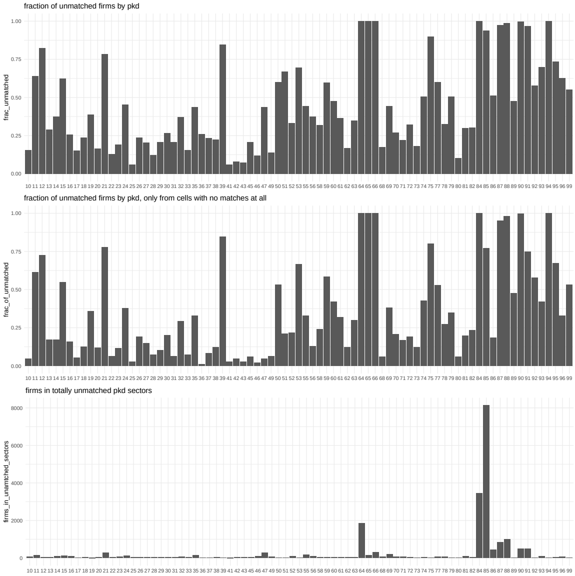
- Problematic sectors:
- beverages (11), pharma(21), metals (24), energy(35), infrastructure construction (42), air transport (51) and some non-market services
- finances (64-66), public administration (84), culture (90), organizations (94) - not present in SF

Fraction of matched firms in cells



- In most cases all SES firms from our defined cells are matched
- relatively small number of cells with partial match
- is a large number of cells with zero matches problematic?

Details of unmatched firms and firms from totally unmatched cells



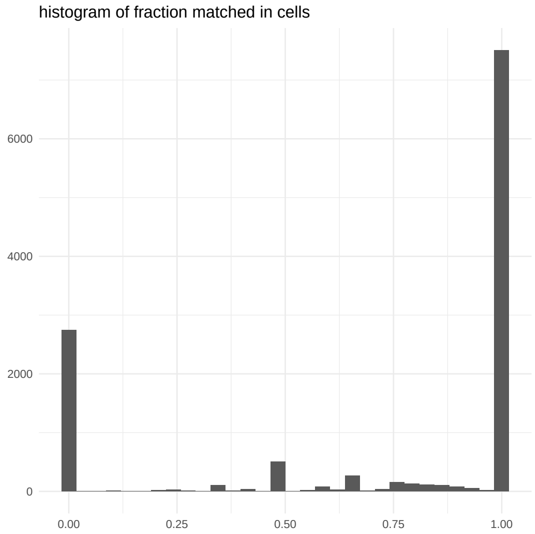
fraction of unmatched firms by pkd

fraction of unmatched firms by pkd, only from cells with no matches at all

firms in totally unmatched pkd sectors

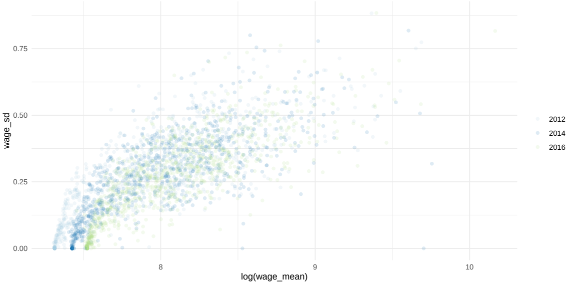
- Most of the unmatched firms are from cells with no matches at all
- Most problematic industries: beverages (11), tobacco(12), pharma(21), recycling (39), finances (64-66) and non-market services (84+, especially 85 - education) - also problematic in terms of size distribution
- We dropped these sectors

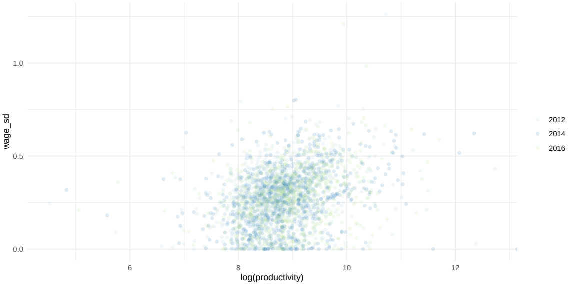
Fraction of matched firms in cells in filtered data



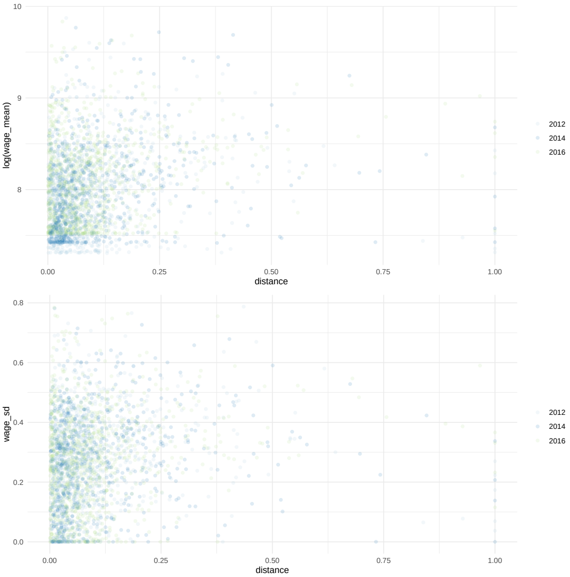
- 25k of matched firms in filtered dataset (vs. 27.3k in unfiltered)
- fraction of matched firms in filtered dataset: 0.75

Mean wages and productivity vs. firm-level wage dispersion





Distance to sectoral productivity frontier and mean wages / dispersion of wages



Conclusions

- Our aim was to match individual employee data from SES to a detailed firm level dataset from FDS
- We used probabilistic record matching methods used mainly to match text records, but we supplement it to account for numeric data
- We carefully accounted for a possible source of differences in the matching metric and set admissible and data-driven thresholds
- We showed that our approach does not induce major selection biases in case of most industries
- We showed that in case of some specific industries no matches have been found
- When dropping these industries the fraction of matched firms rises from 48% to 75%, with a small drop of a number of matched firms

- But still, there are many firms that potentially can be matched
- And the possibilities to do a very detailed research on productivity-wages relationships are really vast and this step is only the beginning of the planned analysis

Literature

Gradzewicz, M., and J. Mućk (2019): “Globalization and the Fall of Markups,” Discussion Paper 304, Narodowy Bank Polski, Economic Research Department.

Sayers, A., Y. Ben-Shlomo, A. W. Blom, and F. Steele (2016): “Probabilistic Record Linkage,” *International Journal of Epidemiology*, 45(3), 954–964.

van der Laan, J. (2018): *Reclin: Record Linkage Toolkit*. R package version 0.1.1.

Winkler, W. E. (2006): “Overview of Record Linkage and Current Research Directions,” Discussion paper, BUREAU OF THE CENSUS.