Adjusting Non-Probability Samples in Job Vacancy Statistics

Statistical Scraping Interest Group Meeting, Vienna, Sept. 16-17 2025

Johanna Bolli-Kemper

Intro

- Complement official statistics about job vacancies (JV) using online vacancy data from job boards or company websites

➔ Representativeness requirement

- Advertising channels of JV in CH:
  - ➢ More than two thirds on firm websites and job boards
  - ➢ ~10% through job boards only, ~15% through firm websites only
- Scraping firm websites and job boards
  - ➢ Covers a large share of the online JV
  - ➢ Could reduce selection bias if way of posting JV differs by e.g. firm size, sector
  - ➢ Requires good deduplication
- BUT: Scraping job boards is cheaper, resulting data is not representative

➔ How do we deal with the NPS nature of job board data?

Swiss Job Market Monitor (SJMM)

– Collects job ads since 2000

Firm sample

- ➢ Probability sample, startified by sector and size
- ➢ 2001–present (refreshed every 10 years)
- ➢ Two functions:

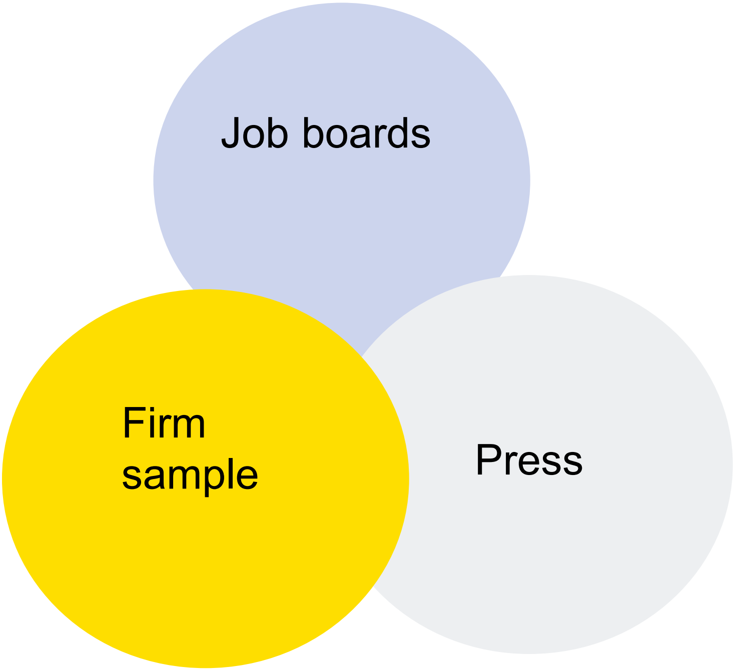
1. Scrape JV from firm websites
2. Survey on firms' posting channels (press, website or job board) and which job boards used for advertising

Job boards

- ➢ Non-probability sample, 2006–present
- ➢ Annual firm survey ensures ~95% coverage

Press archives

- ➢ Non-probability sample
- ➢ Collected retrospectively 1950–2000 and 2001- 2018; discontinued in 2018



Swiss Job Market Monitor (SJMM)

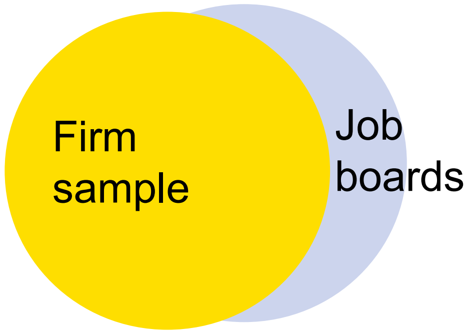
Information extraction

- Fully automated pipline
- Coding of variables with NLP models

Data products

- Quarterly dataset: data collection on a quarterly basis
- ➢ Job index
- ➢ Labor shortage index
- Scientific use file (SUF): subset of quarterly dataset in 1st quarter, more variables, annually published
- **Goal: Increase N of SUF to that of the quarterly dataset**
  - ➢ Fully automated information extraction pipeline reduces cost of coding variables

Current challenge: Handling non-probability samples (NPS)



- **Problem:** Not all JV from firm sample are also contained in job board sample and vice versa

➔ While firm sample is a probability sample, job board sample is an NPS

- Not possible to apply methods of randomization inference
- Estimates based on NPS (without corrections) are usually biased due to selection bias

➢ same applies to press as another source of job ads

- **Central question:** How do we deal with the NPS nature of job board data?
- **Ways to adjust NPS:** All current methods for handling NPS are model-based (e.g. Zhang 2019) and require auxiliary information about the underlying population or a reference sample

How to use NPS to make inferences about a population

Three main approaches (Elliott & Valliant, 2017; Zhang, 2019):

- **Quasi-randomization (QR) approach:**
  - ➢ Estimate a model for the sample inclusion indicator using a reference sample to create pseudo-inclusion probabilities for weighting
  - ➢ All NPS estimates are based on these pseudo-inclusion probabilities
- **Superpopulation (SP) approach:**
  - ➢ Estimate a model for each variable of interest using the NPS
  - ➢ Apply estimated model to the reference sample to predict the variables of interest (mass imputation)
- **Doubly Robust (DR)**
  - ➢ Combine QR and SP
  - ➢ Consistent if either model is correct, more robust to misspecification

How to use NPS to make inferences about a population

Use in official statistics

- SP and DR approaches well suited for official statistics: Few well-defined variables of interest (e.g., company turnover, # of FTEs)
- QR approaches better suited for SUF: goal is not to release just “a single number” but an entire dataset, output are weights used for estimations based on SUF

Key QR methods

- **Post-stratification (PS):** needs only marginal population data (e.g. firm size and sector) to generate pseudo-inclusion probabilities; simple but may be insufficient to correct selection bias
- **Inverse Propensity Weighting (IPW):** requires a reference sample where the inclusion indicator is observed or can be constructed
    - ➢ Enables matching (ideally at unit level of firms, otherwise via strata like size, industry, region)
    - ➢ Estimate pseudo-inclusion probabilities for NPS; their inverse serves as weights in the SUF
    - ➢ BUT: need detailed information about which websites frims advertise on and which firms advertise on job board x
    - ➢ Otherwise, PS and IPW coincide

Conclusion and future challenge

- Still some work to be done to implement the NPS adjustment at SJMM
- PS way of adjusting NPS can be interesting for cases where large JV datasets are uniquely or partially drawn from job boards
- Future challenges:
  - Data integration strategy: once we have managed to make NPS more representative, how to combine probability sample with “adjusted” NPS (including press)
  - Improve infrastructure: deduplication (…)

References

Elliott, M. R. and Valliant, R. (2017). Inference for Nonprobability Samples, Statistical Science, **32**, 249–264. DOI 10.1214/16-STS598

Zhang, L.-C. (2019). On valid descriptive inference from non-probability sample, Statistical Theory and Related Fields, **3**, 103–113. DOI: 10.1080/24754269.2019.1666241