IFC Satellite Seminar on "Granular data: new horizons and challenges for central banks"

Classifying Job Postings into NAICS Codes[1]

Tatjana Dahlhaus, Reinhard Ellwanger, Gabriela Galassi and Pierre-Yves Yanni,
Bank of Canada

[1] This contribution was prepared for the IFC Satellite Seminar held at the ISI 64th World Statistics Congress, co-organised with the Bank of Canada in Ottawa, Canada, on 15 July 2023. The views expressed are those of the authors and do not necessarily reflect the views of the Bank of Canada, the BIS, the IFC or the other central banks and institutions represented at the event.

Classifying Job Postings into NAICS Codes

Tatjana Dahlhaus, Reinhard Ellwanger,
Gabriela Galassi, and Pierre-Yves Yanni

Bank of Canada

BoC-IFC satellite seminar
July 15, 2023

Motivation

- Compared to JVWS data, Indeed job postings data are
  - more timely and frequent
  - more granular
- But they have their own sectoral classification, different from NAICS
- By classifying these job postings into NAICS, we can
  - track vacancies by industries at higher frequency and ahead of official data
  - backcast the missing official data
  - further exploit the granularity of the data
- Methodology to match datasets without common keys

Literature

- **Job posting-company matching:** Van Dijcke, Buckmann, Turrell, and Key (2023)

  Match firm-level balance sheets to analyze effect of leverage on employment; match on name only

- **Technology adoption during recessions:**
  - Hershbein and Kahn (2018), Jaimovich and Siu (2020)
  - Soh, Oikonomu, Pizzinelli, Shibata, and Tavares (2023), Pierri and Timmer (2020), Bellatin, Chu and Galassi (2023)

  Do not consider different types of firms; implications for how permanent are changes.

Roadmap

- Data
- Matching to NAICS
  - Methodology
  - Validation
- Applications
  - Backcasting missing data
  - Digital occupations and tech firms
- Takeaways

Data

Indeed job postings: largest job board in Canada; posting from companies and fetched from other websites (2018-present)

Advan: visits to points-of-interests from cellphone location data (2019-present)

Indeed job postings
job key
company name
city
province
first date
last date

Advan
company name
city
province
NAICS (4-6 digits)

Matching to NAICS

1. Preprocess data
2. Transform company names into numerical vectors
3. Match company names at 3 geographical levels: city, province, national
4. Pick cutoff on matching scores

Matching to NAICS

Preprocess | Embeddings | Matching | Cutoff

- For company names and cities, remove
  - capital letters
  - accents
  - spaces and special characters

- Remove stopwords from company names: *inc, limited, group, . . .*

- Map cities in Indeed to cities in Advan

Matching to NAICS

Preprocess | **Embeddings** | Matching | Cutoff

- We create embeddings for company names using tf-idf:

  $$\textbf{tf-idf} = \text{term frequency (tf)} \times \text{inverse document freq. (idf)}$$
  $$\textbf{idf} = \log(\text{number of documents / document frequency})$$

- In our case:
  - *document* = company name
  - *term* = trigram

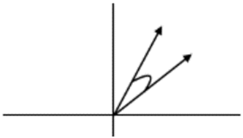
- Example: *pepsico, pepsicola and pepper*

	Term (trigram) frequency									
	pep	eps	psi	sic	ico	col	ola	epp	ppe	per
pepsico	1	1	1	1	1	0	0	0	0	0
pepsicola	1	1	1	1	1	1	1	0	0	0
pepper	1	0	0	0	0	0	0	1	1	1
Document (company name) frequency	3	2	2	2	2	1	1	1	1	1

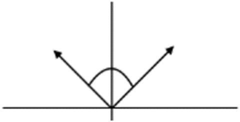
Matching to NAICS

Preprocess | **Embeddings** | Matching | Cutoff

- Cosine similarity (2 dimensions):



similar trigrams:
$\cos \to 1$



different trigrams:
$\cos \to 0$

- In our example:
  - vectors of dimension 10
  - cos(pepsico, pepsicola) $= 0.76$, cos(pepsico, pepper) $= 0.12$
- Our data:
  - 670k company names in Indeed and 640k in Advan
  - vectors of dimension 26k

Matching to NACS

Preprocess | Embeddings | **Matching** | Cutoff

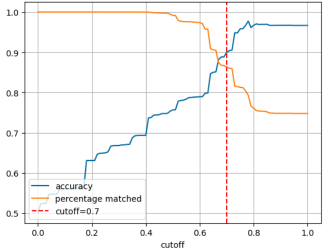
- Find best matches at city, province, national levels based on cosine similarity

- For given cutoffs on cosine similarity,
  1. keep matches above cutoff at **city** level
  2. among companies unmatched at city level,
     keep matches above cutoff at **province** level
  3. among companies unmatched at city and province level,
     keep matches above cutoff at **national** level

Matching to NAICS

Preprocess | Embeddings | Matching | **Cutoff**

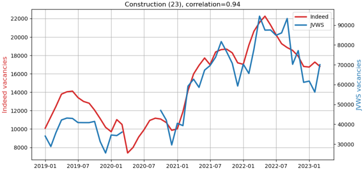
- Vary cutoff on labelled sample of firms matched:

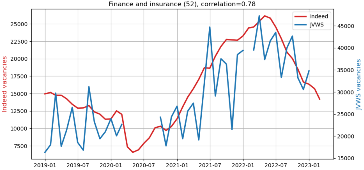


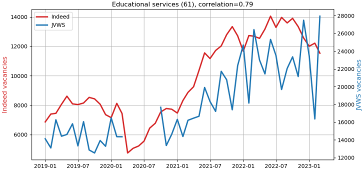
- Cutoff $= 0.7$ chosen to get 90% weighted accuracy
- Matching results:

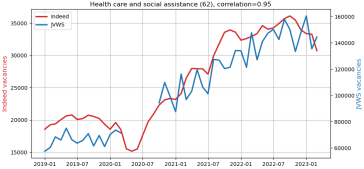
Indeed company name	Advan company name	City	Province	Score	NAICS
Pizza Munno	Pizza Munno	Oshawa	ON	1.0	722511 Full-service restaurants
David's Tea	Les Thés David's Tea	Sherbrooke	QC	0.72	722515 Snack and Nonalcoholic Beverage Bars
Canadian Pacific	Canadian Tire	Cochrane	AB	0.42	unmatched

Matching to NAICS: Validation with JVWS Data







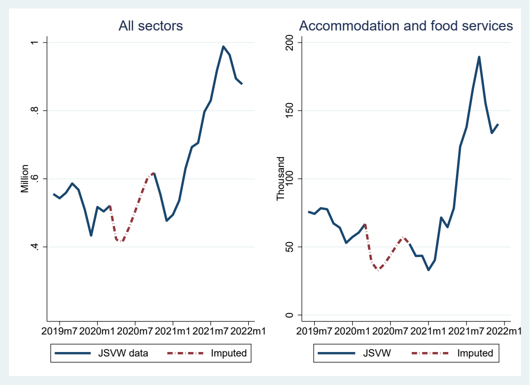


Application 1: Backcasting Missing Data

- JVWS did not collect data from April to September 2020
- Estimate relationship:

$$\Delta v_{t,j}^{jvws} = \alpha + \beta_0 \Delta v_{t-1,j}^{ind} + \beta_1 \Delta v_{t,j}^{ind} + \beta_2 \Delta v_{t+1,j}^{ind} + \varepsilon_{t,j}.$$

- Use the estimated $\widehat{\Delta v_{t,j}^{jvws}}$ to fill in the missing values:



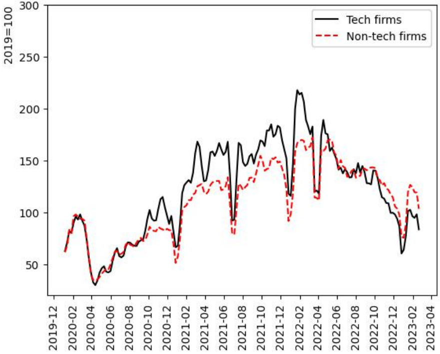
Application 2: Digital occupations and tech firms

- Bellatin, Chu and Galassi (2023) study digital job creation during the COVID-19 pandemic
- Adding tech/non-tech firms dimension:
  1. what type of firms drive the creation of digital vacancies?
  2. digital production vs digital adoption:

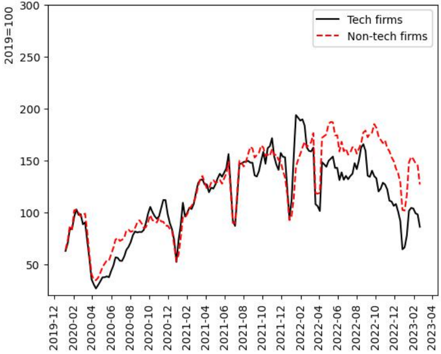
	Digital Jobs	**Non-Digital Jobs**
Tech Firms	Digital production	Digital-adjacent
Non-Tech Firms	Digital adoption	

Application 2: Digital Jobs and Tech Firms

Digital job postings



Non-digital job postings



- disproportionately more digital jobs vacancies in tech industry during pandemic
- relative to non-tech industry, post-pandemic vacancies decreased mainly in non-digital occupations

Takeaways

- Fuzzy matching to classify job postings into NAICS codes
  - high accuracy in matching
  - tracks well vacancies stocks by industry from official statistics
- Newly created dataset
  - complements official data
  - allows to analyze digital production vs digital adoption
- Next steps:
  - obtain Business Register data
  - reduce lag in job postings stock

Representativeness Indeed Job Postings

name	NAICS	cutoffs	fractions	corr.
Utilities	22	0.90, 0.80, 1.00	0.04	0.617
Construction	23	0.85, 0.65, 0.60	0.38	0.96
Manufacturing	31-33	0.90, 0.80, 0.95	0.06	0.951
Wholesale trade	41/42	0.70, 0.70, 0.60	0.22	0.9
Retail trade	44-45	0.90, 0.80, 0.85	0.67	0.877
Transportation and warehousing	48-49	0.70, 0.80, 1.00	0.15	0.86
Information and cultural industries	51	0.80, 1.00, 0.60	0.31	0.858
Finance and insurance	52	0.95, 0.60, 0.75	0.57	0.791
Real estate and rental and leasing	53	0.80, 0.90, 0.60	0.84	0.803
Prof, scientific and technical services	54	0.70, 0.60, 0.70	0.22	0.938
Management of companies and enterprises	55	0.60, 0.90, 0.70	1.42	0.762
Administrative and support	56	0.85, 0.70, 0.60	0.09	0.811
Educational services	61	0.85, 0.75, 0.60	0.64	0.841
Health care and social assistance	62	1.00, 1.00, 1.00	0.16	0.96
Arts	71	0.70, 0.60, 1.00	0.68	0.815
Accommodation and food services	72	0.70, 0.70, 0.60	0.32	0.862
Other services (except public administration)	81	0.60, 0.60, 0.60	0.61	0.939
Public administration	91/92	0.95, 0.65, 0.60	0.19	0.794

Representativeness Indeed Job Postings

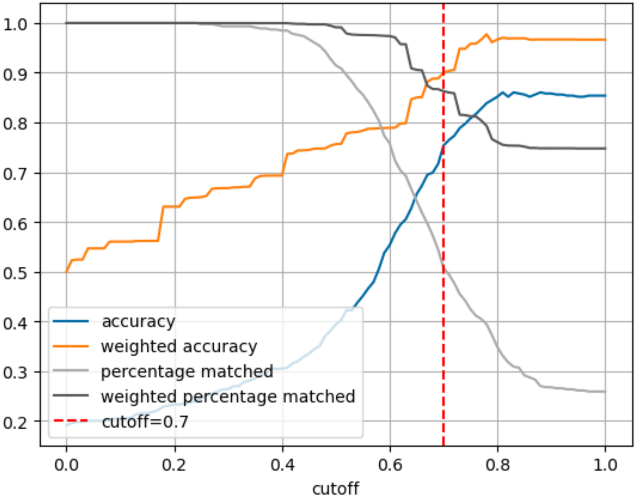
Ratio companies in Advan / Employer businesses from StatCan:

NAICS	11	21	22	23	31-33	41/42	44-45	48-49	51
ratio Advan / StatCan	0	0	0.31	0.27	0.31	0.23	1.15	1.39	0.37

NAICS	52	53	54	55	56	71	72	81
ratio Advan / StatCan	0.54	0.33	0.12	0.04	0.15	1.25	0.97	0.78

source: Business Dynamics measures, Statistiques Canada

Unweighted vs Weighted Accuracy



Application 1: Backcasting Missing Data

- JVWS did not collect data in Q2 and Q3 2020
- Solution:
  - estimate relationship:

$$\Delta v_{t,j}^{jvws} = \alpha + \beta_0 \Delta v_{t-1,j}^{ind} + \beta_1 \Delta v_{t,j}^{ind} + \beta_2 \Delta v_{t+1,j}^{ind} + \varepsilon_{t,j}.$$

  - estimate difference in trend during the period with missing observations

$$\Delta T = (v_{Oct20,j}^{jvws} - v_{Mar20,j}^{jvws}) - \sum_{i=1}^{7} \Delta \widehat{v_{Mar20+i,j}^{jvws}}$$

  - use the estimated $\widehat{\Delta v_{t,j}^{jvws}}$ to fill in the missing values k periods after March 2020:

$$\hat{v}_{Mar20+k,j} = v_{Mar20,j} + \sum_{i=1}^{k} \Delta \widehat{v_{Mar20+i,j}^{jvws}} + k \cdot \Delta T/7,$$