mannheimer zentrum für europäische sozialforschung

Institute for Employment Research

The Research Institute of the Federal Employment Agency

A Comparison of Automatic Algorithms for Occupation Coding

BigSurv 2018

Barcelona

October 26, 2018

Malte Schierholz

Occupations are analyzed within ...

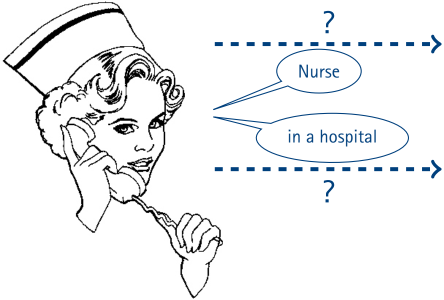
- Economics: individuals' tasks and duties, ...
- Sociology: social status, ...
- Medicine: health risks, ...

How to measure occupations in interviews?
With open-ended questions!

The Coding Problem

Assign verbatim answers to an official classification

Survey Answers



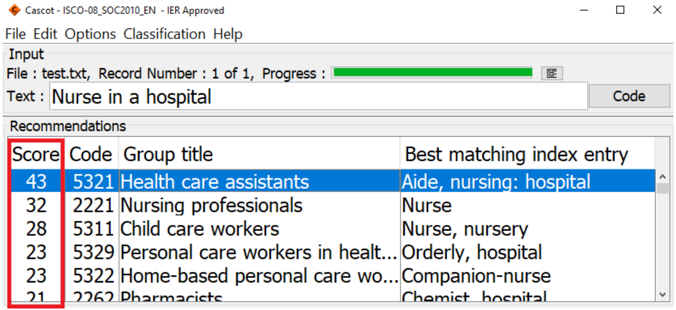
Classification

KldB 2010	Job Title (illustrative)
81302	Nurse
81313	Graduated Nurse
81323	Pediatric Nurse
⋮	⋮
81102	Doctor's Assistant
82102	Geriatric Nurse

→ Coding is mostly manual work and therefore expensive

→ How to automate it?

Modes of Automation



©CASCOT (`https://warwick.ac.uk/fac/soc/ier/software/cascot/`)

Two modes of automation:

- Computer-assisted coding: Human choice, see picture
- Automated coding: Select top-ranked category

→ How to calculate scores / probabilities?

The Challenge

- Manual occupation coding is hard
  - Coders are trained on the classification and coding conventions
    (Geis and Hoffmeyer-Zlotnik (2000), Paulus and Matthes (2013))
  - Coders draw on information from multiple questions
  - Inter-coder reliabilities range between 61-83%
    (Prigge et al. (2014), Schierholz et al. (2018))

- High-dimensional classification problem
  - Our document-term matrix has 54.880 observations and 14.223 columns (words)
  - 1.291 outcome categories
  - $>$ 15 million parameters in a multinomial logistic regression

Ten Algorithms

Comprehensive comparison of ten algorithms:

1. Coding index (exact)
2. Coding index (\w similars) (Elias et al. (2014))
3. Memory-based Reasoning (Creecy et al. (1992))
4. Adapted Nearest Neighbor (Gweon et al. (2017))
5. Multinomial Logit (R-package `glmnet`, Friedman et al. (2010))
6. XGBoost (R-package `XGBoost`, Chen and Guestrin (2016))
7. Coding index + Bayes (fulltext similarity)
8. Coding index + Bayes (substring similarity)
9. Coding index + Bayes (wordwise similarity)
10. Highest probability

Items 1–4: Developed for occupation coding

Items 5–6: General purpose

Items 7–10: Own contribution

Ten Algorithms

	Algorithm	Reference data	Text processing
1.	Coding index (exact)	Coding Index	String identity
2.	Coding index (\w similars)	Coding Index	String similarity
3.	Memory-based Reasoning	Training data	Document-term matrix
4.	Adapted Nearest Neighbor	Training data	Document-term matrix
5.	Multinomial Logit	Training data	Document-term matrix
6.	XGBoost	Training data	Document-term matrix
7.	Coding index+Bayes (fulltext similarity)	Coding Index & Training	String similarity (fulltext similarity)
8.	Coding index+Bayes (substring similarity)	Coding Index & Training	String similarity (substring similarity)
9.	Coding index+Bayes (wordwise similarity)	Coding Index & Training	String similarity (wordwise similarity)
10.	Highest probability	Among algorithms 6-9, select the algorithm with highest probability	

Algorithms for Automatic Coding (1)

First algorithm: **Consult a coding index**

2010 German Classification

(27.575 job titles)

Job title	Code
Säuglings- und Kinderkrankenschwester/-pfleger	81302
Säuglings- und Krankenschwester/-pfleger	81302
Säuglingskrankenschwester/-pfleger	81302
Säuglingsschwester/-pfleger	81302
Schwester/Pfleger (Kinderkrankenpflege)	81302
Schwester/Pfleger (Krankenpflege)	81302
Anästhesieschwester/-pfleger	81313
Dialysefachkraft	81313
Dialyseschwester/-pfleger	81313
Endoskopieassistent/in	81313
Endoskopieschwester/-pfleger	81313
Fachkrankenpfleger/in - Notfallpflege	81313

Source: KldB 2010

2018 US Classification

(6.539 job titles)

Job Title	Code	Category Title
Hospice Registered Nurse	29-1141	Registered Nurses
Obstetrical Nurse	29-1141	Registered Nurses
Oncology Registered Nurse	29-1141	Registered Nurses
PACU Nurse	29-1141	Registered Nurses
Pediatric Registered Nurse	29-1141	Registered Nurses
Post-Anesthesia Care Unit Nurse	29-1141	Registered Nurses
Psychiatric Nurse	29-1141	Registered Nurses
RN	29-1141	Registered Nurses
Triage Registered Nurse	29-1141	Registered Nurses
Certified Registered Nurse Anesthetist	29-1151	Nurse Anesthetists
Certified Registered Nurse Anesthetist (CRNA)	29-1151	Nurse Anesthetists
DNAP	29-1151	Nurse Anesthetists
Doctor of Nurse Anesthesia Practice	29-1151	Nurse Anesthetists
Certified Nurse Midwife	29-1161	Nurse Midwives
Certified Nurse Midwife (CNM)	29-1161	Nurse Midwives
Acute Care Nurse Practitioner	29-1171	Nurse Practitioners
Adult Nurse Practitioner	29-1171	Nurse Practitioners
Cardiology Nurse Practitioner	29-1171	Nurse Practitioners

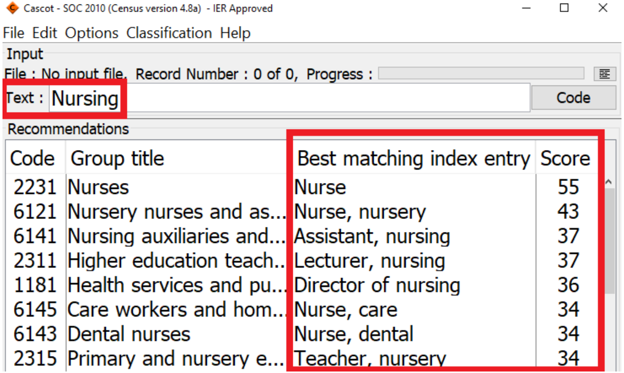
Source: 2018 SOC

Algorithms for Automatic Coding (2)

Second algorithm:

Consult a coding index (based on string similarity)

(Appel and Hellerman (1983), Wenzowski (1988), United Nations Statistical Commission and Economic Commission for Europe (1997), Elias et al. (2014))



©CASCOT (`https://warwick.ac.uk/fac/soc/ier/software/cascot/`)

Ten Algorithms

	Algorithm	Reference data	Text processing
1.	Coding index (exact)	Coding Index	String identity
2.	Coding index (\w similars)	Coding Index	String similarity
3.	Memory-based Reasoning	Training data	Document-term matrix
4.	Adapted Nearest Neighbor	Training data	Document-term matrix
5.	Multinomial Logit	Training data	Document-term matrix
6.	XGBoost	Training data	Document-term matrix
7.	Coding index+Bayes (fulltext similarity)	Coding Index & Training	String similarity (fulltext similarity)
8.	Coding index+Bayes (substring similarity)	Coding Index & Training	String similarity (substring similarity)
9.	Coding index+Bayes (wordwise similarity)	Coding Index & Training	String similarity (wordwise similarity)
10.	Highest probability	Among algorithms 6-9, select the algorithm with highest probability	

Algorithms for Automatic Coding (3–6)

Documents:

$d_1 =$ Nurse in a hospital $\rightarrow$

$d_2 =$ Hospital manager $\rightarrow$

$\vdots$

Document-term matrix:

...	a	hospital	in	manager	nurse	...
...	1	1	1	0	1	...
...	0	1	0	1	0	...
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

Prediction techniques:

Based on vector similarity (k-NN-like):

3. *Memory-based Reasoning*
   (Creecy et al. (1992))
4. *Adapted Nearest-Neighbor*
   (Gweon et al. (2017))

Based on loss minimization:

5. *Multinomial logistic regression*
   (R-package `glmnet`, Friedman et al. (2010))
6. *Gradient boosting with decision trees*
   (R-package `XGBoost`, Chen and Guestrin (2016))

Algorithms for Automatic Coding (3–6)

Documents:

d_1 = Nurse in a hospital →
d_2 = Hospital manager →
⋮

Document-term matrix:

...	a	hospital	in	manager	nurse	...
...	1	1	1	0	1	...
...	0	1	0	1	0	...
⋮	⋮	⋮	⋮	⋮	⋮	⋮

Prediction techniques:

Based on vector similarity (k-NN-like):

3. *Memory-based Reasoning*
   (Creecy et al. (1992))
4. *Adapted Nearest-Neighbor*
   (Gweon et al. (2017))

Based on loss minimization:

5. *Multinomial logistic regression*
   (R-package `glmnet`, Friedman et al. (2010))
6. *Gradient boosting with decision trees*
   (R-package `XGBoost`, Chen and Guestrin (2016))

Algorithms for Automatic Coding (3-6)

Documents:

d_1 = Nurse in a hospital →
d_2 = Hospital manager →
⋮

Document-term matrix:

...	a	hospital	in	manager	nurse	...
...	1	1	1	0	1	...
...	0	1	0	1	0	...
⋮	⋮	⋮	⋮	⋮	⋮	⋮

New problem:

Algorithms 3-6 fail if important words (especially job titles) are not in the training data

- Infrequent jobs (e.g., ornithologist, horseshoer)
- Misspellings

Algorithm 2 already solved the issue. How to combine the different approaches?

Ten Algorithms

	Algorithm	Reference data	Text processing
1.	Coding index (exact)	Coding Index	String identity
2.	Coding index (\w similars)	Coding Index	String similarity
3.	Memory-based Reasoning	Training data	Document-term matrix
4.	Adapted Nearest Neighbor	Training data	Document-term matrix
5.	Multinomial Logit	Training data	Document-term matrix
6.	XGBoost	Training data	Document-term matrix
7.	Coding index+Bayes (fulltext similarity)	Coding Index & Training	String similarity (fulltext similarity)
8.	Coding index+Bayes (substring similarity)	Coding Index & Training	String similarity (substring similarity)
9.	Coding index+Bayes (wordwise similarity)	Coding Index & Training	String similarity (wordwise similarity)
10.	Highest probability	Among algorithms 6-9, select the algorithm with highest probability	

Algorithms for Automatic Coding (7–9)

A novel algorithm relies on the coding index, but uses past data as well.

Key idea:

- Training: Estimate for each entry in the coding index a posterior predictive distribution (Hierarchical Bayesian model)

$$\mathbb{P}(\text{code}_j | \text{index entry})$$

- Prediction: Search for a similar index entry and output its distribution as estimated

Three ways of calculating string similarity (Algorithms 7–9):

7. *Fulltext similarity*: Similar if the verbal answer differs by at most one character
8. *Substring similarity*: Similar if an entry from the coding index is a substring of the verbal answer
9. *Wordwise similarity*: Similar if one word in the verbal answer differs by at most one character

Algorithms for Automatic Coding (7–9)

A novel algorithm relies on the coding index, but uses past data as well.

Key idea:

- Training: Estimate for each entry in the coding index a posterior predictive distribution (Hierarchical Bayesian model)

$$\mathbb{P}(\text{code}_j | \text{index entry})$$

- Prediction: Search for a similar index entry and output its distribution as estimated

Three ways of calculating string similarity (Algorithms 7–9):

7. *Fulltext similarity*: Similar if the verbal answer differs by at most one character
8. *Substring similarity*: Similar if an entry from the coding index is a substring of the verbal answer
9. *Wordwise similarity*: Similar if one word in the verbal answer differs by at most one character

Ten Algorithms

	Algorithm	Reference data	Text processing
1.	Coding index (exact)	Coding Index	String identity
2.	Coding index (\w similars)	Coding Index	String similarity
3.	Memory-based Reasoning	Training data	Document-term matrix
4.	Adapted Nearest Neighbor	Training data	Document-term matrix
5.	Multinomial Logit	Training data	Document-term matrix
6.	XGBoost	Training data	Document-term matrix
7.	Coding index+Bayes (fulltext similarity)	Coding Index & Training	String similarity (fulltext similarity)
8.	Coding index+Bayes (substring similarity)	Coding Index & Training	String similarity (substring similarity)
9.	Coding index+Bayes (wordwise similarity)	Coding Index & Training	String similarity (wordwise similarity)
10.	Highest probability	Among algorithms 6-9, select the algorithm with highest probability	

Ten Algorithms

	Algorithm	Reference data	Text processing
1.	Coding index (exact)	Coding Index	String identity
2.	Coding index (\w similars)	Coding Index	String similarity
3.	Memory-based Reasoning	Training data	Document-term matrix
4.	Adapted Nearest Neighbor	Training data	Document-term matrix
5.	Multinomial Logit	Training data	Document-term matrix
6.	XGBoost	Training data	Document-term matrix
7.	Coding index+Bayes (fulltext similarity)	Coding Index & Training	String similarity (fulltext similarity)
8.	Coding index+Bayes (substring similarity)	Coding Index & Training	String similarity (substring similarity)
9.	Coding index+Bayes (wordwise similarity)	Coding Index & Training	String similarity (wordwise similarity)
10.	Highest probability	Among algorithms 6-9, select the algorithm with highest probability	

Evaluation

Application in survey interviews envisioned:

- Only texts from the first open-ended question are available (unlike current coding situations)
- Let the computer decide what to ask next
  - Reduce respondent burden
  - Increase data quality

1\. What is the occupational activity you do in your main job?

Nursing

Next

2\.

Automated Coding

(=Computer selects the top-ranked category)

81302 (Nurse)

Computer-assisted Coding

(=Human chooses from a list of suggestions)

Welche Tätigkeit führen Sie in Ihrem Beruf hauptsächlich aus?

- ○ **eigenständige Pflege und Betreuung von Patienten und Assistenz bei ärztlichen Maßnahmen**
  Krankenschwester/-pfleger
- ◉ **Unterstützung von Pflegefachkräften bei der Versorgung und Pflege von Patienten in Krankenhäusern oder Pflegeheimen**
  Krankenpflegehelfer/in
- ○ **Betreuung und Pflege von kranken Säuglingen, Kindern und Jugendlichen in bestimmten medizinisch-pflegerischen Fachgebieten**
  Fachkinderkrankenschwester/ -pfleger
- ○ **Werksärzten bei allen regelmäßigen Untersuchungen und Maßnahmen der Mitarbeiter eines Betriebes assistieren**
  Werkpfleger/-schwester

Data

2 German surveys, coded into the latest German occupational classification

- Dataset 1: 55,944 texts from *BIBB/BAuA Employment Survey 2012*
  (Hartmann et al. (2012), Hall et al. (2015))
  - Split in training and test data
- Dataset 2: 1,064 texts from *Selectivity effects in address handling*
  (Schierholz et al. (2018))
  - Test data only

Coders had access to more occupational variables in the second study – higher coding quality?

1. Performance of Automated Coding
(=Computer selects the top-ranked category)

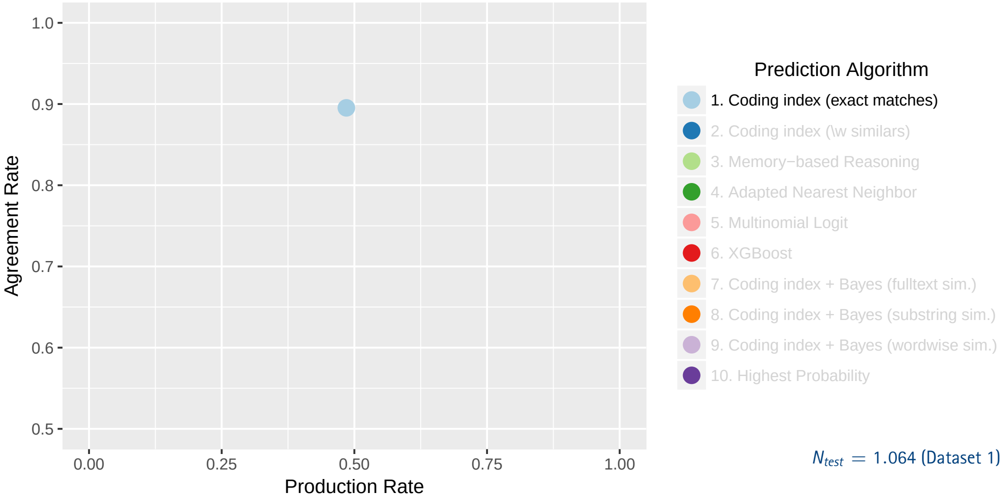
Metrics used:

- Production rate: Proportion of answers that are coded automatically
- Agreement rate: proportion of coded answers that are in agreement with human coder

Trade-Off: Production rate $\leftrightarrow$ Agreement rate

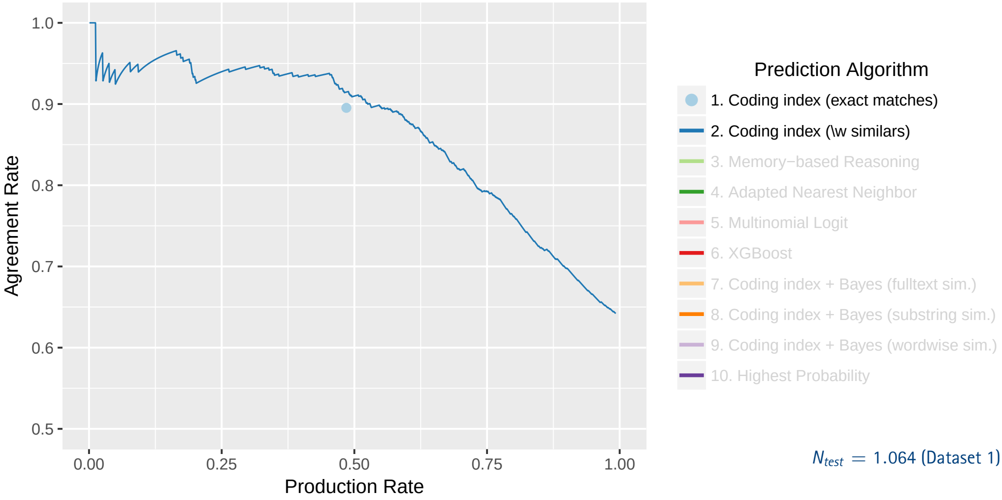
2. Performance of Computer-assisted Coding
(=Human coder chooses from a list of suggestions)

Automated Coding – Results



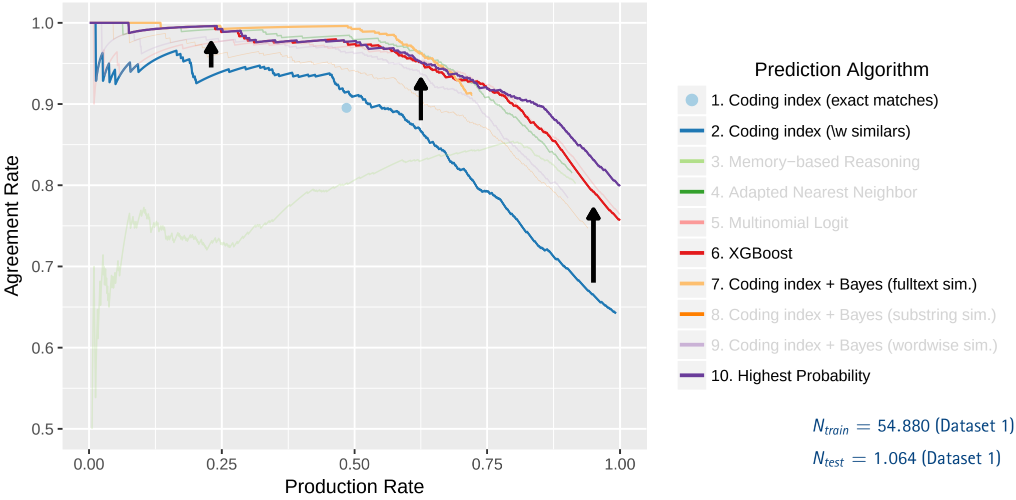
$N_{test} = 1.064$ (Dataset 1)

Automated Coding – Results



$N_{test} = 1.064$ (Dataset 1)

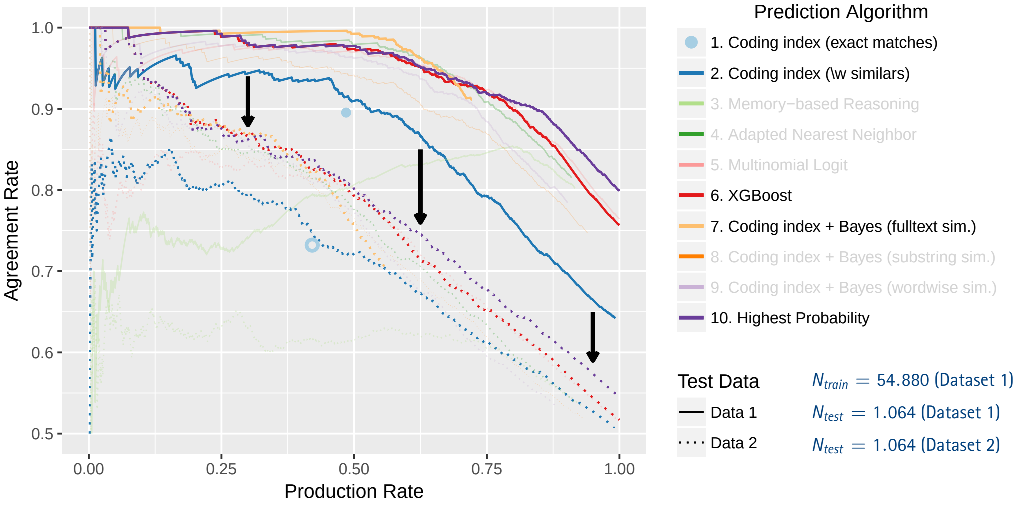
Automated Coding – Results



$N_{train} = 54.880$ (Dataset 1)

$N_{test} = 1.064$ (Dataset 1)

Automated Coding – Results



$N_{train} = 54.880$ (Dataset 1)

$N_{test} = 1.064$ (Dataset 1)

$N_{test} = 1.064$ (Dataset 2)

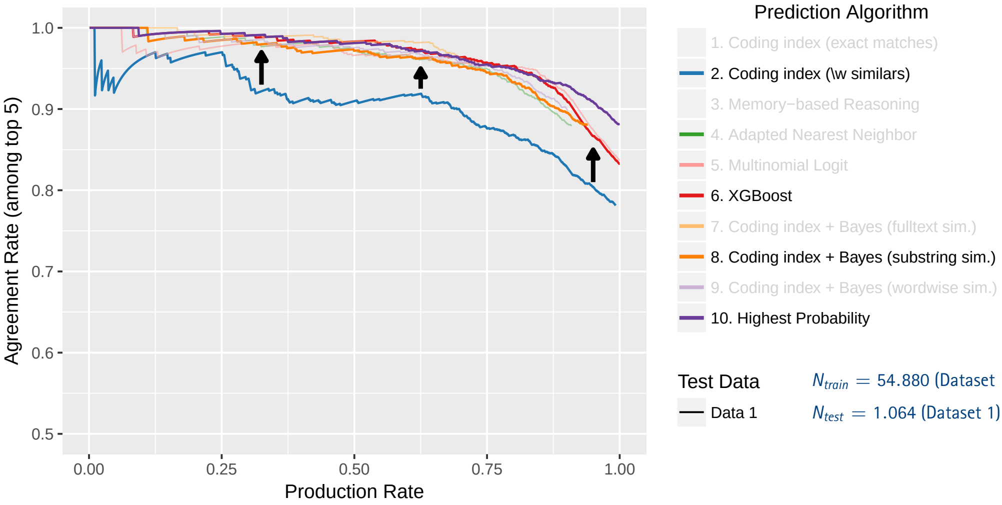
1. Performance of Automated Coding
(=Computer selects the top-ranked category)

2. Performance of Computer-assisted Coding
(=Human coder chooses from a list of suggestions)

Agreement rate is calculated in a different way:

- For automated coding: highest-ranked category in agreement with human coder? ($k = 1$ until now)
- For computer-assisted coding: human-coded category among top k suggestions? ($k = 5$ in the following)

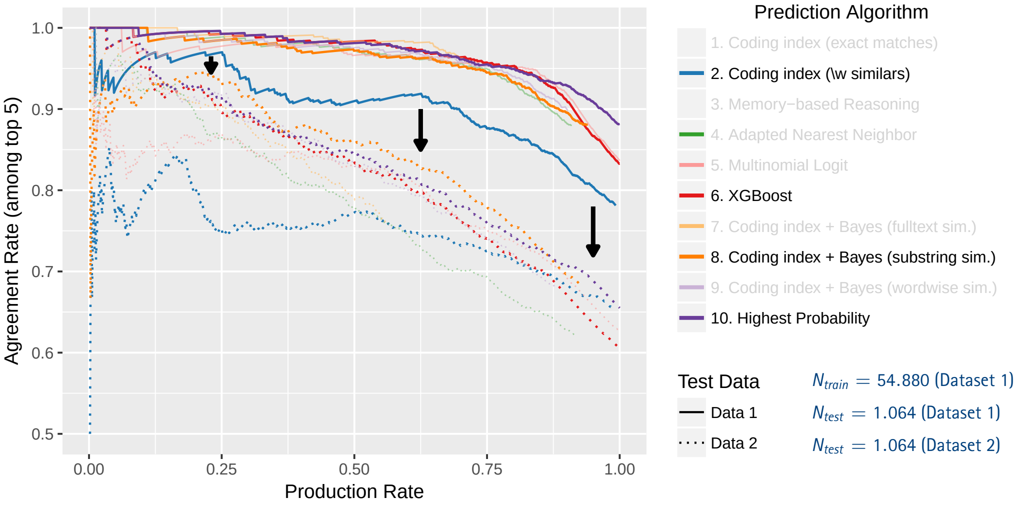
Computer-Assisted Coding – Results



$N_{train} = 54.880$ (Dataset 1)

$N_{test} = 1.064$ (Dataset 1)

Computer-Assisted Coding – Results



$N_{train} = 54.880$ (Dataset 1)

$N_{test} = 1.064$ (Dataset 1)

$N_{test} = 1.064$ (Dataset 2)

Summary

Performance

- Algorithm 2 (Coding index \w similars) is often used in practice but outperformed by other algorithms (e.g. 6. `XGBoost`)
- Combining the coding index and previously coded data is very promising (algorithms 7-10)
  - Best algorithm is problem-dependent

Generalization

- Poor generalization from one dataset to different test data

Application

- Agreement rates are often insufficient for fully automated coding
- Manual verification will usually be necessary, making computer-assisted coding most promising
  - Current plan: Suggest plausible answer options to respondents during the interview

Thank you!

Related publications:

The methods presented in this talk are also implemented in an R-Package:
`https://github.com/malsch/occupationCoding`

Background material:
Schierholz, M; Gensicke, M.; Tschersich, N. and Kreuter, F. (2018). Occupation Coding During the Interview, Journal of the Royal Statistical Society: Series A 181(2): 379–407.

Malte Schierholz
Malte.Schierholz@iab.de

References I

Appel, M. V. and Hellerman, E. (1983). Census bureau experiments with automated industry and occupation coding, *Proceedings of the Survey Research Methods Section: American Statistical Association*, pp. 32–40.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system, *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, ACM, New York, NY, USA, pp. 785–794.
URL: *http://doi.acm.org/10.1145/2939672.2939785*

Creecy, R. H., Masand, B. M., Smith, S. J. and Waltz, D. L. (1992). Trading mips and memory for knowledge engineering, *Commun. ACM* **35**(8): 48–64.

Elias, P., Birch, M. and Ellison, R. (2014). CASCOT International version 5, *User Guide*, Institute for Employment Research, University of Warwick, Coventry. (Available from http://www2.warwick.ac.uk/fac/soc/ier/software/cascot/internat/.).

References II

Friedman, J., Hastie, T. and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent, *Journal of Statistical Software* **33**(1): 1–22.

Geis, A. and Hoffmeyer-Zlotnik, J. H. (2000). Stand der Berufsvercodung, *ZUMA-Nachrichten* **24**(47): 103–128.

Gweon, H., Schonlau, M., Kaczmirek, L., Blohm, M. and Steiner, S. (2017). Three methods for occupation coding based on statistical learning, *Journal of Official Statistics* **33**(1): 101–122.

Hall, A., Siefer, A. and Tiemann, M. (2015). BIBB/BAuA Employment Survey of the Working Population on Qualification and Working Conditions in Germany 2012. vt_1.0, sv_2.0, *Research Data Center at BIBB (ed., data access)*, Federal Institute for Vocational Education and Training, Bonn.
URL: *https://doi.org/10.7803/501.12.1.4.10*

References III

Hartmann, J., Tschersich, N. and Schütz, G. (2012). Die Vercodung der offenen Angaben zur beruflichen Tätigkeit nach der Klassifikation der Berufe 2010 (KldB 2010) und nach der International Standard Classification of Occupations 2008 (ISCO08), *Technical report*, TNS Infratest Sozialforschung, Munich.
URL: *https://metadaten.bibb.de/download/684*

Paulus, W. and Matthes, B. (2013). Klassifikation der Berufe * Struktur, Codierung und Umsteigeschlüssel, *FDZ-Methodenreport 08/2013*, Forschungsdatenzentrum der Bundesagentur für Arbeit im Institut für Arbeitsmarkt- und Berufsforschung, Nuremberg.

Prigge, M., Köhr, M., Pfeiffer, N., Blettner, M., Beutel, M., Wild, P., Münzel, T., Blankenberg, S., Seidler, A., Letzel, S., Latza, U. and Liebers, F. (2014). Codierung der Tätigkeitsangaben im Basiskollektiv der Gutenberg-Gesundheitsstudie under Anwendung der Klassifikation der Berufe KldB 2010 - Darstellung des Vorgehens und der Datenqualität, *Zeitschrift für Arbeitswissenschaft* **68**(3): 153–162.
URL: *https://link.springer.com/article/10.1007/BF03374441*

References IV

Schierholz, M., Gensicke, M., Tschersich, N. and Kreuter, F. (2018). Occupation coding during the interview, *Journal of the Royal Statistical Society: Series A* **181**(2): 379–407.

United Nations Statistical Commission and Economic Commission for Europe (ed.) (1997). *Statistical Data Editing Volume No. 2*, United Nations, New York, chapter 6.

Wenzowski, M. J. (1988). Actr - a generalized automated coding system, *Survey Methodology* **14**(3): 299–307.
URL: *https://www150.statcan.gc.ca/n1/pub/12-001-x/1988002/article/14586-eng.pdf*