The Journal of Open Source Software

occupationMeasurement: A Comprehensive Toolbox for Interactive Occupation Coding in Surveys

Jan Simson [1¶], **Olga Kononykhina** [1], **and Malte Schierholz** [1]

1 Department of Statistics, Ludwig-Maximilians-Universität München, Germany ¶ Corresponding author

DOI: 10.21105/joss.05505

Software

- Review
- Repository
- Archive

Editor: Chris Vernon

Reviewers:

- @welch16
- @danielruss

Submitted: 30 March 2023
Published: 23 August 2023

Summary

People earn a living a multitude of ways which is why the occupations they pursue are almost as diverse as people themselves. This makes quantitative analyses of free-text occupational responses from surveys hard to impossible, especially since people may refer to the same occupations with different terms. To address this problem, a variety of different classifications have been developed, such as the International Standard Classification of Occupations 2008 (ISCO) (ILO, 2012) and the German Klassifikation der Berufe 2010 (KldB) (Bundesagentur für Arbeit, 2011), narrowing down the amount of occupation categories into more manageable numbers in the mid hundreds to low thousands and introducing a hierarchical ordering of categories. This leads to a different problem, however: Coding occupations into these standardized categories is usually expensive, time-intensive and plagued by issues of reliability.

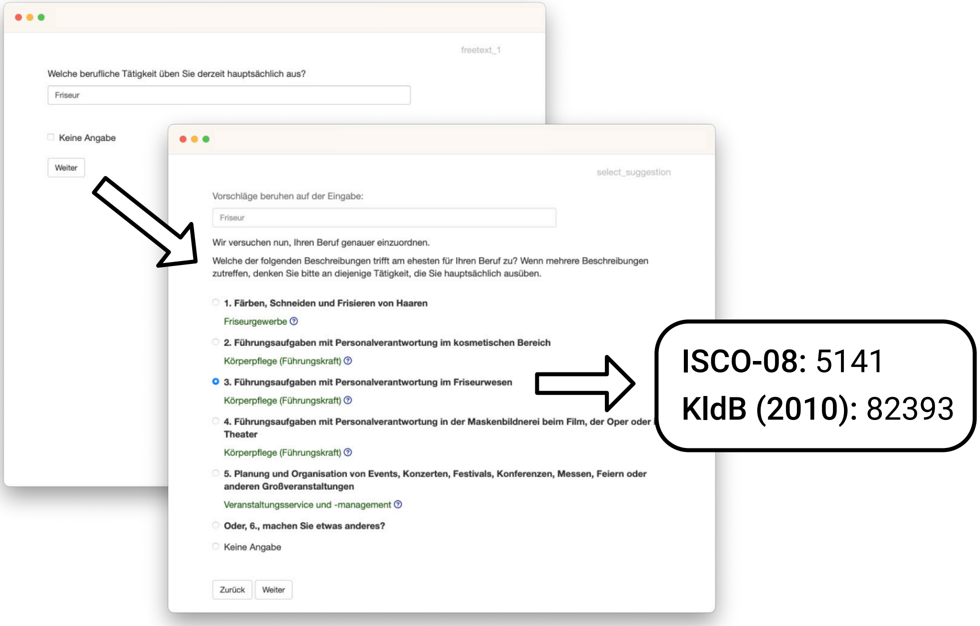
Here we present a new instrument that implements a faster, more convenient and interactive occupation coding workflow where respondents are included in the coding process. Based on the respondent's answer, a novel machine learning algorithm generates a list of suggested occupational categories from the Auxiliary Classification of Occupations (Schierholz, 2018), from which one is chosen by the respondent (see Figure 1). Issues of ambiguity within occupational categories are addressed through clarifying follow-up questions. We provide a comprehensive toolbox including anonymized German training data and pre-trained models without raising privacy issues, something not possible yet with other algorithms due to the difficulties of anonymizing free-text data.

Statement of Need

Assigning occupations to standardized codes is a critical task frequently encountered in research, public administration and beyond: They are used in government censuses (e.g. USA, UK, Germany) and administrative data to better understand economic activity, in epidemiology to estimate exposure to health hazards, and in sociology to obtain a person's socio-economic status, to name a few examples.

To date, the standard approach to coding occupations is to collect one or two free-text responses and later hand-coding these descriptions by trained personnel with a classification manual, possibly assisted by computer software. Since coding typically occurs after data collection, based on the responses only and without the ability to request clarifying information from the respondent, the assignment of categories is often inaccurate. This approach to occupational coding is costly due to the experts' time needed and often suffers from low inter-coder reliability[1].

[1] An international review found rates of agreement between 44% and 89% when different coders code the same answers across different studies (Mannetje & Kromhout, 2003). For a more in-depth discussion of the factors affecting the reliability of occupation coding, see Conrad et al. (2016) and Massing et al. (2019).



freetext_1

Welche berufliche Tätigkeit üben Sie derzeit hauptsächlich aus?

Friseur

□ Keine Angabe

Weiter

select_suggestion

Vorschläge beruhen auf der Eingabe:

Friseur

Wir versuchen nun, Ihren Beruf genauer einzuordnen.

Welche der folgenden Beschreibungen trifft am ehesten für Ihren Beruf zu? Wenn mehrere Beschreibungen zutreffen, denken Sie bitte an diejenige Tätigkeit, die Sie hauptsächlich ausüben.

○ **1. Färben, Schneiden und Frisieren von Haaren**
Friseurgewerbe ⓘ

○ **2. Führungsaufgaben mit Personalverantwortung im kosmetischen Bereich**
Körperpflege (Führungskraft) ⓘ

● **3. Führungsaufgaben mit Personalverantwortung im Friseurwesen**
Körperpflege (Führungskraft) ⓘ

○ **4. Führungsaufgaben mit Personalverantwortung in der Maskenbildnerei beim Film, der Oper oder Theater**
Körperpflege (Führungskraft) ⓘ

○ **5. Planung und Organisation von Events, Konzerten, Festivals, Konferenzen, Messen, Feiern oder anderen Großveranstaltungen**
Veranstaltungsservice und -management ⓘ

○ **Oder, 6., machen Sie etwas anderes?**

○ Keine Angabe

Zurück Weiter

ISCO-08: 5141
KldB (2010): 82393

Figure 1: Typical flow of the interactive application. 1. A respondent provides a free text response describing her occupation. 2. The machine learning model then generates a list of suggested categories, from which the respondent will select one. 3. As a result, the associated occupational category codes from both the German KldB-2010 and the international ISCO-08 are returned.

Given the limitations of manual coding, a technical solution that can generate a suggested code fast enough to elicit immediate feedback from respondents would be a boon. However, implementations of this idea are few, and none are openly available. Technological solutions using machine learning have been proposed (Creecy et al., 1992; Gweon et al., 2017; Russ et al., 2014, 2016; Schierholz & Schonlau, 2021) but face problems obtaining training data and sharing trained models due to privacy issues, as training data often contains sensitive free-text responses that may personally identify respondents.

Our toolbox addresses these issues by implementing an occupation coding workflow during the interview as discussed in Peycheva et al. (2021) and Schierholz et al. (2018). Difficulties of data and model sharing are resolved by using a novel machine learning algorithm specifically crafted to work with anonymized occupational data.

Functionalities

We provide an open-source implementation of a machine learning algorithm for occupation coding with immediate feedback and verification, available as an R (R Core Team, 2019) package on CRAN[2]. An introductory "Getting Started" guide is available for anyone looking to use the package. To make it widely useful, our toolkit can be readily integrated in both self-administered web surveys as well as interviewer-administered (telephone) surveys using the included questionnaires. Programmers can adapt these questionnaires to fit a wide array of requirements. The toolbox includes custom survey software built on top of the shiny (Chang et al., 2023) framework to integrate machine learning predictions into surveys. On the off-chance that further flexibility is needed, we offer direct API access for completely custom

[2] Package "occupationMeasurement" on CRAN: https://cran.r-project.org/web/packages/occupationMeasurement/index.html.

data collection and integration into existing survey software. As we built the toolbox on top of the shiny (Chang et al., 2023) and plumber (Schloerke et al., 2022) frameworks, deployments on the Web are easy. We further provide pre-built container images for even easier deployment in production environments.

The toolbox uses a specifically developed list of occupational task descriptions, the Auxiliary Classification of Occupations, designed to be easier to understand and less ambiguous than existing lists of job titles (Schierholz, 2018). Alongside this list, it provides matching follow-up questions to enable a fine-grained assignment into existing classification systems.

The machine-learning algorithm used in our instrument is able to work with anonymized training data while retaining predictive performance on-par with other machine learning and non-machine-learning algorithms (Schierholz, 2019; Schierholz & Schonlau, 2021). This enables sharing training data as well as trained models, allowing on the one hand out-of-the-box usage of our instrument without the need for labeled data or pre-training, but also the further sharing of newly trained models by users of the toolbox. Anonymized training data and pre-trained models in German are included with the package.

The toolbox is released under the MIT license alongside extensive online documentation. Quality of the software is ensured using automated testing via continuous integration. The toolbox has successfully been piloted with various modes of data collection in collaboration with different German survey institutes.

Related Work

This project is not the first to apply technology to assist in the coding of occupations, although it is the first tool to be released as open-source software and to offer this level of convenience and flexibility. Notable examples of prior work include the WISCO[3] database (Tijdens, 2010), providing a search tree with levels of nested categories of occupations for use in Web surveys. Another prominent tool is CASCOT (Elias et al., 2014), short for Computer Assisted Structured Coding Tool. CASCOT uses a mixture of a coding index, which requires manual editing, text distances and manually specified rules to code responses into occupational categories. A promising tool has also been developed for the US Standard Occupational Classification System (SOC) (U. S. Bureau of Labor Statistics, n.d.), called SOCcer (Russ et al., 2014, 2016). Similar to the software presented here, SOCcer relies on using a machine learning algorithm. Unfortunately, neither SOCcer nor CASCOT are open-sourced, with the former offering coding via a free online version[4] and the latter requiring payment for use.

Acknowledgements

This project is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project numbers 290773872 and 460037581.

References

Bundesagentur für Arbeit. (2011). *Klassifikation der Berufe 2010: Vols. 1 & 2*. Bundesagentur für Arbeit.

Chang, W., Cheng, J., Allaire, J., Sievert, C., Schloerke, B., Xie, Y., Allen, J., McPherson, J., Dipert, A., & Borges, B. (2023). *Shiny: Web application framework for r*. https://shiny.rstudio.com/

[3] The WISCO database used different names over time, but kept the same acronym. The latest description is: "World database of occupations, coded ISCO". The database is available at https://surveycodings.org.

[4] Online version available at: https://soccer.nci.nih.gov/soccer/.

Conrad, F. G., Couper, M. P., & Sakshaug, J. W. (2016). Classifying Open-Ended Reports: Factors Affecting the Reliability of Occupation Codes. *Journal of Official Statistics*, *32*(1), 75–92. https://doi.org/10.1515/jos-2016-0003

Creecy, R. H., Masand, B. M., Smith, S. J., & Waltz, D. L. (1992). Trading MIPS and memory for knowledge engineering. *Communications of the ACM*, *35*(8), 48–64. https://doi.org/10.1145/135226.135228

Elias, P., Birch, M., & Ellison, R. (2014). CASCOT international version 5 user guide. *Institute for Employment Research, University of Warwick, Coventry*. https://warwick.ac.uk/fac/soc/ier/software/cascot/internat/cascot_international_user_guide.pptx

Gweon, H., Schonlau, M., Kaczmirek, L., Blohm, M., & Steiner, S. (2017). Three Methods for Occupation Coding Based on Statistical Learning. *Journal of Official Statistics*, *33*(1), 101–122. https://doi.org/10.1515/jos-2017-0006

ILO. (2012). *International Standard Classification of Occupations 2008 (ISCO-08): Structure, group definitions and correspondence tables*. http://www.ilo.org/global/publications/ilo-bookstore/order-online/books/WCMS_172572/lang--en/index.htm

Mannetje, A. 't., & Kromhout, H. (2003). The use of occupation and industry classifications in general population studies. *International Journal of Epidemiology*, *32*(3), 419–428. https://doi.org/10.1093/ije/dyg080

Massing, N., Wasmer, M., Wolf, C., & Zuell, C. (2019). How Standardized is Occupational Coding? A Comparison of Results from Different Coding Agencies in Germany. *Journal of Official Statistics*, *35*(1), 167–187. https://doi.org/10.2478/jos-2019-0008

Peycheva, D. N., Sakshaug, J. W., & Calderwood, L. (2021). Occupation Coding During the Interview in a Web-First Sequential Mixed-Mode Survey. *Journal of Official Statistics*, *37*(4), 981–1007. https://doi.org/10.2478/jos-2021-0042

R Core Team. (2019). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Russ, D. E., Ho, K.-Y., Colt, J. S., Armenti, K. R., Baris, D., Chow, W.-H., Davis, F., Johnson, A., Purdue, M. P., Karagas, M. R., Schwartz, K., Schwenn, M., Silverman, D. T., Johnson, C. A., & Friesen, M. C. (2016). Computer-based coding of free-text job descriptions to efficiently identify occupations in epidemiological studies. *Occupational and Environmental Medicine*, *73*(6), 417–424. https://doi.org/10.1136/oemed-2015-103152

Russ, D. E., Ho, K.-Y., Johnson, C. A., & Friesen, M. C. (2014). *2014 IEEE 27th International Symposium on Computer-Based Medical Systems (CBMS)*. 347–350. https://doi.org/10.1109/CBMS.2014.79

Schierholz, M. (2018). Eine Hilfsklassifikation mit Tätigkeitsbeschreibungen für Zwecke der Berufskodierung. *AStA Wirtschafts- und Sozialstatistisches Archiv*, *12*(3-4), 285–298. https://doi.org/10.1007/s11943-018-0231-2

Schierholz, M. (2019). *New Methods for Job and Occupation Classification* [PhD thesis].

Schierholz, M., Gensicke, M., Tschersich, N., & Kreuter, F. (2018). Occupation coding during the interview. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *181*(2), 379–407. https://doi.org/10.1111/rssa.12297

Schierholz, M., & Schonlau, M. (2021). Machine learning for occupation coding—a comparison study. *Journal of Survey Statistics and Methodology*, *9*(5), 1013–1034. https://doi.org/10.1093/jssam/smaa023

Schloerke, B., Allen, J., Tremblay, B., Dunné, F. van, Vandewoude, S., & RStudio. (2022). *Plumber: An API generator for r*. https://CRAN.R-project.org/package=plumber

Tijdens, K. (2010). *Measuring occupations in web-surveys: the WISCO database of occupations.* https://dare.uva.nl/search?identifier=2a72aad5-24dc-4891-9bc4-05deddad520b

U. S. Bureau of Labor Statistics. (n.d.). *Standard Occupational Classification System (SOC) System*. https://www.bls.gov/soc/