Descriptor

Patterns

Occupational models from 42 million unstructured job postings

Highlights

- Job titles follow conventions and can be normalized to reduce variation
- We curated a list of 775 skills and estimated their probability by occupation
- Skill probabilities are consistent with the hierarchy of occupational codes
- Job posting counts are an imperfect proxy for job openings reported in survey data

Authors

Nile Dixon, Marcelle Goggins, Ethan Ho, ..., Emma Northcott, Karen Shen, Carrie Yeats

Correspondence

connect@ripl.org

In brief

We present an open data resource and software tool for understanding the associations between occupations, job titles, and skills in the United States labor market. These associations are used in several fields of research, including economics and public health, as well as for practical applications like matching job seekers to available jobs.

Dixon et al., 2023, Patterns *4*, 100757
July 14, 2023
https://doi.org/10.1016/j.patter.2023.100757

CellPress

Patterns

CellPress
OPEN ACCESS

Descriptor

Occupational models from 42 million unstructured job postings

Nile Dixon,[1] Marcelle Goggins,[1] Ethan Ho,[1] Mark Howison,[1,5,*] Joe Long,[1] Emma Northcott,[2,3] Karen Shen,[1,4] and Carrie Yeats[2]

[1]Research Improving People's Lives, 1 Park Row, Suite 401, Providence, RI 02903, USA
[2]National Association of State Workforce Agencies, 444 N. Capitol Street NW, Suite 300, Washington, DC 20001, USA
[3]George Washington University, Trachtenberg School of Public Policy and Public Administration, 805 21st Street NW, Washington, DC 20052, USA
[4]Department of Health Policy and Management, Bloomberg School of Public Health, Johns Hopkins University, 615 N. Wolfe Street, Baltimore, MD 21205, USA
[5]Lead contact
*Correspondence: connect@ripl.org
https://doi.org/10.1016/j.patter.2023.100757

THE BIGGER PICTURE Online job postings offer an abundant and detailed view of occupations and skills in the labor market. However, the variation in how employers refer to and describe job openings makes it difficult to use unstructured job postings for analysis and research. Occupations have long been standardized in the United States into a hierarchy of 867 codes through the Standard Occupational Classification system. By categorizing a sample of 42 M United States job postings into these standardized occupational codes and extracting the skills in each posting, we constructed an open dataset with empirical probabilities for associations among occupational codes, job titles, and skills. We bundled these data in a software tool, called *sockit*, that can analyze new job titles, job descriptions, or resumes.

Proof-of-Concept: Data science output has been formulated, implemented, and tested for one domain/problem

SUMMARY

Structuring jobs into occupations is the first step for analysis tasks in many fields of research, including economics and public health, as well as for practical applications like matching job seekers to available jobs. We present a data resource, derived with natural language processing techniques from over 42 million unstructured job postings in the National Labor Exchange, that empirically models the associations between occupation codes (estimated initially by the Standardized Occupation Coding for Computer-assisted Epidemiological Research method), skill keywords, job titles, and full-text job descriptions in the United States during the years 2019 and 2021. We model the probability that a job title is associated with an occupation code and that a job description is associated with skill keywords and occupation codes. Our models are openly available in the *sockit* python package, which can assign occupation codes to job titles, parse skills from and assign occupation codes to job postings and resumes, and estimate occupational similarity among job postings, resumes, and occupation codes.

INTRODUCTION

Structured occupational codes have been in use in the United States since 1977 with the release of the Standard Occupational Classification (SOC) system,[1] which is now the federal statistical standard for defining occupations.[2] Official statistics on workforce participation from the United States Bureau of Labor Statistics, the United States Census Bureau, and other federal agencies are structured in terms of these codes, of which there are 867 at the most detailed level in the 2018 version. However, the titles and descriptions that workers and employers use for particular jobs vary widely. Likewise, the functional descriptions and skill keywords associated with particular jobs vary, even though there are commonalities in the skills required among jobs within the same occupation or across similar occupations.

Occupational codes are central to many research studies. For example, recent studies of the labor market's response to the COVID-19 pandemic examined the dynamics of supply and demand shocks by occupation[3] and the feasibility of remote work by occupation.[4] Similarly, studies of occupational hazards in the public health literature often use SOC codes to proxy for exposure to hazards, for example in studying the differential risks to healthcare workers during the pandemic.[5]

Assigning SOC codes by hand is time consuming and does not scale to large datasets or to real-time applications. Several tools for automatically assigning SOC codes to job titles are available[6] but are limited by their model transparency and software accessibility. The National Institute for Occupational Safety and Health developed the NIOCCS system based on hand-coded SOC assignments to survey data,[7] but access to the system currently requires account registration and approval.[8] Similarly, the National Cancer Institute created a tool called Standardized Occupation Coding for Computer-assisted Epidemiological Research (SOCcer) by modeling expert-coded job titles,[9] but it is only accessible through a web interface, and results are retrieved later by e-mail.[10] The United States Department of Labor provides another web-based tool, the O*NET Code Connector.[11] Another web-based tool, Occupational Self-Coding and Automatic Recording, requires self-reporting by research participants.[12] There are also commercially licensed options, including the Lightcast Titles API[13] and the O*NET-SOC AutoEncoder.[14] Existing approaches either do not provide the parameters underlying their models, cannot run offline (e.g., to efficiently process large amounts of job title data), or do not adhere to FAIR principles for research software.[15]

We present a reusable data resource and software toolkit that models the occupational structure in unstructured job titles and job descriptions derived from a comprehensive sample of online job postings. There are over 3 million distinct job titles in the approximately 42 million job postings underlying our models. In contrast to existing methods, our model parameters are openly available and reproducible. Our models are pre-packaged in the downloadable *sockit* python package,[16] as well as in a hosted web application,[17] for convenient reuse with minimal dependencies.

Beyond their applications in scientific research, occupation codes also have important practical uses for policy makers and in real-world applications. The use case that motivated this data resource was a practical application to extract structured occupational information from available unstructured data. Specifically, *sockit* was implemented in a recommendation system that helps job seekers discover new careers, recently deployed by labor departments in Rhode Island, Hawai'i, New Jersey, Colorado, and Maryland of the United States.[18] The entry point for job seekers to these applications is a resume upload or manual entry of previous job titles, which are unstructured data. The algorithm for recommending careers, however, requires structured SOC codes and skill keywords that are estimated from the unstructured input using the methods described in this article. With the increasing volume of unstructured job and resume data available online, automatic processing with methods like *sockit* will be increasingly important for a broader range of both research and policy applications.

RESULTS

Our primary data come from the NLx Research Hub,[19] a real-time and historical data warehouse representing the diversity of jobs available in the United States labor market, which is accessible at no cost for approved research projects. Job postings in the Research Hub originate in the National Labor Exchange,[20] which is a partnership between the National Association of State Workforce Agencies[21] and the DirectEmployers Association[22] to collect and distribute online job postings from corporate career websites, state job banks, and the United States federal jobs portal.[23] At the time of writing, the National Labor Exchange advertises that they collect job postings for 300,000 employers with a daily volume of 3.7 million active (both new and existing) job postings.[20]

We accessed 42,298,617 historical records in the NLx Research Hub for the years 2019 (13,241,134 records) and 2021 (29,057,483 records). We chose these two years because they represent the United States labor market before and after the COVID-19 economic crisis but not during the beginning of the COVID economic crisis itself in 2020 Q2. Each record contains unstructured fields for job title, a full-text job description, and structured fields for acquisition date, city, state, and postal code.

We make use of prefix trees (also known as tries) throughout the data processing pipeline. Briefly, a prefix tree is a data structure that allows efficient lookups of strings and is frequently used to solve string searching, spellchecking, and autocompletion tasks. We developed the open-source *wordtrie* python package[24] specifically to implement substring matching in *sockit* and the data processing pipeline described below, but we released it as its own package due to its generality.

Research Hub job postings contain 849,284 distinct job titles after normalization

In practice, job titles are often written as a series of adjectives that add specificity to a principal noun. For example, a "licensed practical nurse" is a specific type of nurse, and a "pizza delivery driver" is a specific type of driver. In these cases, nurse and driver are the principal nouns that encode the most general meaning of the job. Common exceptions to this adjective-noun ordering are supervisory job titles, such as "director of nursing" or "supervisor of delivery drivers," and assistant job titles, such as "special assistant to the vice president." However, we can normalize those types of titles to adjective-noun ordering by pivoting them around the prepositions "of," "for," or "to" so that, for example, "director of nursing" becomes "nursing director."

Based on these insights, we started by identifying suitable principal nouns from existing datasets with job titles. We applied a natural language processing technique called part-of-speech tagging to identify nouns in all of the sample job titles available in the O*NET 27.0 Database,[25] including the 1,016 titles in the "Occupation Data" file and the 52,742 titles in the "Alternative Titles" file, as well as the 6,520 titles in 2018 SOC Direct Match Title File from the United States Bureau of Labor Statistics.[26] We manually reviewed all words identified as nouns and curated them into a list of 2,514 principal nouns. At the same time, we curated a list of 259 unambiguous acronyms by extracting and reviewing all acronyms occurring in parentheses in the job titles,

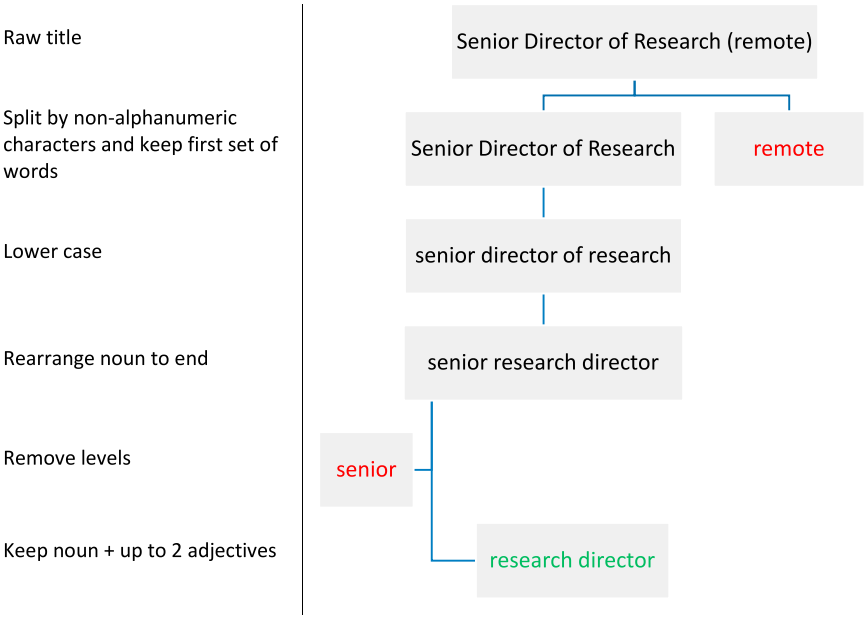

Figure 1. Job title cleaning process in *sockit.title.clean*, illustrated with an example title of "Senior Director of Research (remote)"

e.g., "RN" in "Registered Nurse (RN)," and retaining only the acronyms that mapped to a distinct SOC code in the files above.

Next, we extracted 3,179,805 distinct job titles occurring in the 42,298,617 records from the Research Hub after converting job titles to lowercase, removing extraneous text, and retaining alphabetical characters (implemented in the *sockit.title.clean* method). We further processed these titles to filter employer names using a prefix tree of 999 members of DirectEmployers,[27] United States place names using a prefix tree of state names and abbreviations as well as 330 large cities,[28] and a smaller set of 26 phrases and abbreviations that denote work schedule (e.g. "part time" or "evenings") and often occur in job posting titles. Of the 3,179,805 distinct job titles, 578,745 titles (representing 3,828,432 job postings) had one or more of these employer names, place names, or scheduling terms filtered out, and 433,764 titles (representing 3,138,421 job postings) were normalized to adjective-noun ordering by pivoting around a preposition. We retained 2,605,739 titles (representing 36,951,252 job postings) containing at least one of the 2,514 principal nouns.

Finally, we truncated 944,562 titles (representing 5,999,760 job postings) containing more than three words to retain only the principal noun and up to two preceding adjectives. This process is visualized in Figure 1, with an example title of "Senior Director of Research (remote)." Truncating the number of words represented in each title helps control the long tail of singleton titles corresponding to a single job posting. Table 1 shows how varying the threshold on the number of words affects the counts of distinct titles and singleton titles. While there is no optimal threshold given that every increase in the threshold also increases the number of unique titles and the proportion of singleton titles, moving from a threshold of two to three words results in the largest marginal return in terms of increasing the number of distinct titles and the proportion of non-singleton titles.

Our approach yielded a final list of 849,284 distinct job titles (representing 36,951,252 job postings) in normalized adjective-noun ordering with between one and three words, where the last word is a principal noun. We submitted these titles to the SOCcer web application[10] to obtain probabilities for the 10 most likely SOC 2010 codes associated with each distinct job title. We converted the SOC 2010 codes to SOC 2018 codes using a crosswalk provided by the United States Bureau of Labor Statistics.[29]

We constructed a job title prefix tree by inserting the distinct job titles with their counts weighted by their SOCcer probabilities of each SOC code, excluding probabilities below 0.02. This threshold is meant to control for false positives and reduce the number of SOC codes assigned to each job title. While there is no way to determine an optimal probability threshold since there is no ground truth available, the threshold value of 0.02 controls the number of jobs that would be assigned multiple SOC codes that differ at the 2-digit level (Table 2), which arguably should be applicable to only a small proportion of jobs. With no threshold, approximately two-thirds of job titles would be assigned a second SOC code that differs at the 2-digit level. In contrast, a threshold of 0.10 would have less than 1% of jobs with a second 2-digit SOC code but would eliminate almost three-quarters of job titles. The threshold of 0.02 allows for 5% of jobs to be assigned a second 2-digit SOC code while retaining approximately half of the job titles.

Because the titles are normalized to end with the principal noun, we inserted the titles into the prefix tree in reverse word order (e.g., from right to left). Therefore, the more general principal nouns occur at the top of the tree, and the leaf nodes are the more specific titles that include adjectives. Each leaf node in the tree contains a histogram of SOC code counts, which we aggregated across parent nodes so that we can assign SOC probabilities to partial title matches, all the way down to the root nodes that contain a single principal noun. Figure 2 illustrates the structure of the job title prefix tree with examples of job title families for nurses and drivers. The *sockit* package includes a "title" module that can search for titles within a longer query string in reverse word order so that all matches start from a principal noun.

In practice, we found that management titles containing the principal nouns "manager," "director," "supervisor," "vp," and "president" were difficult to classify correctly with the job title prefix tree because of variation in their adjectives and modifiers. Therefore, we built a separate management title prefix tree that maps 6,150 such titles to the one or two most relevant SOC codes using search results from the O*NET Code Connector.[11]

Term-weighted job descriptions estimate skill probabilities by occupation

To study skills in job descriptions, we began by manually sampling 1,075 skill keywords from CareerOneStop,[30] online worker

Table 1. Counts of distinct and singleton titles after truncating job titles at varying thresholds for the number of words preceding and including the principal noun

Word threshold	2	3	4	5	6
Distinct titles	196,430	849,284	1,236,604	1,313,177	1,324,580
Singleton titles	65,903	366,472	599,222	656,490	666,187

Table 2. Counts of distinct job titles with at least one assigned SOC code from SOCcer after removing SOC codes below the specific probability threshold

Probability threshold	0.00	0.01	0.02	0.05	0.10
Distinct titles	849,284	490,250	402,497	309,864	249,383
… with SOC codes differing at 2-digit level	553,281	46,523	18,887	5,564	1,849

At low thresholds, many titles have multiple SOC codes differing at the 2-digit level.

profiles, and the O*NET 27.0 Database[25] under the "Abilities," "Interests," "Knowledge," "Skills," "Technology Skills," "Tools Used," "Work Activities," "Work Context," "Work Values," and "Work Styles" files. Six reviewers manually edited these keywords and suggested groupings of similar keywords. One of the reviewers used the others' edits and groupings to curate a final list of 755 keywords, and we constructed a skills prefix tree to map the original 1,075 keywords plus 254 alternative forms (e.g., plural vs. singular) to the curated 755 skill keywords.

Next, we counted the occurrence of skill keywords in the 42,298,617 records from the NLx Research Hub using the skills prefix tree and estimated SOC probabilities for their corresponding job titles using the job titles prefix tree. We found that the records contained 26,953,261 distinct job descriptions (see Table S1), and 24,009,146 of those (89.1%) contained at least one skill keyword and had at least one SOC code with ≥ 0.1 probability in their title. We represented these associations as a sparse "job-skill" matrix with the dimensions 24,009,146 × 755 and a sparse (transposed) "SOC-job" matrix with the dimensions 867 × 24,009,146.

Because the skill keywords vary from general to specific and technical, we applied a natural language processing technique called Term Frequency-Inverse Document Frequency (TF-IDF)[31] to reweight the occurrences of skill keywords in the job-skill matrix to better approximate the relevance of individual skill keywords for determining occupation.[32] We calculated the matrix product of the SOC-job matrix and the TF-IDF-weighted job-skill matrix to produce a dense SOC-skill matrix with dimensions of 867 × 775. We normalized the rows of the SOC-skill matrix, which can be interpreted as probability distributions over skills for each SOC code. Figure S1 visualizes the structure of this matrix.

Cosine distance between skill probabilities captures occupational similarity in the SOC code hierarchy

We estimated occupational similarity by computing pairwise distances between vectors of skill probabilities and using the inverse of the distance as a similarity measure. To compare occupations, we computed distances between all pairs of SOC code rows in the SOC-skill matrix to produce a "SOC-SOC" similarity matrix. We tested four distance measures for this matrix: the Euclidean (L^2) metric, the Manhattan (L^1) metric, the cosine metric, and the Kullback-Leibler divergence.[33] We found that the cosine metric best captured the block-diagonal structure of occupations at the 2-digit SOC code level (Figure S2). To quantitatively assess this, we grouped SOC code pairs by whether they share the same first 2 digits and calculated the ratio of the mean similarity score within these two groupings. We found that the highest ratio was for cosine similarity (2.083), followed by Manhattan (1.467), Kullback-Leibler (1.113), and Euclidean (1.099). Therefore, cosine distance, on average, assigns higher similarity between SOC codes within the same 2-digit SOC code level.

Skill distributions occur in many sources, including job descriptions and resumes. Therefore, we extended this method to be able to count skill keywords in arbitrary documents, apply the same TF-IDF transformation from the SOC-skill matrix above, and compare similarity between two documents or between a document and all SOC code rows in the SOC-skill matrix. These functions are provided in the "parse" and "compare" modules (Figure 3) of the *sockit* package.

Research Hub job postings sample approximately 12% of United States job openings

We filtered the NLx Research Hub job postings using their acquisition date by removing jobs that were exact duplicates on job description content within an acquisition month. We aggregated the probability-weighted SOC counts at the month, year, and United States state level. These counts are a proxy for job openings, and we compared the counts both nationally and for the largest five states in the United States against official job opening estimates from the United States Bureau of Labor Statistics' Job Openings and Labor Turnover Survey[34] (JOLTS).

We found that, on average, across all months in 2019 and 2021, there were 8.3 times as many job openings reported in JOLTS as job postings in the NLx Research Hub, suggesting that it represents a 12% sample of job openings in the United States. We scaled the NLx Research Hub counts by a factor of 8.3 and compared it at the month level with the JOLTS estimates and found that these are closely related (Figure 4) and likely reflect the job market recovery to pre-pandemic levels.[35] However, the same comparison for the five largest states showed that California is under-represented, especially in the year 2019 (which is consistent with a known technical issue regarding California's data in the NLx Research Hub), and that New York is consistently over-represented. Therefore, our job posting data appear to be representative of job openings at the national level but not at the level of individual states in the United States.

We also found that occupational representation in job postings differs from actual United States employment by comparing the proportion of job postings at the 2-digit SOC code level with estimates of employment levels in the United States in 2019 and 2021. The employment estimates at the 2-digit SOC code level come from the United States Census Bureau's American Community Survey,[36] including estimates of all employed workers and of non-seasonal full-time workers, and from the United States Bureau of Labor Statistics' Occupational Employment and Wage

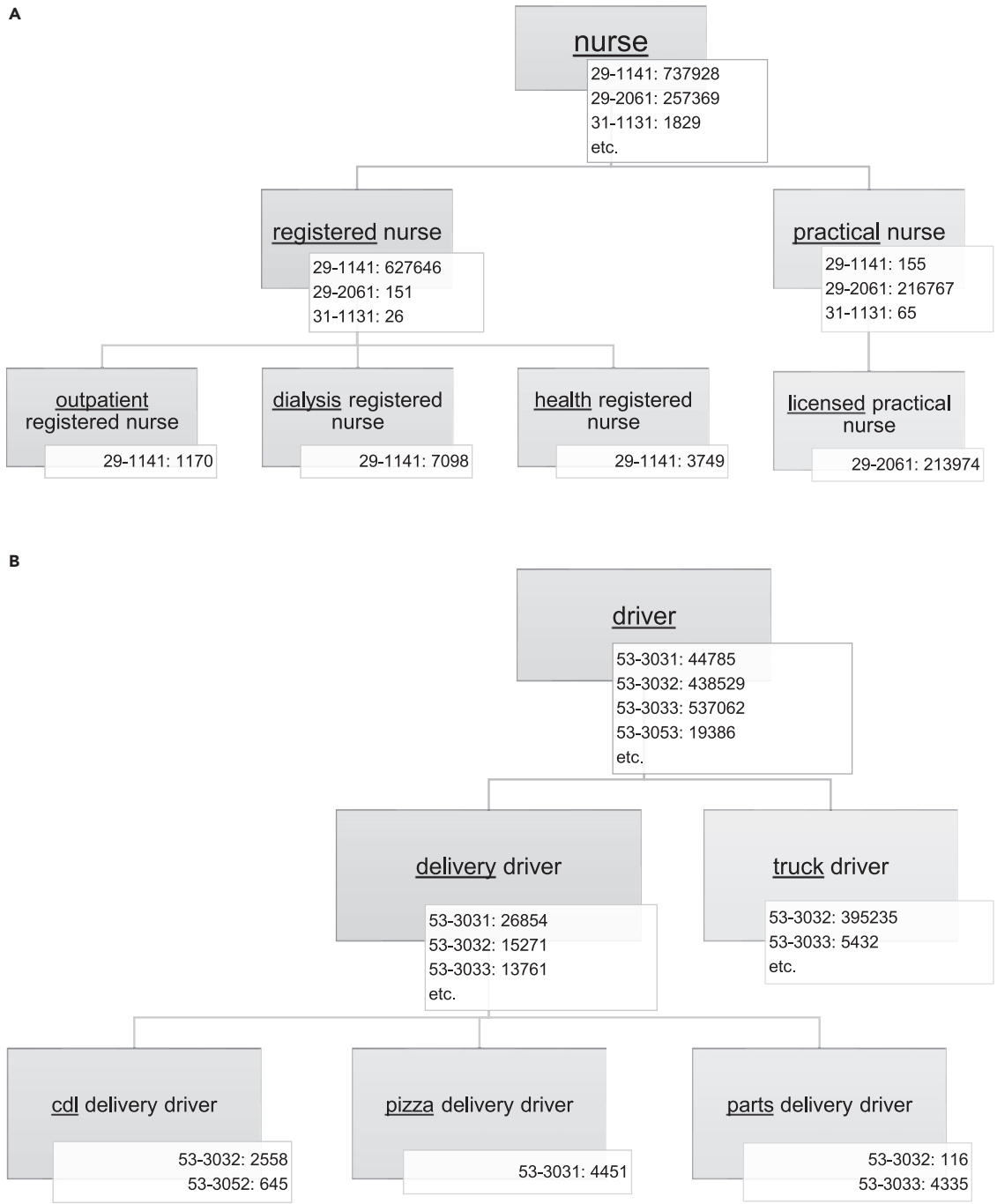

Figure 2. The prefix tree data structure used for substring matching of job titles to SOC code frequencies
Illustrated with job title families for nurses (A) and drivers (B).

Statistics.[37] This comparison examines which occupations in our data are over- or under-represented due to a combination of actual demand in the labor market and potential occupational biases in our job postings. We found that NLx Research Hub job postings are over-represented in computer and healthcare occupations and under-represented in legal, food service, farming, and construction occupations relative to actual employment levels (Figure 5).

Accuracy of matching job titles to SOC codes varies by occupation

We tested the accuracy of estimating SOC codes from job titles and job postings using the title and parse modules in *sockit*. The title module estimates the most probable SOC codes for a job title using the job title prefix tree. The parse module estimates the most similar SOC codes for a job posting from the cosine

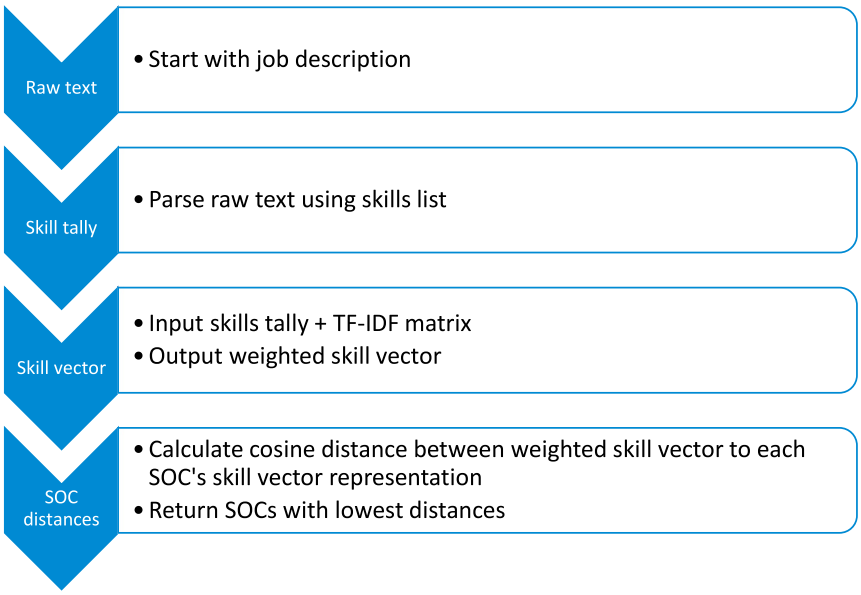

Figure 3. Job description parsing and SOC comparison process in *sockit.parse* and *sockit.compare*

similarity between a TF-IDF-weighted skill keyword vector parsed from the job posting vs. each row in the SOC-skill matrix.

To establish a ground truth for our tests, we used the O*NET 27.0 Database.[25] We tested the 7,541 titles in the "Sample of Reported Titles" file against their corresponding SOC codes. We tested synthetic job postings that we constructed for 818 SOC codes by concatenating all their entries in the "Task Statements" and "Detailed Work Activities" files, which approximate the language used to describe qualifications in a job posting for these occupations.

We defined accuracy as the fraction of cases where the correct SOC code was contained in the three most probable codes (for titles) or in the three most similar codes (for postings) returned by *sockit*. Overall, titles matched at the 6-digit level with 56.7% accuracy and at the 2-digit level with 81.7% accuracy, while postings matched at the 6-digit level with 27.8% accuracy and at the 2-digit level with 78.9%. Accuracy varied by SOC code levels and by the occupational group of the test SOC codes (Figure 6).

DISCUSSION

Job postings contain a rich source of information on the associations between job titles, skill keywords, and occupational codes. In our sample of job postings from the NLx Research Hub, these empirical associations accurately recovered 2-digit SOC codes from job titles, although matching at the 6-digit SOC code level was less accurate for most occupations. This is consistent with previous findings.[6]

Variation in accuracy may be due to varying effectiveness of our title cleaning methods for certain occupations and specialized job titles. Overall accuracy might be improved by supplementing our title cleaning methods and job title prefix tree with information from the job descriptions. Variation may also be due to sampling bias in either the SOCcer model used in our initial estimates or in the reported titles and task statements in the O*NET survey data used for testing. In this case, collecting additional labeled training and testing data would help, for example using an active learning strategy that targets label acquisition according to which occupational groups have lower accuracy.[38]

The distribution of job postings by state in our sample is biased relative to official statistics on job openings by state. Further corrections or supplemental data may be required for job posting frequencies to serve as accurate proxies for actual job openings at the state or city level. However, the overall frequency of job postings in our sample is consistent with a 12% month-to-month sampling rate among national job openings.

Our sample could be biased in terms of occupational representation, although this is more difficult to test. Our comparison of job postings with actual employment levels by occupation is not ideal since it conflates sampling bias with actual demand in the labor market. A preferable comparison would have been between job postings and job openings at the 2-digit SOC level, but JOLTS estimates of job openings are not available at this level. We expect more job postings relative to actual employment levels for occupations that are in high demand, for example in healthcare, where we observe roughly twice as many job postings as currently employed workers (Figure 5). This over-representation in healthcare job postings is greater post-pandemic and could be driven by increased demand for healthcare workers following turnover during the pandemic. Legal, construction, and farming job postings have similar under-representation pre- and post-pandemic, which could be driven by preferences in those industries to post jobs offline or on specialized sites.

A limitation of our use of keyword analysis is that 10.9% of the 26,953,261 distinct job descriptions in our data are dropped because they are concise and list few skills or qualifications. In future work, an alternative approach might examine all occurring unigrams, bigrams, or trigrams that are putative skills and cluster them into a skills taxonomy with topic modeling. This approach might be able to retain all job postings in our data but would also introduce greater model complexity and potential noise from ambiguous job postings that are currently dropped in our analysis. Alternatively, additional keyword analysis could capture educational, licensing, and certification requirements that are sometimes used in place of skills in concise descriptions.

The associations between skills and occupations in our data provide a level of detail not currently available in official statistics. Through natural language processing of skill keywords and their associations with occupational codes, we found that occupations can be modeled as probability distributions over term-weighted skills and that cosine distance between these distributions captures the existing SOC-code hierarchy of occupations.

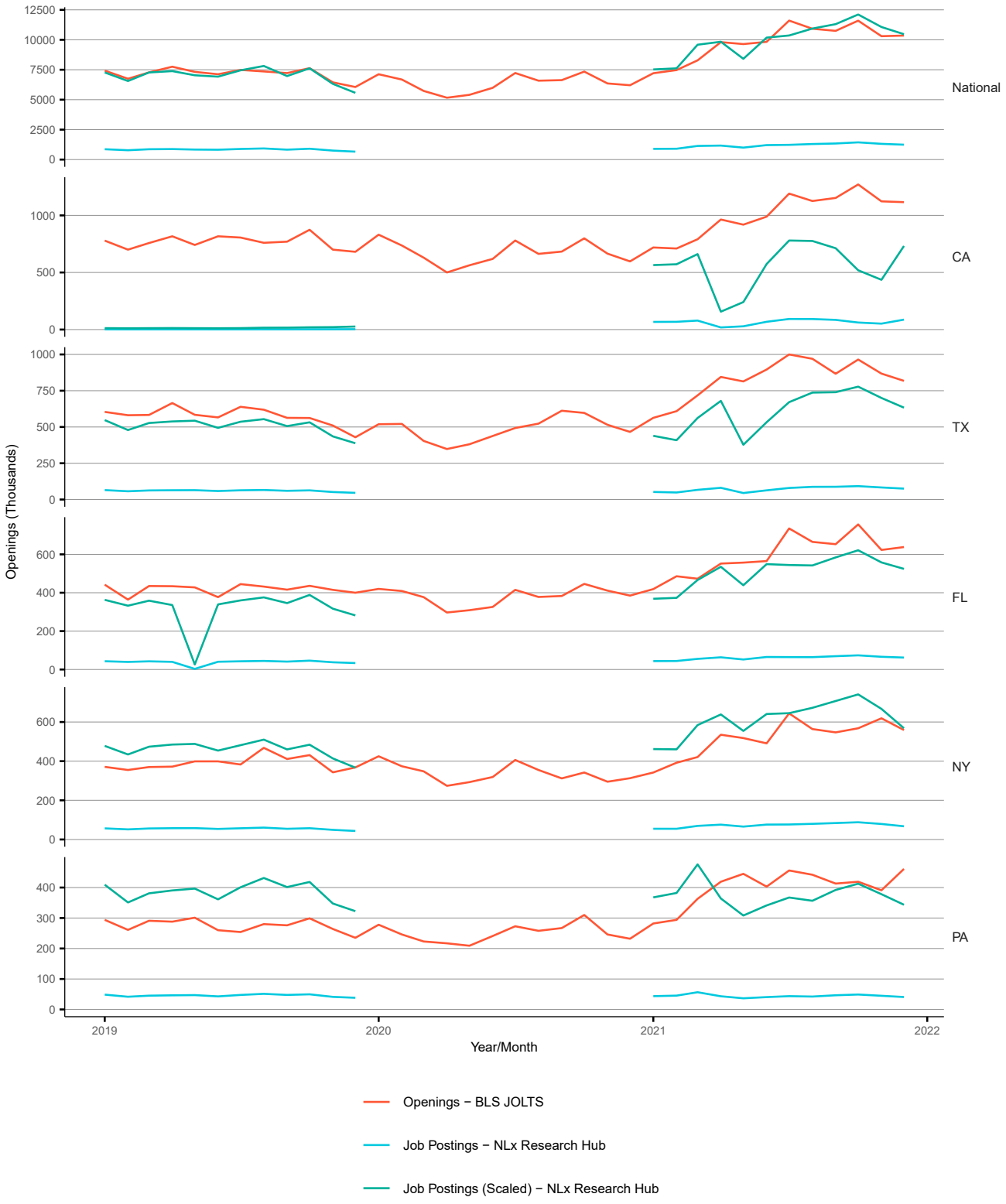

Figure 4. Comparison of Research Hub job postings vs. estimates of job openings across states in the United States for 2019 and 2021

Figure 5. Comparison of the distribution of Research Hub job postings vs. estimates of current employment for occupations at the 2-digit SOC level for 2019 and 2021

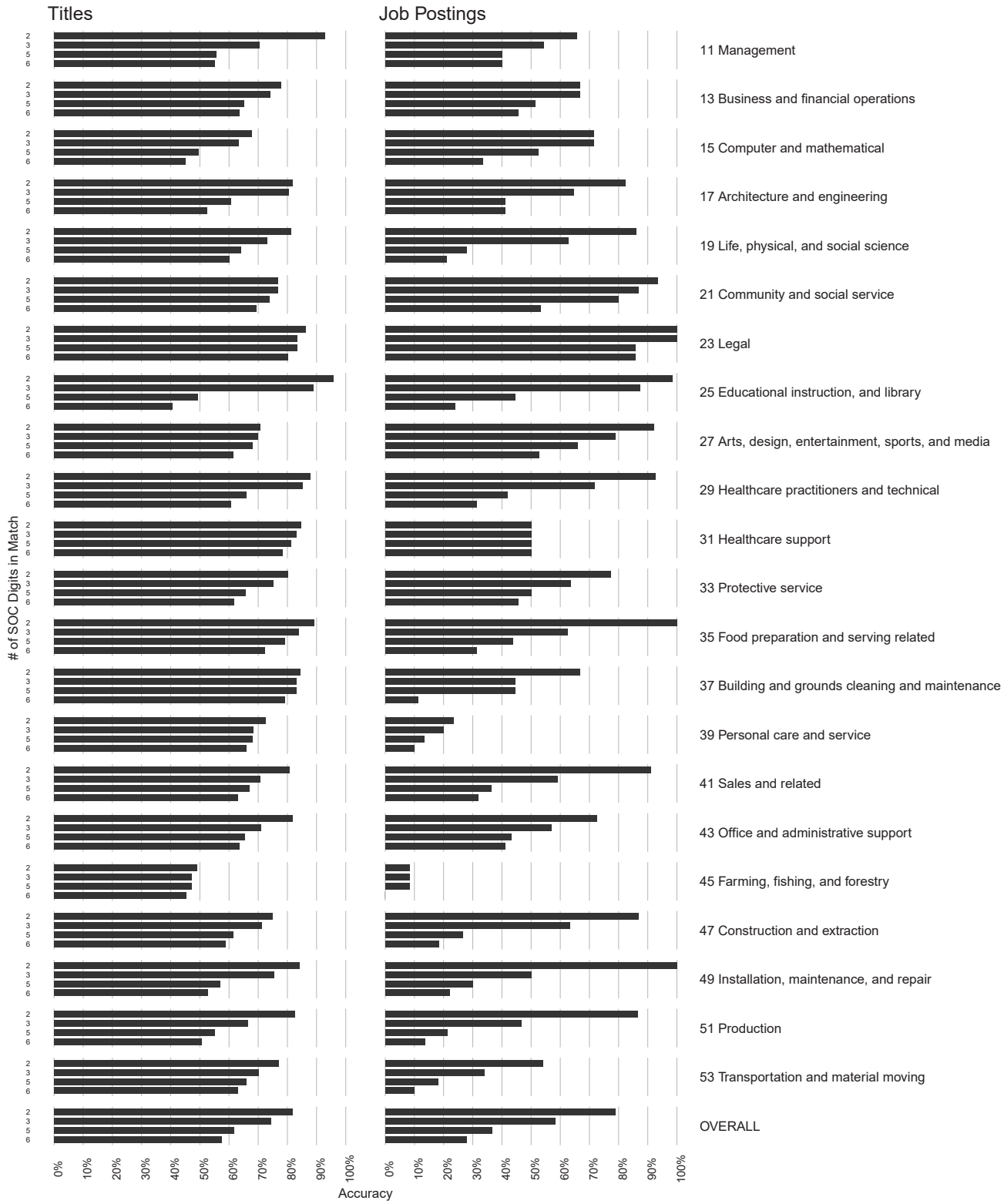

Figure 6. Accuracy of title to SOC models and job postings to SOC models as measured by the percentage of test cases where the three most probable SOC codes match at the 2-, 3-, 5-, or 6-digit SOC level

We have applied these models in a real-world application to recommend career transitions to job seekers based on skill similarity to their previous occupations.[18] Using these methods, researchers can parse unstructured job description, resume, or job title data in order to conduct analyses that rely on structured SOC codes, which could open up new lines of research that were previously not possible.

In future work, we hope to improve our sampling through additional sources of job postings, in particular to address the lower accuracy of under-represented SOC codes in the Research Hub data. Improved sampling may reduce occupational and regional biases and increase the accuracy of matching SOC codes to job titles and approximating job openings from job posting frequencies.

Resource availability

The NLx Research Hub data reported in this study cannot be deposited in a public repository because it is accessible only by authorized users under agreement with the National Association of State Workforce Agencies. For more information, see the NLx Research Hub's request process at https://nlxresearchhub.org/request-nlx-data. Datasets reported in this study that were derived from the NLx Research Hub data have been deposited at Zenodo and are publicly available as of the date of publication.[39] All original code has been deposited at GitHub and Zenodo.[40,41] These datasets and software are openly available for academic use, including reverse engineering and derivative works, under a custom license.

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.patter.2023.100757.

ACKNOWLEDGMENTS

We thank Justine S. Hastings, Eric Chyn, and Seth D. Zimmerman for their feedback. We thank Mintaka Angell, Madelyn Rahn, Amelia Roberts, Sarah White, and April Yee for their manual review of skill keywords. We thank Erin Maness for her development work on the *sockit* web application. The NLx Research Hub development was supported by National Science Foundation award 1937026 to the National Association of State Workforce Agencies. The National Labor Exchange (NLx) Data Trust bears no responsibility for the analyses or interpretations of the data presented here. The opinions expressed herein, including any implications for policy, are those of the authors and not of the NLx Data Trust members. JOLTS data were downloaded using the Public API from the United States Bureau of Labor Statistics.

AUTHOR CONTRIBUTIONS

The authors are listed in alphabetical order (by last name) in the author line. M.G., M.H., and K.S. conceived the methods and approach. E.N. and C.Y. facilitated access to data in the NLx Research Hub and provided feedback on potential applications. M.G., M.H., J.L., and K.S. performed data analysis. N.D., M.G., E.H., and M.H. implemented the methods and data in the *sockit* package. M.G., M.H., and J.L. drafted the manuscript. All authors reviewed the manuscript.

DECLARATION OF INTERESTS

M.H. is currently senior data scientist at Amazon.com, Inc., but conducted this research prior to starting that role.

Received: November 16, 2022
Revised: January 10, 2023
Accepted: April 24, 2023
Published: May 22, 2023

REFERENCES

1. Levine, C., Salmon, L., and Weinberg, D.H. (1999). Revising the Standard Occupational Classification System. Monthly Labor Review (U.S. Bureau of Labor Statistics), pp. 36–45.
2. Theroux, R.P. (2017). Standard occupational classification (SOC) system—revision for 2018. Fed. Regist. *82*, 56271–56273.
3. Dingel, J.I., and Neiman, B. (2020). How many jobs can be done at home? J. Publ. Econ. *189*, 104235. https://doi.org/10.1016/j.jpubeco.2020.104235.
4. del Rio-Chanona, R.M., Mealy, P., Pichler, A., Lafond, F., and Farmer, J.D. (2020). Supply and demand shocks in the COVID-19 pandemic: an industry and occupation perspective. Oxf. Rev. Econ. Pol. *36*, S94–S137. https://doi.org/10.1093/oxrep/graa033.
5. Gibson, D.M., and Greene, J. (2020). Risk for severe COVID-19 illness among health care workers who work directly with patients. J. Gen. Intern. Med. *35*, 2804–2806. https://doi.org/10.1007/s11606-020-05992-y.
6. Buckner-Petty, S., Dale, A.M., and Evanoff, B.A. (2019). Efficiency of autocoding programs for converting job descriptors into standard occupational classification (SOC) codes. Am. J. Ind. Med. *62*, 59–68. https://doi.org/10.1002/ajim.22928.
7. Schmitz, M., and Forst, L. (2016). Industry and occupation in the electronic health record: an investigation of the national Institute for occupational safety and health industry and occupation computerized coding system. JMIR Med. Info *4*, e5. https://doi.org/10.2196/medinform.4839.
8. U.S. Centers for Disease Control and Prevention. NIOSH Industry and Occupation Computerized Coding System (NIOCCS). https://csams.cdc.gov/nioccs/.
9. 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., et al. (2016). Computer-based coding of free-text job descriptions to efficiently identify occupations in epidemiological studies. Occup. Environ. Med. *73*, 417–424. https://doi.org/10.1136/oemed-2015-103152.
10. U.S. National Cancer Institute, SOCcer - Standardized Occupation Coding for Computer-assisted Epidemiological Research. https://soccer.nci.nih.gov/soccer/.
11. U. S. Department of Labor. Employment and Training Administration. O*NET Code Connector. https://www.onetcodeconnector.org/.
12. De Matteis, S., Jarvis, D., Young, H., Young, A., Allen, N., Potts, J., Darnton, A., Rushton, L., and Cullinan, P. (2017). Occupational self-coding and automatic recording (OSCAR): a novel web-based tool to collect and code lifetime job histories in large population-based studies. Scand. J. Work. Environ. Health *43*, 181–186. https://doi.org/10.5271/sjweh.3613.
13. Lightcast™. Titles API. https://api.lightcast.io/apis/titles.
14. R.M. Wilson Consulting, Inc. O*NET-SOC AutoEncoder™. https://www.onetsocautocoder.com/.
15. Barker, M., Chue Hong, N.P., Katz, D.S., Lamprecht, A.L., Martinez-Ortiz, C., Psomopoulos, F., Harrow, J., Castro, L.J., Gruenpeter, M., Martinez, P.A., and Honeyman, T. (2022). Introducing the FAIR Principles for research software. Sci. Data *9*, 622. https://doi.org/10.1038/s41597-022-01710-x.
16. Python Package Index. Sockit. https://pypi.org/project/sockit/.
17. Research Improving People's Lives. Sockit. https://research.ripl.org/#/sockit.
18. Howison, M., and Long, J.. Recommending career transitions to job seekers using earnings estimates, skills similarity, and occupational demand (February 26, 2023). https://ssrn.com/abstract=4371445.
19. U.S. National Labor Exchange. NLx Research Hub. https://nlxresearchhub.org/.

20. U.S. National Labor. Exchange. https://usnlx.com/.

21. National Association of State Workforce Agencies. https://www.naswa.org/.

22. DirectEmployers Association. https://directemployers.org/.

23. U.S.. Office of Personnel Management. USAJOBS. https://www.usajobs.gov/.

24. Python Package Index. WordTrie. https://pypi.org/project/wordtrie/.

25. U.S. Department of Labor. Employment and Training Administration. O*NET 27.0 Database. https://www.onetcenter.org/database.html.

26. U.S. Bureau of Labor Statistics (2018). SOC Direct Match Title File. https://www.bls.gov/soc/2018/home.htm.

27. DirectEmployers Association. Member Companies. https://dejobs.org/member-companies/.

28. Wikipedia. List of United States cities by population. https://en.wikipedia.org/wiki/List_of_United_States_cities_by_population.

29. U.S. Bureau of Labor Statistics. Crosswalks between the 2018 SOC and Systems Used by Other Federal and International Statistical Agencies. https://www.bls.gov/soc/2018/crosswalks_used_by_agencies.htm.

30. U.S. Department of Labor. Employment and Training Administration. CareerOneStop. https://www.careeronestop.org/.

31. Spärck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. J. Doc. *28*, 11–21. https://doi.org/10.1108/eb026526.

32. Wu, H.C., Luk, R.W.P., Wong, K.F., and Kwok, K.L. (2008). Interpreting TF-IDF term weights as making relevance decisions. ACM Trans. Inf. Syst. *26*, 1–37. https://doi.org/10.1145/1361684.1361686.

33. Kullback, S., and Leibler, R.A. (1951). On information and sufficiency. Ann. Math. Stat. *22*, 79–86. https://doi.org/10.1214/aoms/1177729694.

34. U.S. Bureau of Labor Statistics. Job Openings and Labor Turnover Survey. https://www.bls.gov/jlt/.

35. Ramos, M. (2022). Employment recovery continues in 2021, with some industries reaching or exceeding their prepandemic employment levels. Monthly Labor Review (U.S. Bureau of Labor Statistics). https://doi.org/10.21916/mlr.2022.15.

36. U.S. Census Bureau. American Community Survey. https://www.census.gov/programs-surveys/acs/.

37. U.S. Bureau of Labor Statistics. Occupational Employment and Wage Statistics. https://www.bls.gov/oes/.

38. Shani, C., Zarecki, J., and Shahaf, D. (2023). The lean data scientist: recent advances toward overcoming the data bottleneck. Preprint at Zenodo. Commun. ACM *66*, 92–102. https://doi.org/10.1145/3551635.

39. Howison, M. (2022). Replication files for: "Occupational models from 42 million unstructured job postings" [Data set]. Zenodo. https://doi.org/10.5281/zenodo.7319953.

40. Howison, M. (2022). ripl-org/sockit: v0.3.1. Zenodo. https://doi.org/10.5281/zenodo.7606616.

41. Howison, M. (2022). ripl-org/sockit-data: v0.3.1. Zenodo. https://doi.org/10.5281/zenodo.7606613.