SAS® GLOBAL FORUM 2021

Paper ###-2021 (if SAS author, use SAS###-2021 or SAS####-2021)

# Are you Qualified? Examining existing SOC Occupational groups and Assessing the chance of a candidate's suitability for a job

Frank Yeboah, Maruthi Sankar Nanduri, Monish Vallamkonda, Nikhil Gunti

Oklahoma State University

## ABSTRACT

Employee hiring process can be a cumbersome and tedious process with increasing number of job seekers entering the market. Adding to this a lack of transparency in the hiring process can put the candidates in the dark by not letting them know their shortcomings. In this paper, existing Standard Occupation Classification (SOC) groupings from the Bureau of Labor Statistics (BLS) are examined through hierarchical clustering based on descriptors that encompass individual skills and abilities. Later using O*NET occupational data, a hybrid recommendation engine on the LCA disclosure data of 2020 for H1-B and other programs is used to recommend the jobs for a candidate. The results of this paper reveal a gap within the current SOC groups and a need to re-evaluate the occupational groups based on the importance of skills, abilities, and other descriptors rather than only the description of the occupation. The paper also recommends similar suitable occupations for a given person using a recommendation engine.

## INTRODUCTION

It has been estimated that on average, about 250 people apply to any given corporate job posting. Out of this, only 4-6 people are selected for an interview and only 1 person gets the job, often after a very long and tedious hiring process. Most of the time, the candidates are unaware of the reason for rejection, leading to a need to identify important skills and abilities for a particular job and self-evaluate on the corresponding factors of the job. A study of the current Standard Occupational Classification (SOC) System used by the federal agencies for collection and dissemination of job data can be handy in identifying the important skills needed for each job.

The Occupational Information Network (O*NET) program is the primary source of occupational information for the US containing 585 occupation-specific descriptors like skills, abilities, interests, and other distinguishing characteristics for 974 occupations within the US. Based on the Standard Occupational Classification (SOC) codes, the 974 occupations are grouped into 23 major occupational groups. The O*NET database is developed through a continuous data collection program by surveying job incumbents using standard questionnaires that provides ratings for each descriptor in the occupation based on the responses to the surveys.

## PROBLEM

The current classification of the SOC is based on the job description, which is a textual representation. Since job descriptions are not always a unique representation of what is required to do the job, those seeking to apply often do not have a clear understanding of the important skills needed for the job. Therefore, alternate grouping systems developed based on the ratings provided for the descriptors – skills, abilities etc. will be very handy in identifying the most important skills needed for each occupation. Using clustering

algorithms to group occupations having similar skills or abilities, this paper proposes an easier way for identifying the most important skills for each occupation in the O*NET program.

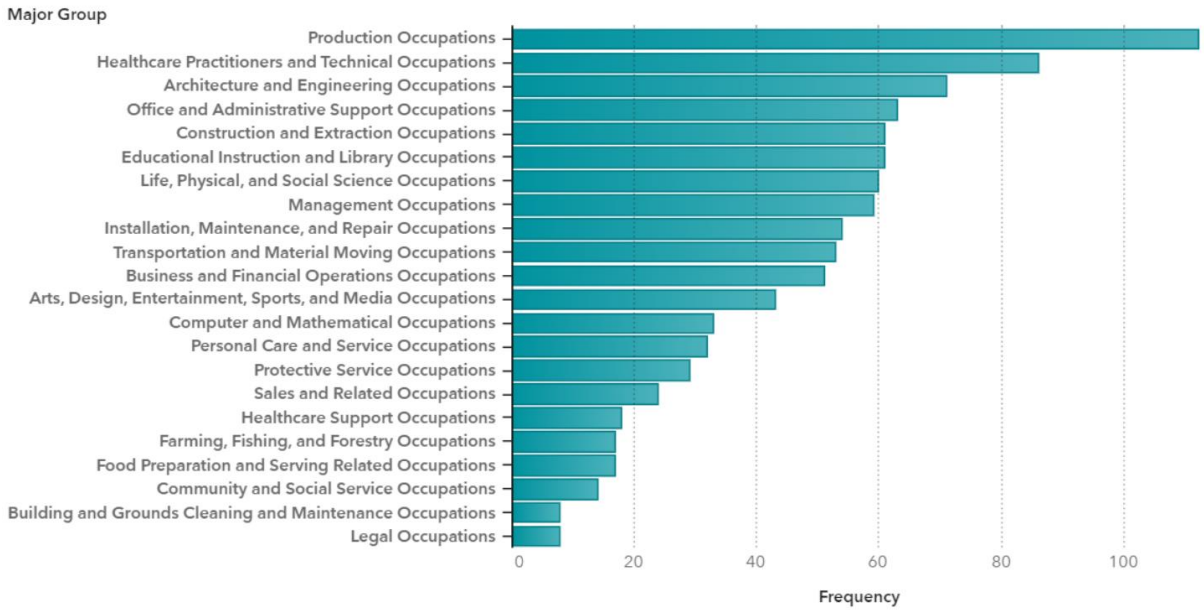



**Figure 1. Distribution of Job titles into 23 major SOC codes**

Another important question is: Can employment rates be increased by bringing candidates closer to more suitable jobs? A recommendation engine that suggests suitable jobs to candidates based on the job description or skills that are required for the job can be useful. The LCA disclosure data for H1-B visas is used to recommend the 10 most suitable jobs to candidates based on the O*NET job description and tasks and data of candidates who have applied to similar jobs and their corresponding profile.

## DATA

There are two data sources used in the following analysis. The first dataset is from the O*NET website[1], which contains 37 tables that encompasses the occupations, education, skills, abilities, interests, work values, work styles, work context, tasks, and other reference data. The second dataset is the LCA disclosure data[2] of 2020 for H1-B visa applications, which is made public by USCIS. This dataset contains information about H1-B applicants and the SOC codes with skills required for the occupation for a given visa application. This dataset contains 97,000 records.

## DATA PREPARATION

The available data (EXCEL format, 37 tables) is converted to SAS tables using a macro function with the PROC IMPORT procedure. Post conversion, 8 tables with descriptors in Abilities, Interests, Knowledge, Skills, Work Activities, Work Styles, Work Values and Work Context are combined to create a master table for the analysis. The master table has 16 columns that include various elements and their ratings on multiple scales and over 671,000 rows representing the occupations listed in the H1-B visa applications. Different scales for measuring different elements are addressed by normalizing the data values that represent the ratings on a scale of 1-10. Post re-scaling, the data is cleaned excluding the columns

[1] O*NET® 25.1 Database https://www.onetcenter.org/database.html

[2] H-1B Employer Data Hub https://www.uscis.gov/tools/reports-and-studies/h-1b-employer-data-hub

like sample size, error, upper bound and lower bound that are unnecessary for the current problem which only needs the score value for a descriptor. Then the table is restructured by pivoting scores for 256 descriptors (columns) for each of the 974 occupations (observations).

Two columns (Apprenticeship and Job-elated Professional certification) with more than 50% of missing values are exempted from the analysis with a final data set of 974 observations and 254 columns available for further analysis. Based on the business requirement of the current problem that requires skills and education for a given job, the LCA disclosure dataset is filtered to take 7 columns and 20k records consisting of the job title, major level of study, minimum education level, specific skills, and job description. With Job title as the common variable, this sample is joined with O*NET occupational data along with the O*NET job description and tasks for each occupation.

# ANALYSIS AND RESULTS

## HIERARCHICAL CLUSTERING

Agglomerative hierarchical clustering is used to evaluate the current SOC groups which are currently organized into a hierarchical structure. New clusters were created via hierarchical clustering using squared Euclidean distance. The within-cluster variance was reduced by setting the Cubic Clustering Criterion (CCC) value to a minimum of 3, since a CCC value greater than 3 indicates good disjoint clusters. The algorithm is implemented using the "Cluster" node in the SAS EM. A maximum of 30 clusters is set to limit the groups to less than 30 that can be comparable to the original 23.

Hierarchical Clustering produced 25 clusters, as compared to the 23 groups of the SOC system. The optimum number of clusters is found at a CCC value of 5.8635. It is observed that 65 variables (descriptors) that are identified as important explain 99.06% of variance in the data. Active Listening and Written Expression are the two main descriptors listed for most of the occupational clusters. After examining the new clusters and comparing them with the current SOC system, multiple discrepancies are observed. For example, "Methane/Landfill Gas Collection System Operators'' and "Methane/Landfill Gas Collection System Technicians" contain similar job descriptions and skill sets but are grouped in separate categories in the SOC structure. As shown in Table 1 below, these two occupations are grouped into the same cluster (Cluster 20) in this analysis mainly because the jobs require similar skill sets.

| Major Group | O_NET_SOC_Code | Title | Cluster |
|---|---|---|---|
| Management Occupations | 11-3051.05 | Methane/Landfill Gas Collection System Operators | 20 |
| Production Occupations | 51-8099.02 | Methane/Landfill Gas Collection System Technicians | 20 |

**Table 1. Different occupations grouped into same cluster**

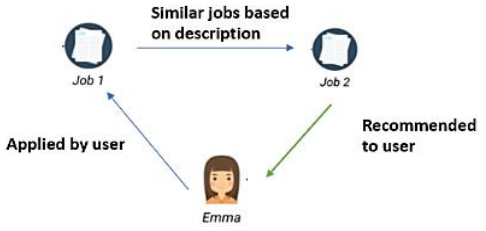
## HYBRID RECOMMENDATION ENGINE

Content-Based Filtering and Collaborative filtering are applied on the LCA dataset for the H1-B Visa applications. The main variables included the O*NET job title, the O*NET Code, H1-B employee's major level of study, employee's minimum educational level & Job description from the O*NET Tasks table.

### Content Based Filtering

For a content-based filtering approach, Job title and its corresponding job description were combined into a single column and a TF-IDF matrix was generated out of it. A linear kernel

function is applied to create similarities between job titles, so that for any given ONET job title, the engine recommends the top 10 similar job titles. These results are explained using for example: Emma, a job applicant is applying to Job 1 shown in Figure 2. Using Content based filtering, Job 2 can be recommended to Emma as an alternate opportunity based on similar skill sets and job requirements, and job descriptions.




| Occupation | Top 10 Recommendations |
|---|---|
| Graphic Designers | Art Directors<br>Desktop Publishers<br>Multimedia Artists and Animators<br>Electronic Drafters<br>Commercial and Industrial Designers<br>Fine Artists, Including Painters, Sculptors<br>Architects, Except Landscape and Naval<br>Manufacturing Engineering Technologists<br>Editors<br>Prepress Technicians and Workers |
| Chemists | Chemical Technicians<br>Quality Control Analysts<br>Medical and Clinical Laboratory Technologists<br>Materials Scientists<br>Food Science Technicians<br>Chemical Engineers<br>Biological Technicians<br>Biochemists and Biophysicists<br>Food Scientists and Technologists<br>Chemical Plant and System Operators |

**Figure 2. Content Based Filtering and sample result**

## Collaborative Filtering

In the collaborative filtering approach, the recommendation engine built is focused on using the employee's highest educational level and the major field of study. A matrix factorization algorithm is used to create K-nearest neighbors for each candidate. From this, top 10 jobs that other applicants with the same educational background and major field of study applied to are known and can be recommended to others. In Figure 3, this concept is explained using the results with two candidates: Emma and Jules, who share similar educational backgrounds and have the same major field of study. Emma applied to Jobs 1, 2 and 3 while Jules applied to Jobs 1 and 2 only. Using the collaborative filtering technique, it is recommended to Jules to apply for Job 3 too as they share similar educational backgrounds.

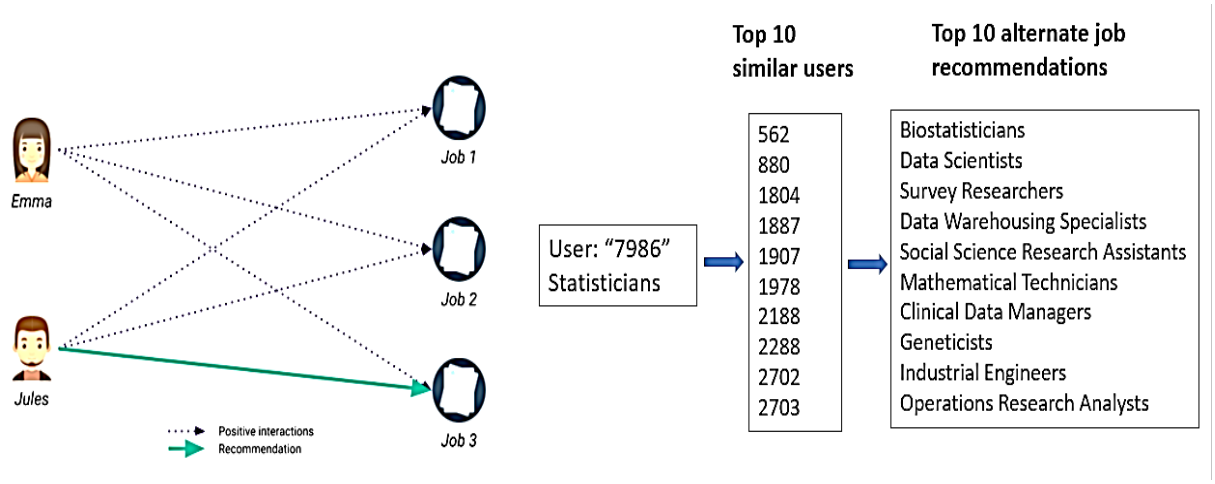



**Figure 3. Collaborative Filtering and sample result**

From Figure 3, Jules, a candidate who applied for a statistician job is recommended to apply to 10 other alternate jobs like Biostatistician etc. based on the jobs of similar applicants with a similar educational profile (Emma).

## GENERALIZATIONS

Results in this paper are beneficial to both job seekers and employers in the hiring process. Hiring companies can leverage the clustering algorithm approach proposed in this paper to assess the skills and other descriptor ratings of applicants prior to an initial resume review. This provides a quick screening of the candidate by validating the applicant's profile with the important skills for a given position and reducing human error during the resume reviews. Job seekers benefit by knowing the 10 required skills for a job and improving those skills to increase the chances of getting hired. The recommendation engine approaches help the applicants by suggesting suitable jobs. The Bureau of Labor Statistics can also utilize the results of the clustering algorithm in this research to re-evaluate their SOC grouping criteria and produce a more representative occupational grouping based on the ratings of skills and abilities. Such an analysis can be incorporated into a periodic review process to keep the SOC grouping valid and up to date with emerging skills, positions, etc.

## CONCLUSION

This paper attempts to provide methodologies and tools to address the gap between job seekers and job opportunities. The paper provides a proof-of-concept for reviewing and updating the SOC occupational groups, by grouping them based on skills and abilities, instead of using only job descriptions. The proof of concept with the hybrid recommendation engine assists applicants by connecting them to positions that match their skillsets. The clustering results and recommendation engine also serves as a potential way for reducing the duration needed to screen thousands of job applicants and thereby providing a platform for hiring managers to improve their resume reviews.

## REFERENCES


Website H-1B Employer Data Hub (https://www.uscis.gov/tools/reports-and-studies/h-1b-employer-data-hub ) (Accessed November 2020)

Website O*NET® 25.1 Database https://www.onetcenter.org/database.html (Sept 2020)

Website 2020 HR Statistics Article: https://zety.com/blog/hr-statistics

Website National Research Council. 2010. A Database for a Changing Economy: Review of the Occupational Information Network (O*NET). Washington, DC: The National Academies Press. https://doi.org/10.17226/12814


## CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:


Frank Yeboah
Oklahoma State University
frank.yeboah@okstate.edu

Maruthi Sankar Nanduri
Oklahoma State University
mnandur@okstate.edu

Monish Vallamkonda
Oklahoma State University
monish.vallamkonda@okstate.edu

Nikhil Gunti
Oklahoma State University
nikhil.gunti@okstate.edu

# APPENDIX

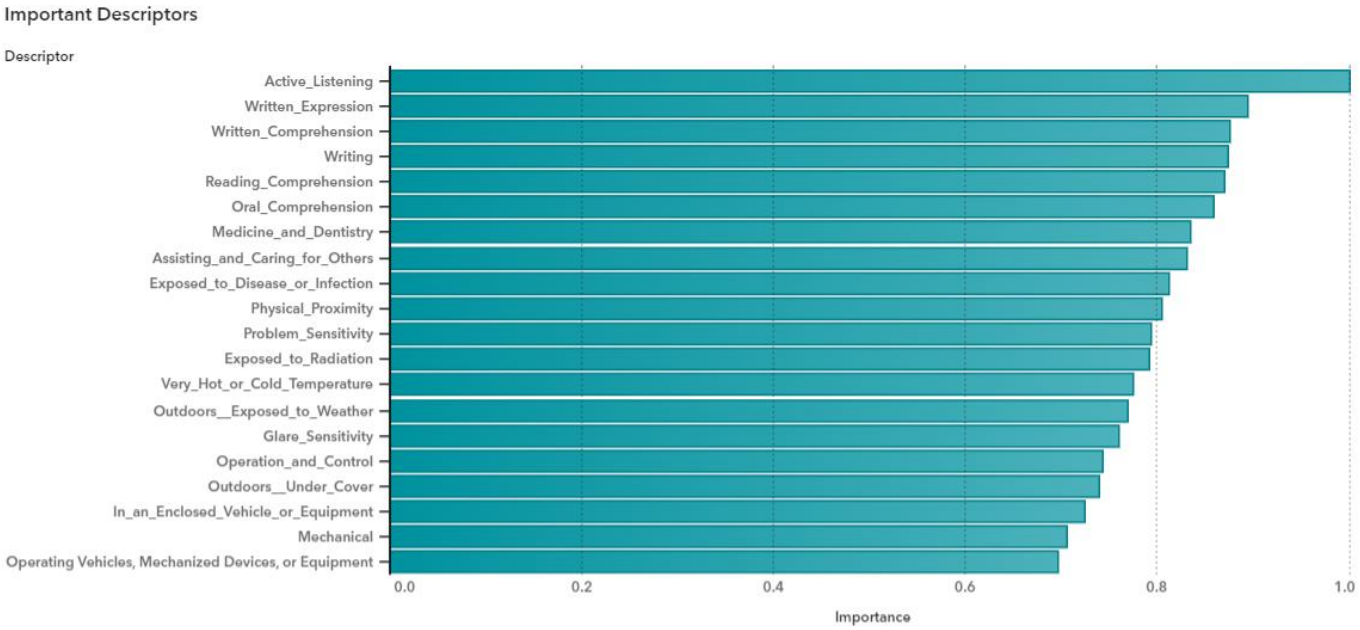



**Figure 4. Top 10 descriptors in the data**

| Major Group | O_NET_SOC_Code | Title | Cluster |
|---|---|---|---|
| Management Occupations | 11-3051.05 | Methane/Landfill Gas Collection System Operators | 20 |
| Production Occupations | 51-8099.02 | Methane/Landfill Gas Collection System Technicians | 20 |
| Management Occupations | 11-9121.02 | Water Resource Specialists | 5 |
| Architecture and Engineering Occupations | 17-2081.01 | Water/Wastewater Engineers | 5 |
| Management Occupations | 11-3021.00 | Computer and Information Systems | 7 |
| Computer and Mathematical Occupations | 15-1199.09 | Information Technology Project Managers | 7 |

**Table 2. An Example of the new Cluster groupings**

To be able to use the proposed applicant skill ratings and the new cluster groupings, suppose a candidate (Nandi) is applying for a job as a statistician as shown in Table 3 below. The top 10 ratings for the skills on Nandi's resume are then compared with the average ratings for the top 10 skills statisticians possess, which is grouped in cluster 5. Since there are three main skillsets (Data analysis, Mathematical reasoning and Programming) where Nandi rates lower than the average statistician in cluster 5, our recommendation will be for Nandi to improve these three skills before applying for the statistician job in order to be competitive. So, improving in these areas would help Nandi increase her chances of getting hired.

| Skills or Abilities | Applicant Skill Ratings | Average Statistician skill ratings |
|---|---|---|
| Analyzing_Data_or_Information | 8.5 | 9.20 |
| Mathematical_Reasoning | 9 | 9.38 |
| Programming | 5 | 6.58 |
| Processing_Information | 10 | 10.00 |
| Estimating_the_Quantifiable | 8.5 | 8.29 |
| Systems_Analysis | 8.5 | 8.46 |
| Number_Faculty | 9 | 7.91 |
| Interacting_with_Computers | 9 | 8.83 |
| Analytical_Thinking | 8 | 5.59 |
| First_Interest_High_Point | 9 | 5.23 |

**Table 3. Comparison of Cluster results with Test data of an applicant**

| Variable | Worth | Rank |
|---|---|---|
| Analyzing_Data_or_Information | 0.073986 | 1 |
| Mathematical_Reasoning | 0.069130 | 2 |
| Programming | 0.065685 | 3 |
| Processing_Information | 0.063685 | 4 |
| Estimating_the_Quantifiable | 0.059851 | 5 |
| Systems_Analysis | 0.058349 | 6 |
| Number_Faculty | 0.057571 | 7 |
| Interacting_with_Computers | 0.056610 | 8 |
| Analytical_Thinking | 0.055767 | 9 |
| First_Interest_High_Point | 0.054917 | 10 |

**Table 4. Top 10 skills and abilities in cluster 5**