RESEARCH & POLICY

INSIGHTS

Occupations by Skills Clusters for the U.S.: Methodological Framework and Experiments

Publication 112

October 2024

Authors

Indraneel Kumar, PhD
Karen Ivanna Siller
Mark C. White, PhD
Benjamin St. Germain
Andrey Zhalnin, PhD
Bertin Mbongo

PURDUE UNIVERSITY | Center for Regional Development

Economic Clusters for the 21st Century

Disclaimer: This article is prepared by Purdue Center for Regional Development using Federal funds under award #ED23RNA0G0174 from the Economic Development Administration, U.S. Department of Commerce. The statements, findings, conclusions, and recommendations are those of the coauthors and do not necessarily reflect the views of Economic Development Administration or the U.S. Department of Commerce.

Meet the Authors

INDRANEEL KUMAR, PHD

Director | Purdue Center for Regional Development

pcrd.purdue.edu/about-us/our-team/indraneel-kumar

KAREN IVANNA SILLER

1st Year Doctoral Student | Purdue Agricultural Economics

linkedin.com/in/ivanna-carrillo-siller-7b277522b

MARK C. WHITE, PHD

Clinical Associate Professor | University of Illinois Urbana Champaign

ace.illinois.edu/directory/whitemc

BENJAMIN ST. GERMAIN

GIS Analyst | Purdue Center for Regional Development

pcrd.purdue.edu/about-us/our-team/benjamin-st-germain

ANDREY ZHALNIN, PHD

GIS & Data Analyst | Purdue Center for Regional Development

pcrd.purdue.edu/about-us/our-team/andrey-zhalnin

BERTIN MBONGO

GIS Analyst | Purdue Center for Regional Development

pcrd.purdue.edu/about-us/our-team/bertin-mbongo

Occupations by Skills Clusters for the U.S.: Methodological Framework and Experiments

// ABSTRACT

Occupations and skills analysis have become as important as the analysis of industries for economic and workforce development in the U.S. This research and policy insight article explores analytical methods to develop occupations by skills clusters for the U.S. by using public sources of data. The article presents different methods and results, and ways to identify an optimal number of occupations by skills clusters. It concludes by presenting policy implications and practical applications for such databases including future directions for the research.

Why are occupations and skills important to the regional economy?

A general saying is that if the economy is a "coin," industries and occupations make the two sides of that coin. This notion of "industries and occupations" making two sides of the same coin is more relevant now especially for regional economic development practitioners and researchers. Previous research such as Kadokawa (2011) found availability of labor and technical skills in the second tier of significant decision-making variables for industries to relocate or move to a specific location. The top reasons for relocating to a specific place included proximity to related firms, headquarters, and research institutions (Kadokawa 2011). However, the International Economic Development Council (IEDC) states that "skilled workforce" is the top most reason why a business would locate at a specific place.[1] During a recent survey in 2023, site selectors ranked availability or potential of skilled workforce as the "top reason to make a place attractive for a new business or establish an industry."[2] They also found that in the post-pandemic period, a large proportion of corporate decision makers are considering reshoring of manufacturing, which is creating more demand for talent and skilled labor force.[3]

Feser (2003) had presented that both industries and occupations are two complementary views of the same regional economy. Industries show the production side of the economy, whereas occupations show the job-related activities of the labor force or workers engaged in those industries. Researchers elucidated that occupational analysis was equally important as industrial analysis in determining competitiveness of regions. Koo (2005) presented that "worker quality and local knowledge bases" were important assets for economic competitiveness as they strengthened the regional comparative advantages. Until the early 2000s, significant research on industry clusters had happened including three major efforts to define the benchmarked industry clusters for the U.S. by Feser and Bergman (2000), Porter (2003), and Feser (2005). Similar efforts were not made to define occupational clusters despite Thompson and Thompson (1987) making the case to use industries and occupations to develop the cross-hair targeting for regional economic development. Similar to industry clusters, occupation clusters are defined as groups of occupations sharing some common characteristics, such as education, training, skills, etc.

What is O*Net?

Occupational characteristics are researched by Occupational Information Network (O*Net), which is an occupational program sponsored by the U.S. Department of Labor (DOL) and Employment and Training Authorization (ETA).[4] On behalf of DOL and ETA, the O*Net collects data about occupations focusing on the knowledge, skills, and abilities needed to perform different tasks, duties, and

[1] https://www.iedconline.org/clientuploads/Downloads/Key_Strategies/IEDC_Why_and_Impact_Workforce_Development.pdf.

[2] https://siteselection.com/issues/2023/jan/where-o-where-have-the-laborers-gone.cfm

[3] Ibid.

[4] https://www.onetcenter.org/

responsibilities required for a particular job.[5] As part of the program, O*Net also provides career planning web-based tools for aspirant and existing labor force, such as My Next Move[6] to serve the general population and My Next Move for Veterans[7], and other information[8] for state, local, and nonprofit workforce development departments and agencies. Currently, the O*Net database is at version 28.3, and it has as many as 873 different types of occupations which varies from helpers (brick masons, roofers, etc.) and clergy to climate change policy analysts. The O*Net compiles occupational descriptors of knowledge, skills, and abilities comprised of several hundred variables that enable the analysis of occupations from different perspectives.

// LITERATURE REVIEW

"Labor" is an important factor of production for a regional or a national economy. For example, the Cobb-Douglas production function uses "labor," "capital," and a "technology" coefficient as the factors for the economic output.[9] The labor force and workers generally represent the human capital because of their education, knowledge, skills, abilities, and expertise. As mentioned previously, the availability of a skilled labor force is an important parameter for new businesses and industries searching for sites in communities and regions. Romer (1990) introduced a seminal idea that technological change innovated by people can induce economic growth or the "stock of human capital determines the rate of growth." Innovation in industrial products and processes do not happen because of exogenous factors, but people taking intentional actions for technological advancements to benefit from market incentives (Romer, 1990). The endogenous growth theory implies that economic growth could result from innovations in technology, products, and processes within the regional or national economy.[10] Note that the model for economic output postulated by Romer (1990) included four inputs which were capital, labor, human capital, and the level of technology. The occupations and their characteristics especially education, knowledge and skills represent the human capital aspects or the quality of the labor force. Hence, investing in improving the quality of human capital such as education, skill-based training, vocational training, etc., became important for regional and national economic development.

Recognizing the distinction between what businesses make and what workers do is an important step toward understanding a regional economy (Thompson and Thompson, 1987). The former relies on industry-based employment and wage data (i.e., NAICS[11]-based data), while the latter often requires occupational data (i.e., SOC[12]-based data) and these different perspectives offer different insights.

[5] https://www.onetcenter.org/overview.html

[6] https://www.mynextmove.org/

[7] https://www.mynextmove.org/vets/

[8] https://www.onetonline.org/

[9] https://spureconomics.com/cobb-douglas-production-function/

[10] https://www.brown.edu/Departments/Economics/Faculty/Peter_Howitt/publication/endogenous.pdf

[11] North American Industry Classification System

[12] Standard Occupation Classification

There are many forms of interdependence that bind regional economic clusters and occupational clusters that provide a tool for understanding shared labor pools (Renski, Koo, and Feser, 2007).

Workforce capacity can dictate regional economic opportunities and where those regions fit within the broader economy. Examining these regional economies through an occupational lens effectively complements more widely used NAICS-based industry clusters approaches. Occupational cluster schemes often utilize the U.S. Bureau of Labor Statistics' (BLS) national industry-occupation staffing patterns and information drawn from the O*Net database. For instance, Feser (2003) created a conceptual framework to identify **knowledge-based occupational clusters** as a tool to help economic development researchers and practitioners identify the competitive advantages within their regional workforce.

Subsequent efforts to operationalize occupational clusters have used similar data sources and methodologies (e.g., Koo, 2005; Nolan et al, 2011; Slaper, 2014), but pursued different goals. For instance, Nolan et al (2011) created a framework that included 15 knowledge-based occupational clusters, thereby prioritizing more knowledge-intensive occupations. By contrast, Chrisinger, Fowler, and Kleit (2012) developed cluster definitions that also included occupations more commonly found in lower wage, locally serving industries (e.g., personal healthcare and assistance, medical and social assistance, and hospitality and personal services, etc.). Other researchers sought to appeal to a much broader audience by creating occupational cluster definitions that cover much of the overall workforce (e.g., Slaper, 2014).

No single approach, however, provides all the information necessary to inform regional development efforts. Consequently, there remains strong practical reasons to examine clusters from a variety of perspectives; perspectives that yield different insights and may reveal connections and potential opportunities that are not obvious. For instance, Markusen and Barbour (2003) noted that engineers in Southern California's aerospace industry also found opportunities in the sportswear industry, due to their knowledge of different materials. Similarly, the demand for woodworking skills allowed displaced furniture workers in North Carolina to find work in the state's boatbuilding industry. As a result, cluster initiatives—large scale investments and activities that grow and leverage the regional competitive advantage arising from a unique concentration of skills or activities—can change a region's economic trajectory. These initiatives, therefore, can help regions transition away from existing clusters that are losing their relative competitiveness, and instead build more competitive clusters that offer greater future growth potential (Donahue, Parilla, and McDearman, 2018).

Workforce analysis can inform many regional decisions

Workforce analysis—including, but not limited to analyzing occupational clusters—has become more essential to community and economic development (Nolan et al, 2011). This information can inform how:

- Economic development organizations promote their regions and build competitive advantage;
- Corporate leaders make site location decisions; and/or
- Post-secondary institutions plan educational programming.

In such instances, the occupational analysis provides a top-down view on the relative workforce strengths of a region. However, regional actors often need more targeted occupational analyses that may instead start with an individual occupation or industry and then identify related occupations or industries with similar workforce needs. These more targeted approaches may better serve the needs of individual jobseekers and counselors. Similarly, bottom-up approaches that focus more on the workforce needs of firms in specific supply chains (e.g., aircraft manufacturing) or market areas (e.g., renewable energy) can also inform broader regional analysis. Once defined, researchers can place these unique occupational clusters within the broader geographic distribution of talent and find other labor markets that possess workers with similar skill sets.

Occupational cluster frameworks offer many applications, but the inclusion of new labor market information tools and resources will bolster these frameworks moving forward. Emerging resources related to online job postings or industry recognized credentials will allow researchers to both ask and answer more specific questions related to skills and certifications and provide more current information. The responsibility for incorporating these analytical tools and building a more robust framework falls primarily on researchers, scholars, and consultants, but to be useful researchers must convey this information so that individual practitioners, employers, or educators can effectively use it to make more data-driven decisions (Kahlaf, Michaud, and Jolley, 2021).

// EXPLORATORY DATA ANALYSIS

It is mentioned previously that O*Net version 28.3 has 873 different types of occupations. Each occupation has information on the “level” and “importance” of 35 different types of skills, which are divided into seven groups. These groups include:

1. Content group includes skills such as reading comprehension;
2. Process group includes active learning;
3. Social group includes negotiation;
4. Complex problem-solving group includes the complex problem-solving skills;

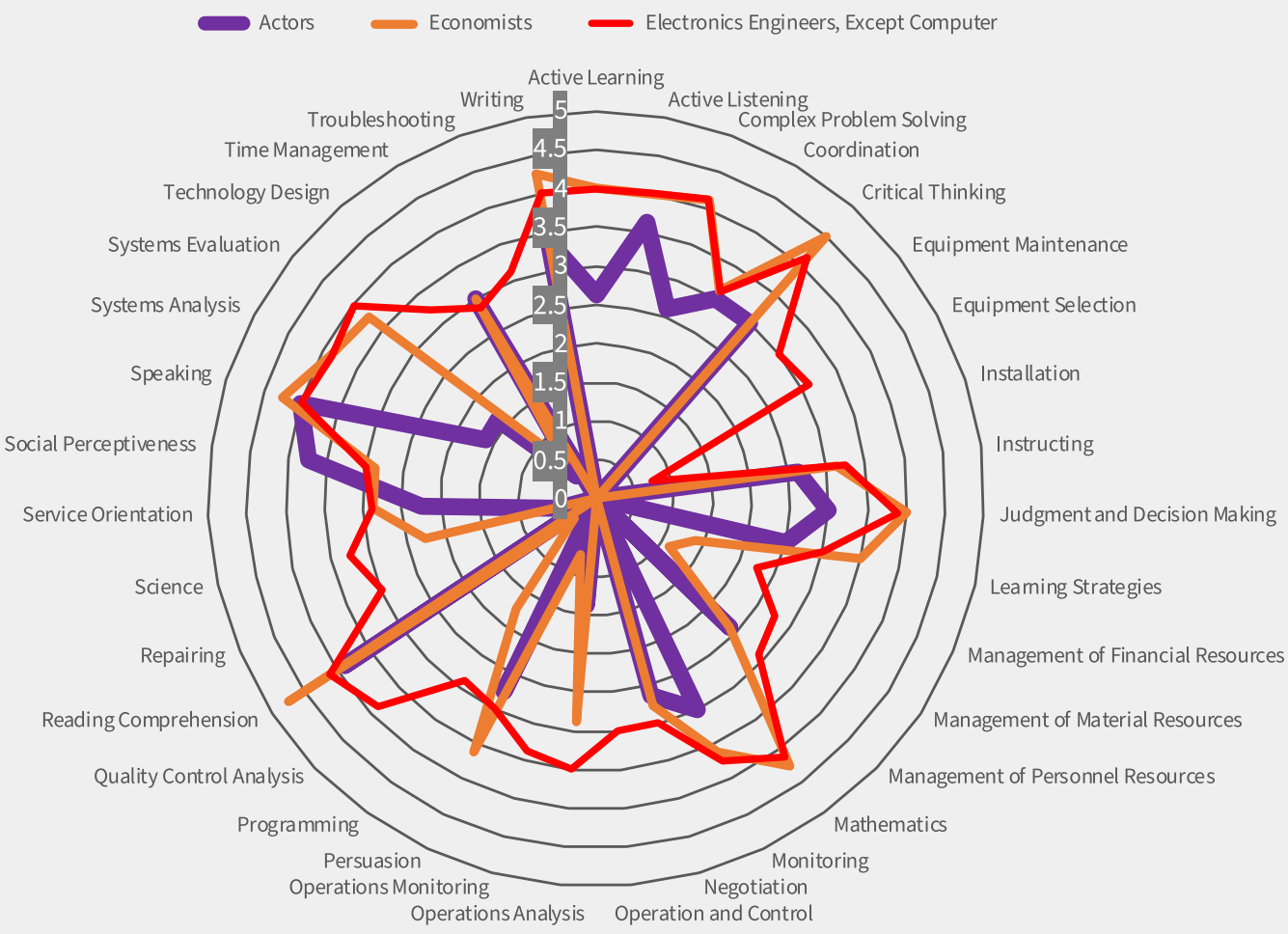
5. Technical group includes programming;

6. Systems group includes skills such as judgment and decision making; and

7. Resource management group includes time management.

In total, O*Net collects data for 70 skill-related variables, which include 35 levels and 35 importance variables for each of the 873 occupations. The amount of a particular skill needed in an occupation or a job is shown by "level," which varies from 0 (minimal or nonexistent) to 7 (proficient/expertise). The degree of requirement of a particular skill in an occupation is shown by "importance," which varies from 1 (not important) to 5 (extremely important). The data can be used to develop a skill-level or a skill-importance chart to compare occupations. For example, Figure 1 compares three different occupations—actors, economists, and electronics engineers except computer in 35 different skill levels. An economist requires higher levels of critical thinking and mathematical skills than an actor. An electronics engineer needs higher levels of equipment maintenance, equipment selection, and installation skills than economists or actors. **Figure 1** is visualizing the skill-levels of individual occupations compiled by the O*Net Program.

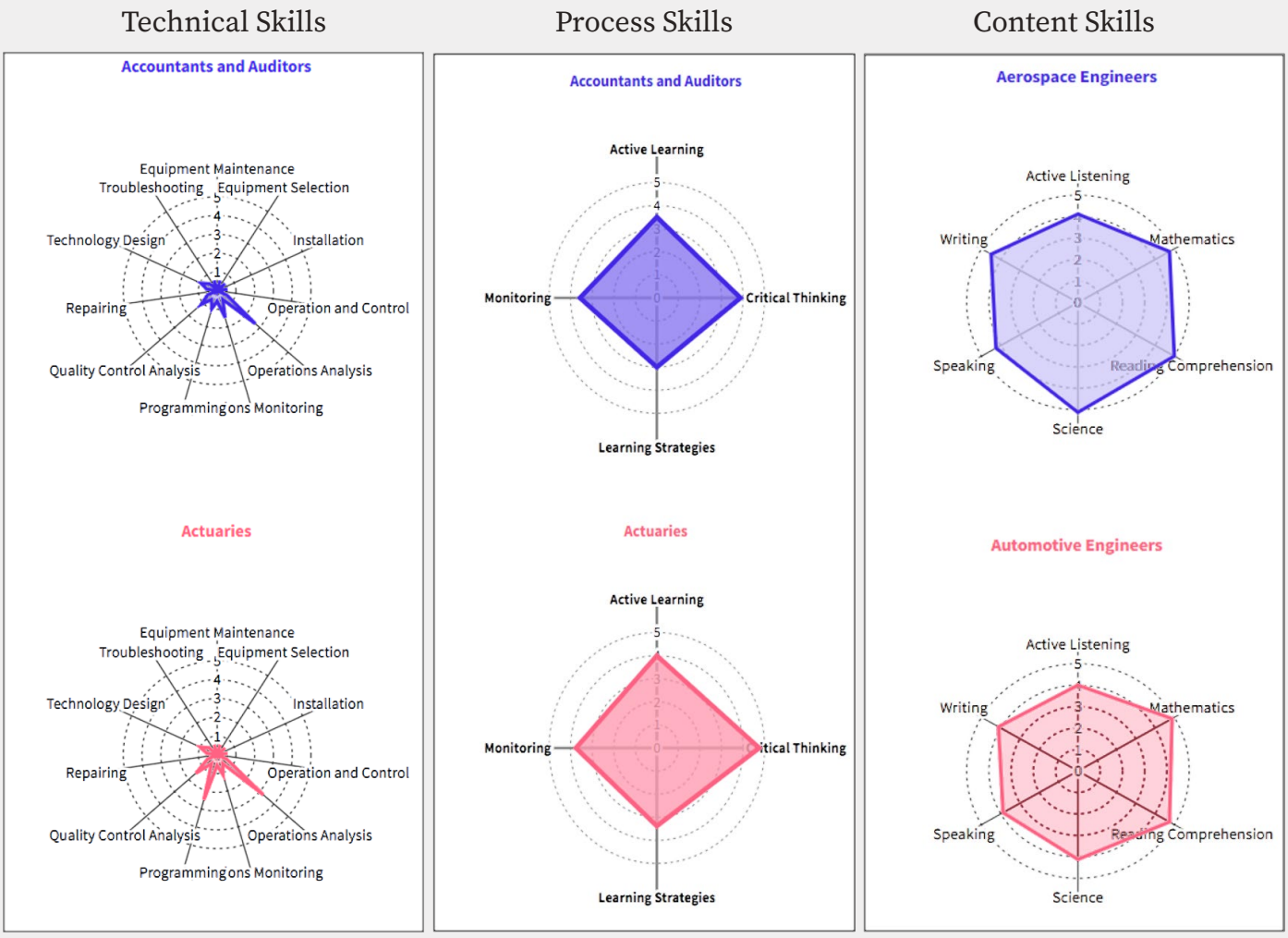
Figure 1. *Comparison of Three O*Net Occupations by 35 Skill-Levels*



Source: Prepared by PCRD using O*Net data.

The skill-level data are visualized by using online visualization tools. **Figure 2** shows visualizations of occupations by technical skill-levels; occupations by process skill-levels; and content skill-levels for various occupations. It is to be noted that O*Net occupations are at a more detailed level than the Standard Occupation Classification (SOC). For example, the O*Net version 28.3 identifies 37 different engineering occupations. In contrast, SOC identifies only 18 different engineering occupations.[13]

Figure 2. *Online Visualization of Skill-levels by Different Groups*



Source: Technical skills: https://public.flourish.studio/visualisation/18887283/
Process skills: https://public.flourish.studio/visualisation/18980940/
Content skills: https://public.flourish.studio/visualisation/18981453/

[13] https://www.bls.gov/oes/current/oes_nat.htm

Table 1 shows the descriptive statistics of 35 different skill-levels and skill-importance variables, respectively. The coefficient of variation (CV)[14] for skill-levels is 52.9%. Similarly, the CV for skill-importance is 33.6%. The CV shows that the skill-level values have more variations across the 873 occupations compared to the skill-importance values. Similarly, the inter quartile range (IQR)[15] for skill-level is 1.75 and IQR for skill-importance is 1.12. The IQR of skill-level data is higher than the skill-importance data. The variation and range statistics show that skill-level data are heterogenous and more dispersed, and hence suitable for further analysis compared to the more homogeneous skill-importance data. The descriptive statistics such as standard deviation reveal that skill-level data can capture the variations and differences between occupations better than the skill-importance data.

Table 1. *Descriptive Statistics of Skill-levels Data by Occupations*

Variables	Skill-level Values	Skill-importance Values
Average	2.38	2.59
Max	6.00	5.00
Min	0.00	1.00
Std. Deviation	1.26	0.87
3rd Quartile	3.25	3.12
2nd Quartile	2.62	2.75
1st Quartile	1.50	2.00

The 873 occupations x 35 skills matrix are incomprehensible for workforce development and policy analysis purposes. For example, if a scatter plot is developed for skill-level by skill-Importance by 873 occupations. One will need 35 different scatter plots for 35 different types of skills to study similarities and dissimilarities between occupations. Such analyses are confusing and less useful. If occupations are clustered or grouped by similarity of skill-level, it could become comprehensible and useful for planning applications as there will be fewer occupational clusters. The skills-based occupation clusters could be useful to determine career pathways and career ladders including vertical and lateral career movements and transitions undertaken by workers and professionals. Occupational clusters by skills can also be used to study the geographic concentration or specialization of skills in the region. The next section explores the methodology for clustering, which falls under the unsupervised machine learning process.

[14] Coefficient of variation or CV= $\sigma / \mu \times 100$. It is the ratio of standard deviation to the mean multiplied by 100.

[15] Inter quartile range (IQR) is Q3 (Quartile 3) minus Q1 (Quartile 1).

// METHODOLOGY

The methodology is divided into four sections. The first section presents unsupervised machine learning and its connection to clustering processes. The second explores agglomerative versus divisive clustering method. Section three delves into the hierarchical clustering method followed by a discussion on distance metrics. Section four looks into the specification methods to identify the optimal number of clusters.

Unsupervised Machine Learning

Unsupervised learning is a type of machine learning that learns from data without the use of a training dataset and human supervision. Unlike supervised learning, unsupervised machine learning models are given unclassified data, and they can discover patterns and statistics without any explicit guidance, or without any prior assumption of statistical distribution.[16] Clustering is considered as a type of unsupervised learning problem. The unsupervised clustering algorithms are classified generally into four types: (1) Exclusive clustering, (2) Overlapping clustering, (3) Hierarchical clustering, and (4) Probabilistic clustering.[17] The goal of these algorithms is to find a structure in a collection of unlabeled data and organize it into groups whose members are similar in some particular way. A cluster therefore collects similar objects together and separates dissimilar objects to other clusters (Mishra, 2024).

Agglomerative versus Divisive Clustering

Agglomerative and divisive are two primary types of hierarchical clustering methods. The first one, agglomerative hierarchical clustering, has been the popular approach to construct a fixed number of classification schemes in research and practice. This method, merges or agglomerates clusters at each step of the algorithm, constructing an output from n objects by developing a set of n-1 or a smaller number of partitions. The algorithm starts with the fine partition of n clusters, and ends with the trivial partition of one cluster, creating a bottom-up approach and a partition tree referred to as a dendrogram, which shows the topology[18] or the steps of how subsets are merged into a cluster (Murtagh and Contreras, 2012). In other words, this procedure creates a new cluster by merging the two closest clusters. The closeness is determined by computing the dissimilarity or distance between the two subclusters. The second method, the divisive hierarchical clustering, is a top-down approach, starting with the whole sample in a unique cluster and splitting it into two subclusters, which in turn are split up again and so on. Thus, at each step the two new clusters are formed by partitioning the former cluster (Roux, 2018).

[16] https://www.seldon.io/supervised-vs-unsupervised-learning-explained

[17] https://cloud.google.com/discover/what-is-unsupervised-learning

[18] Topology is a branch of mathematics that studies geometric and spatial properties and interrelationships of the objects. A dendrogram is a tree topology, which is one of the six types of network topology. The other five topologies are bus, ring, star, mesh, and hybrid.

Ward Agglomerative Hierarchical Clustering and Distance Measures

Kononenko and Kukar (2007) presented that hierarchical clustering methods were of three types, which included single linkage, average linkage, and the Ward's method. The Ward's method creates clusters by combining objects in groups such as there is minimum variance within the group or cluster, and maximum variance between the groups or clusters. Note that all of the three methods can be applied on multivariate or multidimensional data. The clustering method splits or merges subclusters based on certain measures for distances.

The three primary measures of distance include Euclidean, Manhattan, and Minkowski as shown by **Equations 1, 2, and 3** (Kopczewska, 2022). There are other kinds of distance metric concepts available, such as Mahalanobis distance, Cosine distance, Gower distance, Hamming distance, Cophenetic distance, and Levenshtein distance (Kopczewska, 2022). For the this study, Euclidean and Manhattan distances were used because previous research had used mainly both types of distance metrics. Also, the R functions were readily available in various packages. In Equations 1, 2, and 3, x_i and y_i represent i skill parameter for occupations x and y, respectively. Here, n is the number of skills, which varies from 1 to 35.

$$\text{Euclidean distance} = \sqrt[2]{\sum_{i=1}^{n}(x_i - y_i)^2} \quad (1)$$

$$\text{Manhattan distance} = \sum_{i=1}^{n}|x_i - y_i| \quad (2)$$

$$\text{Minkowski distance} = \sum_{i=1}^{n}(|x_i - y_i|^p)^{1/p} \quad (3)$$

Specification Tests

In cluster analysis, a significant challenge is to determine the optimal number of clusters because the dendrogram is simply a tree topology type revealing all the bifurcations from n clusters to one cluster.[19] The three main statistical specification tests for identifying the optimal number of clusters include Average Silhouette, Elbow, and Gap Statistic methods. The average silhouette method explores the quality of clustering by focusing on how well placed an object is within its own cluster versus other clusters.[20] A high value of average silhouette shows better clustering results. The elbow method works on minimizing the total of within cluster variances.[21] **Equation 4** shows the elbow method where C_k is the k^{th} cluster and W symbolizes within cluster variation.[22] The gap statistic method is more versatile than the average silhouette and elbow methods. The gap statistic compares

[19] Agglomerative hierarchical clustering process starts by assuming that each of the n objects is a cluster by itself. It merges occupations based on minimization of variance within and maximization of variance between clusters. It ends by putting all the n objects into one cluster.

[20] https://uc-r.github.io/kmeans_clustering#silo. It is the average of silhouette coefficients of objects varying from -1 to +1.

[21] Ibid.

[22] Ibid.

total intra-cluster variations of different number of clusters k to reference distributions generated by using the Monte Carlo simulation. The method works to maximize the difference in variations from reference data compared to the observed data. **Equation 5** shows the maximization function of the gap statistics method where E_n is the expectations from reference distribution, W_k is the cluster variation, and k is the number of clusters.[23]

$$\text{Elbow method} = \textit{minimize} \sum_{k=1}^{k} W(C_k) \tag{4}$$

$$\text{Gap statistic} = \textit{maximize}\ Gap_n(k) = E_n^* log(W_k) - log(W_k) \tag{5}$$

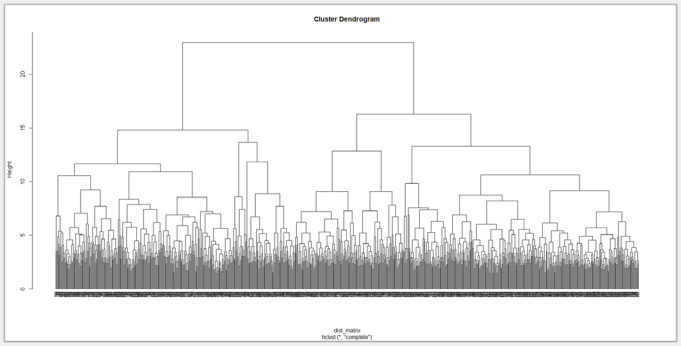
// RESULTS

Various types of distance measures such as Euclidean, Manhattan, Minkowski, etc., are discussed previously in the methodology. Similarly, various agglomerative clustering methods such as Ward, Complete, Average, and Single are also described previously. In the descriptive analysis, it was ascertained that skill-level data are more heterogenous than the skill-importance data. However, both skill-level and skill-importance data are used to run the agglomerative hierarchical clustering in R because of high correlations, and various dendrograms are created to study the pattern and sequence of clustering.

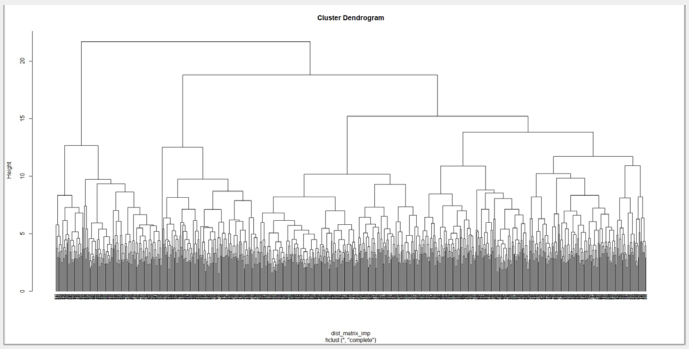
Figure 3 shows four dendrograms for skill-level and skill-importance by using Ward and Complete clustering algorithms on the Euclidean distance matrices. **Figure 4** shows four dendrograms for skill-level and skill-importance by using Ward and Complete clustering algorithms on the Manhattan distance matrices. It is evident that the choice of distance matrix can make a difference in the partitioning of occupational clusters as revealed by the dendrograms. For example, there are distinct differences between Euclidean and Manhattan for skill-level clustering using the Ward algorithm. Similarly, dendrograms for skill-level versus skill-importance data are quite different even if the same distance and clustering methods are used. A dendrogram can be visualized as a tree like structure where 873 individual occupations are at one end, and occupations are grouped based on closeness, eventually making one large group or cluster comprised of all the occupations. A tanglegram can be used to compare two different tree like structures or the two dendrograms. **Figure 5** shows the tanglegram between skill-level and skill-importance data where both use Euclidean distance and Ward clustering algorithm. The Entanglement coefficient is 0.2. The Entanglement coefficient varies from 1 to 0 with lower coefficient values showing the better alignment. In this case, it reveals distinct differences between the two dendrograms despite using the same distance and clustering methods.

[23] Ibid.

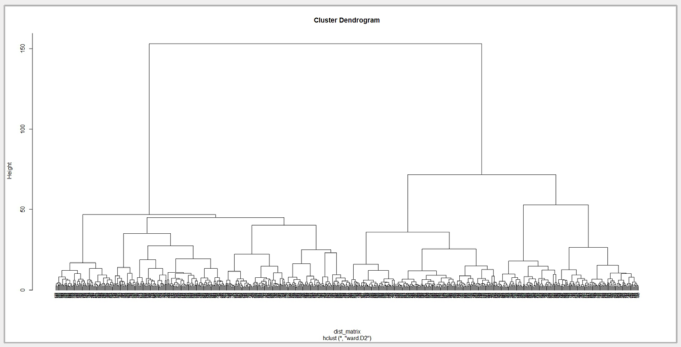
Figure 3. *Skill-level and skill-importance dendrograms 1*



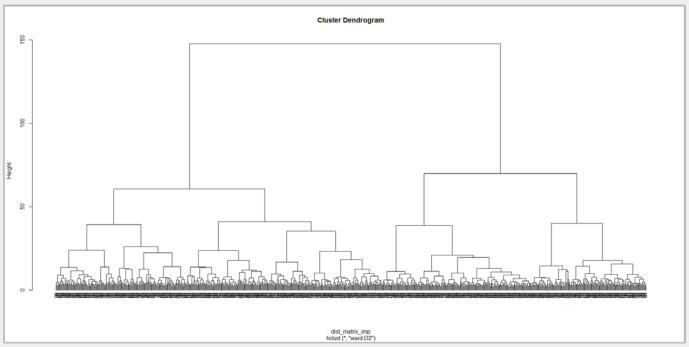
Distance Matrix Euclidean (Complete): Levels



Distance Matrix Euclidean (Complete): Importance

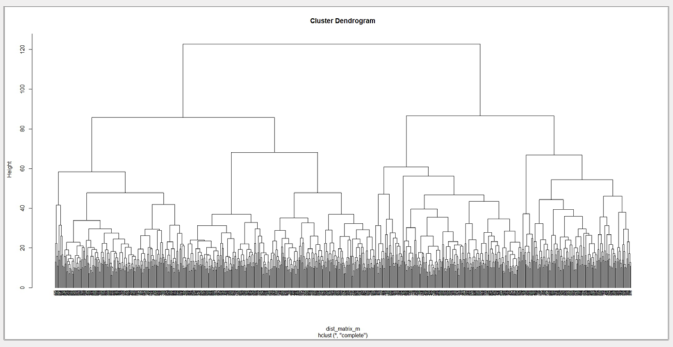


Distance Matrix Euclidean (Ward): Levels

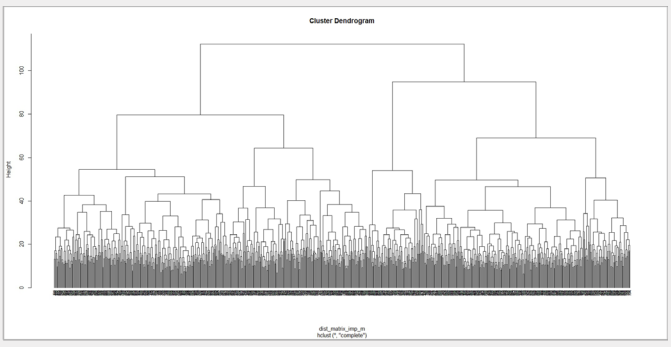


Distance Matrix Euclidean (Ward): Importance

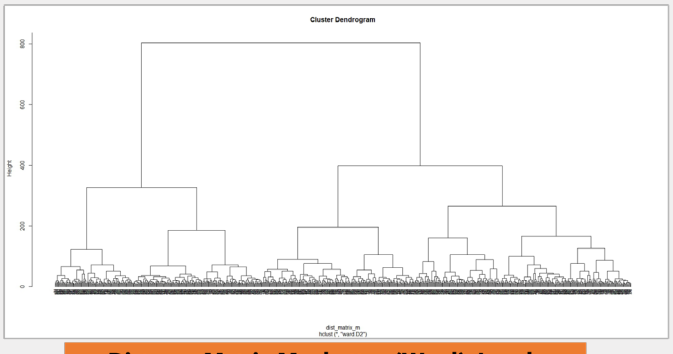
Figure 4. *Skill-level and skill-importance dendrograms 2*



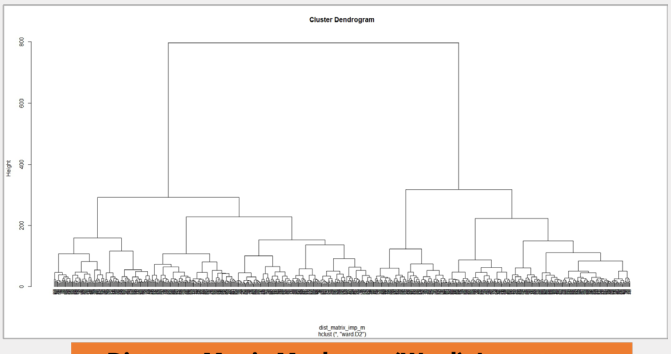
Distance Matrix Manhattan (Complete): Levels



Distance Matrix Manhattan (Complete): Importance

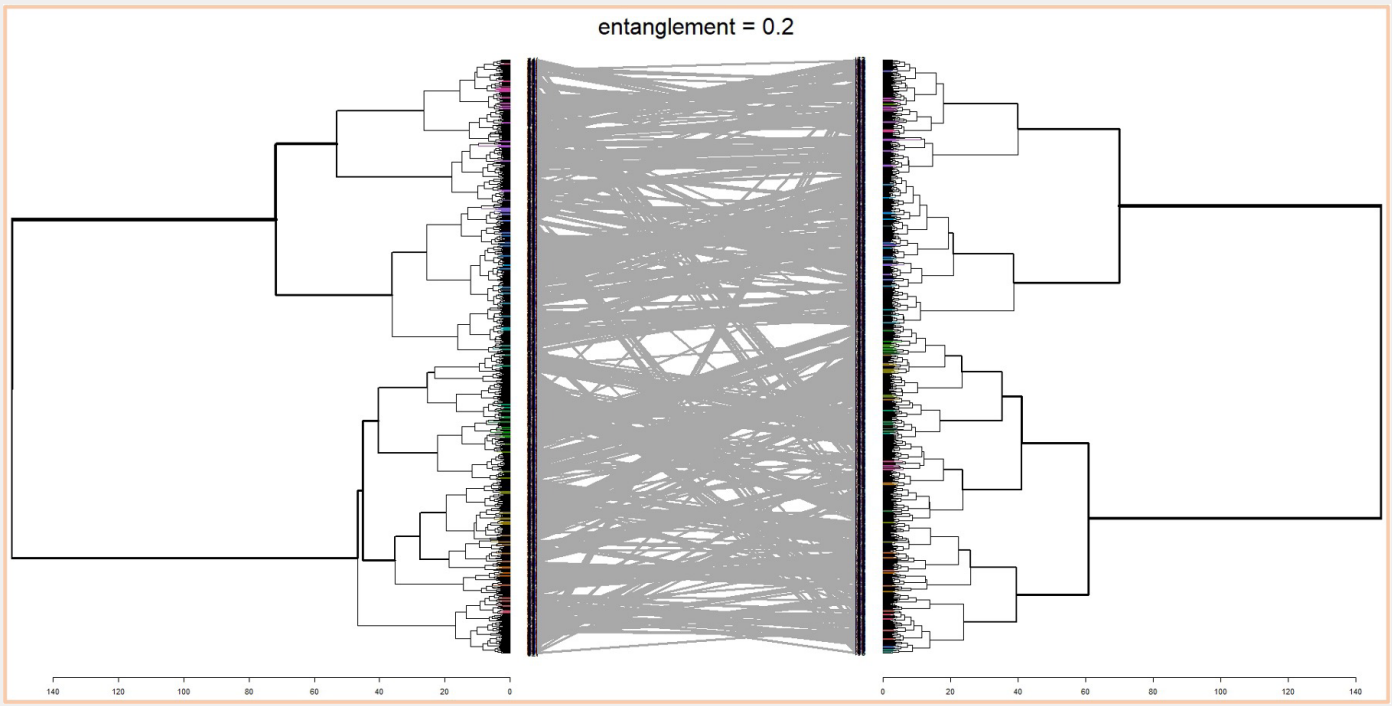


Distance Matrix Manhattan (Ward): Levels



Distance Matrix Manhattan (Ward): Importance

Figure 5. *Tanglegram between skill-level and skill-importance data*



Agglomerative coefficient values can be used to compare different hierarchical clustering methods and algorithms. **Tables 2 and 3** show the Agglomerative coefficient values for skill-importance and skill-level by using Euclidean versus Manhattan distances, and Ward versus Complete algorithms. It is evident that the Ward hierarchical clustering method outperforms the Complete hierarchical clustering method regardless of data type and distance measures. Similarly, the agglomerative coefficient for skill-level is the highest when Manhattan distance and Ward hierarchical clustering method are used. The same combination is the highest for skill-importance data as well. Table 3 shows that the agglomerative coefficient value for skill-level plus Manhattan distance plus Ward is 0.985608. The agglomerative coefficient value for skill-level plus Euclidean distance plus Ward is 0.982337. Because the difference is almost negligible, the second option of skill-level plus Euclidean distance plus Ward hierarchical clustering algorithm is used to develop the occupation by skills clusters for the U.S. As mentioned previously that descriptive statistics revealed skill-level data outperformed the skill-importance data in heterogeneity. The tanglegram revealed that the scale of the difference was smaller but distinct.

The optimal number of clusters becomes a judgement call in any kind of agglomerative hierarchical clustering analysis. If the dendrogram is cut at a higher height, it will produce lesser number of clusters which will not capture the fine differences between occupations. If a dendrogram is cut at a lower height, it will produce a larger number of clusters which could become unwieldy for practitioners. Three specification tests, Average Silhouette method, Elbow method, and Gap Statistics

method are applied to determine the optimum number of clusters. **Figure 6** shows the Average Silhouette, Elbow, and Gap Statistics results. It is evident that the Average Silhouette and Elbow methods do not reveal a useful number of clusters. However, the Gap Statistics method reveals that 44 occupation clusters could be the optimal number of clusters for the given skill-level data.

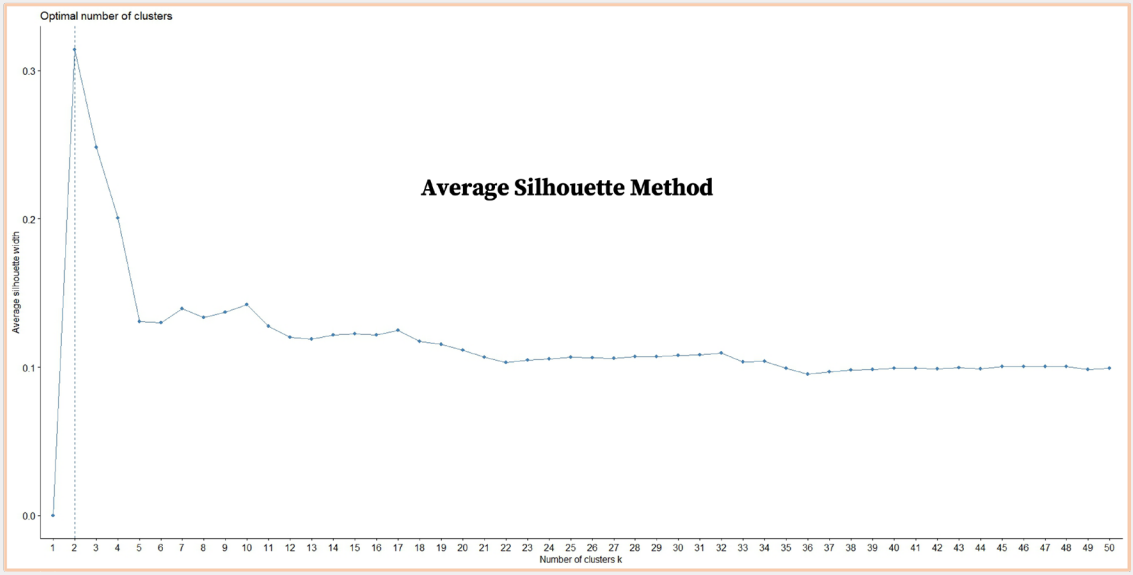
Table 2. *Agglomerative Coefficient Values for Skill-Importance*

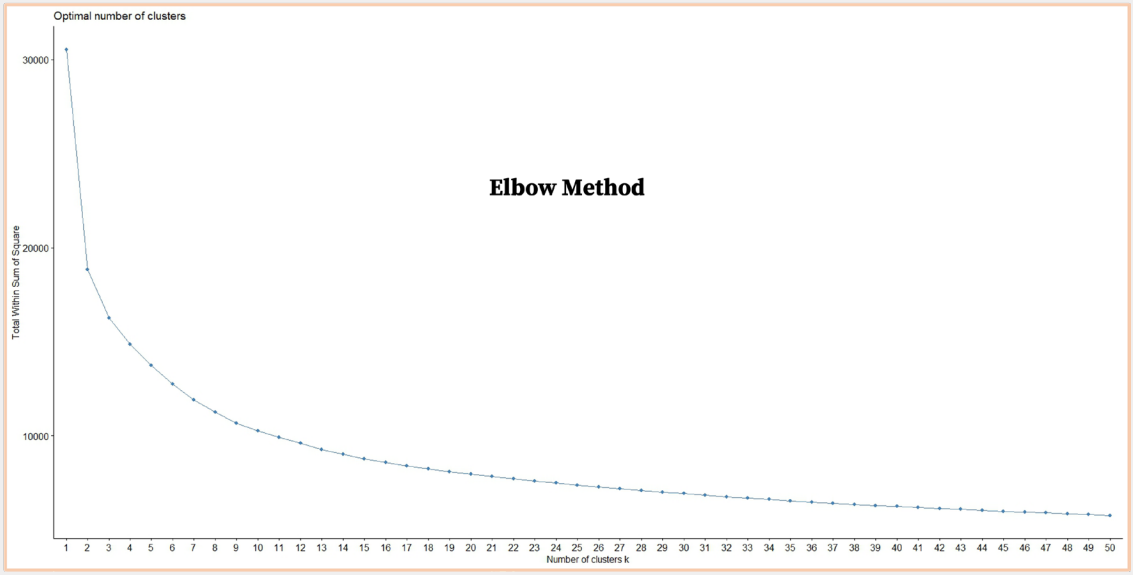
	Data	Distance Matrix	Clustering Method	Agglomerative Coefficient
1	Skill Importance	Manhatten	Ward	0.983872
2	Skill Importance	Euclidean	Ward	0.979579
3	Skill Importance	Manhatten	Complete	0.885707
4	Skill Importance	Euclidean	Complete	0.860330

Table 3. *Agglomerative Coefficient Values for Skill-Level*

	Data	Distance Matrix	Clustering Method	Agglomerative Coefficient
1	Skill Level	Manhatten	Ward	0.985608
2	Skill Level	Euclidean	Ward	0.982337
3	Skill Level	Manhatten	Complete	0.905542
4	Skill Level	Euclidean	Complete	0.882193

Figure 6. *Specifications Tests for Optimal Number of Clusters*





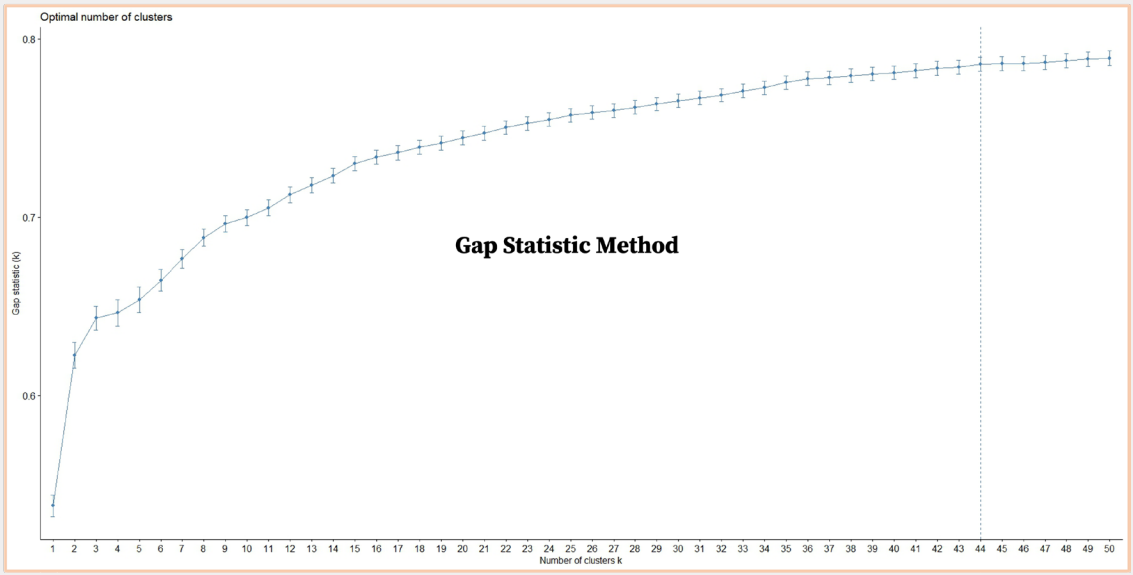


Figure 7 shows the dendrogram with different colored cut-lines on the x-axis for the 44 occupational clusters. **Figure 8** shows the number of occupations in each of the 44 clusters. In case of Clusters #35 and #44, both contain only four occupations. On the other hand, Cluster #13 has the maximum number of 53 occupations. Remaining 41 occupation clusters have less than 40 occupations. On average, an occupation cluster has about 20 occupations. The median number of occupations is 18, and the mode value is 14. It is evident that the agglomerative hierarchical clustering of skill-level data with Euclidean distance and Ward hierarchical clustering algorithm provides a diverse configuration of occupation clusters. At the same time, there is an opportunity to study the occupations within each cluster and assess if the clusters can be combined. For example, small clusters with four occupations can be merged to larger clusters with similar occupations. Hence, this analysis has provided the first set of 44 occupation clusters by skills.

Figure 7. *Dendrogram for 44 Occupation Clusters*

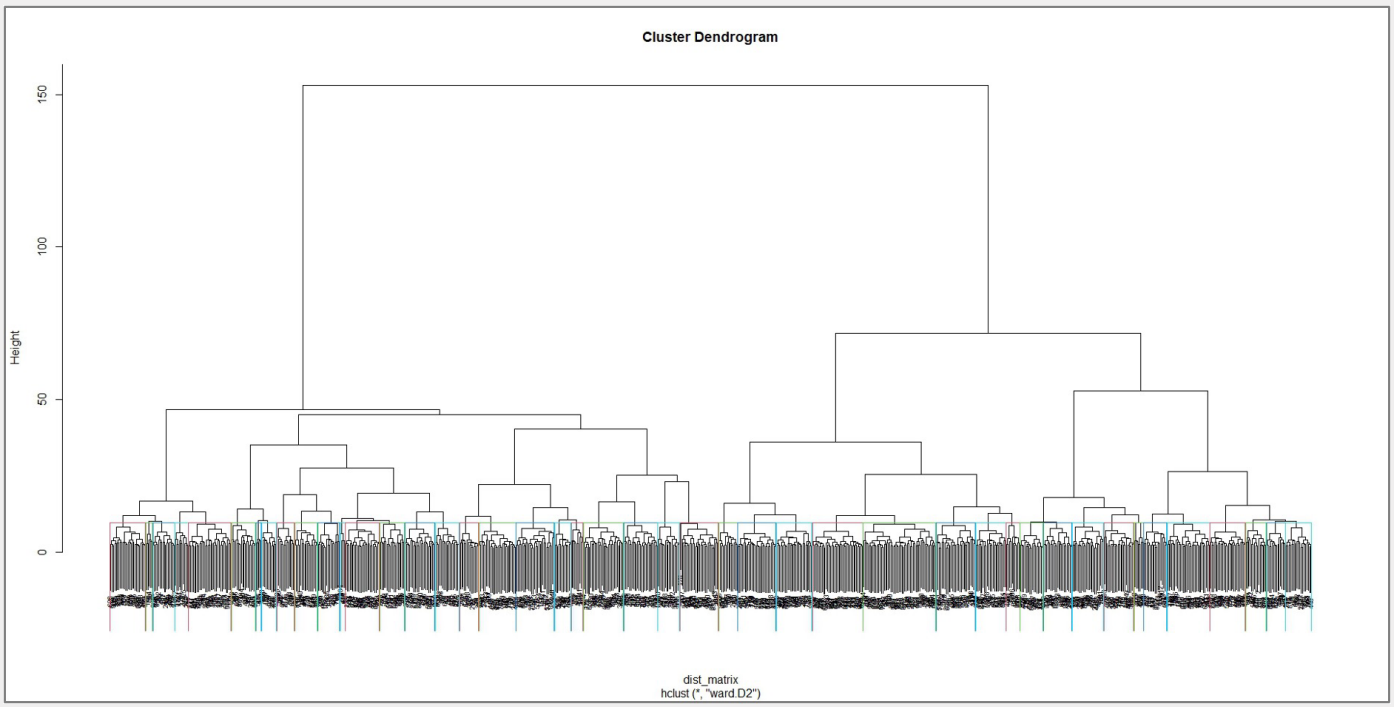
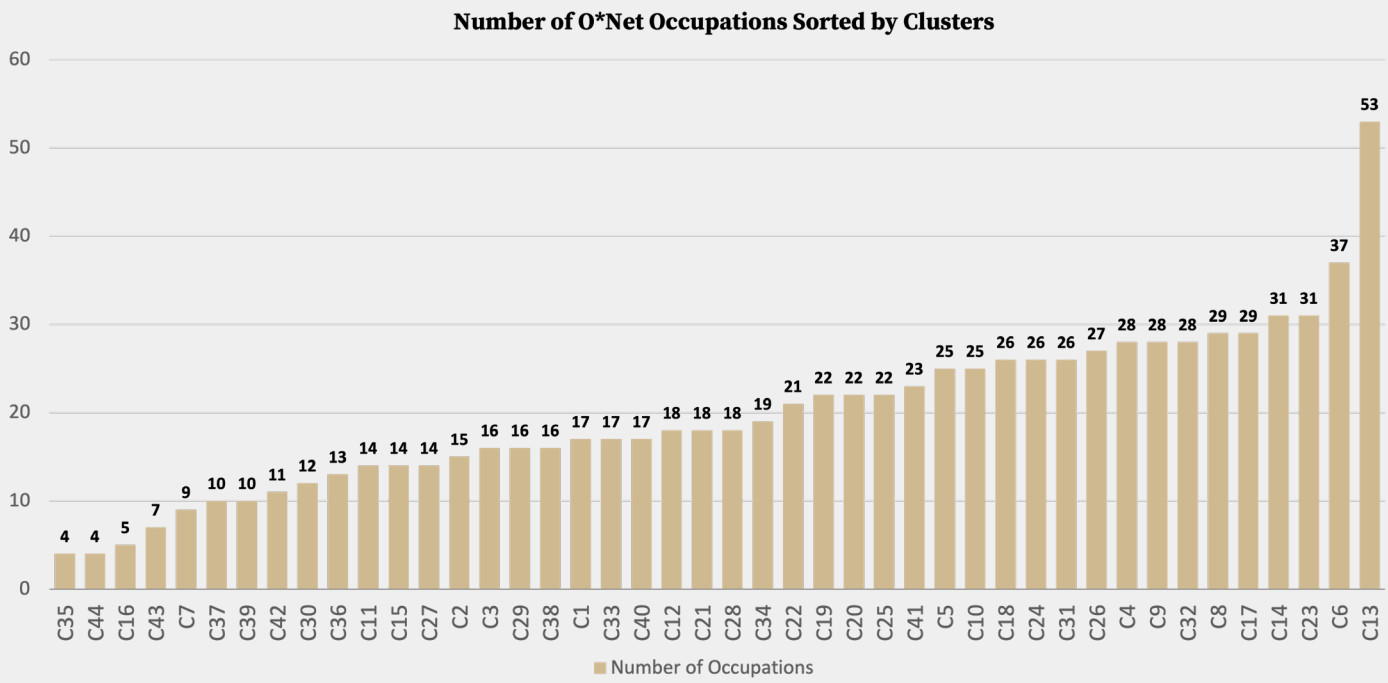


Figure 8. *Number of Occupations in 44 Occupation Clusters*



// DISCUSSIONS AND CONCLUSIONS

The analysis has provided a preliminary set of 44 occupation clusters by skills, which can be useful for the workforce development agencies, planning departments, and human resource (HR) departments of industries and businesses. Occupation clusters by skills can provide useful information for both, public and private agencies. Because these are benchmarked clusters, the cluster configurations and definitions can be used to compare peer, aspirant, and competitive regions. Note that occupations identified by O*Net are not directly comparable to the occupations included in SOC (Standard Occupation Codes), however the O*Net to SOC Taxonomy[24] information is available for research purposes and public use. The next step in this research is to finalize the occupation cluster definitions, and receive feedback from practitioners and researchers.

In the current labor market conditions, skills are important because employers have been valuing skills as much as the educational attainment. Hence, badges, micro-credentials, and online certifications have proliferated because they can make job seekers competitive in the labor market. Occupation clusters by skills can highlight groups of occupations that have similarities in terms of 35 different skills. Although, the 35 different skills cover the entire spectrum of hard-and-soft skills, they are broader skills. The inventory of skills has been expanding day-by-day as employers seek advanced skills and certifications especially in the areas of computer science, artificial intelligence, data science, semiconductors, etc. Hence, micro-credentials and badges are valued by both, the job

[24] https://www.onetcenter.org/taxonomy.html

providers and the job seekers. There are opportunities to apply this framework on more detailed and nuanced skills databases, which are the emerging areas of research.

Automation and computerization to replace human capital have been an impending challenge for both job seekers, employed workforce, and workforce development professionals (Kumar et al., 2020). For example, the dockworkers of the east coast ports were on strike in fall 2024 against automation, especially the use of automated cranes and driverless trucks.[25] Occupations by skills clusters can shed light on automation propensities of the groups of occupations. There is also an opportunity to map the occupations by skills cluster to study the spatial distribution and specialization of the groups of occupations. Similar to the clustering of occupations by skills, a clustering of skills by occupations is also feasible. A temporal analysis of skills by occupations clusters can reveal emerging changes in the nature of occupations and job activities. The analysis can also be extended to the larger group of skills provided by proprietary sources. Hence, there are several emerging research opportunities in the area of occupations and skills clusters.

[25] https://www.cnn.com/2024/10/02/business/dock-workers-strike-automation-nightcap/index.html

// REFERENCES

Chrisinger, C. K., Fowler, C. S., and Kleit, R. G. (2012). Shared Skills: Occupation Clusters for Poverty Alleviation and Economic Development in the U.S. Urban Studies, 49(15), 3403–3425. https://doi.org/10.1177/0042098011433489

Donahue, R., Parilla, J., and McDearman, B. (2018). Rethinking Cluster Initiatives. The Brookings Institution Metropolitan Policy Program.

Feser, E. J. (2003). What Regions Do Rather than Make: A Proposed Set of Knowledge-based Occupation Clusters. Urban Studies, 40(10), 1937–1958. https://doi.org/10.1080/0042098032000116059

Feser, Edward. (2005). Benchmarking Value Chain Clusters for Applied Regional Research. Regional Economics Applications Laboratory, University of Illinois Urbana-Champaign (UIUC).

Feser, Edward and Edward Bergman. (2000). National Industry Cluster Templates: A Framework for Applied Regional Cluster Analysis. Regional Studies. 34(1): 1-19.

Kadokawa, Kauzo. (2011). Applicability of Marshall's Agglomeration Theory to Indus-trial Clustering in the Japanese Manufacturing Sector: An Exploratory Factor Analysis Approach. *The Journal of Regional Analysis and Policy*. 41(2): 83-100.

Khalaf, C., Michaud, G., and Jolley, G. J. (2021). How to Assess the Transferability of Worker Skills: A Hybrid Clustering Approach [Pdf]. 731.0 kB. https://doi.org/10.22004/AG.ECON.339948

Kononenko, I., and Kumar, M. (2007). Machine Learning and Data Mining: Introduction to Principles and Algorithms. Woodshed Publishing. Cambridge: UK.

Koo, J. (2005). How to Analyze the Regional Economy With Occupation Data. Economic Development Quarterly, 19(4), 356–372. https://doi.org/10.1177/0891242405279910

Kumar, I., Beaulieu, L., Zhalnin, A., and Song, C. (2020). Occupational Competitiveness Analysis of the U.S. Transportation and Logistics Cluster. Transportation Research Record. Vol. 2674(1). https://doi.org/10.1177/0361198120901677

Markusen, A. and Barbour, E. (2003). California's Occupational Advantage. Public Policy Institute of California. Working Paper No. 2003.12.

Murtagh, F., and Contreras, P. (2012). Algorithms for Hierarchical Clustering: An Overview. WIREs Data Mining Knowledge Discovery. 2, 86-97.

Nolan, C., Morrison, E., Kumar, I., Galloway, H., and Cordes, S. (2011). Linking Industry and Occupation Clusters in Regional Economic Development. Economic Development Quarterly, 25(1), 26–35. https://doi.org/10.1177/0891242410386781

Porter, Michael. (2003). The Economic Performance of Regions. Regional Studies. 37 (6-7): 549-578.

Renski, H., Koo, J., and Feser, E. (2007). Differences in Labor versus Value Chain Industry Clusters: An Empirical Investigation. Growth and Change, 38(3), 364–395. https://doi.org/10.1111/j.1468-2257.2007.00375.x

Romer, Paul. (1990). Endogenous Technological Change. Journal of Political Economy. Vol. 98, 5:71-102.

Roux, M. (2018). A Comparative Study of Divisive and Agglomerative Hierarchical Clustering Algorithm. Journal of Classification. 35, 345-366.

Slaper, T. F. (2014). Clustering Occupations. Indiana Business Review. https://www.ibrc.indiana.edu/ibr/2014/summer/article2.html

Thompson, W. R., and Thompson, P. R. (1987). National Industries and Local Occupational Strengths: The Cross-Hairs of Targeting. Urban Studies, 24(6), 547–560. https://doi.org/10.1080/00420988720080781

Unsupervised Clustering: A Guide. Mishra, S. (2024). Available at https://builtin.com/articles/unsupervised-clustering

Publication 112 October 2024

RESEARCH & POLICY

INSIGHTS

Discover More Findings at:

www.pcrd.purdue.edu/publications

Center for Regional Development

Report Design by Ryan Maluchnik, Purdue Center for Regional Development