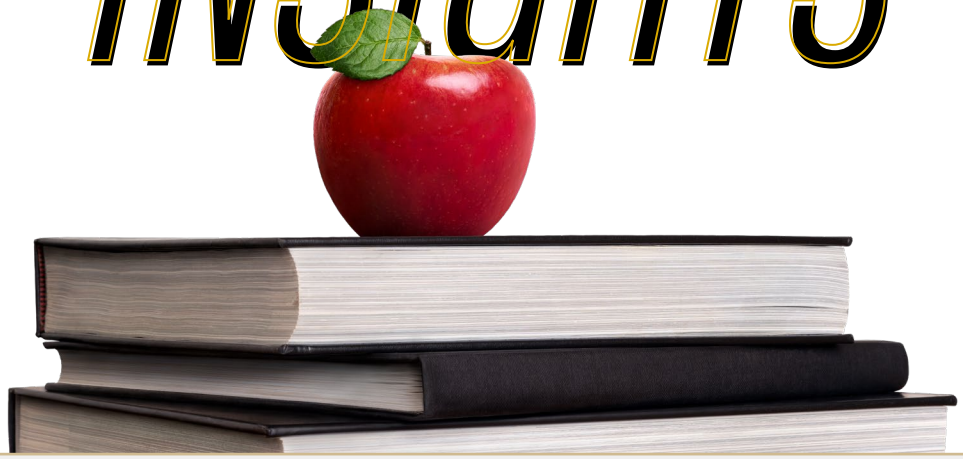
RESEARCH & POLICY

INSIGHTS



Knowledge Economy and Knowledge-based Occupation Clusters for the U.S.

Publication 117

September 2025

Co-Authors

Indraneel Kumar, PhD
Taoyi Sun
Benjamin St. Germain
Andrey Zhalnin, PhD
Bertin Mbongo

PURDUE UNIVERSITY | Center for Regional Development

Economic Clusters for the 21st Century

Disclaimer: This technical report 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 author(s) and do not necessarily reflect the views of Economic Development Administration or the U.S. Department of Commerce.

ABSTRACT

The global economy is moving away from production and industrial economy to services, and further to the knowledge economy. Data and metrics are needed to understand the knowledge workers' component of the regional labor force. Knowledge-based occupation clusters provide a framework to understand the knowledge workers. These clusters are comprised of occupations requiring above high school education, better training, and preparation. This technical and policy insights report describes the methodology and statistical tests in detail, and lays out the groundwork for how workforce practitioners can use these data. It is anticipated that the knowledge-based occupation clusters will complement the industry clusters to understand the regional competitiveness.

INTRODUCTION

This technical and policy insight report explores occupational cluster definitions based on "knowledge" domain of the occupations. Human capital development through talent pipeline and workforce development is an important component of regional economic development. The knowledge-based occupation clusters provide opportunities to assess concentrations of knowledge-workers in the region. These occupation clusters complement the skills-based occupation clusters and industry clusters that are developed as part of the Economic Development Administration (EDA) Updating Industry Cluster Grant 2023-2025. These data would enable regions to visualize and analyze specializations not only in industry clusters but also for skills and knowledge occupation clusters. Every region is unique because of the combinations of assets, and regional strengths and gaps. It is anticipated that regions are also unique in their combinations of specializations in knowledge, skills, and industry clusters. The knowledge-based occupation clusters enable to assess the knowledge-workers in the region.

"Knowledge capital" has unique economic characteristics because unlike other goods and products, it does not diminish due to consumption (Hogan, 2011). Similarly, several users or knowledge-workers can use the same knowledge simultaneously at different locations without affecting the stock. The use of knowledge, under the right circumstances, further increases the knowledge by increasing expertise and know-how in the knowledge-workers (Hogan, 2011). As the global economy shifts from industrial to a knowledge economy, the economic principles are changing. In a knowledge economy, the knowledge is both an input and an output; principles of diminishing returns and economies of scale do not apply; and human capital and knowledge-

[1] https://dl.acm.org/doi/abs/10.1145/3517745.3561441

[2] https://www.sciencedirect.com/science/article/pii/S0143622814000782?via%3Dihub

workers become a significant part of the production process.[1] Scholars have asserted that industrialization has changed the "capital-to-labor" ratios, increasing the labor productivity manifolds (Abis and Veldkamp, 2022). Similarly, in a knowledge economy, "data-to-labor" ratio would change significantly with each unit of labor consuming and processing larger amounts of data and information (Abis and Veldkamp, 2022). Characteristics of the knowledge economy include "knowledge as a factor of production," globalization, attracting and retaining talent, open innovation, borderless collaboration, and changes in the management style from people-management to knowledge-management (Hawamdih, Kim and Wang, 2023). Scholars have suggested that the metrics of a knowledge economy include industry-based measures, such as knowledge-intensive industries or occupation-based measures, such as knowledge-workers (Hogan, 2011). A key difference is that not all people employed in a knowledge-intensive industry are knowledge workers. Similarly, knowledge-workers can be distributed into both, knowledge-intensive and other kinds of industries.

The knowledge-based occupation cluster(s) is a way of measuring knowledge-workers in the regional economy. Here, occupations requiring educational attainment of associates, bachelor's, master's, and doctoral and professional degrees are clustered based on the knowledge domains and knowledge characteristics of the occupations. The value of knowledge-workers in economic development of regions is revealed by Annalee Saxenian's seminal work on social and professional networks of information technology (IT) workers in Silicon Valley, California (Saxenian, 1996). These informal but close-knit networks between IT professionals are facilitated by the nonhierarchical business culture and open innovation, accelerating the economic growth in Silicon Valley (Saxenian, 1996). This shows the important role of a knowledge-worker in today's knowledge economy. The knowledge-based occupation clusters can enable regions to discover specializations or concentrations of knowledge-workers within their jurisdictions. A concentration of knowledge-workers can enhance economic competitiveness of the regions.

LITERATURE REVIEW

Romer (1990) discovered that the stock and quality of human capital and its capacity for invention motivated by market incentives, determined the economic growth and productivity in regions. A significant insight was that the economic growth and productivity could be accelerated endogenously through quality human capital without any exogenous stimuli (Romer, 1990). This seminal research laid the ground work for knowledge workers and the skilled labor force as the key drivers of regional economic development globally. Occupational clusters offer a useful framework for understanding a region's workforce capacity and economic opportunities by clearly distinguishing what businesses make and what workers do (Thompson and Thompson, 1987).

[1] Knowledge Economy, Office of the University Economist, Arizona State University, https://economist.asu.edu/p3-productivity-prosperity-project/knowledge-economy.

These occupational constructs often rely on the U.S. Bureau of Labor Statistics' (BLS) national industry-occupation staffing patterns and data from Occupational Information Network (O*Net) database. They provide a valuable complement to the more commonly used NAICS[2]-based industry cluster[3] approaches, which primarily focus on what businesses do.

Additionally, workforce analysis including occupational clusters has become central to regional planning, informing how economic developers promote regions, how businesses choose locations, and/or how educators plan post-secondary programs (Kumar et al., 2024). Occupational clusters group occupations based on skills or knowledge characteristics (e.g., Feser, 2003; Haas et al., 2001; Khalaf et al., 2021, Koo, 2005; Nolan et al., 2011; Slaper, 2014), enabling the regions to uncover strengths and weaknesses in human capital and workforce capacity. This allows policymakers and practitioners to align the efforts for workforce development with the economic development opportunities in their regions.

To further explore what workers performed, beyond what businesses made or industries manufactured, occupations by skills and knowledge clusters could be utilized to understand embedded competitiveness in terms of human capital including long-term innovation potentials. The skills-clusters provide insights into "what skills workers have," whereas the knowledge-clusters help identify "what knowledge workers have." See Kumar et al. (2024) for the methodological framework for defining skills-based occupation clusters, which can inform the strengths and gaps in the skills of the available labor force in the regions. In comparison, knowledge-based occupation clusters[4] are particularly useful to assess the strengths and gaps in knowledge embedded within the regional workforce.

Methodologies for categorizing occupations for knowledge-based occupation clusters have been developed in previous research by Feser (2003), Koo (2005), Nolan et al. (2011) and Slaper (2014). These methodologies provide valuable insights into:

i) regional strengths and competitiveness, particularly in specific knowledge domains to align with the human capital and labor market trends, and

ii) economic development strategies, especially fostering specific industries and shaping workforce and talent development policies.

Specifically, Feser (2003) proposed a framework to identify occupations by knowledge clusters, emphasizing the importance of understanding relationships among occupations for meaningful regional economic analysis. Based on Feser (2003) and Koo (2005) further extended this methodology by developing practical tools for regional analysis, occupation cluster analysis, and occupation-based industry-targeting analysis. However, both only focused on knowledge-intensive and high-technology industries. Further, Nolan et al. (2011) offered a broad applicable tool that integrated occupation and industry clusters in a region. The authors emphasized occupations from

[2] North American Industry Classification System.

[3] See a report by PCRD for a comprehensive literature review for industry cluster and related topics: https://pcrd.purdue.edu/wp-content/uploads/2025/04/Literature-Review_04-21-2025.pdf.

[4] Occupations by knowledge clusters are referred to as knowledge-based occupational clusters in Feser (2003).

Job Zones 3, 4 and 5 of O*Net that required higher level of preparation[5] for an occupation. These occupations demand progressively greater levels of knowledge, preparation, and training including alignment of education and training with the workforce needs. While the knowledge clusters approach might have excluded a large number of the U.S. workforce (those in Job Zones 1 and 2), it allowed for a more targeted focus on the knowledge attributes of the occupations. In contrast, Slaper (2014) expanded the focus of job zones by calculating cluster-level Job Zone averages, showing how clusters aligned with industry presence and workforce needs. However, because the goal of this report is to identify "what workers know" and support education and training planning, workforce development, and regional innovation strategies, we utilize the Job Zones 3, 4 and 5 in the occupational cluster analysis.

The previous research on knowledge-based occupations clusters has used Ward's clustering algorithm developed by Ward (1963) to identify clusters in large multivariate datasets of occupations. It is a hierarchical and agglomerative clustering technique that groups data by minimizing within-cluster variance by grouping observations that have the smallest total differences based on measurements of knowledge characteristics. This is widely used in occupational research because it offers interpretable results, supports policy applications, and doesn't require pre-specifying the number of clusters. One limitation of this algorithm is that it is sensitive to outliers and can produce biased results if there are non-random patterns of missing data or if the underlying assumptions of normality and equal variances are violated. In this context, outliers can be addressed either by removing them, as suggested by Nolan et al. (2011), or by reviewing preliminary results for consistency and reasonableness, as recommended by Slaper (2014) to ensure they are not influenced by outlier observations.

DATA PREPARATION AND EXPLORATORY DATA ANALYSIS

The data on knowledge characteristics[6] are available from the O*Net (Occupational Information Network). For each occupation, O*Net collects information on 33 different types of knowledge domains. The examples of knowledge domains include sociology and anthropology, production and processing, physics, medicine and dentistry, mathematics, mechanical, fine arts, engineering and technology, etc. O*Net divides the 33 knowledge domains into 10 groups, which include arts and humanities, business and management, communications, education and training, engineering and technology, health services, law and public safety, manufacturing and production, mathematics and science, and transportation. O*Net identifies two parameters, importance and level, for each knowledge domain for each occupation by surveying people employed in specific occupations

[5] A Job Zone is a group of occupations that are similar in: i) how much education people need to do the work, b) how much related experience people need to do the work, and c) how much on-the-job training people need to do the work (Source: O*Net, more information can be found on https://www.onetonline.org/help/online/zones).

[6] https://www.onetonline.org/find/descriptor/browse/2.C/2.C.7/2.C.1/2.C.9/2.C.3/2.C.5/2.C.8/2.C.2/2.C.4.

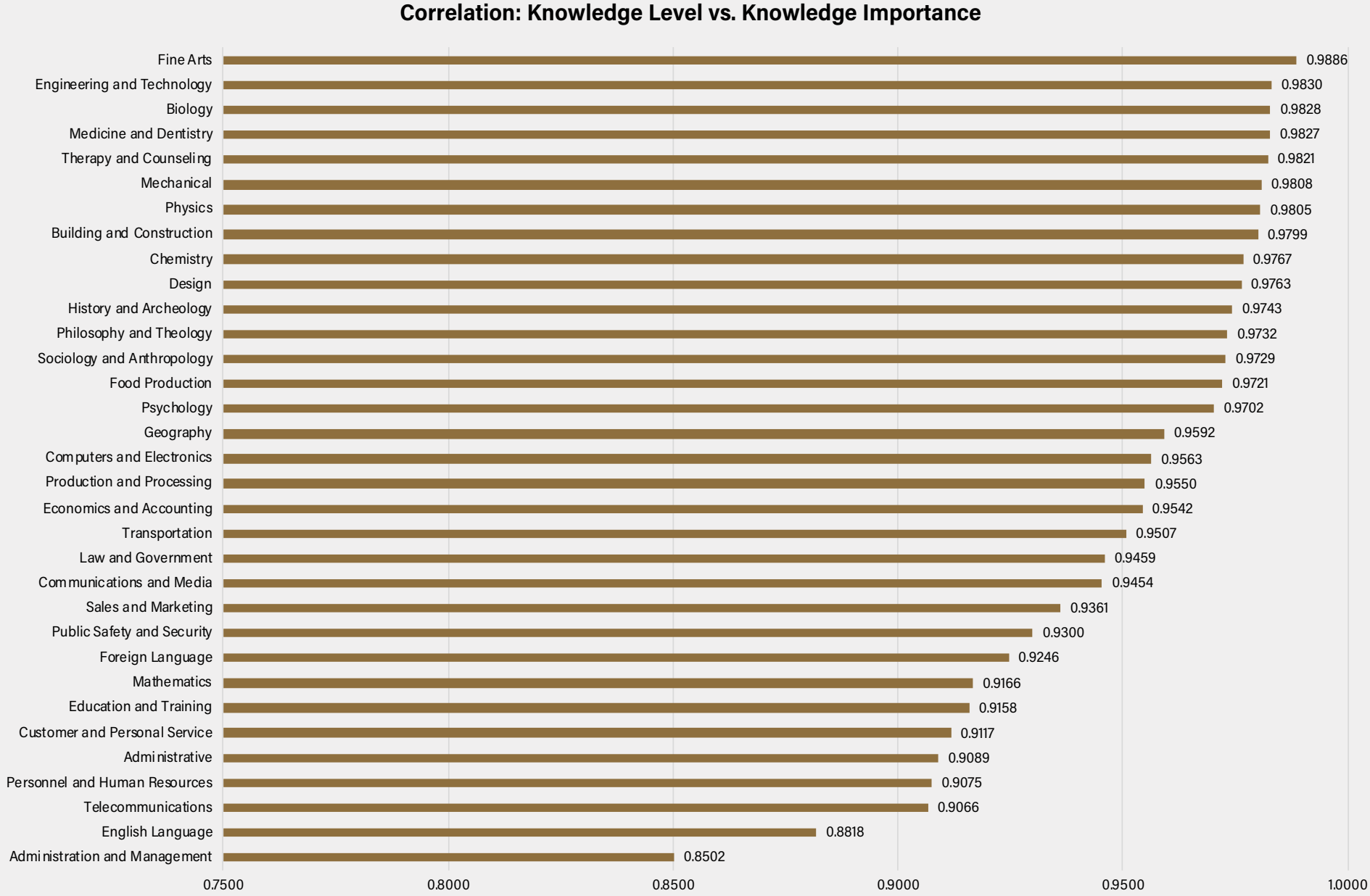
or jobs. The amount of the knowledge needed is shown by "Level," which varies from 0 (minimal or not-needed) to 7 (proficient or expert). The degree of importance for a particular knowledge to perform the occupational task is shown by "Importance," which varies from 1 (unimportant) to 5 (extremely important). For the data available from O*Net Version 28.3, the minimum and maximum value for level was 0 and 6.95, respectively. Whereas, the minimum and maximum values for importance was 1 and 5, respectively. The importance and level of knowledge characteristics include two matrices of dimensions 33 by 873 (knowledge domains by occupations).

The correlations between knowledge-level and knowledge-importance of 33 knowledge domains based on data from 873 occupations vary between 0.85 for administration and management to 0.99 for fine arts. Engineering and technology, biology, medicine and dentistry, therapy and counseling, mechanical, and physics have correlation values of 0.98, respectively. This means for majority of occupations the knowledge domains are not only important but also needed at higher levels. Refer to Figure 1 for correlation values for all the 33 knowledge domains. The descriptive statistics reveal that the average correlation value is 0.95 with a low Coefficient of Variation of 3.56%. Most of the STEM domains revealed high correlation values between knowledge-importance and knowledge-level. This indicates that either level or importance can be used as a characteristic to group the occupations. Refer to Figure 2 for the three tables of descriptive statistics for knowledge-level, knowledge-importance, and correlations between level and importance. The knowledge-level has a Coefficient of Variation (CV) value of 69.8% compared to 40.8% value for the knowledge-importance. This indicates that even though level and importance values are correlated by knowledge domains, the knowledge-level has more variation than knowledge-importance. The Inter Quartile Range (IQR) is also higher at 2.31 for knowledge-levels than 1.41 for knowledge-importance. Overall, knowledge-levels have higher relative variability around the mean because of the higher CV, and also higher absolute variability because of the higher IQR or more spread of the middle 50% of the distribution. Since the objective is to identify distinct clusters of occupations based on knowledge domains, the knowledge-level is used as the representative data.

O*Net classifies occupations into job zones based on the level of preparation needed to perform that particular occupation. The job zones are divided from 1, little or no preparation needed to 5, which means extensive preparation needed. The preparation means levels of education, experience, and on/off the job training needed to perform the work.[7] Table 1 shows the distribution of job zones into the number of occupations. For knowledge-based occupation clusters, Job Zones 3, 4 and 5 are selected which contain occupations requiring medium, considerable, and extensive preparation. Hence, the cluster analysis is performed for **63.8% (557 out of 873 occupations)** of the total occupations. This ensures that occupations requiring associates, bachelor's, master's, doctoral or professional degrees are selected for cluster analysis. Based on the literature review, the knowledge-level are squared to use further in cluster analysis.

[7] https://www.onetonline.org/find/zone?z=0

FIGURE 1. Correlations between Knowledge-level and Knowledge-importance



Source: O*Net 28.3 data processed by authors

FIGURE 2. Descriptive Statistics: Knowledge-level, Knowledge-importance, and Correlation between level and importance

Knowledge-Level	Value	Knowledge-Importance	Value	Correlations of Level & Importance	Value
Average	2.11	Average	2.27	Average	0.95
Max	6.95	Max	5.00	Max	0.99
Min	0.00	Min	1.00	Min	0.85
Std. Deviation	1.47	Std. Deviation	0.93	Std. Deviation	0.03
3rd Quartile	3.17	3rd Quartile	2.90	3rd Quartile	0.98
2nd Quartile	1.88	2nd Quartile	2.11	2nd Quartile	0.96
1st Quartile	0.86	1st Quartile	1.49	1st Quartile	0.92
Coefficient of Variation (CV)	69.84%	Coefficient of Variation (CV)	40.83%	Coefficient of Variation (CV)	3.61%
Inter Quartile Range (IQR)	2.31	Inter Quartile Range (IQR)	1.41	Inter Quartile Range (IQR)	0.05

Source: O*Net 28.3 data processed by authors

TABLE 1. Occupational Distribution by Job Zones

Job Zone	Description	Number of Occupations	% Age
1	Little or no preparation needed	31	3.6%
2	Some preparation needed	285	32.6%
3	**Medium preparation needed**	**207**	23.7%
4	**Considerable preparation needed**	**202**	23.1%
5	**Extensive preparation needed**	**148**	17.0%
	TOTAL	873	100.0%

The occupations are based on version O*Net 28.3

The cluster analysis is for 557 (63.8%) occupations from Job Zones 3, 4 and 5

METHODOLOGY AND INTERMEDIATE RESULTS

Clustering is a way of classifying or making groups from objects by utilizing data related to their characteristics and features. As per Everitt et al. (2011), classification is a common methodology used in various disciplines such as biology for classifying organisms and species, market research for developing consumer segments, and psychology for classifying patients. The clustering algorithms can be grouped into two major types, partitioning and hierarchical methods, which include several specific algorithms (Kaufman and Rousseeuw, 1990). Partitioning method can divide the data with *n* objects into *k* groups, however the user is required to provide the *k*. Hierarchical method explores different configurations of groups or *k*, and provides an option for a user to either choose *k* or select an optimal value of *k* based on specific tests. Partitioning methods include Partitioning Around Medoids (**PAM**), Clustering Large Applications (**CLARA**), and **Fuzzy Analysis** (Kaufman and Rousseeuw, 1990). The **PAM** method focuses on identifying *k* representative objects or medoids, and clusters are formed by grouping objects to their nearest medoids. **CLARA** is similar to **PAM** in that it searches for optimal locations of medoids. However, it is specifically developed for very large datasets, and applies **PAM** on samples instead to the entire data. **Fuzzy Analysis** algorithm provides membership coefficients for each object to each representative medoid, and hence generates large output datasets. A user is expected to choose the group with the highest membership coefficient for a particular object or choose several groups having higher membership coefficients for the same object. **PAM** and **CLARA** generate mutually exclusive groups or clusters whereas, **Fuzzy Analysis** generates overlapping groups or clusters.

Hierarchical methods explore groups or clusters *k* of all sizes varying from *n* or all objects to 1 and everything else in between (Kaufman and Rousseeuw, 1990). Unlike partition methods, a user can choose *k* in hierarchical methods as well as select an optimal *k* or number of clusters based on statistical tests. Two major hierarchical methods include agglomerative and divisive hierarchical clustering, and both classify a collection of objects into groups albeit in opposite directions (Kaufman and Rousseeuw, 1990). Agglomerative hierarchical methods include the Agglomerative Nesting (**AGNES**) and Divisive Analysis (**DIANA**), and two subbranch within Diana known as Monothetic Analysis (**MONA**) and **Polythetic method** (Kaufman and Rousseeuw, 1990; Everitt et al., 2011). In the **AGNES** algorithm, each object is considered a cluster by itself in the beginning, and two closest objects are merged or fused into one cluster, and the process continues until all objects are merged into one trivial cluster (Kumar et al., 2024). The hierarchical agglomerative clustering can vary depending on different types of linkage. These include single linkage or the nearest neighbor; complete linkage or the farthest neighbor; group average linkage; centroid linkage; weighted average linkage; median linkage; and **Ward's** method where variances within the group is minimized and between the groups are maximized (Everitt et al., 2011). The **DIANA** algorithm begins with all objects in a single cluster and consecutive splitting of each cluster subsequently in each step, ending when each object is a cluster by itself (Kaufman and Rousseeuw, 1990). It is computationally intensive if all combinations of *n* objects or $2^{n-1}-1$ possibilities are attempted (Everitt et al., 2011). The **MONA** algorithm or Monothetic Analysis simplifies computational burden by using only one variable for splitting the clusters at each step.

In contrast, the **Polythetic Method** uses all variables at each step. The clustering methods merge or split occupational subclusters based on distances that specifically measured as Euclidean[8], Manhattan[9], or Minkowski[10] (Kumar et al., 2024; Kopczewska, 2022).

It is evident from Exploratory Data Analysis that knowledge-levels have more variability. The level values are squared prior to applying the clustering method. It is anticipated that squaring the knowledge-level values should expand the distances and facilitate the identification of the clusters (PCRD, 2009). The Agglomerative Hierarchical Clustering with Euclidean distances and Ward clustering method is used to develop occupational clusters based on squared knowledge-levels. This is based on the Agglomerative Coefficient values calculated for different combinations of distances and clustering methods as shown in Table 2. **Euclidean + Ward + AGNES** (Agglomerative Hierarchical) provides the 2nd highest Agglomerative Coefficient value of 0.95 as per Table 2. **Manhattan + Ward + AGNES** has the 1st rank with a value of 0.96 as per Table 2. Prior research on clustering of occupations had used the agglomerative hierarchical method with the Ward algorithm (Feser, 2003; Koo, 2005; and Nolan et al., 2011). It is evident from agglomerative coefficient values in Table 2 that other combinations of clustering methods cannot produce stronger cluster structure.

Three different specification tests, **Elbow**, **Average Silhouette**, and **Gap Statistic** were used to identify the appropriate number of clusters with limited results. Figures 3, 4, and 5 show results from **Elbow**, **Average Silhouette**, and **Gap Statistic** methods, respectively. The **Elbow** method minimizes the within cluster variance based on Equation 1, where W is the variance and C_k is the k_{th} cluster. **Silhouette score** for an occupation is based on the difference between inter-cluster and intra-cluster distances, where the inter-cluster means the nearest cluster, and distances to all occupations in the nearest cluster and all occupations within the cluster is measured. The Average **Silhouette score** is the mean of **Silhouette scores** of all the occupations, providing a significance value for the number of clusters. Equations 2 and 3 show the Silhouette score and **Average Silhouette score**, respectively. In Equation 2, S_i is the **Silhouette score** for an occupation i, b_i is distance to the nearest cluster, and a_i is distance within the cluster. If S_i is close to +1 clustering is robust, if it is close to 0 means there are no clusters, and less than 0 means misspecification. In Equation 3, S is the mean of Silhouette scores of all occupations in a given cluster and n is the number of clusters.

The **Gap Statistic** method attempts to maximize the difference between total intra cluster or within cluster dispersion to the expected[11] dispersion from a reference distribution generated randomly with null clusters (Kumar et al., 2024; Tibshirani, Walther and Hastie, 2001). The number of clusters giving the maximum difference as per the maximizing equation should be the optimal clusters as per the Gap Statistic method. Equation 4 shows the intra cluster or within cluster variation, which is the dispersion. In Equation 4, x_i and x_j are two occupations, n_r is the number of occupations in cluster r, k is the number of clusters. Equation 4 measures first the intra cluster or within cluster

[8] Euclidean distance = $\sqrt[2]{\sum_{i=1}^{n}(X_i - Y_i)^2}$; where i is the knowledge parameter for occupations *X* and *Y.*

[9] Manhattan distance = $\sum_{i=1}^{n}|X_i - Y_i|$

[10] Minkowski distance = $\sum_{i=1}^{n}(|X_i - Y_i|^p)^{1/p}$

[11] Expected value means average of mean of reference distributions based on *k*, where *k* is the number of clusters.

dispersion or distance followed by the sum of dispersion of all the clusters shown by W_k. Equation 5 shows the **Gap Statistic** maximization function, where E_n is the reference distribution expectation and W_k is the sum of cluster variation or dispersion, and k is the number of clusters (Kumar et al., 2024). The **Elbow** specification test shown in Figure 3 is not insightful, however, the **Average Silhouette** scores in Figure 4 reveal maximum scores for number of clusters at 23 and 50. **The Gap Statistic** method in Figure 5 reveals an optimum number of clusters at 47. The next step entails testing the number of clusters suggested by the Gap Statistic method.

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

Silhouette for point i = $S_i = {b_i - a_i}\Big/{max\{a_i, b_i\}}$ (2)

Average Silhouette = $S = \frac{1}{n}\sum_{1}^{n} S_i$ (3)

Sum of squares = $W_k = \sum_{r=1}^{k} \frac{1}{2n_r} \sum_{i,j \in C_r} (x_i, x_j)$ (4)

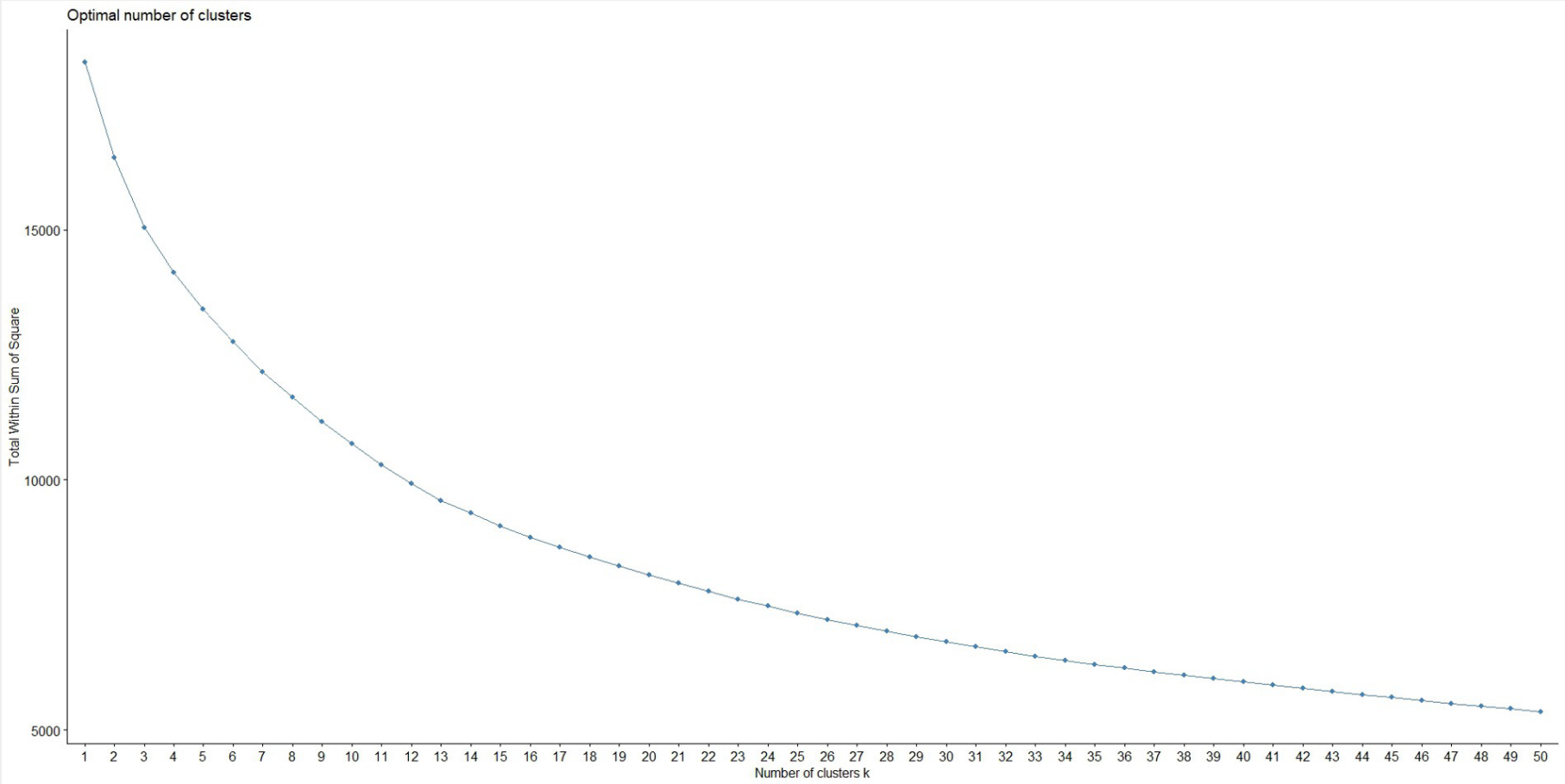
Gap Statistic = $maximize\ Gap_n(k) = E_n^* log(W_k) - (W_k)$ (5)

TABLE 2. Agglomerative Coefficient Values based on Knowledge-level Squared Data (Job Zone 3, 4 and 5)

SL. No.	Distance Matrix	Clustering Method	Agglomerative Coefficient	Rank	Cluster (Knowledge-Level)	Type
1	**Manhattan**	**Ward**	0.9576704	1	"Agnes" function on skill level squared	Agglomerative hierarchical
2	**Euclidean**	**Ward**	**0.948199**	**2**	"Agnes" function on skill level squared	Agglomerative hierarchical
3	**Manhattan**	**Complete**	0.8415293	3	"Agnes" function on skill level squared	Agglomerative hierarchical
4	**Euclidean**	**Complete**	0.8356035	4	"Agnes" function on skill level squared	Agglomerative hierarchical
5	**Euclidean**	**Average**	0.7317411	5	"Agnes" function on skill level squared	Agglomerative hierarchical
6	**Manhattan**	**Average**	0.7125245	6	"Agnes" function on skill level squared	Agglomerative hierarchical
7	**Euclidean**	**Single**	0.6676242	7	"Agnes" function on skill level squared	Agglomerative hierarchical
8	**Manhattan**	**Single**	0.5336212	8	"Agnes" function on skill level squared	Agglomerative hierarchical

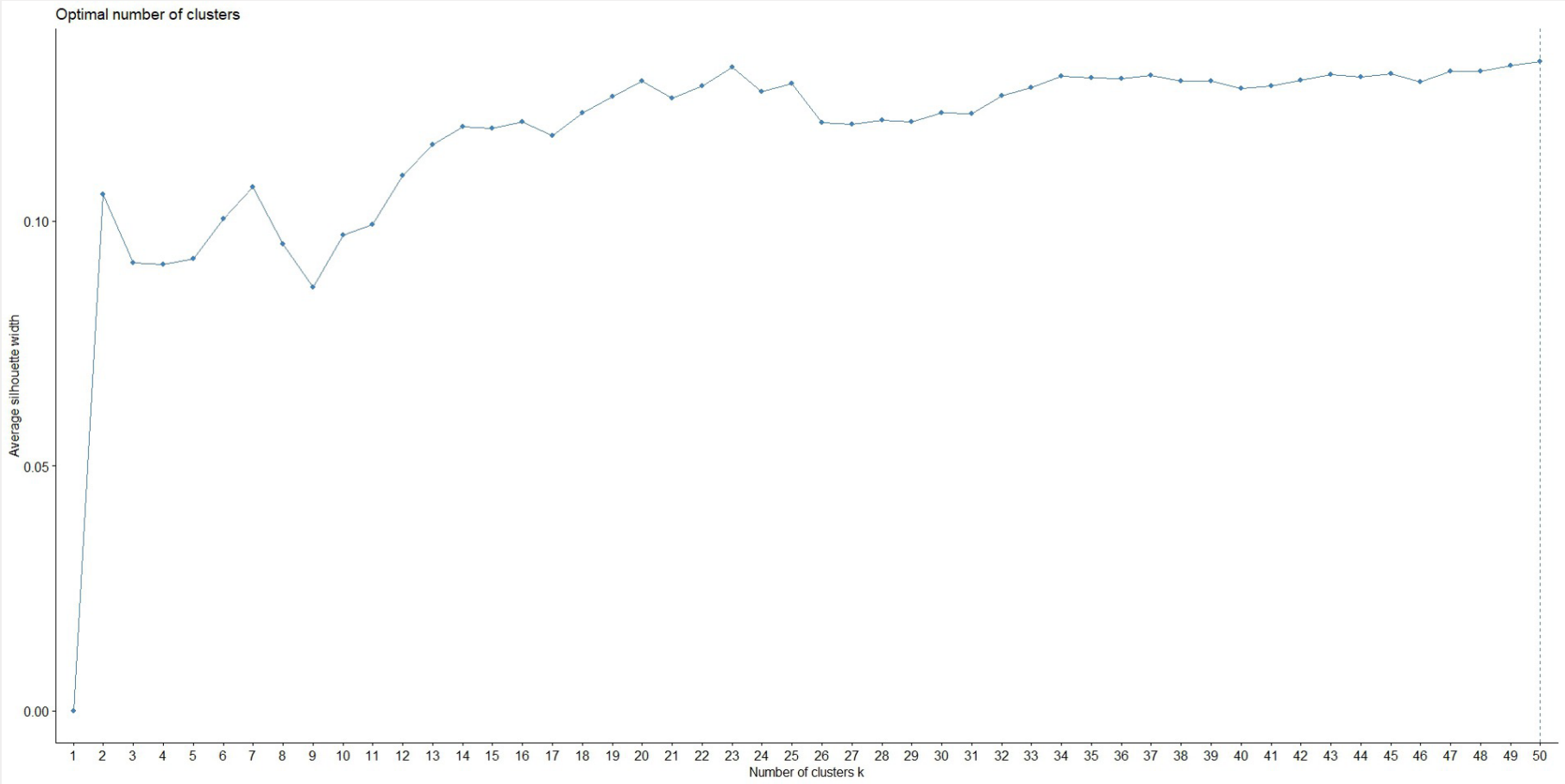
Source: Developed by PCRD.

FIGURE 3. Elbow Specification Test for Number of Clusters (Job Zones 3, 4, and 5; Knowledge-level Squared)



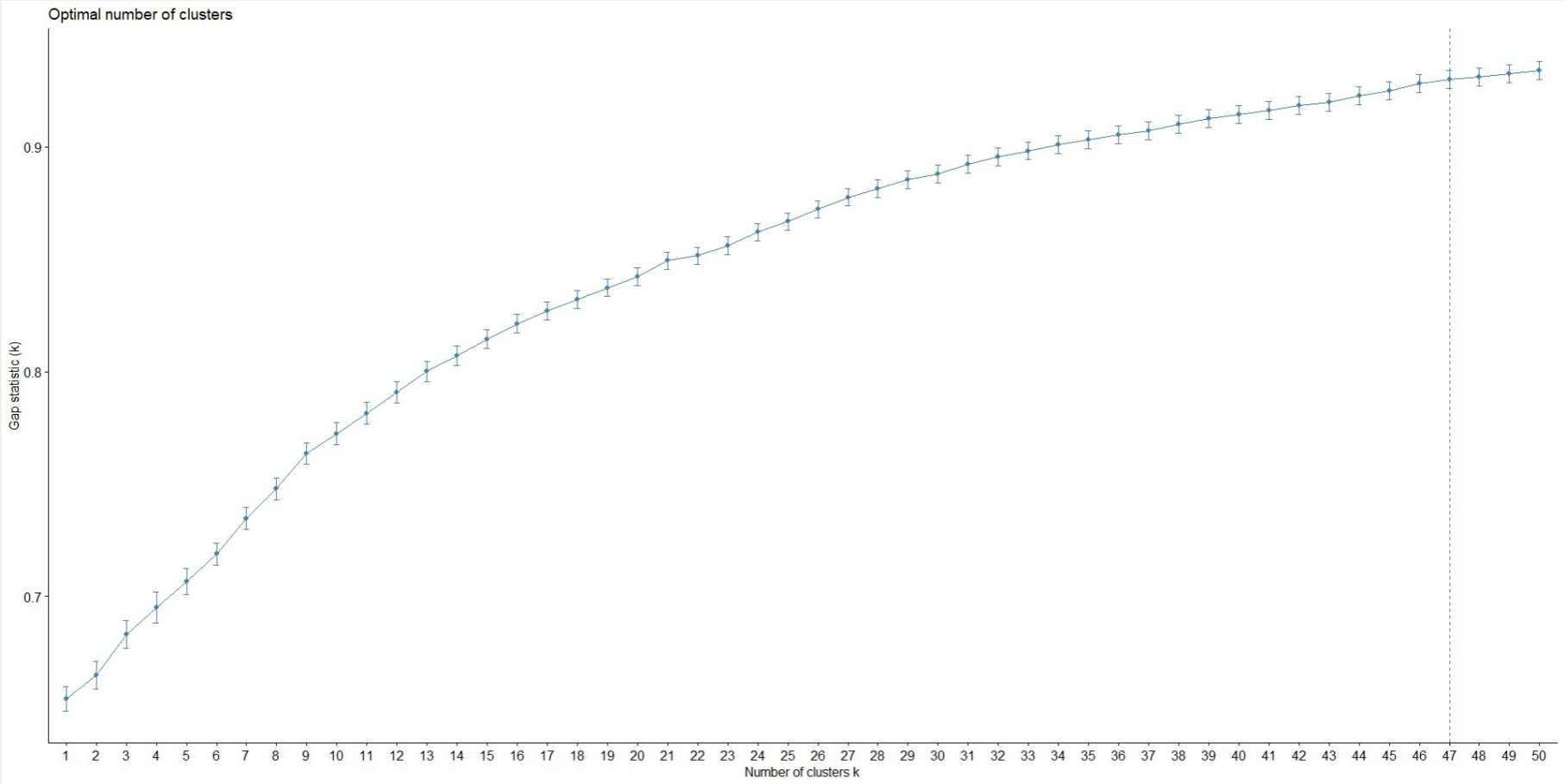
Source: Developed by PCRD.

FIGURE 4. Average Silhouette Specification Test for Number of Clusters (Job Zones 3, 4, and 5; Knowledge-level Squared)



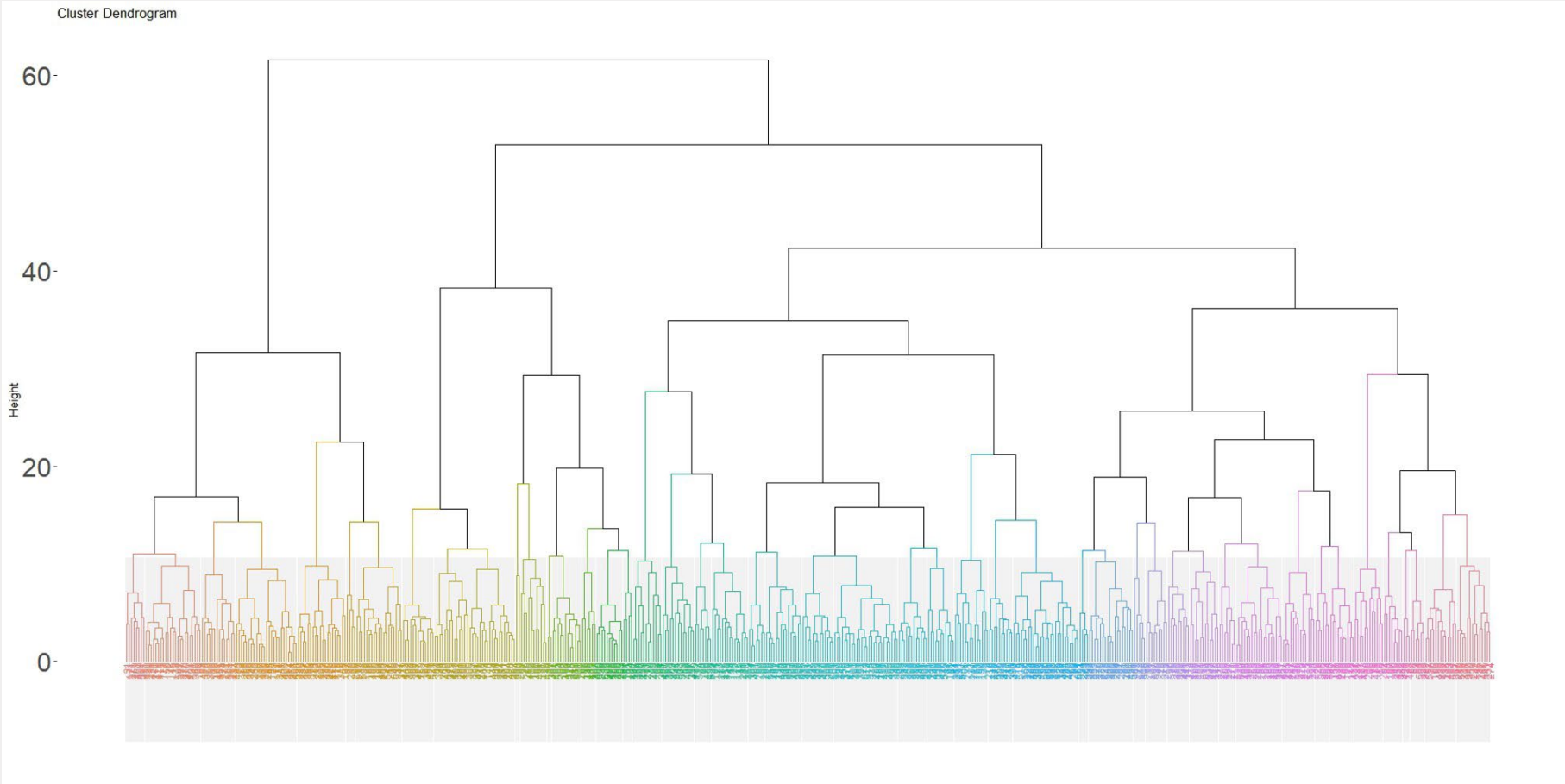
Source: Developed by PCRD.

FIGURE 5. Gap Statistic Specification Test for Number of Clusters (Job Zones 3, 4, and 5; Knowledge-level Squared)



Source: Developed by PCRD.

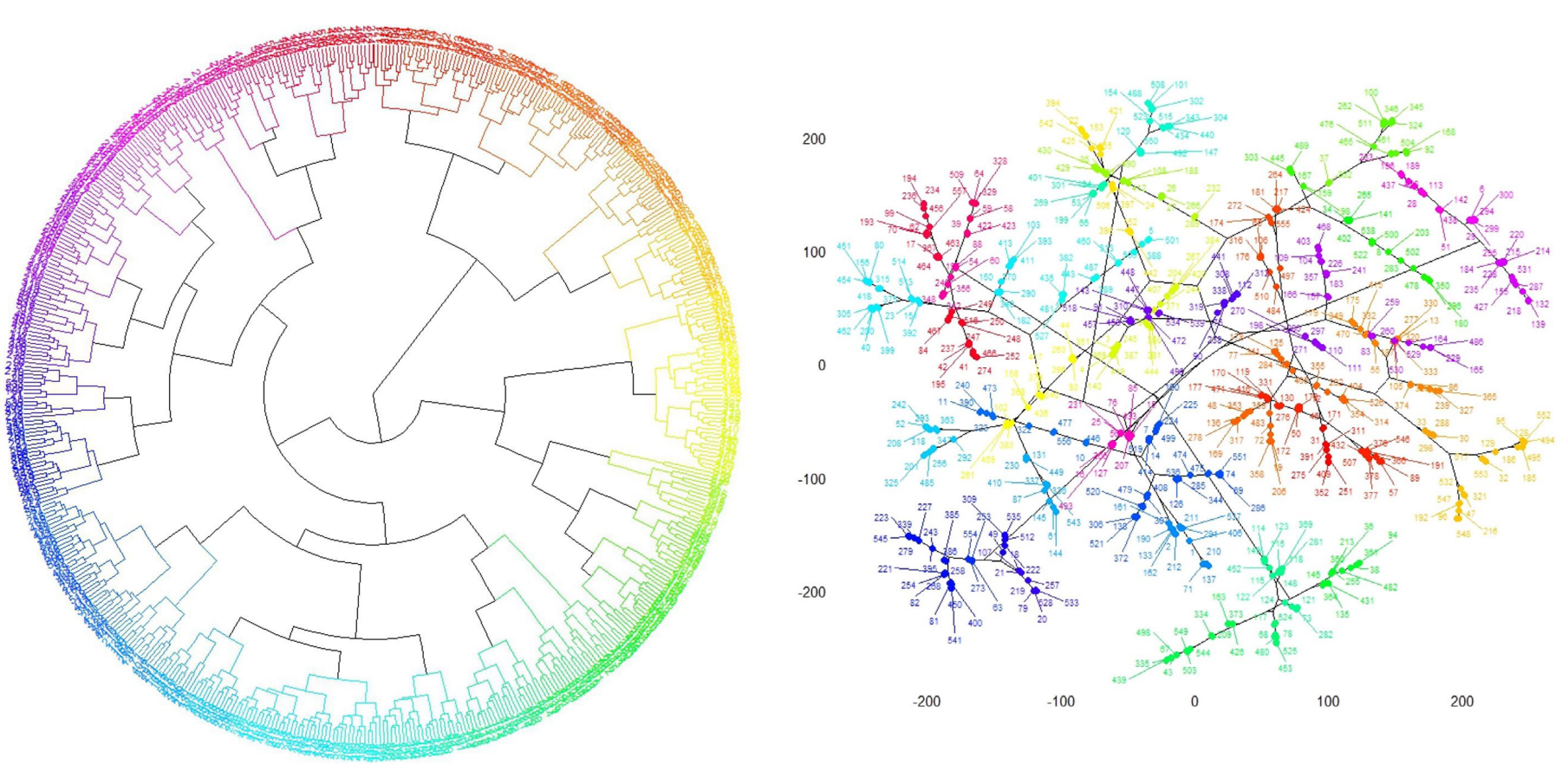
FIGURE 6. Dendrogram of 47 Occupation Clusters for 557 Occupations



Note: Knowledge-levels squared; Job Zones 3, 4 and 5.

Source: Based on Gap Statistic Method. Processed by PCRD.

FIGURE 7. Circular and Tree Dendrograms of 47 Occupation Clusters for 557 Occupations



Note: Knowledge-levels squared; Job Zones 3, 4 and 5.

Source: Based on Gap Statistic Method. Processed by PCRD.

Figures 6 and 7 show the dendrograms for 47 knowledge-based occupation clusters derived from a total of 557 occupations from Job Zones 3, 4 and 5. For an economic development practitioner, 47 clusters covering only 63.8% of the total occupations is not practicable. Quite a few clusters had less than 10 occupations and a few clusters had only two occupations. The largest cluster had 27 occupations. Clusters with a small number of occupations cannot be useful for talent pipeline and workforce development strategies. Hence, the next step was to assess the Mean Silhouette Scores for different number of knowledge clusters based on Silhouette Score method. Table 3 shows the Mean Silhouette Scores for number of knowledge clusters varying from 15 to 47. It is evident that Silhouette Scores is the highest for 23 clusters followed by 47 clusters with the second-highest score. This provided the statistical basis to create 23 knowledge-based occupation clusters. Figure 8 shows the Silhouette Plots created for 23 and 47 clusters, respectively. Figure 9 shows the dendrogram for 23 clusters.

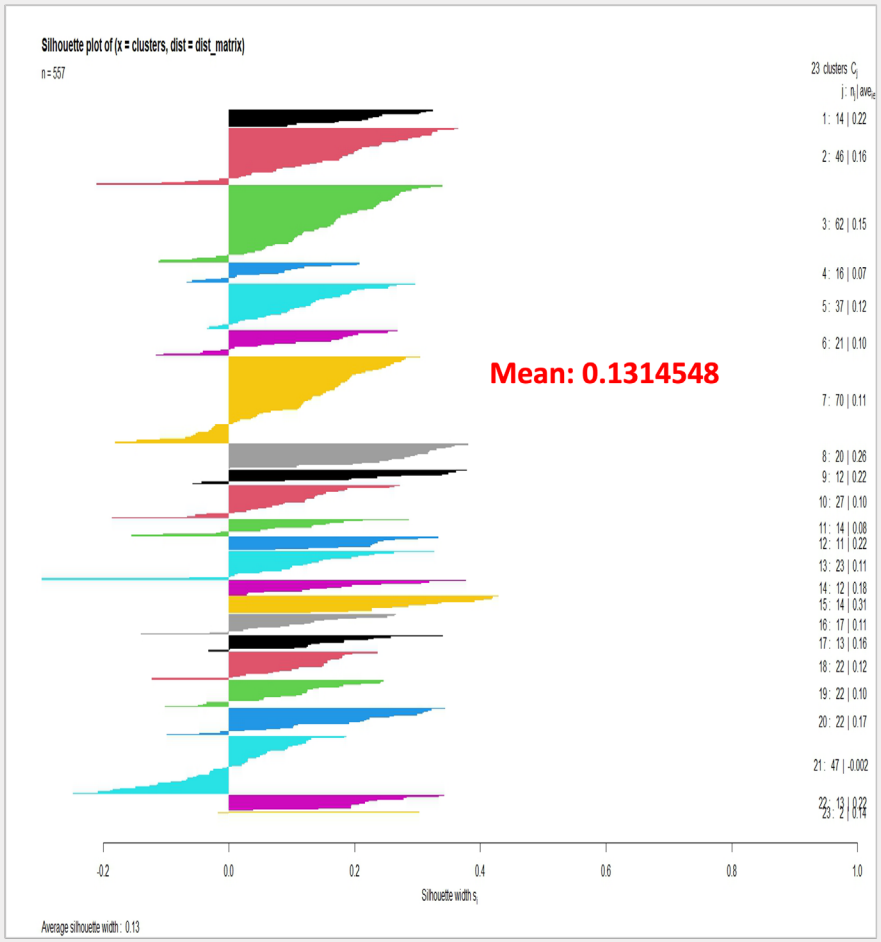
Broadly, the clustering methods described above falls under **Unsupervised Machine Learning**, where algorithms do not require prior training dataset nor assume any prior statistical distribution (Kumar et al., 2024). The clusters or groups are formed solely based on statistical analyses of the given data without any prior assumption.

TABLE 3. Silhouette Score Table of Knowledge-levels Squared (Job Zones 3, 4 and 5)

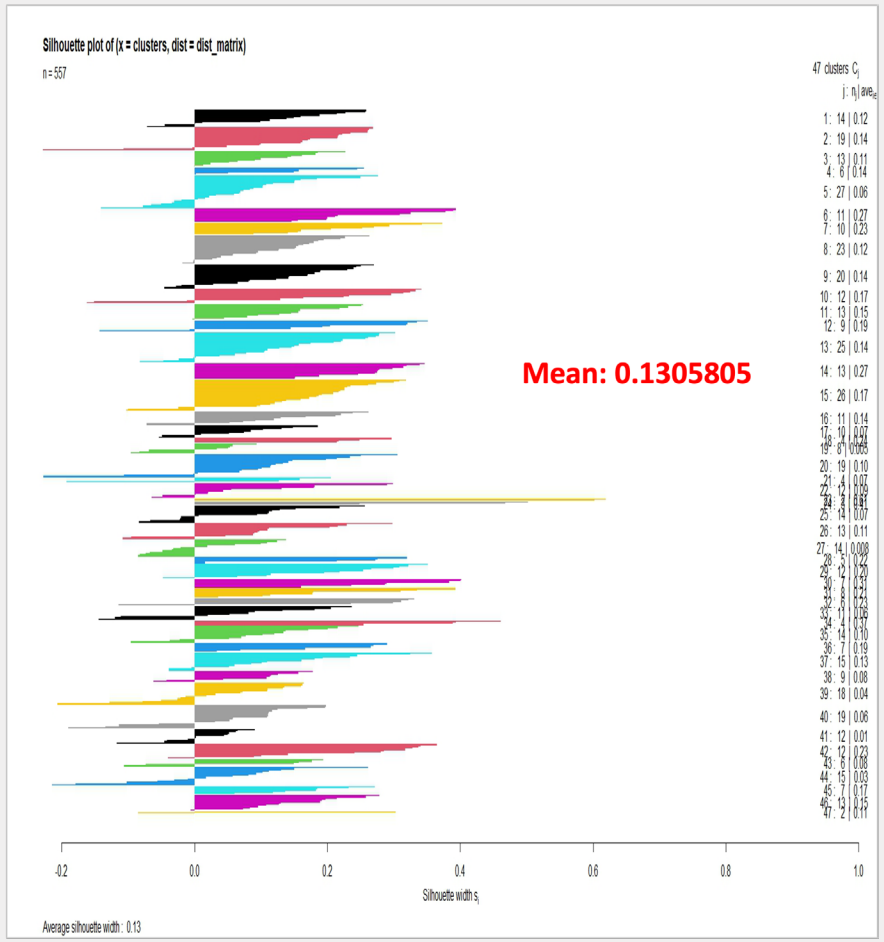
Number of knowledge clusters	Mean Silhouette Score
15	0.1189706
17	0.1174186
18	0.12220657
19	0.1254772
20	0.1286605
23	0.1314548
24	0.1264017
25	0.1280238
30	0.1220087
35	0.1291828
40	0.1270338
45	0.1300078
47	0.1305805

Source: Processed by PCRD.

FIGURE 8. Silhouette Plots of Knowledge-levels Squared



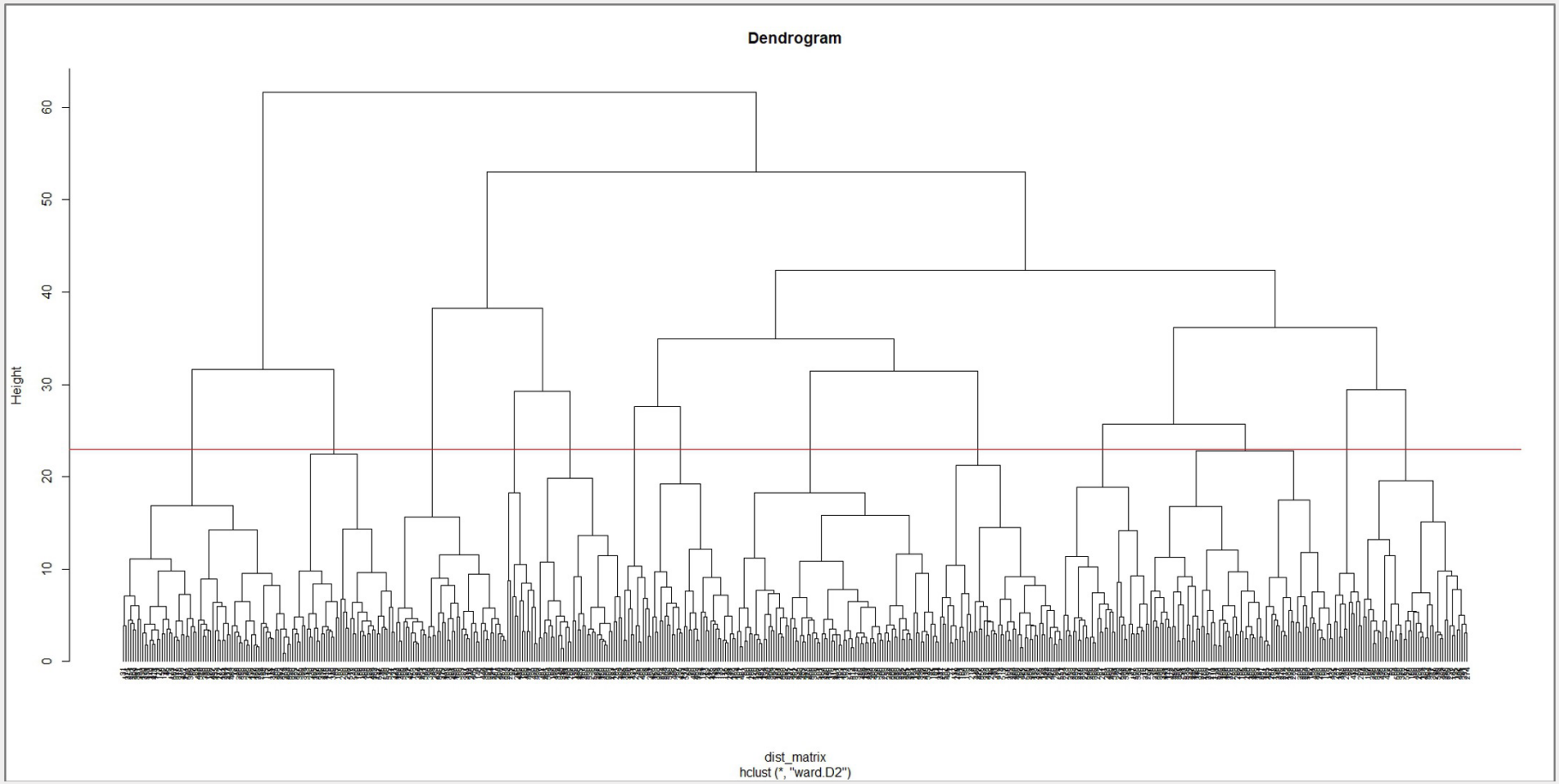
23 Clusters



47 Clusters

Source: Based on Average Silhouette Method. Processed by PCRD.

FIGURE 9. Dendrogram Cut at 23 Knowledge-based Occupation Clusters for 557 Occupations



Note: Knowledge-levels squared; Job Zones 3, 4 and 5.

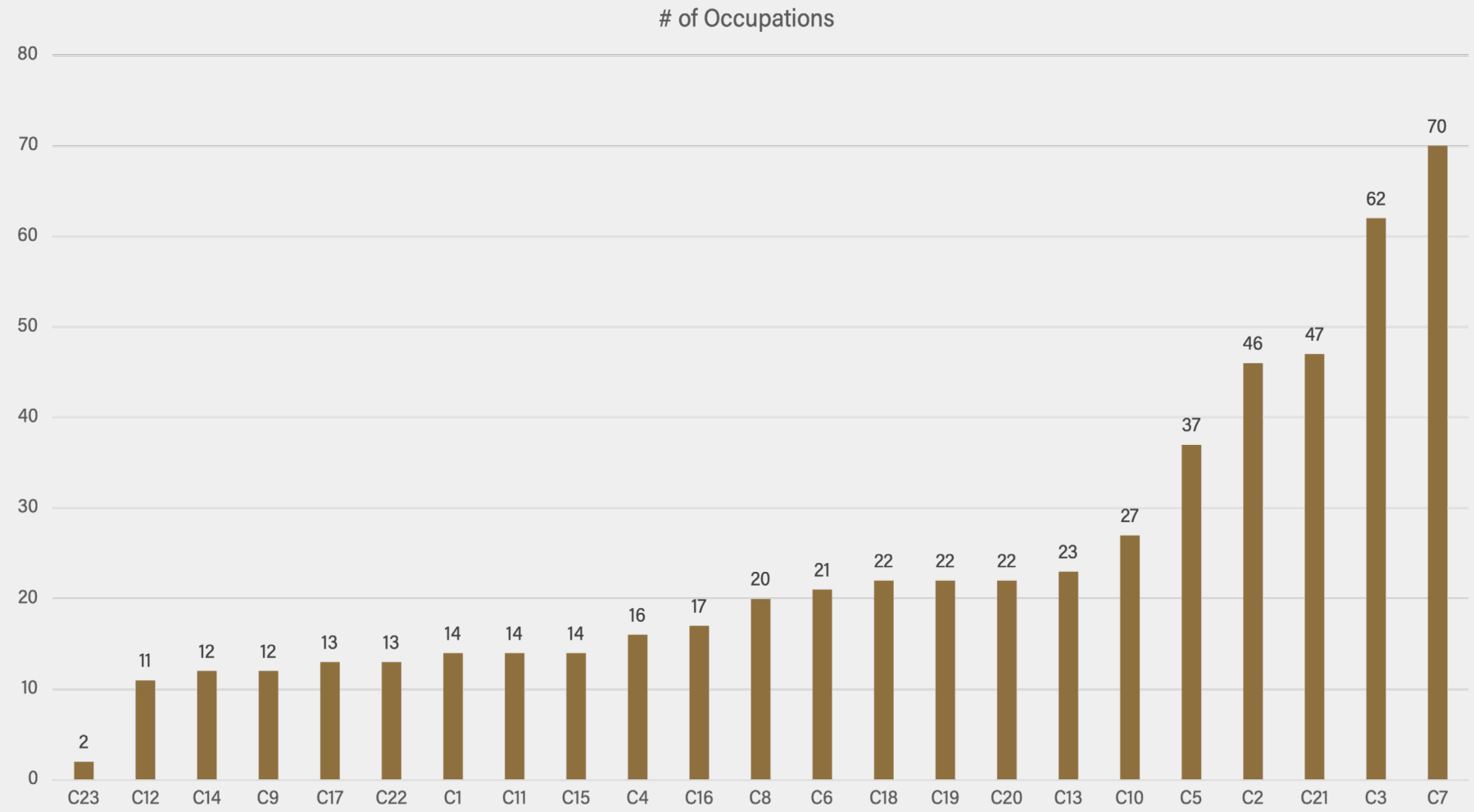
Source: Based on Gap Statistic Method. Processed by PCRD.

FINAL RESULTS AND DISCUSSIONS

Figure 10 shows the frequency of occupations in each of the 23 occupation clusters. It is evident that Cluster 23 (C23) has only two occupations, which include foreign language and literature teachers postsecondary, and interpreters and translators. Such small clusters are not useful for developing strategies, and hence those occupations were merged into other relevant clusters. Similarly, C17 included 13 occupations related to arts and media which were merged into C14 to create one large cluster for Arts, Media, Communications and Creative Professions. Finally, the original 23 clusters were reduced to 21 knowledge-based occupation clusters. Table 4 shows the final 21 knowledge-based occupation clusters with the number of occupations and job zones for those occupations in each cluster. It is evident that the smallest cluster C21, Public Safety and Law Enforcement Professions, has only 11 occupations and the largest cluster C7, Engineering and Technical Trade Technicians, has 65 occupations at the Standard Occupation Code (SOC) 6-digit level. The remaining clusters have occupations varying between 11 and 65. The 21 knowledge-based occupation clusters cover 64% of the total occupations in a labor market. They are comprised of knowledge workers or occupations requiring an associate's, bachelor's, master's, doctoral, or professional degrees. It is anticipated that 21 clusters can be managed by the workforce development practitioners, providing insights into the knowledge worker component of the regional labor force.

The knowledgeable and skilled workforce is one of the top assets sought by industries in their siting decisions. This includes business retention and expansion as well as recruitment of new industries into the region. The site selectors and local economic development organizations need data to assess the regions in terms of human capital competitiveness. The knowledge and skills-based occupation clusters provide data and metrics to help regions uncover the makeup of their knowledge workers and skilled labor force. The knowledge-based occupation clusters can provide insights into concentrations of specific types of knowledge workers in the region. The regional competitiveness in terms of knowledge workers is an important aspect of the knowledge economy. As the global economy shifts from production to services and then to the knowledge economy, regions need to understand their competitive and comparative advantages. Knowledge-based occupation clusters provide insights into such unique advantages for the regions.

FIGURE 10. Occupational Frequency of 23 Knowledge-based Occupation Clusters



Source: Processed by PCRD.

TABLE 4. 21 Knowledge-based Occupation Clusters

Knowledge Cluster #	Knowledge Occupation Cluster Name	Number of Occupations	Job Zones
C1	C1: Finance, Economics, and Accounting Professions	16	3,4,5
C2	C2: Healthcare and Medical Science Practitioners and Specialists	46	3,4,5
C3	C3: Healthcare and Medical Science Technicians	40	3,4,5
C4	C4: Legal, Environmental Compliance and Enforcement Professions	14	3,4,5
C5	C5: Business Management and Client Services Professions	31	3,4
C6	C6: Education, Religion, History, and Cultural Professions	18	3,4,5
C7	C7: Engineering and Technical Trade Technicians	65	3,4,5
C8	C8: Engineering and Advanced Technology Professions	31	3,4,5
C9	C9: Agriculture, Food and Natural Sciences Professions	21	3,4,5
C10	C10: Environmental and Earth/Geo Science Professions	18	3,4,5
C11	C11: Transportation and Logistics Professions	16	3,4,5
C12	C12: Humanities, Sciences and Engineering Post Secondary Education and Knowledge Creation	40	3,4,5
C13	C13: Architecture, Landscape and Energy Professions	13	3,4,5
C14	C14: Arts, Media, Communications and Creative Professions	33	3,4,5
C15	C15: Mental Health and Social Service Professions	14	4,5
C16	C16: Microbiological, Physical, Mathematical, and Statistical Sciences Professions	20	4,5
C17	C17: Office, Publishing, Information Management and Curation Support Professions	16	3,4,5
C18	C18: Operations, Production and Safety Management Professions	16	3,4
C19	C19: Information Technology and Telecommunication Professions	22	3,4,5
C20	C20: Vocation, Management, Community and Childcare, and Personal Care Professions	56	3,4,5
C21	C21: Public Safety and Law Enforcement Professions	11	3,4

Source: Processed by PCRD.

REFERENCES

Abis, Simona and Veldkamp, Laura. (2022). The Changing Economics of Knowledge Production. National Bureau of Economic Research. https://conference.nber.org/confer/2021/YSAIf21/AV2021.pdf.

Everitt, Brian S., Landau, Sabine, Leese, Morven, and Stahl, Daniel. (2011). Cluster Analysis. John Wiley & Sons, Ltd. 5th edition. UK.

Feser, Edward. J. (2003). What regions do rather than make: A proposed set of knowledge-based occupation clusters. *Urban Studies*. 40(10), 1937-1958.

Haas, Carl, T., Rodriguez, Ana Maria, Glover, Robert, & Goodrum, Paul M. (2001). Implementing a multiskilled workforce. *Construction Management & Economics*. 19(6), 633-641.

Hawamdih, Suliman, Kim, Jeonghyun, and Wang, Xin. (2023). The Knowledge Economy. Foundations of the Information and Knowledge Professions. University of North Texas. https://openbooks.library.unt.edu/information-knowledge-professions/chapter/chapter-4-the-knowledge-economy/.

Hogan, Timothy. (2011). An Overview of the Knowledge Economy, With A Focus On Arizona. W. P. Carey School of Business. Arizona State University.

Kaufman, Leonard, Rousseeuw, Peter, J. (1990). Finding Groups in Data: An Introduction to Cluster Analysis. Wiley Interscience. Hoboken: New Jersey.

Khalaf, C., Michaud, G., & Jolley, G. J. (2021). How to assess the transferability of worker skills: A hybrid clustering approach. *Journal of Regional Analysis and Policy*. 51(1), 67-78.

Koo, Jun. (2005). How to analyze the regional economy with occupation data. *Economic Development Quarterly*. 19(4), 356-372.

Kopczewska, Katarzyna. (2022). Spatial Machine Learning: New Opportunities for Regional Development. *The Annals of Regional Science*. 68: 713-755.

Kumar, Indraneel, Siller, Karen, I., White, Mark, C., St. Germain, Benjamin, Zhalnin, Andrey, Mbongo, Bertin. (2024). Occupations by Skills Clusters for the U.S.: Methodological Framework and Experiments. Economic Clusters for the 21st Century. https://pcrd.purdue.edu/wp-content/uploads/2024/11/Occupations-by-Skills-Clusters-for-the-U.S.pdf.

Nolan, Christine, Morrison, Edward, Kumar, Indraneel, Galloway, Hamilton, & Cordes, Sam. (2011). Linking industry and occupation clusters in regional economic development. *Economic Development Quarterly*. 25(1), 26-35.

Purdue Center for Regional Development (PCRD). (2009). Crossing the Next Regional Frontier: Information and Analytics Linking Regional Competitiveness to Investment in a Knowledge-based Economy. https://pcrd.purdue.edu/wp-content/uploads/2021/10/crossing_regional_frontier_full_report.pdf.

Romer, Paul. (1990). Endogenous technological change. *Journal of Political Economy*. Vol. 98, 5:71-102.

Saxenian, Annalee. (1996). Regional Advantage: Culture and Competition in Silicon Valley and Route 128. Harvard University Press.

Slaper, Timothy F. (2014). Clustering occupations. Indiana Business Review. 89(2), 7-12.

Tibshirani, Robert, Walther, Guenther, and Hastie, Trevor. (2001). Estimating the number of clusters in a data set via the Gap Statistic. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*. 63:2, pp. 411-423.

Thompson, Wilbur R., & Thompson, Philip R. (1987). National industries and local occupational strengths: The cross-hairs of targeting. *Urban Studies*. 24(6), 547-560.

Ward, Jr., Joe. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*. 58(301), 236-244.

Publication 117 September 2025

RESEARCH & POLICY

INSIGHTS

Discover More Insights at:

www.pcrd.purdue.edu/publications

Center for Regional Development

Report Cover by Ryan Maluchnik, Purdue Center for Regional Development