DEFINING U.S. INDUSTRY CLUSTERS

METHODOLOGICAL INSIGHTS FROM INTER-INDUSTRY TRANSACTIONS ANALYSIS

PURDUE UNIVERSITY | Center for Regional Development

Technical Report

Defining U.S. Industry Clusters: Methodological Insights from Inter-Industry Transactions Analysis

Co-Authors

Indraneel Kumar, PhD; Rachel Zhang, PhD; Ben St. Germain, Andrey Zhalnin, PhD; and Bertin Mbongo

FEBRUARY 2026

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

Understanding how industry sectors are interrelated through supply chains and buying relationships is critical for policies related to regional economic development. In particular, the sectoral interrelationships provide insights to the taxonomy for industrial clusters. This study explores a methodological framework to inform taxonomy for industrial clusters based on their input-output (IO) linkages. The two major methods applied include the agglomerative hierarchical clustering and the principal component analysis (PCA), and the results are presented, respectively. The report also evaluated and discussed the feasibility and benefits of each method. This report concludes with industry cluster taxonomies resulted from both the methods including policy implications and some practical applications.

1.0 Introduction

This research and policy insight report explores the U.S. inter-industry relationships from two methodological perspectives, which include network and statistical analyses of the U.S. Input Output (IO) transactions. The network analysis applies graph-theoretic techniques on the inter-industry transactions or dollar flows between industry sectors to study the structure of the national economies. The network analysis values not only the magnitude of the dollar transactions between industries but also the directions of the transactions. In other words, network analysis accounts for the “purchasing (buying)” and “supplying (selling)” transactions patterns between industries. Within the statistical analysis, the focus is on data partitioning and data reduction techniques, particularly hierarchical clustering as a data partitioning method, and principal components analysis as a data reduction method. The report explains these methodologies in detail and applies the methods on the IO data and tables available for the U.S. In doing so, the research highlights unique aspects of these methods in context of uncovering the U.S. inter-industry relationships and informing the delineation of the industry cluster definitions. The hierarchical analysis is a clustering technique where the focus is on identifying groups or clusters within a large dataset. The principal components analysis reduces the dimensions by grouping the variables such that a significant part of the variance in the original data can be explained properly by the groups or principal components.

Deciphering how industries are interrelated through supplier and buyer relationships is critical for understanding the structural economy of the U.S. The information can help shape effective policies for the national and regional economic development. Sectoral interdependencies captured through the IO relationships form one of the methods to identify industry clusters, which could, in turn, inform and influence decisions for economic development such as infrastructure investments, workforce development programs, and innovation policy. A well-defined taxonomy of industry clusters can reveal how industry sectors are related, and how they could form regionally competitive industry clusters. It also reveals how value chains are structured, and where sectoral synergies or vulnerabilities exist forming strong and weak linkages between industry sectors.

In the literature on industry clusters, researchers have introduced a range of quantitative approaches to uncover the structure of these interrelationships. Hierarchical clustering is frequently used to group industries based on their similarity in input or output profiles. Principal Component Analysis (PCA) serves as a method of reducing dimensionality and uncovering the latent structures that define major economic functions or themes. In parallel, network analysis has been employed to visualize and analyze the complex web of inter-industry flows, with a focus on connectivity, centrality, and community detection. Each method provides unique insights into the structure of the economy and reflects different assumptions about how industries are linked.

This study systematically compares two widely used methods, agglomerative hierarchical clustering and PCA, to generate industry groupings from the national IO tables. The study also presents the insights obtained from the network analysis. The aim is to understand how methodological choices affect the interpretation of inter-industry relationships. The placement of several key industries, such as the semiconductor and battery-related sectors is used to assess how the methods capture the functional roles and economic context of high-impact industries. The comparative analysis provides practical insights for researchers and policymakers seeking to define industry clusters more effectively, while also clarifying the trade-offs between interpretability, dimensionality, and structural nuances.

The remainder of this report is structured as follows: Section 2 reviews the literature of main methods used in exploring interindustry relationships. Section 3 introduces the data and results from exploratory data analysis. Section 4 outlines the methodological framework for the network, hierarchical clustering and PCA methods. Section 5 presents the clustering results from different methods. Finally, sections 6 and 7 include discussions and conclusion with key takeaways and suggestions for future research. The report ends with Section 8 containing the references.

2.0 Literature Review

The network analysis is an appropriate method to study the structural characteristics of an economy, especially when input-output (IO) table is used (Xu and Liang, 2019). Scholars argue that the “classic IO analysis” might uncover direct and indirect relationships and impacts in aggregate terms, however, only network analysis can reveal the importance of a particular chain of related industries or interindustry relationships to the overall economy (Xu and Liang, 2019). Economic systems based on inter-dependencies evolve into a complex system, where network analysis can provide new insights to the researchers and practitioners (Schweitzer et al., 2009). Note that social network analysis (SNA) is rapidly emerging as an important area of research and the SNA methods are applicable for economic analysis. For example, Bergman and Feser (1999) identifies network analysis as an emerging area of research, which can be applied in the context of industry clusters. DePaolis et al. (2022) revisited the well-researched subject of identifying key sectors in the regional economy by using network analysis of the IO data. The random walk-based measure proposed by the researchers revealed the "domino effect" or how fast the shocks would reach a sector (DePaolis et al., 2022). Compared to the output and employment multiplier, network analysis, especially random walk-based measure, could reveal the systemic structural weaknesses in the regional economy (DePaolis et al., 2022). An important area of research in cluster analysis has been identifying the core sectors or driver industries or chain of industries within the industry cluster definitions, and network analysis can help uncover the driver industries.

The network analysis of IO table and other economic systems borrow methods from the SNA. The SNA methods have some important indicators that measure relationships between two actors, industries, trade partners, or economic agents. The two fundamental elements in SNA are "node" or the individual agent such as an industry sector, and "ties" or the interrelationships such as trade, economic value of supplying and purchasing transactions, meeting frequency, etc. Hence, economic networks are extensions of social networks and SNA can be used to analyze the economic networks.

The literature related to agglomeration economies have explored grouping or clustering of industries and businesses for a long period. The scholars have explored the IO and graph theoretic methods previously. Researchers have distinguished between "industry clusters" and "industrial complexes" as both are based on the network of goods and services flows, however, industrial complexes demonstrate a stronger spatial presence and concentration, whereas industry clusters are focused more on the interrelationships between industry sectors at the national-level (Czamanski and Ablas, 1979). Hence, industry clusters could be considered as "aspatial," whereas industrial complexes have "spatial" characteristics (Czamanski and Ablas, 1979). Feser and Bergman (2000) and Feser (2005) explored interindustry relationships by using the national IO table, hence economic relationships became paramount in their research than the geographic colocation or proximity. In comparison, Porter (2003) focused on the geographic colocation by estimating Pearson's correlation of jobs for pairs of industry sectors at the U.S. state level. Here, the assumption was that if two industry sectors collocated spatially, the economic relationships existed between both the sectors. Delgado et al.

(2016) applied both methodologies to develop the latest set of industry cluster definitions.

During early 1970s, scholars applied Pearson and Rank Order correlation coefficients, factor analysis, and hierarchical cluster analysis to define meaningful industry clusters by using the Standard Industry Classification (SIC) at the level of metropolitan areas available from the U.S. Census Bureau (Bergsman et al., 1972 and Bergsman et al., 1975). The factor analysis provided overlapping industry clusters, whereas hierarchical clustering provided mutually exclusive industry cluster definitions. However, some of the high correlations between industry sectors were meaningless and could be artifacts of the urbanization or urban attraction processes (Bergsman et al., 1975; Czamanski and Ablas, 1979). Streit (1969) explored both, spatial colocation of industry sectors by using correlation coefficients and economic linkages by using output and intermediate inputs from the input output (IO) table. Equation 1 shows the correlation measure and Equation 2 shows the economic linkages based on Streit (1969). In Equation 1, i and j are industry sectors, g is region, x is employment, and r is the correlation coefficient. In Equation 2, O_{ij} and O_{ji} are sales from industry i to j and vice versa; O_i and O_j are outputs of i and j; and I_i and I_j are the inputs (Czamanski and Ablas, 1979). Note that O are outputs, and I are intermediate inputs emphasizing importance of the interindustry linkages (Streit, 1969).

$$r_{ij} = {Cov\left(x_{ig}, x_{jg}\right)} \Big/ {\sigma_{x_{ig}} \sigma_{x_{jg}}} \qquad (1)$$

$$L_{ij} = L_{ji} = \frac{1}{4}[O_{ij}\left(\frac{I}{\sum_i O_i} + \frac{I}{\sum_j I_j}\right) + O_{ji}\left(\frac{I}{\sum_j O_j} + \frac{I}{\sum_i I_i}\right)] \qquad (2)$$

Roepke et al. (1974) extended the analysis of the IO table to determine industrial complexes in Ontario, a province of Canada. Despite IO analysis being an “aspatial” method, it could reveal industrial complexes and spatial aspects, if instead of national, a provincial IO table is used. The author applied three types of analysis on the provincial IO table of 44 x 44 sectors. R-mode factoring was done on aij or purchase coefficients to group industry sectors with similar patterns of inputs, whereas Q-mode factoring was done on aji or sales coefficients to group sectors with similar patterns of sales or markets (Roepke et al., 1974). The third analysis was conducted on the sum of technical coefficients aij and aji based on Equation 3 (Roepke et al., 1974).

$$b_{ij} = b_{ji} = a_{ij} + a_{ji} \qquad (3)$$

Bergman and Feser (1999) explained the IO-based coefficients developed by Stan Czamanski in Czamanski (1974). The intermediate transactions of the IO table contain both, purchases and sales between industries i and j, which can be expressed by four coefficients (Czamanski, 1974; Bergman and Feser, 1999). The coefficients proposed by Czamanski (1974) accounted for both direct and indirect linkages. The maximum of the four coefficients values was selected followed by a factor analysis of the symmetric matrix by assuming that important backward and forward linkages were captured in the process (Czamanski, 1974). As per Bergman and Feser (1999), the symmetric matrix of maximum values partially captured the direct and indirect relationships.

The agglomeration of related economic activities is an inherent characteristic of regional economic geographies. More than a century ago, Marshall (1920) identified three distinct drivers behind industry and business agglomeration: input–output (IO) linkages or suppliers' and buyers' relationships, labor market pooling and knowledge spillovers—all of which contribute to the cost reductions and productivity gains for firms. Since then, the research on agglomeration has expanded this framework to include additional drivers such as local market demand, specialized institutions, the organizational structure of regional business and social networks. As a result, industry clusters often comprise industries connected through a range of linkages, including knowledge, skills, inputs, outputs, and market demand. **Table 2.0.1** below summarizes the key works in the literature that uses input-output table to cluster industries.

Table 2.0.1 Key methods used in IO analysis for industry clusters

Method	Definition	Strengths	Weakness	Articles
Hierarchical clustering	Constructs a nested tree of industries based on their pairwise similarity or dissimilarity.	Intuitive, dendrogram visually reveals nested relationships.	Does not allow overlapping industries. Sensitive to distance metric and linkage method. Selecting of clusters might be subjective.	(Feser, 2005; Delgado et al., 2016; Hill & Brennan, 2000; Argüelles et al., 2014)
Principal Component Analysis	Extracts latent industry groupings and allows for overlapping industries. Produce groups with complementary relationships (Slater, 1977).	Reveals latent structures of economic dynamics. Allows overlapping industries. Loadings provides a ranking of industries.	Cannot be interpreted directly (e.g., PCs are not "clusters"). Selecting the number of PCA could be subjective.	(Feser & Bergman, 2000; O'hUallachain, 1984; Roepke et al., 1974; Vom Hofe & Bhatta, 2007)
Network analysis	Identifies closely connected groups that reflects economic linkages. Different from pairwise similarities, network analysis finds direct and indirect inter-industry dependencies.	Captures complex relationships such as hubs or bridges. Can identify key industries based on centrality.	I-O matrix needs to be converted into a network or node and link data first. Relatively more complicated, difficult to compute.	(Giuliani, 2013; Montresor & Marzetti, 2009; Nuss et al., 2016; Titze et al., 2011)
Q-Analysis	A mathematical framework based on algebraic topology. It identifies clusters based on the shared participation of industries (e.g., co-supplying or co-demanding)	Can identify higher-order relationship than pairwise relations. Produces a hierarchical structure.	Q-analysis needs a binary matrix, so a continuous I-O table must be converted by using thresholds. This loses variation in magnitudes and nuances. More complex method.	(Atkin, 1974; Sonis & Hewings, 1997, 2000)

3.0 Data Sources and Exploratory Data Analysis

An input-output (IO) table is a structured representation of the economic transactions between industry sectors and major consumers, such as households and government within an economy. The IO table has generally four components which include interindustry transactions, value added, nonmarket transfer such as intergovernmental transfers, and the final demand comprised of consumption by different consumer groups. The interindustry transactions capture how the output of one industry sector serves as an input for another, illustrating the flow of goods and services in dollars across industry sectors. In this table, each row indicates the industry that sells goods or services, while each column represents the industry that buys them. The row sum shows the output whereas the column sum shows the input.

This study uses the 2017 input-output table derived by using the Bureau of Economic Analysis (BEA) data, tracking the transactional dollar values among industries. The processed data include 391 industries with NAICS codes and industry names. BEA provides IO data as Supply and Use tables. The Supply table has commodities [C] in rows and industries [I] in columns. It can be transposed to develop the Make table with industries [I] in rows and commodities [C] in columns. The Use table has commodities [C] in rows and industries [I] in columns. If Z is a national matrix for IO then Z can be obtained as:

$Z = M * \hat{Q}^{-1} * U$; where $\widehat{\mathrm{Q}}^{-1}$ is an inverse of the diagonalized commodity output matrix $T = M \bullet (diag(O_c))^{-1} \bullet U$; where T is the transactions matrix, $\mathrm{O_c}$ is the commodity output matrix. Bergman and Feser (1999) used the transactions matrix formula T to arrange the matrices in the following way. Note that in multiplication of two matrices, internal elements cancel each other. Hence, the final result is a table of industries in rows and industries in columns, $(I \times \cancel{C})(\cancel{C} \times C)(C \times I) = (I \times \cancel{C})(\cancel{C} \times I) = (I \times I)$.

Below is an illustration of a simplified IO table. S_i represents the outputs (i.e., the total sales by industry i), and B_i represents the inputs (i.e., the total purchase by industry i).

Input-output Table	i	j	Row Sum (Total sales)
i	(i, i)	(i, j)	S_i
j	(j, i)	(j, j)	S_j
Column Sum (Total Buying)	B_i	B_j	

To capture the relative strength of buy-sell linkages between industries, this study follows Czamanski (1971) and compute four ratio-based measures that express transactions as proportions. The first two ratios measure proportionately how much an industry sells to or buys from another, scaled by its own total sales (row total) or total purchases (column total), respectively, of the industry in question:

$$Sell_{ij} = \frac{IO_{ij}}{S_i} \tag{4}$$

$$Buy_{ij} = \frac{IO_{ji}}{B_i} \tag{5}$$

The other two ratios evaluate the same transactional value, but relative to the trading partner's scale, normalized by either the partner's row total (sales) or column total (purchases):

$$Buy_{ji} = \frac{IO_{ji}}{B_j} \tag{6}$$

$$Sell_{ji} = \frac{IO_{ij}}{S_j} \tag{7}$$

These four ratio measures incorporate the actual dollar value of transactions, providing a more nuanced view of inter-industry linkages than counting the number of common buyers or suppliers. The latter method, used by Feser (2005), focused on the common economic relationship, but excluded the intensity and consideration of directionality of the economic transactions. Practically, the calculated ratio measures also introduce greater variations across industries, which enhances the ability to differentiate industry clusters in subsequent analyses.

Next, the four buy-sell linkage ratios either by taking their mean or maximum for each industry pair are calculated as follow:

$$R_{max} = \max\{Buy_{ji}, Sell_{ji}, Buy_{ij}, Sell_{ij}\} \tag{8}$$

$$R_{mean} = \mathrm{E}\{Buy_{ji}, Sell_{ji}, Buy_{ij}, Sell_{ij}\} \tag{9}$$

In the literature, it is more common to take the maximum than mean of the buy-sell ratios (Czamanski, 1971; Feser, 2005). A few reasons contribute to this: the maximum emphasizes the strongest single connection between two industries—highlighting dominant trade flows or dependencies. Maximum ratio can also improve the sparce matrix and have less 0 in the matrix. In contrast, the mean provides a more balanced view, accounting for all types of transactional linkages rather than focusing solely on the largest one, but the mean could also be sensitive to extreme values and outliers. The mean reflects overall integration, while the maximum identifies the most influential relationships. Choosing between the two depends on whether the analysis prioritizes comprehensive linkage strength or key strategic ties for clustering purposes.

Table 3.0.1 summarizes the statistics of four buy-sell ratios. The mean buy-sell linkages are small (all below 0.003), while max linkages show more dispersion. As further analysis requires variations in the data, this study chooses R_{max} for further analysis.

Table 3.0.1 Descriptive Statistics of Four Buy-Sell Linkages

	$Sell_{ij}$	$Sell_{ji}$	Buy_{ij}	Buy_{ji}	Mean	Max
Mean	0.0025	0.0026	0.0028	0.0023	0.0028	0.0073
Max	0.9837	0.9941	0.8343	1.0000	0.6898	1.0000
Min	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Std. Dev	0.0197	0.0202	0.0201	0.0132	0.0149	0.0335
1st Quartile	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
2nd Quartile	0.0000	0.0000	0.0000	0.0000	0.0001	0.0003
3rd Quartile	0.0003	0.0004	0.0002	0.0006	0.0011	0.0030

4.0 Methodology

The methodology includes three methods. The first introduces methods of network analysis. The second presents the process of hierarchical clustering. The third looks into clustering using PCA method.

4.1 Social Network Analysis

Based on Jackson (2008), Degree Centrality measures the connectedness of a node where a node with degree centrality of 1 is connected to every other node in the network. The Closeness Centrality can be measured as how close a node is to any other node in the network by taking an inverse of the average distance between two nodes (Jackson, 2008). The Degree Centrality measures magnitude or the number of connections without emphasizing the power of the node, whereas in Closeness Centrality, the assumption is that the nearer and distant nodes have the same weight (Jackson, 2008). The author suggests that Closeness Centrality can be improved by applying a decay parameter where the unconnected nodes have the decay parameter set to infinity so that the inverse or the centrality value becomes 0 (Jackson, 2008). Another important measure in social network analysis is Betweenness Centrality, which measures how many geodesics or the direct paths between two nodes are passing through the node of interest (Jackson, 2008). The value of Betweenness Centrality is that it can identify critical nodes for the network. If such nodes are removed, the entire network can disintegrate into parts (Jackson, 2008). The three centrality measures discussed above are based on the magnitude, length, and denseness of ties. The centrality measure which can integrate “prestige and power” of the node is known as the Eigenvector Centrality (Jackson, 2008). The Eigenvector Centrality can give value to a node not only based on how many other nodes are connected, but also the importance and rank of the connecting nodes (Jackson, 2008; Iacobucci et al., 2017). It is not important “how many you know” but “who you know,” and hence the emphasis is on the importance of the node (Jackson, 2008; Iacobucci et al., 2017). As per Katz (1953), the power will be based on how many walks of 1 can be made to reach other nodes. The most centralized or powerful node is when all other nodes can be reach through walk of 1 (Katz, 1953). Equation 10 shows the coefficient for Degree Centrality, where $d_i(g)$ is the number of connected nodes and (n-1) is all possible connections to the nodes (Jackson, 2008). Equation 11 is Closeness Centrality, where (n-1) is the total possible connections to nodes and *l* is the number of 1 links to reach from *i* to *j* (Jackson, 2008). Equation 12 is Betweenness Centrality, where p(kj) is all shortest or geodesic paths connecting k and j (Jackson, 2008). Equation 13 is Eigenvector Centrality where walks of lengths 1, 2, 3, etc., are summed up by assuming that walk of length 1 has value a (1 > a > 0), and hence walk of length 2 is a^2 and length 3 is a^3 (Jackson, 2008). This way, walks of lengths 2 and 3 have less numeric values than 1.

$$\text{Degree Centrality} = {}^{d_i(g)}\!\big/_{(n-1)} \tag{10}$$

$$\text{Closeness Centrality} = {}^{(n-1)}\!\big/_{\sum_{j\neq i} l(i,j)} \tag{11}$$

$$Ce_i^B(g) = \sum\nolimits_{k\neq j: i\notin\{k,j\}} {}^{{}^{P_i(kj)}/_{P(kj)}}\!\Big/_{{}^{(n-1)(n-2)}/_2} \tag{12}$$

$$Pk\,(g, a) = agII + a^2g^2II + a^3g^3II + a^4g^4II + \ldots\ldots\ldots \tag{13}$$

Note that the centrality measures mentioned above emphasize the quantitative or physical connections, and lack the qualitative aspects. Moreover, the network represents the Dyadic or mutually paired relationships. In contrast, Granovetter (1973) presented the seminal work on the “Strength of Weak Ties,” where secondary social relationships outperformed primary relationships in getting a better economic outcome. Such analysis of socioeconomic systems is still evolving.

4.2 Hierarchical Clustering

Hierarchical clustering is an unsupervised machine learning[1] method used to group similar observations or items into clusters based on their characteristics. It’s useful

[1] Unsupervised machine learning refers to algorithms that learn patterns from data without labeled outcomes or predefined categories. The goal is to identify underlying structures or groupings within the data itself, such as clusters or associations.

when the natural groupings or clusters formed by the data are informational for the objectives of the research.

This process works by either agglomerative (bottom-up) or divisive (top-down) method approaches (Boley, 1998; Sokal & Sneath, 1963; Ward Jr., 1963). In agglomerative clustering, each observation starts in its own cluster, and pairs of clusters are merged step by step based on a similarity measure (e.g., Euclidean distance) and a linkage criterion (e.g., single, complete, average, or Ward's method) until all items belong to one big cluster. The results are often visualized using a dendrogram, which helps identify meaningful groupings at different levels of the hierarchy.

In contrast, 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 (Boley, 1998; Roux, 2018). Thus, at each step the two new clusters are formed by partitioning the former cluster.

There are five common methods for linkages in hierarchical clustering: single linkage, complete linkage, average linkage, weighted average linkage, and the Ward's method (Kononenko & Kukar, 2007).

- Single linkage, also known as the nearest neighbor method, calculates the shortest distance between any two observations belonging to different clusters. At each iteration, it merges the pair of clusters with the smallest minimum distance. This method is effective for identifying elongated or chain-like clusters. For example, industries with strong sequential buy-sell relationships (e.g., agriculture → food processing and manufacturing → wholesale → retail) may form a linear chain-like structure. In contrast, dense mutual buying and selling relationships across industries (not just along a linear chain, but in multiple directions across the group) might not be well captured by single linkage. Also, single linkage is highly sensitive to noise and outliers, often resulting in unbalanced or loosely connected clusters due to the "chaining effect."
- Complete linkage, or the furthest neighbor method, merges clusters based on the maximum distance between observations in different clusters. This approach tends to produce compact and evenly connected clusters, such as industrial groups with strong and reciprocal linkage across all members. Complete linkage ensures all within-cluster distances remain relatively small, and avoid including industries that are loosely connected to the core. A potential drawback is that complete linkage can exaggerate the effects of outliers and might split natural clusters if a single distant point causes large inter-cluster distances.
- Average linkage, also referred to as the Unweighted Pair Group Method with Arithmetic Mean (UPGMA), merges clusters based on the average distance between all pairs of observations from two clusters. It strikes a balance between

the extremes of single and complete linkage, often resulting in moderately cohesive clusters. In other words, average linkage generates clusters that are less tightly bound than those formed by complete linkage but more cohesive than the elongated, chain-like structures found in single linkage. Average linkage is especially useful when neither chaining nor very tight clusters are desired. However, it may be computationally intensive on large datasets due to the number of distance calculations required.

- Weighted average linkage (WPGMA) modifies the average linkage approach by assigning equal weight to the distances between clusters, regardless of the number of elements in each cluster. Unlike average linkage, it does not account for cluster size, which can introduce bias when clusters vary in size. This method is more appropriate in cases where clusters are expected to grow at a consistent rate, such as evolutionary biology or balanced experimental groupings.
- The Ward's method (1963) 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. This method is especially effective when industries form mutually reinforcing subgroups with comparable levels of buy-sell intensity. Because Ward's method discourages chaining and rewards overall cohesion, it works best when industries are densely linked in all directions, rather than arranged in a linear or peripheral fashion.

4.2.1 Number of Clusters

After performing hierarchical clustering and generating a dendrogram, this study determined the optimal number of clusters using three commonly adopted techniques: the elbow method, the silhouette method, and the gap statistic. Each method offers a different perspective on cluster groupings and provides complementary insights.

(1) Elbow method (Thorndike, 1953): This approach examines the total within-cluster sum of squares (WCSS) as a function of the number of clusters k. The WCSS is defined as the sum of squared distances between each point and its assigned cluster centroid. As k increases, WCSS decreases, but the marginal gain diminishes. The "elbow" point—where the rate of decrease sharply changes—suggests a suitable trade-off between the number of clusters and explained variance. While intuitive, this method is sensitive to scale and may not yield a clearly identifiable elbow in some datasets (Ketchen & Shook, 1996).

(2) Silhouette method (Rousseeuw, 1987): This technique evaluates cluster cohesion and separation using the silhouette coefficient $s(i)$ for each observation i, defined as:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \qquad (14)$$

Where $a(i)$is the average distance from ii to all other points in its own cluster, and $b(i)$ is the lowest average distance to points in any other cluster (i.e., the nearest neighboring cluster). The average silhouette score across all observations indicates the overall clustering quality, with higher values suggesting well-defined, separated clusters. This method provides both a quantitative metric and a visual tool (silhouette plot), though it can be computationally intensive for large datasets.

(3) Gap statistics: Proposed by Tibshirani et al. (2001), the gap statistic compares the observed within-cluster dispersion W_k with its expected value under a null reference distribution. It is computed as:

$$\text{Gap(k)} = E * [\log(W_k)] - \log(W_k) \tag{15}$$

where $E * [\log(W_k)]$ is the expected log $(WCSS)$ under a reference distribution (e.g., uniform). A larger gap indicates that the observed clustering structure is more distinct than would be expected by chance. This method accounts for data geometry but requires repeated sampling, making it more computationally intensive. In practice, the expected dispersion is estimated through bootstrap resampling[2], and the number of bootstrap iterations can be customized to balance computational cost and stability of the gap estimate.

4.3 Principal Component Analysis

The PCA is a statistical technique used to reduce the dimensionality of a dataset while preserving as much variation as possible (Hotelling, 1933). By transforming the original correlated variables into a new set of uncorrelated variables, called principal components (PCs), PCA captures the underlying structure of the data in a more compact form. Each principal component is a linear combination of the original variables and is ordered by the amount of variance it explains, with the first few components often capturing most of the variation in the dataset. This makes PCA particularly useful in exploratory data analysis, dimension reduction, or data preprocessing for multicollinearity (Jolliffe, 2002).

In the context of regional economics and clustering, PCA can be used not only to reduce dimensionality but also to reveal latent group structures in the data. When interpreting PCA for clustering purposes, each principal component can be viewed as representing a distinct "dimension of clustering" or theme. For example, in industrial input-output

[2] Bootstrap resampling is a statistical technique where multiple samples are drawn (with replacement) from the original or reference dataset to estimate the variability of a statistic. In the gap statistic, it is used to simulate the reference distribution of within-cluster dispersion under the assumption of no clustering structure.

analysis, one component may capture healthcare-related linkages, while another captures energy-intensive production. Industries with high loadings on a particular principal component can be interpreted as belonging to the same economic cluster, thus allowing PCA to identify groupings based on shared economic characteristics and structural linkages. Note that each industry can have non-zero loadings across multiple principal components. Hence, PCA allows for a non-mutually exclusive or overlapping clustering approach where entities can belong to multiple thematic clusters simultaneously.

This flexibility contrasts with hierarchical clustering, where each industry is assigned to one and only one cluster, and hence producing mutually exclusive clusters. While hierarchical clustering produces mutually exclusive groups based on similarity in input-output patterns, PCA acknowledges that industries may have roles and linkages in multiple economic subsystems. This overlapping cluster structure is often more realistic, especially in complex economic networks where the same industry can serve diverse functions, such as production, distribution, and consumption, and can sell to as well as buy from different industries. For example, a transportation industry might support both healthcare and tourism clusters, which PCA can capture through high loadings on multiple components. Similarly, industries in logistics clusters can be assemblers, packers, distributors and transporters simultaneously. Therefore, PCA provides a complementary perspective to hierarchical clustering, offering a nuanced view of industrial linkages that accommodates multidimensional roles within the economy.

4.3.1 The number of principal components

To determine the appropriate number of principal components (PCs), three common approaches were employed: the scree plot, parallel analysis, and a variance threshold rule.

The scree plot (Cattell, 1966) identifies the turning point where the added explanatory power of additional components diminishes. It displays the eigenvalues of the principal components in descending order, which helps identify the point at which the additional explanatory power of further components becomes marginal. This method is straightforward and easy to apply, but could be subjective since there might not be a clear “elbow.”

Parallel analysis (Horn, 1965) compares the eigenvalues of the actual data to those of randomly simulated datasets to determine statistically significant components. The components whose eigenvalue exceed the corresponding percentile from the null distribution are selected. This method is statistically grounded because it controls for spurious variance explained by chance. However, it is relatively more computationally intensive as it requires resampling or simulation.

Lastly, it is common in practice to select a cumulative variance threshold (e.g., 60%) to ensure the selected components collectively explain a sufficient portion of the total variance of the dataset. This rule is simple and interpretable, however, the threshold could be arbitrary, which might lead to redundant components or missing nuances (Jolliffe, 2002).

4.3.2 Additional specification

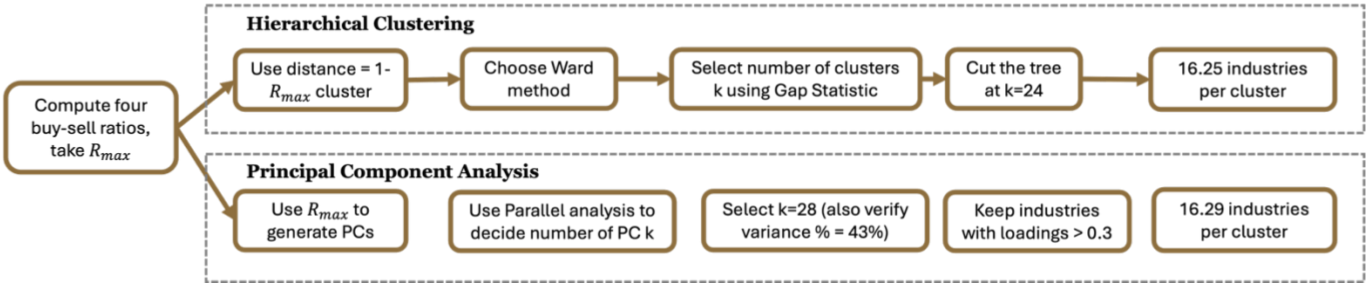
In addition to the number of principal components, a few additional parameters need to be decided. First, as PCA is sensitive to scale, the input matrix should be standardized. It is necessary to ensure each variable has a mean of zero and unit variance (Jolliffe, 2002).

Second, rotation makes the component loadings more interpretable. The main purpose of rotation is to simplify the structure so that each variable loads strongly on as few components as possible. There are two broad types of rotation: orthogonal (uncorrelated components) and oblique (correlated components) rotations (Fabrigar & Wegener, 2012). Choosing between orthogonal and oblique rotations depends on whether interpretability through uncorrelated components or theoretical realism by allowing component correlations is the priority. There are four common rotation methods:

- Varimax is the most common orthogonal rotation—it assumes that components are uncorrelated and aims to simplify interpretation by maximizing the variance of squared loadings within each component, making the structure more interpretable.
- Quartimax, another orthogonal rotation, simplifies the variables by attempting to load each variable highly on a single component, but it may produce less distinct components than varimax.
- Equimax balances the goals of varimax and quartimax, trying to simplify both variables and components simultaneously, though it can be harder to interpret due to this trade-off. In contrast, oblique rotations such as promax or oblimin allow components to be correlated, which may be more realistic in cases where underlying factors are believed to be related.
- Oblique rotations often yield a more accurate reflection of the data's structure but complicate interpretation because both loadings and component correlations need to be considered.

Last but not least, a loading cutoff is necessary in PCA or factor analysis to focus on the most meaningful relationships between variables and components, reducing noise and improving interpretability. It is typically applied after rotation, when examining the component loading matrix to identify which variables significantly contribute to each principal component. The threshold determines which loadings are considered "strong enough" to indicate association with a component. Common cutoffs range from 0.3 to 0.5, depending on the context: a higher threshold (e.g., 0.5) yields more distinct clusters but may exclude relevant variables, while a lower threshold (e.g., 0.3) includes more variables but may introduce weaker associations. The choice depends on the balance between interpretability and comprehensiveness.

Figure 4.3.2.1 Statistical Analytical Procedures



Compute four buy-sell ratios, take R_{max}
Hierarchical Clustering
Use distance = 1- R_{max} cluster
Choose Ward method
Select number of clusters k using Gap Statistic
Cut the tree at k=24
16.25 industries per cluster
Principal Component Analysis
Use R_{max} to generate PCs
Use Parallel analysis to decide number of PC k
Select k=28 (also verify variance % = 43%)
Keep industries with loadings > 0.3
16.29 industries per cluster

5.0 Results

5.1 Buy-sell linkages

Based on input-output table, this study computes four buy-sell ratios for each pair of industries. Taking Semiconductor machinery manufacturing (NAICS: 333242) as the example, its highest $R_{max} = 0.2077$, which is the sales ratio to itself. The second highest R_{max} (0.0376) for semiconductor machinery manufacturing is other industrial machinery manufacturing (NAICS: 33329A).

5.1.1 Hierarchical Clustering

Since hierarchical clustering works with distance measure, this study computes $distance = 1 - R_{max}$. To evaluate the performance of different hierarchical clustering methods, this study calculated the agglomerative coefficient for each method. The agglomerative coefficient measures the strength of the clustering structure, with values closer to 1 indicating that clusters are well-formed throughout the agglomeration process. Formally, it quantifies the average dissimilarity of each observation to the first cluster it is merged into, relative to its dissimilarity to the final cluster.

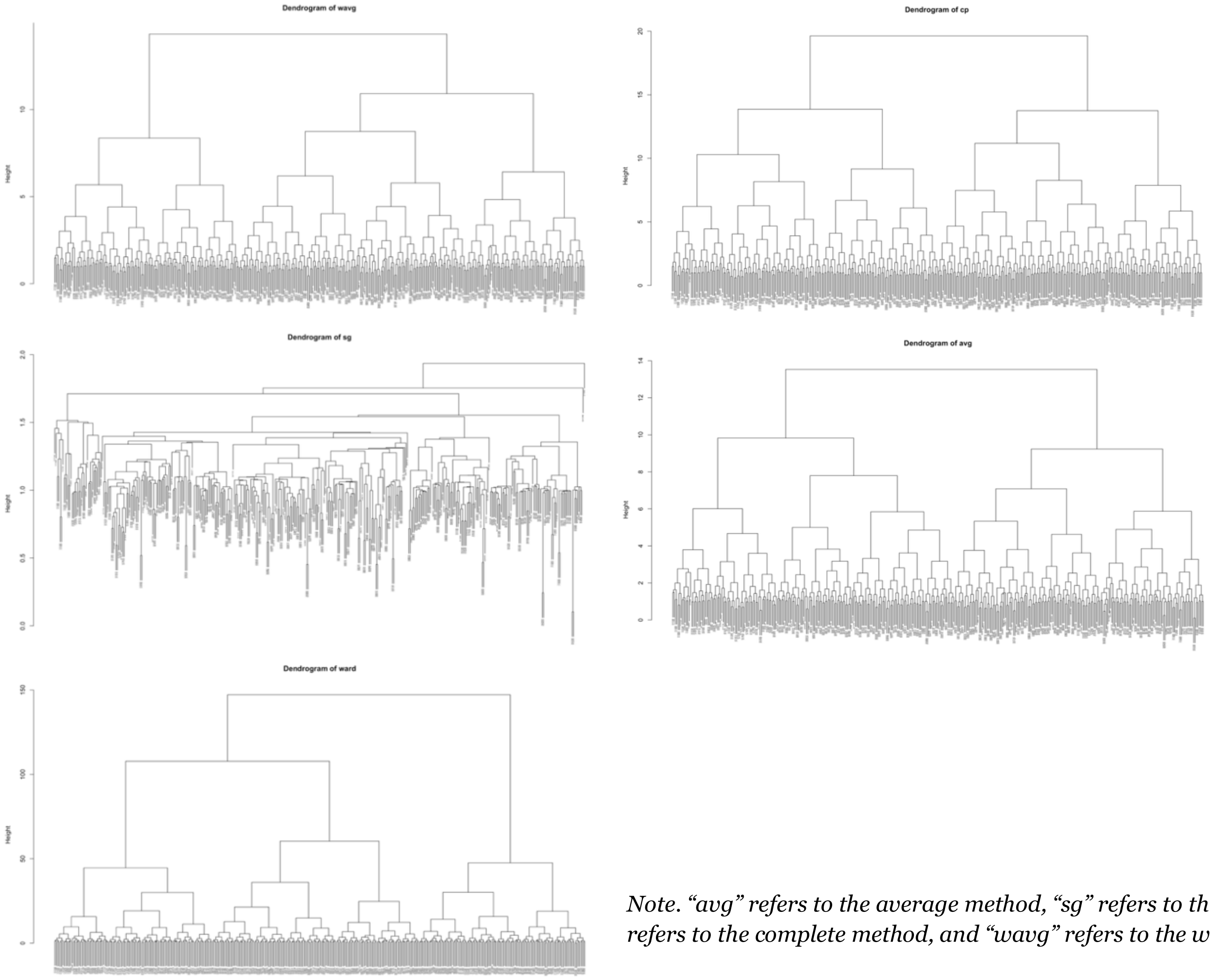
As shown in **Table 5.1.1.1**, the Ward method yielded the highest agglomerative coefficient (0.9938), suggesting that it had the most cohesive clustering structure among the other methods tested. In contrast, the single linkage method had a much lower coefficient (0.2754), indicating poor clustering performance likely due to the chaining effect. The average, weighted average, and complete linkage methods performed moderately well, with coefficients above 0.93, but still below the Ward's method.

Table 5.1.1.1 Agglomerative Coefficients for Hierarchical Methods

Method	**Agglomerative Coefficients**
Ward	0.9938359
Average	0.931351
Single	0.2754142
Complete	0.9515483
Weighted Average	0.9329424

Figure 5.1.1.2 shows the dendrograms using Ward clustering and other methods. At the bottom of the dendrogram, all the industries are included. Overall, Ward method has the most cohesive clustering structure, and it will be used in our subsequent analysis.

Figure 5.1.1.2 Dendrograms of Hierarchical Clustering Results

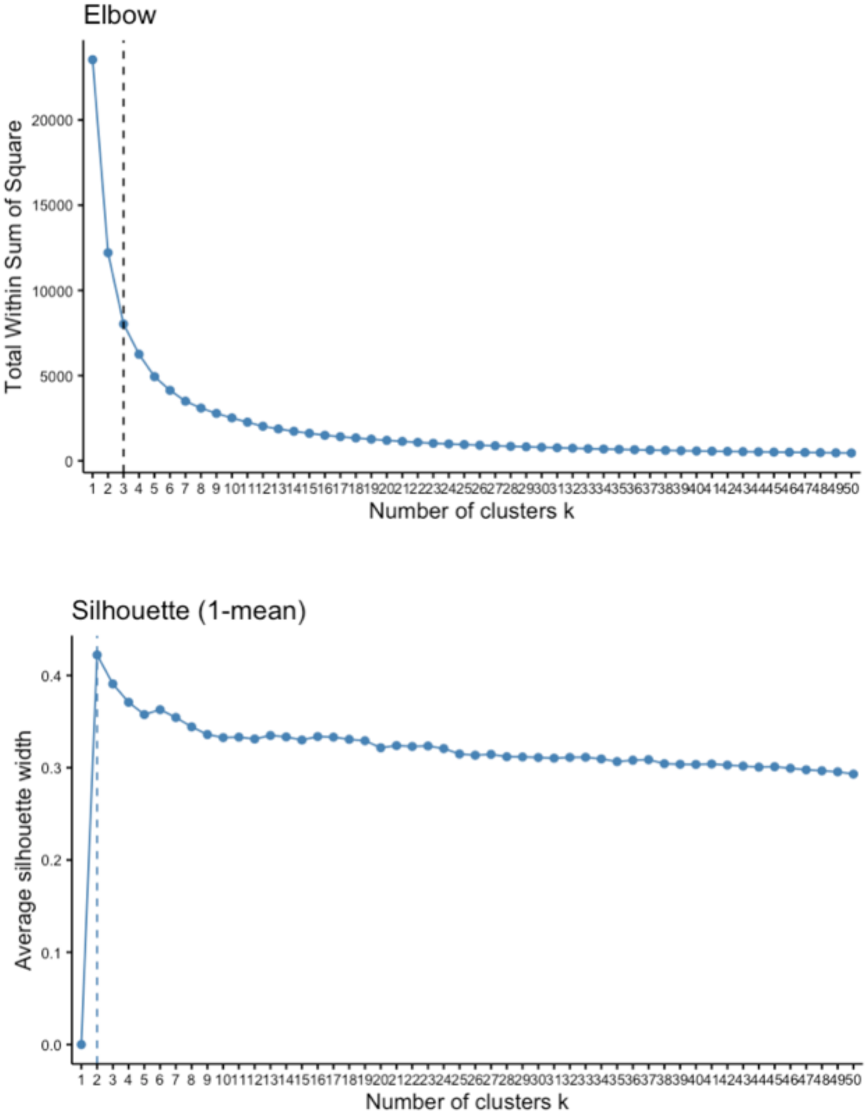


Note. "avg" refers to the average method, "sg" refers to the single method, "cp" refers to the complete method, and "wavg" refers to the weighted average method.

The optimal number of clusters is a critical decision. It decides where the dendrogram is cut. If the dendrogram is cut at a higher height, small number of clusters are generated where finer differences between industries get ignored. If a dendrogram is cut at a lower height, it will produce a large number of clusters that could make clusters less interpretable for practitioners. Three specification tests, elbow method, average silhouette method, and gap statistics are employed to determine the optimal number of clusters.

As **Figure 5.1.1.3** shows, Elbow and Silhouette methods return a suggested value of k = 3 and k = 2, which are too low for practitioner's use.

Figure 5.1.1.3 Specification Tests for Hierarchical Clustering


Elbow

Silhouette (1-mean)

Gap method allows for various maximization methods (Maechler et al., 2025):

- *Firstmax:* Selects the first local maximum of the gap function. This is a simple heuristic that may work well if the gap curve is clearly peaked.
- *Tibs2001SEmax*: The default rule proposed by Tibshirani et al. (2001), which selects the smallest k such that: $Gap(k) \geq Gap(k+1) - s_{k+1}$, where $s_{k+1}$is the standard error of : $Gap(k+1)$. This method balances model simplicity with robustness.
- *firstSEmax*: Chooses the first k for which the gap is within one standard error of the global maximum. This approach tends to favor simpler models while remaining statistically defensible.
- *globalSEmax*: Identifies the first k whose gap is within one standard error of the global maximum gap value, which can be more conservative than *firstSEmax*.

In addition, the bootstrap method can be applied to improve the robustness of the estimation. Bootstrap is a resampling technique used to estimate the sampling distribution of a statistic by repeatedly drawing samples with replacements from the observed data. Each resample (called a bootstrap sample) is the same size as the original dataset and allows for duplication of observations. By calculating the gap statistic across many bootstrap samples, the results are more stable as it leverages the empirical distribution of the data itself.

Table 5.1.1.4 shows the suggested optimal k under combination of maximization method and bootstrap numbers. As *firstSEmax* provides a good balance between interpretability and fit, and larger bootstrap numbers provide more stable result, k=24 is chosen as the optimal number of clusters. **Figure 5.1.1.5** shows the result of Ward clustering with different colored branches for the 24 clusters.

Table 5.1.1.4 Bootstrap sample

Method	Bootstrap sample	Optimal K
firstmax	50	23
firstmax	100	28
firstmax	200	20
firstmax	300	25
Tibs2001SEmax	50	20
Tibs2001SEmax	100	19
Tibs2001SEmax	200	23
Tibs2001SEmax	300	22
firstSEmax	50	28
firstSEmax	100	24
firstSEmax	200	22
firstSEmax	300	24
globalSEmax	50	50
globalSEmax	100	50
globalSEmax	200	47
globalSEmax	300	48

Figure 5.1.1.5 Dendrogram of 24 Industrial Clusters using Ward's method

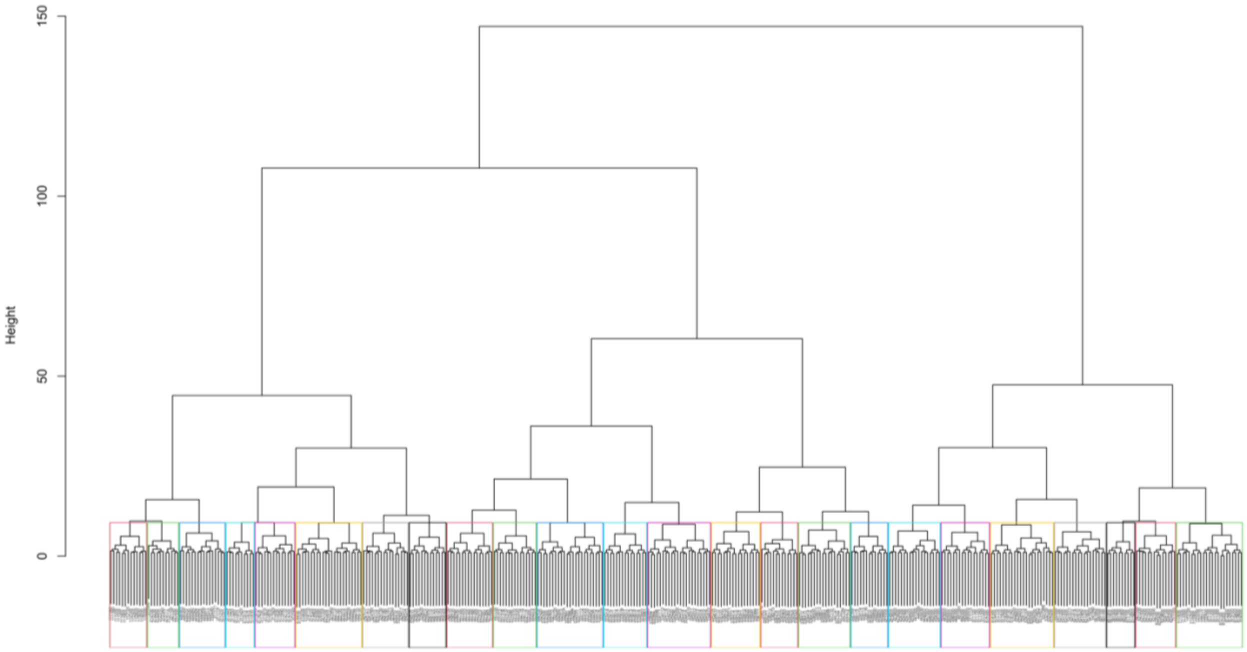
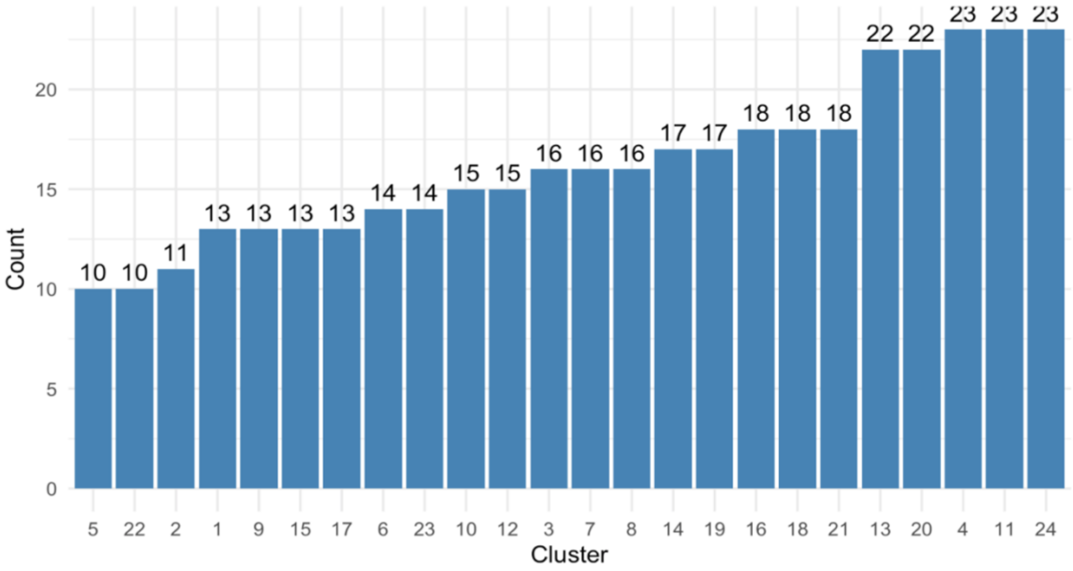


Figure 5.1.1.6 shows the number of industry sectors in each of the 24 clusters. Cluster 5 and 22 have the minimum number of industries (both have 10 industries). Cluster 4, 11, and 24 have the maximum number of industry sectors. On average, each cluster has around 16 industry sectors. Note that Agglomerative Hierarchical Clustering provides mutually exclusive clusters or one particular industry sector appears in only one industry cluster.

Figure 5.1.1.6 Number of industry sectors by individual cluster

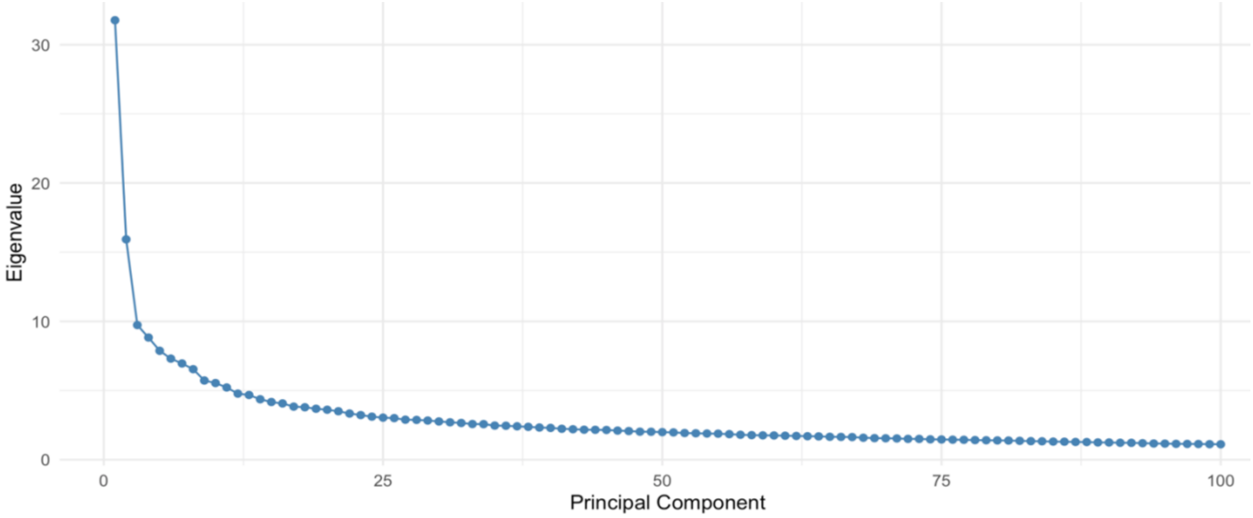


5.1.2 Principal Component Analysis

To decide the number of principal components, a combination of scree plot, explained variances, and parallel analysis are conducted.

The scree plot provides a visual snapshot of the eigenvalues associated with the individual principal components. It helps identify the point where the explained variance starts to level off—commonly referred to as the "elbow"—which suggests a natural cutoff for the number of principal components to retain as results. The observed eigenvalues represent the amount of variance in the original data that each principal component can explain. In contrast, parallel analysis adds a comparison line based on randomly generated data, allowing us to retain only those components whose eigenvalues exceed what would be expected by chance. The scree (**Figure 5.1.2.1**) suggests k=3, which is not practical in industry clustering, especially for practitioners.

Figure 5.1.2.1 Scree plot



To supplement the scree plot, a table of eigenvalues and cumulative variance explained offers a more detailed numerical summary. It shows how much variance each principal component accounts for, both individually and cumulatively, providing further justification for the number of components chosen. The referred explained variance for k=28 is 45.9%, which is lower than the common rule of 60%. To reach 60% of explained variance, 52 principal components are needed.

Table 5.1.2.2 Eigen value and cumulative variance of principal components

PC	Eigenvalue	Proportion Variance	Cumulative Variance
1	30.929	0.079	0.079
2	17.792	0.046	0.125
3	10.319	0.026	0.151
		... ...	
25	3.092	0.008	0.436
26	2.962	0.008	0.444
27	2.945	0.008	0.452
28	2.919	0.007	0.459
		... ...	
52	1.909	0.005	0.6
53	1.868	0.005	0.605
75	1.429	0.004	0.696
76	1.418	0.004	0.7
77	1.402	0.004	0.703

Finally, parallel analysis suggests the number of principal components, which compares the eigenvalues from the observed data to those generated from randomly simulated datasets of the same size. Specifically, the function generates a distribution of eigenvalues from 100 randomly permuted datasets, and retains only the components whose eigenvalues from the actual data exceed the 95th percentile of the simulated eigenvalues. This approach guards against overfitting by ensuring that only components explaining more variance than would be expected by chance are retained. It is a more statistically rigorous alternative to visual methods like the scree plot. The parallel analysis suggests number of clusters or k=28 for PCA. Note that PCA provides clusters that are overlapping and not mutually exclusive enabling that same industry sector can appear in more than one industry cluster definition.

Based on the above, this study selects k = 28 as it offers a reasonable balance between capturing sufficient variance and maintaining interpretability. Choosing a cluster count in the range of 25–30 is also common practice in empirical and practical applications, allowing for meaningful differentiation without over-fragmentation.

5.2 Network Analysis

The BEA 2022 IO transactions table consisted of 71 nodes or industry sectors and 4,183 edges or dollar flow transactions between industry sectors. Note that an IO table is a directed graph where dollar flows from industry 1 to 2 and industry 2 to 1 can vary. The Graph Visualization and Manipulation (Gephi) software is used for the network analysis of 2022 IO table. Only flows $1 million or greater are retained in the Edges table. All negative, 0, and decimal values are removed from the Edges table, if present. Gephi requires first uploading of the Node and then Edges csv files and ensuring that IDs of the sectors are matched properly. **Table 5.2.1** shows the top flows of $150 billion or more transactions between industry sectors in 2022. Oil and gas extraction to the petroleum products has the maximum transaction value in 2022. Insurance activities have the second highest transactions. Farming provides a large value of intermediate transactions of $308 billion of commodities to the food and beverage industries. Chemical industries and other real estate dollar transactions have the fourth and fifth largest transaction values at the national level.

Table 5.2.1 Top IO 2022 Network Flows ($150 Billion or more)

Source Industry Description	Target Industry Description	Value ($2022 Billions)
Oil and gas extraction	Petroleum and coal products	$494.55
Insurance carriers and related activities	Insurance carriers and related activities	$418.46
Farms	Food and beverage and tobacco products	$308.29
Chemical products	Chemical products	$217.16
Other real estate	Other real estate	$216.30
Motor vehicles, bodies and trailers, and parts	Motor vehicles, bodies and trailers, and parts	$205.23
Miscellaneous professional, scientific, and technical services	Miscellaneous professional, scientific, and technical services	$200.59
Food and beverage and tobacco products	Food and beverage and tobacco products	$197.09
Federal Reserve banks, credit intermediation, and related activities	Housing	$163.09
Administrative and support services	Other real estate	$162.36
Other real estate	Other retail	$160.91
Petroleum and coal products	State and local general government	$156.50

Source: Processed by author using BEA 2022 IO data

Table 5.2.2 shows the network metrics estimated by Gephi, which provides a glimpse of the 2022 linkages. The average degree of the network graph is 59, which is the sum of indegree or edges coming into the node and outdegree, or edges going out from the node. A higher value of average degree shows denser connections or economic linkages within the network. The graph density is 0.842 (84.2%), which shows the proportion of connected nodes. If all nodes were connected to each other, the graph density value should have been 1 (100%). Even at the national level, there are few unconnected nodes, and removal of small transaction values of $1 million or less could have affected this metric. Gephi allows changing the resolution of the modularity, where lower modularity resolution can provide a larger number of communities and vice versa. For modularity resolution of 0.4, Gephi reveals modularity class of 13, which means that 13 communities or clusters are partitioned in the network. Gephi uses Blondel et al. (2008) to derive communities or partitions a network into smaller subnetworks. The algorithm starts by assigning each node or a sector to its own community, followed by assigning two nodes into one community and recalculating the metric (Blondel et al., 2008). The algorithm works through modularity optimization where the modularity metric is estimated for subnetworks or communities, and iterated until the metric cannot be improved further (Blondel et al., 2008; Mesa-Arango and Kumar, 2017). Note that Blondel method is also known as Louvain method for modularity-based community detection.

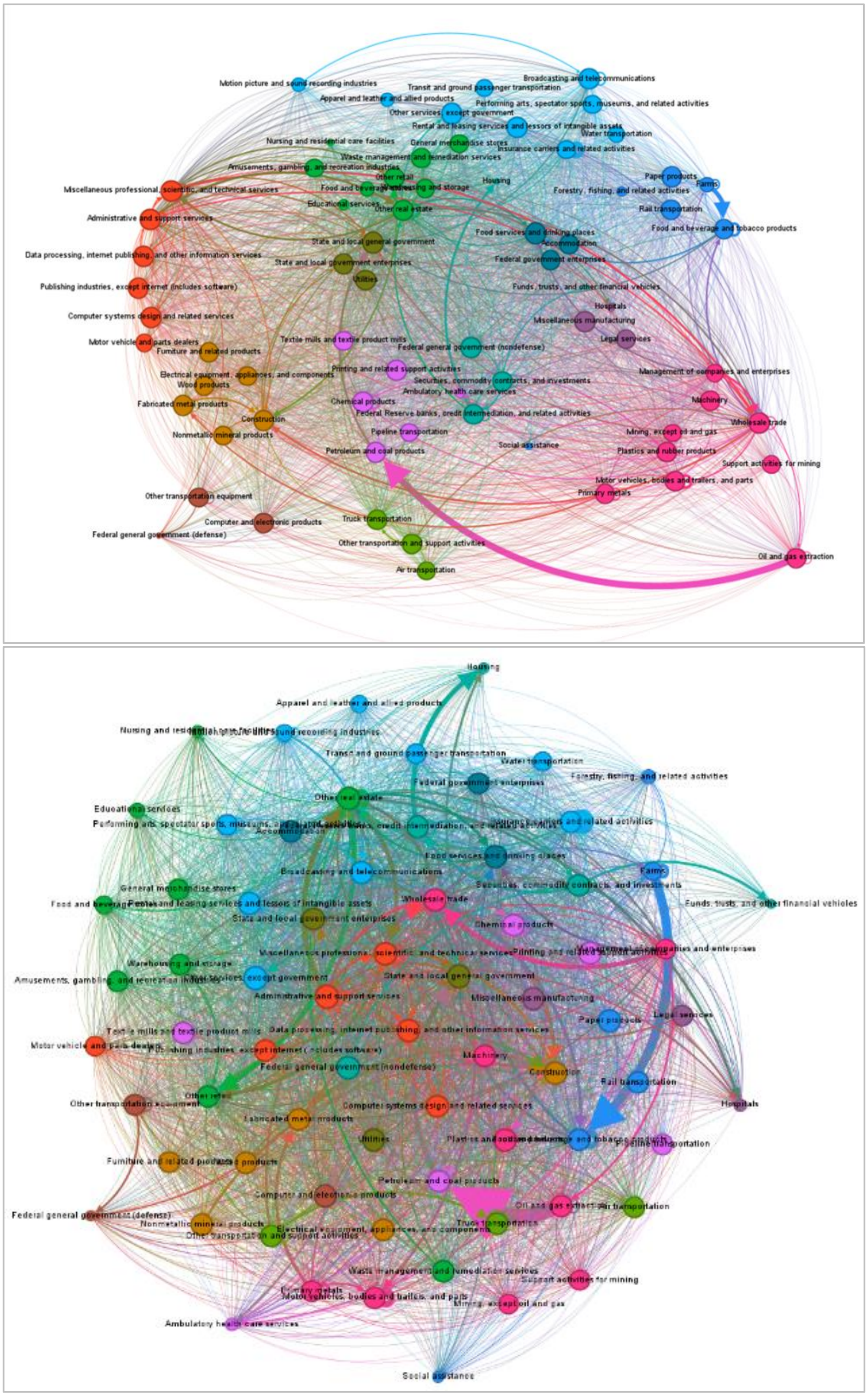
Gephi allows for visualization of the economic network where communities can be represented in different colors. **Figure 5.2.3** visualizes the network by using the default method with some adjustments and the Fruchterman Reingold forced-layout algorithm.

Table 5.2.2 Network Metrics for Economic IO 2022

Variable	Value	Explanation
Average Degree	58.9	Degree of a node is the sum of indegree (how many edges are incoming from other nodes) and outdegree (how many edges are outgoing to other nodes).
Avg. Weighted Degree	286,471.5	Average sum of weights of the edges of a node.
Network Diameter	2	Average graph-distance between all pairs of nodes when connected nodes have graph distance of 1. The diameter is the longest graph distance between any two nodes in the network.
Graph Density	0.842	Measures how complete is the network. A network with all possible edges is 1.
Modularity (0.4 resolution)	0.178	Detection of communities in the network based on algorithm by Blondel et al.
Modularity Class	13	13 communities are identified.
Avg. Clustering Coefficient	0.853	Small world effect or how embedded nodes are in their neighborhood or community.
Average Path Length	1.147	The average graph-distance between all pairs of nodes.

Source: Processed by author using BEA 2022 IO data and Gephi

Figure 5.2.3 Economic IO 2022 Flow Networks


Motion picture and sound recording industries
Broadcasting and telecommunications
Transit and ground passenger transportation
Apparel and leather and allied products
Performing arts, spectator sports, museums, and related activities
Other services, except government
Rental and leasing services and lessors of intangible assets
Water transportation
Nursing and residential care facilities
General merchandise stores
Insurance carriers and related activities
Waste management and remediation services
Amusements, gambling, and recreation industries
Other retail
Housing
Paper products
Farms
Miscellaneous professional, scientific, and technical services
Food and beverage stores
Warehousing and storage
Forestry, fishing, and related activities
Educational services
Other real estate
Rail transportation
Administrative and support services
Food and beverage and tobacco products
Food services and drinking places
State and local general government
Accommodation
Data processing, internet publishing, and other information services
State and local government enterprises
Federal government enterprises
Utilities
Publishing industries, except internet (includes software)
Funds, trusts, and other financial vehicles
Hospitals
Computer systems design and related services
Miscellaneous manufacturing
Legal services
Textile mills and textile product mills
Motor vehicle and parts dealers
Federal general government (nondefense)
Furniture and related products
Electrical equipment, appliances, and components
Printing and related support activities
Management of companies and enterprises
Wood products
Securities, commodity contracts, and investments
Ambulatory health care services
Fabricated metal products
Chemical products
Machinery
Federal Reserve banks, credit intermediation, and related activities
Construction
Wholesale trade
Nonmetallic mineral products
Pipeline transportation
Mining, except oil and gas
Social assistance
Petroleum and coal products
Plastics and rubber products
Support activities for mining
Other transportation equipment
Motor vehicles, bodies and trailers, and parts
Primary metals
Truck transportation
Computer and electronic products
Federal general government (defense)
Other transportation and support activities
Oil and gas extraction
Air transportation
Housing
Apparel and leather and allied products
Nursing and residential care facilities
Motion picture and sound recording industries
Transit and ground passenger transportation
Water transportation
Forestry, fishing, and related activities
Federal government enterprises
Other real estate
Educational services
Performing arts, spectator sports, museums, and related activities
Accommodation
Federal Reserve banks, credit intermediation, and related activities
Insurance carriers and related activities
Food services and drinking places
Broadcasting and telecommunications
Farms
General merchandise stores
Securities, commodity contracts, and investments
Food and beverage stores
Rental and leasing services and lessors of intangible assets
Funds, trusts, and other financial vehicles
Wholesale trade
State and local government enterprises
Chemical products
Miscellaneous professional, scientific, and technical services
Printing and related support activities
Management of companies and enterprises
Warehousing and storage
Amusements, gambling, and recreation industries
Other services, except government
State and local general government
Administrative and support services
Miscellaneous manufacturing
Paper products
Legal services
Textile mills and textile product mills
Data processing, internet publishing, and other information services
Motor vehicle and parts dealers
Publishing industries, except internet (includes software)
Machinery
Federal general government (nondefense)
Construction
Rail transportation
Other retail
Other transportation equipment
Computer systems design and related services
Hospitals
Fabricated metal products
Utilities
Plastics and rubber products
Food and beverage and tobacco products
Pipeline transportation
Furniture and related products
Wood products
Petroleum and coal products
Computer and electronic products
Oil and gas extraction
Air transportation
Federal general government (defense)
Nonmetallic mineral products
Electrical equipment, appliances, and components
Truck transportation
Other transportation and support activities
Waste management and remediation services
Support activities for mining
Primary metals
Motor vehicles, bodies and trailers, and parts
Mining, except oil and gas
Ambulatory health care services
Social assistance

Source: Processed by author using Gephi

6.0 Discussions

6.1 Two clustering and network methods

This section examines inter-industry relationships using input-output (IO) linkages and compares two widely used analytical methods: agglomerative hierarchical clustering (specifically Ward's method) and principal component analysis (PCA). Both approaches aim to uncover the structural patterns embedded in the economy, yet they operate from fundamentally different perspectives and yield distinct clustering results.

Hierarchical clustering offers several advantages. It is transparent and intuitive, as it relies on distance-based metrics and produces a tree-like (dendrogram) representation that visually captures the nested structure of industry relationships. This nested hierarchy aligns well with economic intuitions, for example, some industries are subgroups within broader sectors. Because it is based on overall similarity in IO profiles, industries placed in the same cluster tend to share similar economic functions or roles. However, a key limitation is that hierarchical clustering is mutually exclusive or assigns each industry to only one cluster, making it difficult to represent cases where an industry plays multiple important roles across different value chains or production domains.

PCA, while traditionally used for dimensionality reduction, can also be applied to clustering by interpreting the principal components as latent economic dimensions. These components highlight patterns of co-variation among industries and can be thought of as underlying "axes" of economic activities. A major strength of PCA in this context is its ability to allow for overlapping clusters or an industry sector may have high loadings on multiple clusters or components, reflecting its presence in more than one functional cluster. This flexibility is particularly relevant for industries with diverse inputs or outputs and wider interdependencies. However, PCA-based clustering also presents challenges: the components are abstract and less intuitive to interpret than explicit groupings, and the results are sensitive to the number of components retained. Not all industries will appear meaningfully in the top components, which can be limiting for comprehensive classification.

In summary, both methods are effective but reflect different facets of inter-industry relationships. Hierarchical clustering emphasizes similarity in overall transactional structure and produces clear, exclusive groupings. PCA captures deeper, latent dimensions of economic activities and allows for overlapping roles, though at the cost of interpretability. This report integrated insights from both approaches to inform a more nuanced definition of industry clusters, recognizing the complexity and multi-dimensionality of economic interdependencies.

The network analysis of economic IO data can reveal important insights including demonstrating a way to identify communities based on the network metrics. However, the lack of granular IO data at the annual level is a major limitation. Overall, this study has focused on using the public sources of data, and not the proprietary sources of economic data in the U.S. BEA provides more detailed 2017 IO data, however, there have been major changes in NAICS definitions from 2017 to 2022. For example, two sectors of storage and primary battery manufacturing are merged from 2017 to 2022, which could affect temporal data for battery manufacturing and identification of a subcluster. Major changes are observed in the retail sectors as electronic shopping and mail-order houses selling via e-commerce did not exist in 2017, but were added in 2022 NAICS codes. Digital technology is getting integrated with agriculture and manufacturing including the emerging use of sensors, and AI and data science in manufacturing. Whether these trends can be captured via industry codes or occupations codes remain open for discussions and further research. Similar to agglomerative hierarchical clustering, the network algorithms also identify communities and clusters by using mutually exclusive method. There are advanced methods for network analysis that can identify overlapping communities or clusters which could be areas for further application and research.

6.2 Key industries

This study further analyzes four key industries: semiconductor industry (NAICS: 334413 and 333242), batteries (335911), manufacturing (NAICS start with 31, 32, and 33), and professional services (start with 54).

The semiconductor and related device manufacturing (NAICS 334413) is comprised of establishments primarily engaged in manufacturing semiconductors and related solid-state devices. Examples of products made by these establishments are integrated circuits, memory chips, microprocessors, diodes, transistors, solar cells and other optoelectronic devices. The semiconductor and machinery manufacturing (NAICS 333242) consists of firms primarily engaged in manufacturing wafer processing equipment, semiconductor assembly and packaging equipment, and other semiconductor making machinery. The semiconductor industry is clustered with different industries under the two approaches. In hierarchical clustering, the semiconductor machinery manufacturing (NAICS 333242) cluster with other manufacturing industries (all NAICS code start with 32 or 33). For example, power boiler and hear exchanger manufacturing (332410) and machine shops (332710) are in the same cluster. In PCA, however, the industries in the same cluster of semiconductor machinery manufacturing are more diverse. Most of the industries are professional services (NAICS code start with 54 and 56). One potential reason is most professional services have high value of loadings and have dominated the ranking. Feser (2005) suggested applying weights to enabling sectors to discount those industries. In addition, in PCA, the semiconductor machinery manufacturing only appears in one PC due to loadings cutoff threshold of 0.3, otherwise it would also appear in another PC with other manufacturing industries.

The battery industry (Storage/Primary Battery Manufacturing, NAICS 335911/335912) shows broadly consistent but slightly different clustering outcomes across the two approaches. Under hierarchical clustering, the battery industry is clustered with 16 other industries. This cluster connects core electrical manufacturing with consumer electronics, automotive applications, and industrial control systems, reflecting a technologically integrated supply chain. Under PCA, it is clustered with specialized materials and electrical components cluster, centered around nonferrous metals processing and energy/electrical infrastructure components. While storage battery manufacturing is included in the cluster, primary battery manufacturing was eliminated due to low loadings.

For professional service industries (NAICS code starting with 54), the clustering results show difference in outputs. Under hierarchical clustering, all professional services are grouped into one single cluster. In contrast, PCA clustering reveals a greater heterogeneity, with professional services spread across four distinct clusters. For example, veterinary services (NAICS 541940) is grouped with healthcare-related industries such as hospitals and pharmaceutical manufacturing. Some services (e.g., legal services, specialized design) appear across multiple clusters, suggesting that they serve multiple roles in the economy and share varying degrees of linkage with different industry groups.

7.0 Conclusions

This report explores inter-industry economic relationships using IO linkages. Two primary analytical methods are employed: agglomerative hierarchical clustering and principal component analysis (PCA). The statistical methods are complemented by the network analysis method. By applying these approaches to the national IO data, the study uncovers patterns of industrial interdependence and identifies clusters of industries that share similar production and consumption structures. The analysis includes a close examination of key sectors such as the semiconductor and battery industries, highlighting how their positions within the economy vary depending on the method used. The report concludes with a discussion of the relative strengths and limitations of each method, offering practical guidance for researchers and policymakers seeking to interpret industrial clustering through the lens of IO data.

8.0 References

Argüelles, M., Benavides, C., & Fernández, I. (2014). A new approach to the identification of regional clusters: hierarchical clustering on principal components. *Applied Economics*, *46*(21), 2511–2519.

Atkin, R. H. (1974). *Mathematical structure in human affairs*. Heinemann Educational London.

Bergman, E., & Feser, E. (1999). Industrial and Regional Clusters: Concepts and Comparative Advantages. Reprint. Edited by Scott Loveridge and Randall Jackson. WVU Research Repository, 2020.

Bergsman, J., Greenston, P., and Healy, R. (1972). The Agglomeration Process in Urban Growth. Urban Studies, 9, 3. Pp. 263-288.

Bergsman, J., Greenston, P., and Healy, R. (1975). A Classification of Economic Activities Based on Location Patterns. Journal of Urban Economics. 2. Pp. 1-28.

Blondel, V. D., Guillaume, J. L., Lambiotte, R., and Lefebvre, E. (2008). Fast Unfolding of Communities in Large Networks. Journal of Statistical Mechanics: Theory and Experiment. doi:10.1088/1742-5468/2008/10/P10008.

Boley, D. (1998). Principal direction divisive partitioning. *Data Mining and Knowledge Discovery*, *2*, 325–344.

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research*, *1*(2), 245–276.

Czamanski, S. (1974). Study of Clustering of Industries. Institute of Public Affairs, Dalhousie University, Halifax, Canada.

Czamanski, S. (1971). Some Empirical Evidence of the Strengths of Linkages between Groups of Related Industries in Urban-Regional Complexes. *Papers in Regional Science*, *27*(1), 137–150.

Czamanski, S., and Ablas, L. A. deQ. (1979). Identification of Industrial Clusters and Complexes: A Comparison of Methods and Findings. Urban Studies. 16, pp. 61-80.

Delgado, M., Porter, M. E., & Stern, S. (2016). Defining clusters of related industries. *Journal of Economic Geography*, *16*(1), 1–38.

DePaolis, F., Murphy, P., & Kaluza, M. C. D. (2022). Identifying key sectors in the regional economy: A network analysis approach using input-output data. Applied Network Science. 7:86.

Fabrigar, L. R., & Wegener, D. T. (2012). *Exploratory factor analysis*. Oxford University Press.

Feser, E. J. (2005). Benchmark value chain industry clusters for applied regional research. *Regional Economics Applications Laboratory, University of Illinois at Urbana-Champaign*.

Feser, E. J., & Bergman, E. M. (2000). National industry cluster templates: A framework for applied regional cluster analysis. *Regional Studies*, *34*(1), 1–19.

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

Giuliani, E. (2013). Network dynamics in regional clusters: Evidence from Chile. *Research Policy*, *42*(8), 1406–1419.

Granovetter, M. S. (1973). The Strength of Weak Ties. American Journal of Sociology. 78, 6. Pp. 1360-1380.

Hill, E. W., & Brennan, J. F. (2000). A methodology for identifying the drivers of industrial clusters: The foundation of regional competitive advantage. *ECONOMIC DEVELOPMENT QUARTERLY*, *14*(1), 65–96. https://doi.org/10.1177/089124240001400109.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, *30*(2), 179–185.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, *24*(6), 417.

Huallachain, B. O. (1984). The identification of industrial complexes. *Annals of the Association of American Geographers*, *74*(3), 420–436.

Iacobucci, D., McBride, R., Popovich, D.L., & Rouziou, M. (2017). In Social Network Analysis, Which Centrality Index Should I Use? Theoretical Differences and Empirical Similarities among Top Centralities. *Journal of Methods and Measurements in the Social Sciences*. 8, 2. P. 72-99.

Jackson, M. O. (2008). Social and Economic Networks. Princeton University Press: New Jersey, U.S.

Jolliffe, I. T. (2002). *Principal component analysis for special types of data*. Springer.

Katz, L. (1953). A New Status Index Derived from Sociometric Analysis. *Psychometrika*. 18, 1.

Ketchen, D. J., & Shook, C. L. (1996). The application of cluster analysis in strategic management research: an analysis and critique. *Strategic Management Journal*, *17*(6), 441–458.

Kononenko, I., & Kukar, M. (2007). *Machine learning and data mining*. Horwood publishing.

Liang, S., Zhang, T. Z., Wang, Y. F., & Jia, X. P. (2012). Sustainable urban materials management for air pollutants mitigation based on urban physical input-output model. *ENERGY*, *42*(1), 387–392. https://doi.org/10.1016/j.energy.2012.03.038.

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., & Hornik, K. (2025). *cluster: Cluster Analysis Basics and Extensions* (2.1.8.1). R Foundation for Statistical Computing. https://CRAN.R-project.org/package=cluster.

Marshall, A. (1920). *Principles of economics*. Springer.

Mesa-Arango, R., and Kumar, I. (2017). Hierarchical Value Chains Encompassing Freight Transportation and Logistics Sectors in the United States. Transportation Research Record. 2609. Pp. 1-10.

Montresor, S., & Marzetti, G. V. (2009). Applying social network analysis to input–output based innovation matrices: An illustrative application to six OECD technological systems for the middle 1990s. *Economic Systems Research*, *21*(2), 129–149.

Nuss, P., Chen, W. Q., Ohno, H., & Graedel, T. E. (2016). Structural Investigation of Aluminum in the U.S. Economy using Network Analysis. *ENVIRONMENTAL SCIENCE & TECHNOLOGY*, *50*(7), 4091–4101. https://doi.org/10.1021/acs.est.5b05094.

O'hUallachain, B. (1984). The Identification of Industrial Complexes. *Annals of the Association of American Geographers*. 74(3): 420-436.

Roepke, H., Adams, D., & Wiseman, R. (1974). A new approach to the identification of industrial complexes using input-output data. *Journal of Regional Science*, *14*(1).

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*, 53–65.

Roux, M. (2018). A Comparative Study of Divisive and Agglomerative Hierarchical Clustering Algorithms. *Journal of Classification*, *35*(2), 345–366. https://doi.org/10.1007/s00357-018-9259-9.

Schweitzer, F., Fagiolo, G., & White, D. R. (2009). Economic networks: What do we know and what do we need to know? Advances in Complex Systems. Vol. 12, Nos. 4&5, 407-422.

Slater, P. B. (1977). The determination of groups of functionally integrated industries in the United States using a 1967 interindustry flow table. *Empirical Economics*, *2*(1), 1–9.

Sokal, R. R., & Sneath, P. H. A. (1963). *Principles of numerical taxonomy*.

Sonis, M., & Hewings, G. J. D. (1997). Symposium: Theoretical and Applied Input-Output Analysis: A New Synthesis Part I: Structure and Structural Changes in Input-Output Systems. *Studies in Regional Science*, *27*(1), 233–256.

Sonis, M., & Hewings, G. J. D. (2000). Introduction to input-output structural q-analysis. *Regional Economics Applications Laboratory*.

Streit, M. E. (1969). Spatial Associations and Economic Linkages Between Industries. Journal of Regional Science. 9, 2.

Thorndike, R. L. (1953). Who Belongs in the Family? *Psychometrika*, *18*(4), 267–276. https://doi.org/DOI: 10.1007/BF02289263.

Tibshirani, R., Walther, G., & Hastie, T. (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), 411–423. https://doi.org/10.1111/1467-9868.00293.

Titze, M., Brachert, M., & Kubis, A. (2011). The identification of regional industrial clusters using qualitative input–output analysis (QIOA). *Regional Studies*, *45*(1), 89–102.

Vom Hofe, R., & Bhatta, S. D. (2007). Method for Identifying Local and Domestic Industrial Clusters Using Interregional Commodity Trade Data. *Industrial Geographer*, *4*(2).

Ward Jr., J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, *58*(301), 236–244.

Xu, M., & Liang, S. (2019). Input-output networks offer new insights of economic structure. *Physica A,* 527, 121178. https://doi.org/10.1016/j.physa.2019.121178.