Structural analysis of regional economies based on I/O data

Lawrence Ho
lh811@scarletmail.rutgers.edu
Rutgers University
New Brunswick, New Jersey, USA

Ryan Vukicevic
rcv35@scarletmail.rutgers.edu
Rutgers University
New Brunswick, New Jersey, USA

Paolo Gervasoni
pg681@scarletmail.rutgers.edu
Rutgers University
New Brunswick, New Jersey, USA

Abstract

Using input–output economic data compiled by the University of Groningen's Growth and Development Centre, we compute Leontief inverse matrices to capture domestic intersectoral dependencies within national economies. These matrices are used as structural representations of economic production networks. We apply both supervised and unsupervised machine learning methods to investigate whether similarities and differences in sectoral output structures correspond to geographic or political regional groupings. Dimensionality reduction, K-nearest neighbors classification, decision tree models, and K-means clustering are implemented to analyze patterns across countries. Our results evaluate the extent to which Leontief inverse matrices reliably group regional economic characteristics and assess their effectiveness for classifying national economies. This study demonstrates how classical input/output economic models can be integrated with modern machine learning techniques for comparative structural economic analysis.

Keywords

Input/output economics, Leontief inverse, machine learning, economic networks, clustering

ACM Reference Format:
Lawrence Ho, Ryan Vukicevic, and Paolo Gervasoni. 2025. Structural analysis of regional economies based on I/O data. In *Proceedings of 01:198:439 Data Science Course Project (CS439 Course Project).* ACM, New York, NY, USA, 11 pages.

1 Introduction

Understanding the structural differences between national economies is central to comparative economic analysis and policy evaluation. Traditional macroeconomic indicators such as GDP or aggregate output lack nuanced information about interactions between industries that shape underlying economic structure. Input/output (I/O) models provide a framework for capturing these inter-industry relationships by describing how the output of one sector is used as an input by others within an economy.

Understanding the structure of economies has numerous practical applications beyond pure economic theory. In finance, knowledge of inter-industry dependencies can inform investment decisions and risk assessment, as shocks in one sector may propagate through supply chains affecting other industries and financial markets. Policymakers can use structural insights to design targeted interventions, such as stimulus packages or trade policies, that account for the cascading effects of changes in one sector on the broader economy. Similarly, businesses can leverage input/output analysis for strategic planning, identifying critical suppliers, potential drawbacks, and opportunities for diversification. Overall, analyzing economic structures provides a foundation for forecasting, resilience assessment, and informed decision-making across public and private sectors.

In this project, we analyze the structural composition of 25 economies over 35 years using input/output tables and their corresponding Leontief inverse matrices. For each country and year, we construct a technical coefficient matrix A, where each entry represents the input requirement from one sector per unit of output of another. The standard Leontief production model is given by

$$x = Ax + y,$$

where x denotes the vector of total sectoral output and y represents final demand. Solving for total output yields

$$x = (I - A)^{-1}y,$$

where $L = (I - A)^{-1}$ is the Leontief inverse. This matrix captures both direct and indirect production requirements across sectors and allows us to determine the total output x that each sector must produce in order to satisfy a given vector of final demand y. In this sense, the Leontief inverse provides a complete structural view of the economy by linking final demand to the production levels required in each sector.

We compute Leontief inverse matrices for each country using input/output data compiled by the University of Groningen's Growth and Development Centre. These matrices serve as structural representations of national production networks, encoding the flow of intersectoral dependencies. By treating each Leontief matrix as a high-dimensional feature representation, we apply machine learning techniques to compare economic structures across countries.

Specifically, this study addresses the following research questions:

- What are the structural similarities and differences in sectoral outputs across national economies?
- Can supervised methods, such as K-nearest neighbors (KNN) and decision trees, differentiate countries regionally based on their Leontief matrices?
- How reliable are Leontief inverse matrices for classifying or grouping countries by geographic or political regions?
- Can unsupervised clustering methods, such as k-means, reproduce the same regional distinctions observed using supervised classification?
- Which sectors or inter-industry relationships are most influential in distinguishing regional economic structures?

Through these questions, the project investigates the extent to which classical input–output economic models can be integrated with modern machine learning tools for comparative structural

CS439 Course Project , Rutgers University, New Brunswick, NJ
2025.

analysis. The results have potential applications in regional economic policy, investment risk assessment, and strategic business planning, providing a framework for understanding how intersectoral dependencies shape national and regional economic behavior.

2 EDA

2.1 Dataset Overview

This study uses input-output data from the initial release of the World Input-Output Database (WIOD), compiled by the University of Groningen's Growth and Development Centre.[1] The dataset provides detailed inter-industry transaction tables for 25 countries covering the period from 1965 to 2000. Each table represents a snapshot of economic activity for a specific country and year.

The WIOD tables record interactions between industries classified according to the International Standard Industrial Classification (ISIC). In each table, rows correspond to the production of a sector, while columns correspond to output of a sector. The values in the matrix represent monetary flows between sectors, expressed in millions of U.S. dollars. These intersectoral transactions form the intermediate demand matrix, which captures how outputs from one industry are used as inputs by others within the same economy.

In addition to intermediate sector–sector transactions, the dataset includes components of final demand. These consist of household consumption expenditure (C), government consumption expenditure (G), gross fixed capital formation (I), and changes in inventories (ΔI). For each sector, total output is computed as the sum of intermediate demand and final demand components, providing the basis for constructing Leontief Inverse matrices. Besides the computation of the (L) matrix, there was not much preprocessing done to the data.

As part of the exploratory data analysis (EDA), we examine the distribution of sectoral outputs, the relative magnitude of intermediate versus final demand, and cross-country differences in sectoral composition. These exploratory steps help verify data consistency, identify dominant industries, and assess structural differences across economies prior to model construction. The cleaned and validated tables are then used to compute each country's input-output coefficient matrix A and corresponding Leontief inverse matrix $L = (I - A)^{-1}$, which serve as the primary inputs for subsequent machine learning analysis.

2.2 EDA and Visualizations

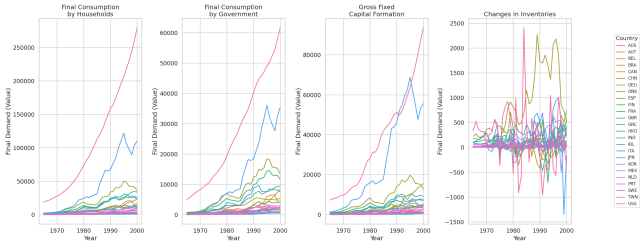

Figure 1: Mean sectoral output over time by geographic region, averaged across countries within each region.

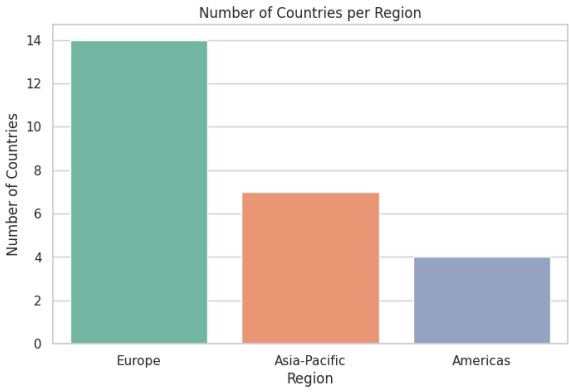

Figure 2: Mean sectoral output over time by geographic region, averaged across countries within each region.

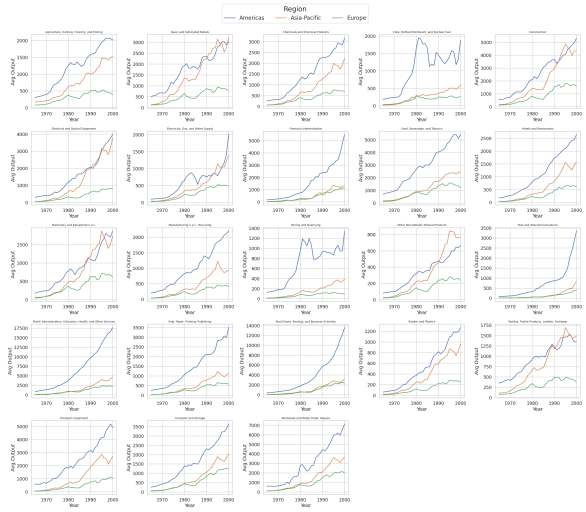

Figure 3: Mean sectoral output over time by geographic region, averaged across countries within each region.

To gain an initial understanding of the dataset, we construct several exploratory visualizations focusing on final demand components and regional composition. In particular, we examine trends in components of GDP, defined as $Y = C + I + G + NX$, where C or Consumption is denoted by household consumption, G denotes government expenditure, and I or Investments is represented by gross fixed capital formation/(ΔI), across the 25 countries over the period 1965–2000.

Figure 1 presents time-series plots of aggregate final demand components. Across all countries, C, G, and I exhibit a clear upward trend over time. This pattern is consistent with long-run economic growth driven by increases in productivity, income, and capital accumulation. These trends are observed without adjusting for inflation and therefore reflect growth in nominal values. Changes in inventories (ΔI) display greater variability over time, reflecting

[1] https://www.rug.nl/ggdc/valuechain/long-run-wiod

short-run market dynamics, demand fluctuations, and inventory management behavior rather than steady long-term growth.

In addition to temporal trends, we visualize the geographic distribution of countries in the dataset.

Figure 2 illustrates the regional composition of the data, which includes 14 European economies, 7 economies from the Asia-Pacific region, and 4 from the Americas. A bar graph summarizing regional representation highlights the imbalance toward European countries, which is an important consideration when interpreting regional clustering and classification results. Although Australia is classified as part of the Asia-Pacific region in this analysis, it can be more accurately categorized as Oceania; however, it is the only country in Oceania and is grouped within the Asia-Pacific region to prevent class imbalances.

We visualize the evolution of mean sectoral output over time, grouped by geographic region. Figure 3 shows the average output across sectors for each region, calculated by aggregating the sector-level output within each country and then averaging with countries in the same region. This representation allows for a comparison of broad structural output patterns while reducing country-specific noise. The visualization reveals systematic differences in the scale and growth trajectories of sectoral output across regions. European economies exhibit relatively stable and gradual growth patterns, while Asia-Pacific economies display more pronounced growth over time, consistent with industrialization and export-oriented development during the latter half of the sample period. Economies in the Americas show extreme growth patterns over time that consistently dominate the other two regions.

Although these trends primarily reflect differences in scale and growth rather than detailed economic structure, they provide important contextual insight into regional economic evolution.

We also examine summary statistics of intersectoral transaction magnitudes across countries. Figure 4 displays the median, mean, minimum, and maximum values of the input/output flows for each country, given by the intermediate demand or I/O matrices over the period 1965-2000 and expressed in billions of U.S. dollars. For each country, these statistics are calculated by aggregating sector-sector transaction values across all years, providing a compact representation of the scale and dispersion of inter-industry flows.

The visualization highlights substantial differences in the magnitude of inter-sector transactions across countries. Large economies such as the United States, Japan, and Germany consistently exhibit higher mean and maximum flow values, indicating more extensive inter-industry activity and greater economic scale. In contrast, smaller economies display more compressed distributions with lower maximum values. The gap between minimum and maximum values within countries also reflects differences in sectoral concentration and the presence of dominant industries, particularly in advanced economies. Using the USA as an example, after 1980, the lower bound of inter-sector interactions began decreasing consistently. This is indicative of a specializing economy where certain sectors no longer are major suppliers to other industries, reflecting increased sectoral concentration, outsourcing of intermediate production, or a shift toward service-oriented activities that rely less on domestic inter-industry inputs. These scale differences underscore the importance of using normalized structural representations, such as technical coefficient matrices and Leontief inverse matrices, in analysis of economic structure. Normalization ensures that machine learning models capture underlying relationships rather than absolute size effects, allowing for meaningful comparisons across countries of varying economic scale.

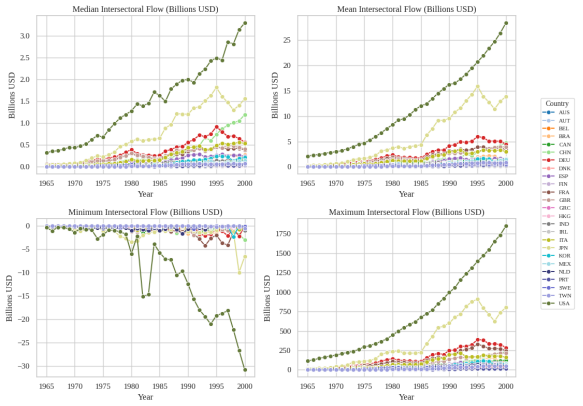

Figure 4: Summary statistics (minimum, median, mean, and maximum) of inter-sector transaction values by country, aggregated over 1965–2000.

3 Classification Methods and Results

3.1 KNN Model Fit for Each Year

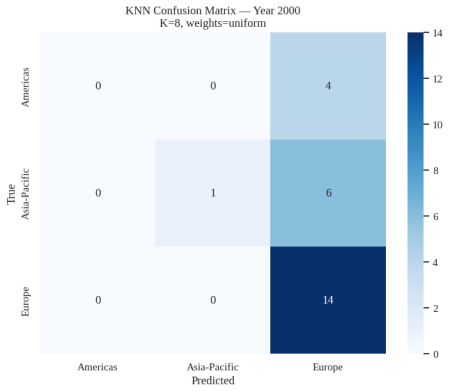

Figure 5: Confusion matrix of the KNN model trained and evaluated on the year 2000. The matrix illustrates the model's classification performance across the three geographic regions.

We fit a K-Nearest Neighbors (KNN) model separately for each year in the dataset to classify countries by region using their flattened Leontief matrices. Models trained on individual years yielded relatively low accuracy. The small sample size of 25 countries limits the model's ability to capture distinguishing structural features, and class imbalances often lead the model to predict the dominant region.

Cross-validation results confirm these limitations: the average CV accuracy for the optimal yearly models never exceeded approximately 85%, with many years achieving only 70–75%. Figure 5 shows the confusion matrix for the KNN model trained and evaluated on the year 2000, illustrating the model's classification performance and the impact of class imbalances. Visualizations of additional cross-validation results are provided in our accompanying notebook and slides.

All models used an 80/20 train/test split.

3.2 KNN Model Fit on All Available Data

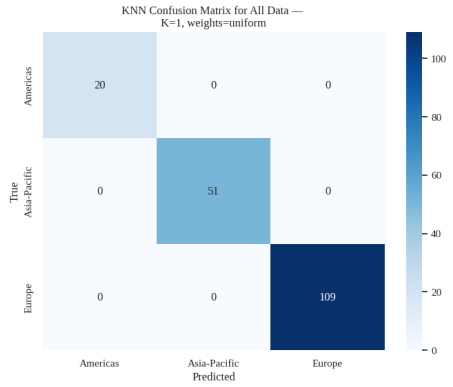

Figure 6: Confusion matrix of the KNN model trained on all data and evaluated on the Test Set, illustrating classification performance across the three geographic regions.

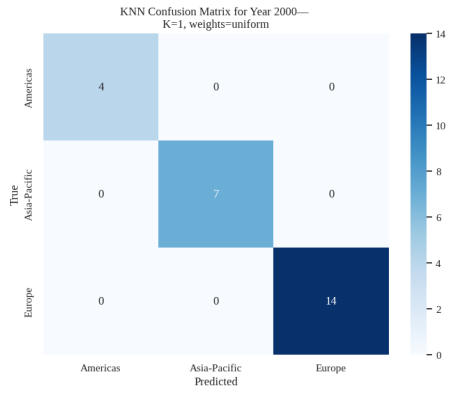

Figure 7: Confusion matrix of the KNN model trained on all data and evaluated on the year 2000, illustrating classification performance across the three geographic regions.

We trained a K-Nearest Neighbors (KNN) model using all available data from 1965–2000, with an 80/20 train/test split. The model achieved perfect classification accuracy on the test set, indicating that the Leontief matrices contain strong structural signals that distinguish regional economies.

Evaluating the model on individual years also produces perfect accuracy; Figure 6 shows the confusion matrix for the year 2000, further supporting the presence of systematic structural differences across regions. However, we observe that the optimal number of neighbors k tends to be low, suggesting potential issues related to the high dimensionality of the feature space relative to the small sample size.

These results support the presence of structural differences across regional economies (via the (L matrices), while also highlighting the challenges of modeling high-dimensional economic data using Euclidean distances.

3.3 Dimensionality

Each Leontief inverse matrix in our dataset is square, representing 23 sectors and their sector-sector interactions. When flattened for machine learning, each matrix becomes a feature vector of size 1×529 (since $23 \times 23 = 529$). With 25 countries and 35 years of data, this results in a high-dimensional dataset containing 529 features per country per year.

In such high dimensions, Euclidean distances become less robust because points tend to be "close together," reducing the sensitivity of distance-based models such as KNN. To address this issue, we apply Principal Component Analysis (PCA) to reduce dimensionality while retaining the majority of structural information. This approach allows us to maintain model robustness and facilitate more effective classification and clustering of the economic structures.

3.4 PCA on All Available Data and Model Fitting on Reduced Dimensionality

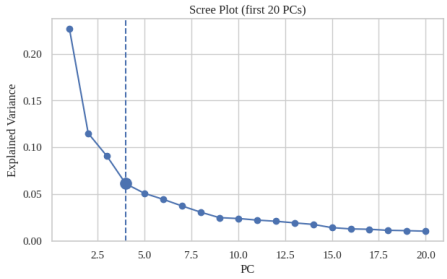

Figure 8: Scree plot with optimal PC=5

Using the same training set employed for the previous KNN models, we performed Principal Component Analysis (PCA) on the flattened Leontief inverse matrices. Figure 8 presents the scree plot showing the cumulative explained variance captured by the first 20 principal components.

To identify the optimal number of components, we applied the Kneedle algorithm, which detects the "elbow point" where the cumulative variance curve transitions from steep to noticeably flatter. In this case, the Kneedle method indicates that the first 4 principal components capture the most meaningful structural variation, after which additional components provide diminishing marginal gains.

Projecting the data onto these 4 principal components reduces the dimensionality while retaining the dominant patterns in the Leontief structure. This reduced representation is then used to

train KNN and other machine-learning models, mitigating issues associated with high-dimensional Euclidean distance metrics.

After fitting the KNN classifier on the reduced feature space, we evaluated it on the same test set used earlier. The resulting confusion matrix (Figure 9) shows two misclassified instances which is slightly worse than the perfect accuracy achieved using the full dataset, but, still demonstrating strong regional separability. The robustness of the classifier under such an aggressive reduction in dimensionality suggests that major structural differences across regions exist within respective Leontief Inverse matrices. Testing again on a subset of 2000 data, we obtain perfect metrics similar to what we saw in the previous high dimensional model.

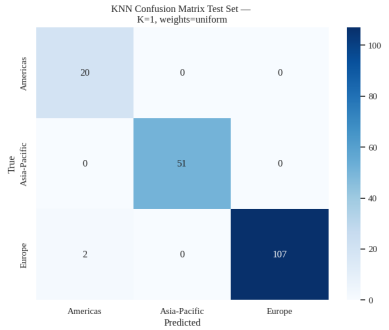

Figure 9: Confusion matrix of the KNN model trained on the reduced PCA representation (4 principal components) and evaluated on the Test set.

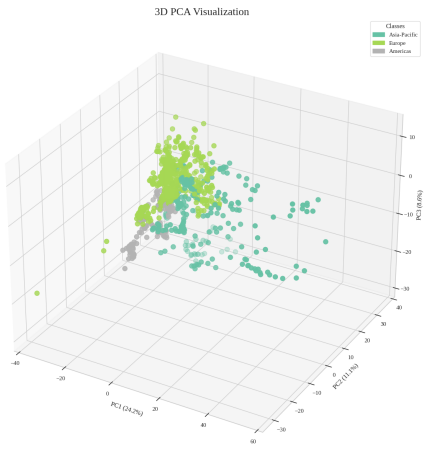

Figure 10: 3D PCA projection that uses the first three principal components. Regional clusters show partial separation, which indicates high dimensional structural differences across economies.

To visualize the reduced feature space, we projected the data into two and three principal components. The 2D PCA plot shows noticeable overlap between regions, which indicates that the strongest variance components alone do not fully separate regional structures. In contrast, the 3D projection (Figure 10) reveals clearer, although still imperfect, cluster separation. This suggests that regional differences are present but inherently multidimensional, which is consistent with the strong classification performance observed when using four principal components.

3.5 Forecast-ability of L Matrices

We trained a KNN model on data from 1965–1990 to forecast regional classifications for years beyond 1990. If the model can still differentiate regional economies, this supports the existence of persistent structural differences across regions. It is important to note that applying this approach as a form of time-series forecasting relies on strong assumptions, including the stability of structural relationships over time and the lack of disruptions in typical intersectoral patterns.

The model achieved the following metrics:

- Train Accuracy: 0.9985
- Test Accuracy: 0.948

Figure 11 shows the confusion matrix of the forecast evaluation. While overall performance is strong, a few misclassifications occur, raising questions about whether class imbalance or unusual economic events drive errors.

Figure 11: Confusion matrix for the forecasting experiment using the 1965–1990 training period.

3.6 Investigating Misclassification

To understand the sources of forecast errors, we examined the country–year observations that were misclassified. Figure 12 visualizes these instances. Key observations include:

- **Korea (1998).** Foreign investment collectively pulled out, leading to a sharp contraction (-6.7% GDP), the worst in modern Korean history.
- **Taiwan (1990s).** Massive political overturning and democratization may have affected key sectors, particularly government spending.
- **Canada (late 1990s).** Freshly out of an early 1990s recession, Canada likely experienced a temporary post-recession expansion before stabilizing.
- **Greece (1999).** Implementation of labor laws affecting foreign nationals and high national debt may have influenced sector-to-sector behavior.

These observations suggest that unusual economic events or structural transitions can temporarily disrupt regional economic patterns. Despite these exceptions, the strong overall classification accuracy demonstrates that the Leontief matrices capture persistent structural differences across regions.

Misclassified future-year cases:

	Country	Year	True Region	Predicted Region
0	CAN	1995	Americas	Europe
1	CAN	1996	Americas	Europe
2	CAN	2000	Americas	Europe
3	GRC	1999	Europe	Americas
4	KOR	1998	Asia-Pacific	Europe
5	TWN	1993	Asia-Pacific	Europe
6	TWN	1994	Asia-Pacific	Europe
7	TWN	1995	Asia-Pacific	Europe
8	TWN	1996	Asia-Pacific	Europe
9	TWN	1997	Asia-Pacific	Europe
10	TWN	1998	Asia-Pacific	Europe
11	TWN	1999	Asia-Pacific	Europe
12	TWN	2000	Asia-Pacific	Europe

Figure 12: Visualization of misclassified country–year observations in the forecast evaluation.

3.7 PCA on Forecasting Data and Model Fitting on Reduced Dimensionality

To address the same dimensionality concerns as earlier (3.4) for the forecasting experiment, we applied PCA with the same methodology in 3.4 to the same forecasting data. Figure 13 presents the scree plot showing the cumulative variance captured by the first 20 principal components.

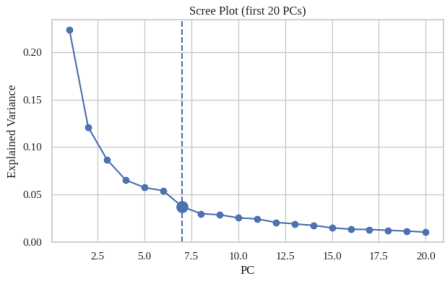

Figure 13: Scree plot with optimal PC=7

Once again for visualization purposes, we projected the data into two and three PC's. Plots are similar with 3D projections (Figure 14) showing clearer cluster separation between the three regions.

After reducing dimensionality, we refit the KNN classifier to predict regional classifications for years beyond 1990. As seen in

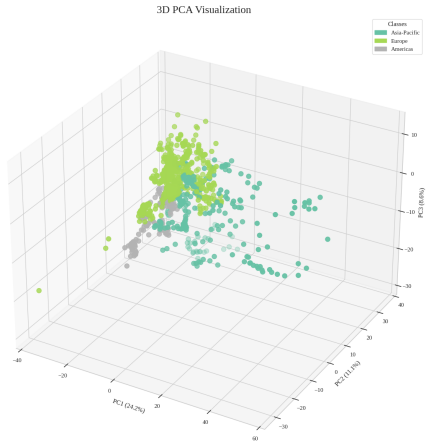

Figure 14: 3D PCA projection that uses the first three principal components. Regional clusters show partial separation, which indicates high dimensional structural differences across economies.

(Figure 15), the model achieved a training accuracy of 1.0 but a test accuracy of only 0.864.

The large gap between training and test performance indicates overfitting and suggests that aggressive dimensionality reduction via PCA removed variance that was critical for distinguishing regions. This loss of predictive power is particularly evident in the model's tendency to overpredict the Europe class, reflecting both the class imbalance in the dataset and the inability of the reduced feature space to fully capture the structural differences between regions in unseen data. These results also highlight the trade-off between dimensionality reduction for robustness as predictive power dropped off after PCA removed variance critical for differentiating regions.

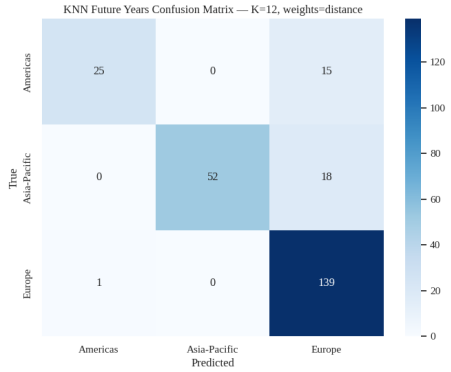

Figure 15: Confusion matrix for the forecasting experiment with reduced dimensions using the 1965–1990 training period.

Future Years Evaluation Metrics

Accuracy: 0.864

Precision (macro): 0.9232
Recall (macro): 0.7869
F1 Score (macro): 0.8337

The reduction in recall relative to precision highlights how PCA compresses minority-class structure more aggressively, limiting the model's ability to identify smaller or more nuanced regional groups. This trade-off demonstrates the tension between reducing dimensionality for robustness and retaining the variance necessary for regional discrimination.

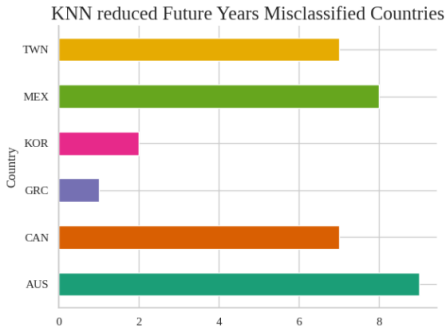

Figure 16: Countries the reduced model misclassified

3.8 Takeaway

There are inherent regional differences in economic structure. This is reflected in the strong performance of the classifiers that use the full Leontief inverse matrices. However, these differences appear to depend on very nuanced and specific industry and sector compositions. The PCA results support this interpretation. After dimensionality reduction, the model performs slightly worse, which indicates that the most important signals for distinguishing regions are embedded in fine-grained sector interactions. Since PCA generalizes feature variance across components to reduce dimensionality, it removes some of the precise information that drives accurate classification.

3.9 Decision Tree Analysis

To improve interpretability, we trained a Decision Tree classifier on all available data to identify the specific sector-to-sector interactions that distinguish the three regions. Although the resulting tree is moderately complex, the model achieves strong and balanced performance. The training accuracy is 1.0, and the test accuracy is 0.983, suggesting that the tree captures stable structural patterns rather than overfitting noise.

Feature importances are computed as the total reduction in impurity (e.g., Gini) contributed by each feature. For each split in the tree, the decrease in impurity is attributed to the splitting feature, and these reductions are summed across all splits. Higher feature importance values indicate that a feature plays a larger role in guiding the tree's decisions.

Figure 17 presents the feature importance plot, highlighting the sector interactions that contribute most to regional differentiation. These results align with earlier findings that fine-grained structural relationships drive classification performance.

Economies with lower combined input dependencies from Construction and Wholesale/Retail Trade & Repairs are largely classified as the Americas, reflecting less reliance on upstream inputs from these sectors. Further splits show that low dependencies from Electricity, Gas, and Water Supply relative to Textiles and Mining refine this classification, indicating minimal reliance on utilities in manufacturing processes. European economies are characterized by moderate inter-dependencies among Mining, Construction, and Chemicals sectors, suggesting a balanced network of upstream inputs. In contrast, Asia-Pacific economies are distinguished by low input dependencies from Chemicals, Utilities, and Services, highlighting structural patterns where certain sectors rely less on upstream inputs. Overall, the tree captures interpretable structural patterns in sectoral interactions, with each leaf representing a group of economies exhibiting similar intersectoral dependency structures rather than merely comparable output shares.

A full-depth version of the decision tree is provided in the accompanying code notebook for completeness.

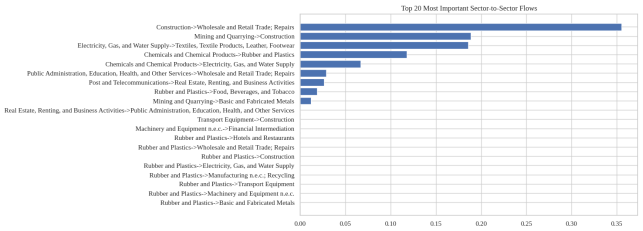

Figure 17: Feature importance plot from the Decision Tree classifier. The most influential sector-to-sector interactions for differentiating regions are highlighted.

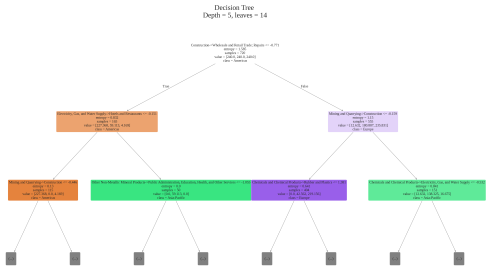

Figure 18: Condensed version of the Decision Tree classifier. This simplified tree highlights the main decision paths used to classify regions based on sectoral structure.

4 Classification Summary + Introduction to Clustering

Across the analysis, several key patterns emerge regarding regional economic structure. Models trained on individual years using K-Nearest Neighbors (KNN) struggled to reliably classify countries, largely due to the small sample size (25 countries) and class imbalances. Visually, cross-validation accuracy for single-year models

rarely exceeded 85%, and many years achieved only 70–75% (Figure 5). This confirms that annual samples alone are insufficient for robust classification.

Training on all available data from 1965–2000, however, produced nearly perfect classification accuracy (Figures 6 and 7), demonstrating that Leontief inverse matrices contain strong and persistent structural signals distinguishing regional economies. Dimensionality remains a challenge: each flattened inverse matrix has 529 features, and Euclidean distances in such high-dimensional spaces are less discriminative. PCA-based reduction mitigates this issue by projecting data onto the first four principal components, retaining the dominant structural variance. Classification on the reduced dataset remains highly accurate, with only minor misclassifications (Figure 9), suggesting that the primary structural differences across regions are captured by the top principal components. The 3D PCA projection (Figure 10) illustrates partial cluster separation, reinforcing that regional differentiation is inherently multidimensional.

Forecasting experiments using data from 1965–1990 to predict post-1990 classifications further confirm persistent structural patterns. While overall performance remains strong (Figure 11), misclassifications often correspond to countries experiencing unusual economic events, such as Korea in 1998 or Greece in 1999 (Figure 12). PCA-based dimensionality reduction for forecasting improves robustness but reduces predictive power (Figures 13, 14, 15, 16), highlighting a trade-off between generalization and preservation of fine-grained structural signals.

Decision Tree analysis complements the KNN findings by providing interpretable insights into sector-to-sector interactions driving regional differences. Feature importance scores (Figure 17) show which sectoral dependencies most influence classification, and the condensed decision tree (Figure 18) reveals the key decision paths distinguishing regions. Across the Americas, Europe, and Asia-Pacific, distinct patterns of sectoral input dependencies emerge, reinforcing that regional differences are nuanced, multidimensional, and embedded in specific industry structures.

In summary, the classification analysis indicate that regional economic structures exhibit strong, persistent differences that are detectable through Leontief inverse matrices. While single-year models face limitations, combining multi-year data and careful dimensionality reduction captures the essential structural patterns. These results demonstrate both the power of structural representations for economic classification and the importance of addressing high-dimensionality in model design.

Having established that regional economic structures are distinguishable using supervised learning methods, we now explore whether these patterns emerge without relying on pre-defined labels. Specifically, we apply unsupervised clustering to the flattened Leontief inverse matrices and their PCA-reduced representations. This approach allows us to assess whether inherent structural differences naturally group countries into similar regional clusters, providing a complementary perspective on the patterns observed in the KNN and Decision Tree analysis.

5 Clustering Methods and Results

5.1 K-Means Clustering Analysis

To explore whether unsupervised methods can recover the regional structure observed in supervised models, we applied K-Means clustering with $K = 3$, corresponding to the three geographic regions: Europe, Asia-Pacific, and the Americas. Clustering was performed on:

- All available data (1965–2000)
- Forecasting subset (1965–1990 training data)

Cluster labels were assigned based on the majority vote of countries within each cluster. Confusion matrices were used to evaluate how well the clusters correspond to the actual regions.

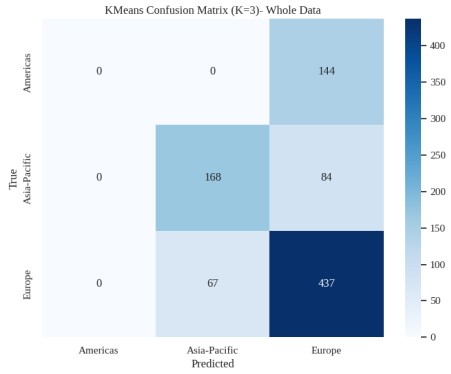

Figure 19: Confusion matrix for K-Means clustering on all available data. Clusters are labeled by majority vote to correspond to the three geographic regions.

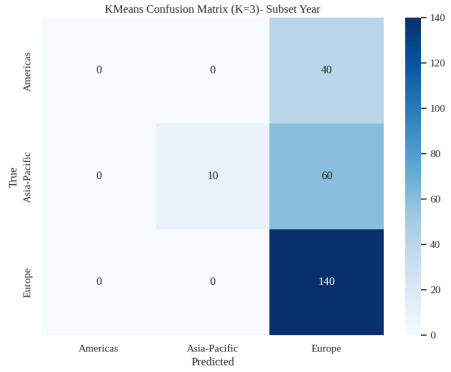

Figure 20: Confusion matrix for K-Means clustering on the forecasting subset (1965–1990 training data). Clusters are labeled by majority vote.

We also applied K-Means clustering to the PCA-reduced dataset using the optimal principal components identified in earlier analysis. Despite the dimensionality reduction, the clustering did not produce strong separation among all three regions, unlike the clearer distinctions observed with the KNN models.

Clustering into three groups shows limited sensitivity, which is expected given the aggregated nature of clusters. Theoretically, if the three regions (Europe, Asia-Pacific, Americas) were well-separated, class imbalance effects would be minimal, as the three classes would naturally form distinct clusters. Achieving clean distinctions may require finer regional breakdowns (e.g., East Asia, Eurasia, Western Europe, North America). Data limitations highlighted in the EDA also constrain the ability to fully separate the regions.

Nonetheless, some separation between Asia-Pacific and Europe is visible. A potential approach to improve clustering performance includes distinguishing East (Asia-Pacific) from West (Americas & Europe) through targeted feature engineering or hierarchical clustering methods.

5.2 Exploration of East/West Regions

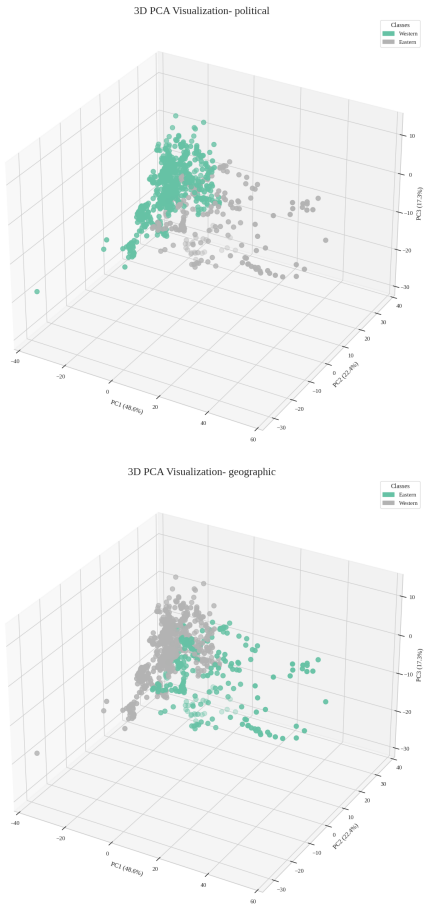

Figure 21: 3D PCA projection showing separation of East vs. West classes, both for political and geographic definitions.

To assess broader structural patterns, we explored clustering and classification using East vs. West groupings rather than the three original regions. The silhouette score measures how similar a point is to its own cluster compared to the nearest other cluster. Values near 1 indicate well-separated clusters, while values close to 0 suggest overlapping clusters. Using this metric, the data naturally splits into two clusters:

- Optimal K by silhouette: 2
- Supports an East vs. West distinction rather than three distinct groups

Geopolitical considerations were accounted for:

- Australia is geographically Eastern but often classified as Western; both perspectives are considered in clustering.
- South America, despite being a separate continent, historically has strong European influence and is treated as Western.

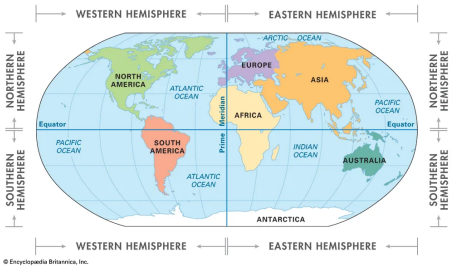

Figure 22: Illustration of East vs. West regions used for clustering and classification.

We refit earlier classification models with these new East/West labels. Key observations include:

- KNN performs well on this broader classification, but forecasting accuracy is affected by class imbalance.
- Figures 21 shows PCA effectively separates the new East/West classes, unlike the nuanced three-region distinctions.
- K-Means captures broader separations better than specific regional distinctions due to its focus on aggregate patterns.

EDA of the new classes highlights strong class imbalance, which may impact model performance. Figure 23 illustrate country counts by geographic and political definitions, respectively.

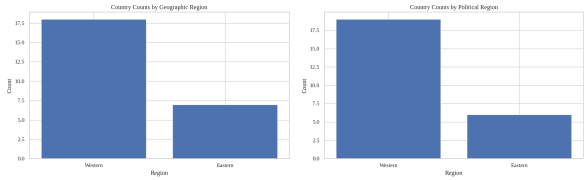

Figure 23: Bar graphs of country counts by geographic (left) and political (right) East/West classifications.

Decision Trees trained on the new East/West labels differ substantially from previous regional trees:

- Political and geographic trees have completely different structures. (Figure 24)
- Top splits are not consistent across the two trees.
- Feature importance distributions are extremely different for both Political and Geographic labels.

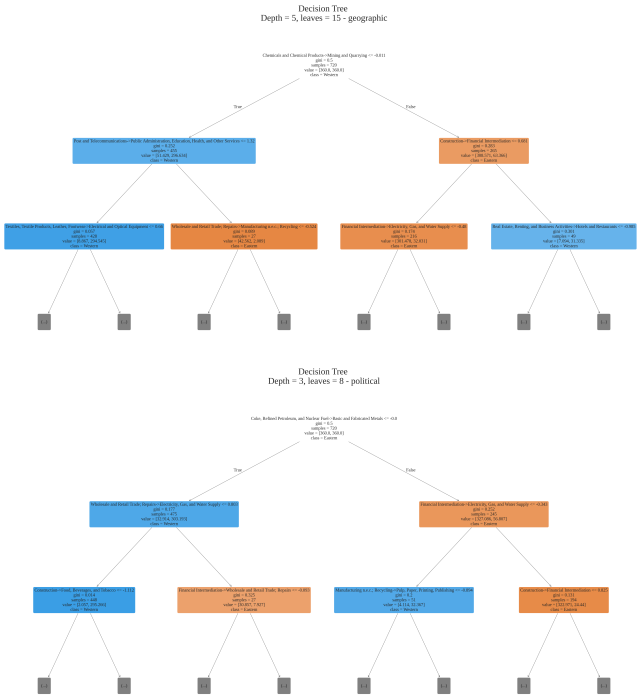

Figure 24: Decision Trees for East/West classification: geographic (left) vs political (right).

Full visualizations and additional analysis are available in the accompanying notebook.

5.3 Clustering of East/West Regions

To explore broader structural patterns, we clustered the data into two classes representing East and West regions, using both political and geographic labels.

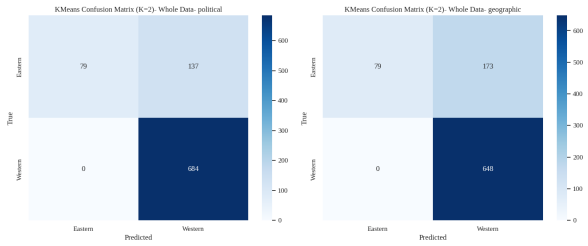

Figure 25: K-Means clustering confusion matrices for Whole Data. Left: Political labels, Right: Geographic labels.

5.3.1 *Clustering Results: Whole Data.*

- **Political Labels:**
  - Accuracy: 0.848
  - Precision (macro): 0.917
  - Recall (macro): 0.683
  - F1 Score (macro): 0.722
- **Geographic Labels:**
  - Accuracy: 0.808
  - Precision (macro): 0.895
  - Recall (macro): 0.657
  - F1 Score (macro): 0.680

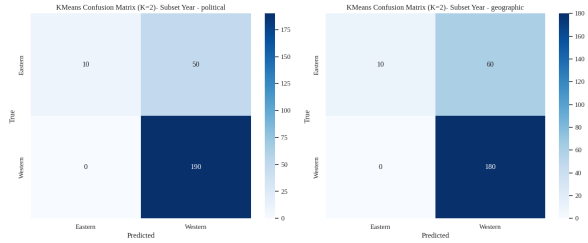

Figure 26: K-Means clustering confusion matrices for Forecasting Subset (1965–1990). Left: Political labels, Right: Geographic labels.

5.3.2 *Clustering Results: Forecasting Subset.*

- **Political Labels:**
  - Accuracy: 0.800
  - Precision (macro): 0.896
  - Recall (macro): 0.583
  - F1 Score (macro): 0.585
- **Geographic Labels:**
  - Accuracy: 0.760
  - Precision (macro): 0.875
  - Recall (macro): 0.571
  - F1 Score (macro): 0.554

6 Clustering Summary

K-Means clustering tends to favor large, structurally similar clusters, resulting in a dominant Western cluster. Average accuracy is influenced by class imbalance, as the model often predicts the Western class by default and only assigns observations to the Eastern class when highly confident. Consequently, recall for Eastern or otherwise structurally distinct groups is relatively low, whereas precision remains high. Smaller or structurally distinct regions, such as Asia-Pacific and Eastern countries, are not well separated when using the raw Leontief inverse matrices. This pattern is consistent across both political and geographic labels. Overall, K-Means captures some broad East versus West divisions, rather than the finer distinctions observed in the earlier supervised classification, reflecting the aggregate nature of the algorithm. Class imbalances heavily impact recall, and although some patterns are identified, the clusters do not fully capture the underlying East/West structural differences as effectively as supervised classification methods.

7 Conclusion and Future Steps

The analysis reveal clear structural differences in economic systems as captured by the Leontief inverse matrices. K-Nearest Neighbors, a local classification method, effectively identifies nuanced regional distinctions and differentiates smaller regions. In contrast, K-Means clustering, which relies on aggregate patterns, struggles to separate

regions with subtle differences and tends to form large dominant clusters under class imbalance. Consequently, K-Means is more suited to capturing general economic structural groupings rather than specific geopolitical or geographic regions, which explains the persistent Western cluster in the results.

7.1 Future Steps

Classification: Expanding the dataset to include comparable time-series data for additional countries within each regional class would allow testing the robustness of the trained models on unseen countries. Current dataset limitations, particularly the small sample size, prevented this, but a larger dataset would enable stronger validation of model generalization.

Clustering: Analysis of the within-cluster sum of squares (WCSS) suggests an optimal $K = 7$, indicating the potential for more detailed structural groupings. These clusters could capture either general patterns of economic structure or finer regional subdivisions beyond the coarse East–West or Europe–Asia–Americas distinctions. Figure 27 presents the WCSS scree plot used to assess cluster compactness and identify the optimal number of clusters.

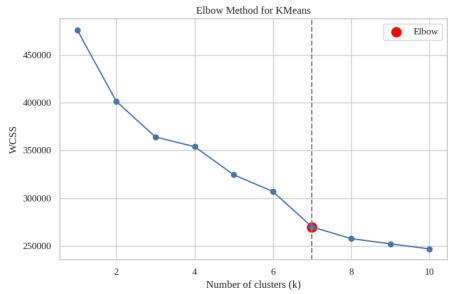

Figure 27: WCSS scree plot used to determine the optimal number of clusters.

8 Citations and Bibliographies

References

[1] GeeksforGeeks. 2025. *Curse of Dimensionality in Machine Learning*. GeeksforGeeks. Last updated 23 July 2025. Available at: https://www.geeksforgeeks.org/machine-learning/curse-of-dimensionality-in-machine-learning/ (accessed Dec. 11, 2025). :contentReference[oaicite:0]index=0

[2] D. T. Coe and S.-J. Kim, editors. 2001. *Korean Crisis and Recovery: Papers Presented at a Conference held in Seoul, Korea, May 17–19, 2001*. Seoul, Korea.

[3] P. Woltjer, R. Gouma, and M. P. Timmer. 2021. *Long-run World Input-Output Database: Version 1.1 Sources and Methods*. GGDC Research Memorandum 190. Groningen Growth and Development Centre, University of Groningen. Available at: https://www.rug.nl/ggdc/html_publications/memorandum/gd190.pdf (accessed Dec. 11, 2025).

[4] Investopedia. 2025. *Economic Cycle*. Available at: https://www.investopedia.com/terms/e/economic-cycle.asp (accessed Dec. 11, 2025).

[5] V. Satopaa, J. Albrecht, D. Irwin, and B. Raghavan. 2011. *Finding a "Kneedle" in a Haystack: Detecting Knee Points in System Behavior*. Simplex11 Proceedings. Available at: https://raghavan.usc.edu/papers/kneedle-simplex11.pdf (accessed Dec. 11, 2025).

[6] "World Input-Output Database, 2021 Release, 1965-2000 Long-run WIOD." 2021. Dataverse NL. Available at: https://dataverse.nl/dataset.xhtml?persistentId=doi:10.34894/A7AXDN (accessed Dec. 11, 2025). :contentReference[oaicite:4]index=4

[7] Google Colaboratory. 2025. *Input–Output analysis notebook (io.ipynb)*. Available at: https://colab.research.google.com/notebooks/io.ipynb (accessed Dec. 11, 2025).

[8] University of Maryland, Department of Mathematics. 2023. *Leontief Input–Output Model*. Available at: https://www.math.umd.edu/~immortal/MATH401/book/ch_leontief.pdf (accessed Dec. 11, 2025).

[9] The Pandas Development Team. 2025. *Pandas Documentation*. Available at: https://pandas.pydata.org/docs/ (accessed Dec. 11, 2025).

[10] M. Waskom et al. 2025. *Seaborn: Statistical Data Visualization — Heatmap*. Available at: https://seaborn.pydata.org/generated/seaborn.heatmap.html (accessed Dec. 11, 2025).

[11] M. Waskom et al. 2025. *Seaborn: Statistical Data Visualization — Lineplot*. Available at: https://seaborn.pydata.org/generated/seaborn.lineplot.html (accessed Dec. 11, 2025).

[12] M. Waskom et al. 2025. *Seaborn Tutorial: Color Palettes*. Available at: https://seaborn.pydata.org/tutorial/color_palettes.html (accessed Dec. 11, 2025).

[13] M. Waskom et al. 2025. *Seaborn Objects: Plot.label Method*. Available at: https://seaborn.pydata.org/generated/seaborn.objects.Plot.label.html (accessed Dec. 11, 2025).

[14] Matplotlib Development Team. 2025. *Matplotlib Font Manager API Documentation*. Available at: https://matplotlib.org/stable/api/font_manager_api.html (accessed Dec. 11, 2025).

[15] Scikit-learn Developers. 2025. *Plot Feature Importances Using Forest of Trees*. Available at: https://scikit-learn.org/stable/auto_examples/ensemble/plot_forest_importances.html (accessed Dec. 11, 2025).

[16] Scikit-learn Developers. 2025. *sklearn.cluster.KMeans — scikit-learn clustering API*. scikit-learn.org. Available at: https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html (accessed Dec. 11, 2025). :contentReference[oaicite:0]index=0

[17] Scikit-learn Developers. 2025. *sklearn.metrics — scikit-learn performance metrics API*. scikit-learn.org. Available at: https://scikit-learn.org/stable/api/sklearn.metrics.html (accessed Dec. 11, 2025). :contentReference[oaicite:1]index=1

[18] Scikit-learn Developers. 2025. *sklearn.decomposition.PCA — scikit-learn decomposition API*. scikit-learn.org. Available at: https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html (accessed Dec. 11, 2025). :contentReference[oaicite:2]index=2