In Search of Empirical Economic Groups: Performing Machine Learning Clustering on the Panel Study of Income Dynamics (PSID)

Domenic Mastromatteo

University of British Columbia

Author Note

This paper was prepared for ECON 490, taught by Dr. Gallipoli, at the University of British Columbia.

Date of submission: April 2025.

In Search of Empirical Economic Groups: Performing Machine Learning Clustering on the Panel Study of Income Dynamics (PSID)

Abstract

This paper explores the application of unsupervised machine learning techniques to economic data, specifically clustering methods on a subset of the Panel Study of Income Dynamics (PSID) dataset. By leveraging clustering algorithms such as k-means, DBSCAN, and spectral clustering, I investigate whether meaningful groupings emerge from the selected economic variables which include a variety of types of income, wealth, and consumption. I assess clustering validity using visualization techniques like t-SNE and PCA, along with manual and indirect evaluation methods. My findings highlight the benefits and challenges of clustering in high-dimensional economic data and suggest further improvements in feature selection, hyperparameter tuning, and validation techniques for future research.

Introduction

Economic theory often relies on categorizing individuals into groups, such as income classes, high and low types, or demographic segments, to better understand financial behaviors and policy impacts. Such groups can then be used for goals such assisting profit maximization through price discrimination (in the case of firms), or to examine the effects of inequality (in the case of policy makers). This paper investigates whether unsupervised machine learning (ML) techniques, specifically clustering algorithms, can uncover natural groupings in high-dimensional economic data without predefined labels.

The research question driving this study is: *Can unsupervised clustering methods reveal meaningful patterns in economic data from the Panel Study of Income Dynamics (PSID)?* This question is significant to the field because the existence of empirically validated clusters could assist in improved theoretical analysis, policy-making, and corporate decision-making by aligning theory with reality.

To answer this question, I applied a range of popular clustering techniques:

k-means, k-medians (PAM), agglomerative, DBSCAN, and spectral clustering to a dataset of 146 economic features derived from the 2019 PSID. I used dimensionality reduction techniques, including t-distributed stochastic neighbor embedding (t-SNE) and principal component analysis (PCA), both for analysis and to visualize the results. Clustering performance was assessed through internal validation metrics, visual inspection, and manual evaluation of cluster interpretability.

My findings indicate k-means did not perform well on this task, with low Silhouette scores (a validation metric commonly used for k-means) on all cluster quantities tested. I hypothesize this is likely due to the lack of spherical clusters in the data, perhaps due to the high degree of inter-feature correlation. DBSCAN, when tuned correctly, likely performed better in capturing the irregular cluster structures, though validation remained challenging due to a lack of appropriate internal validation measures and poor visual results when plotted with t-SNE. Spectral clustering provided the most visually congruent results with t-SNE, providing the most easy to interpret results and highlighting its potential for economic analysis. Manual validation was performed on the spectral clustering results, producing a range of economically interpretable results. The lack of ground truth labels across the study complicated validation, making manual and indirect evaluation methods necessary and hindering my ability to effectively analyze all clusterings created.

These findings highlight both the potential and limitations of unsupervised clustering in economic research. While machine learning methods can reveal structure in high-dimensional data, meaningful interpretation requires careful feature selection, hyperparameter tuning, and validation techniques beyond standard clustering metrics. Future research should explore alternative distance metrics, refined feature engineering, and more robust validation approaches to enhance the reliability of clustering in economic studies.

Literature Review

Unsupervised machine learning, particularly clustering, has slowly gained traction in economic research as a tool for uncovering latent patterns in large datasets. This section reviews existing literature on clustering techniques in economics, challenges associated with high-dimensional data, and validation methods for unsupervised learning.

Clustering in Economic Research

Economists have long sought to group individuals to simplify analyses. Traditional approaches rely on predefined classifications, such as income quintiles or demographic segments. However, these methods are inherently one dimensional and may fail to capture the complexity of economic decision-making. Recent studies have suggested machine learning-based clustering as a data-driven alternative (Athey and Imbens, 2019). To that end, clustering techniques such as k-means and DBSCAN have been increasingly used for consumer segmentation (Kansal et al., 2018; Paramita and Hariguna, 2024; Tabianan et al., 2022), however their use —particularly for other purposes —remains small and limited relative to the size of the field.

It is unclear why clustering has not made a big splash in economics yet, despite the impact it has had in other fields, particularly biology (Linderman and Steinerberger, 2017). Einav and Levin, 2014 covered the rise of empiricism and big data within economics and suggested that the greater focus on predictive models within machine learning, and therefore difficulty in establishing causality is one reason why economists may shy away from machine learning techniques.

Challenges of High-Dimensional Clustering

A key difficulty in applying clustering algorithms to economic data is the high dimensionality of features. "The Curse of Dimensionality" is a term coined by Bellman et al., 1957 while discussing challenges in dynamic programming, but has now found widespread use in machine learning (particularly clustering) to address the issues with "filling" high dimensional volumes due to their exponential growth. As the feature-space

becomes larger, more data points are required to reduce the sparsity in the space, the gathering of which may be particularly challenging for economics where high quality survey data is relatively rare. Outside of "The Curse", traditional clustering methods, such as k-means, assume spherical clusters and can struggle in complex feature spaces (Hastie et al., 2017). Dimensionality reduction techniques, such as principal component analysis (PCA) and t-distributed stochastic neighbor embedding (t-SNE), have been developed to visualize and interpret clustering results (van der Maaten and Hinton, 2008). However, while some proponents argue for the usefulness of t-SNE in clustering, the extent to which dimensionality reduction preserves meaningful structure remains a contested question within machine learning (Linderman and Steinerberger, 2017).

Clustering Validation and Interpretability

A persistent challenge in economic clustering research is the validation of results. Unlike supervised learning, where labeled data enables straightforward evaluation, clustering requires alternative assessment methods. Internal validation metrics, such as the Silhouette score (Rousseeuw, 1987), can provide numerical benchmarks, but they often assume specific geometric properties that may not hold in general economic datasets. External validation, which compares clusters to known classifications, may be suitable for some economic analysis such as customer segmentation, but is difficult when no predefined labels exist (Hennig, 2015).

Relevance to This Study

This study builds on the existing literature by applying a wide variety of unsupervised clustering techniques to the Panel Study of Income Dynamics (PSID) dataset, a widely used longitudinal economic dataset. By exploring different clustering methods and validation approaches, this research contributes to the discussion on the practical utility of machine learning to solve problems in economics. Furthermore, it highlights and explores key methodological concerns, such as hyperparameter selection, feature selection, and interpretability, which must be addressed for clustering to become a

reliable tool in economic analysis.

Data

This study utilizes the 2019 Panel Study of Income Dynamics (PSID), a longitudinal survey conducted by the University of Michigan (Panel Study of Income Dynamics, 2025). The PSID is one of the most comprehensive household surveys available, tracking economic, demographic, and financial characteristics of U.S. families over multiple decades. The dataset includes information on income, wealth, consumption, employment, education, and household composition. The full list of selected variables is available in Tables 1 through 4 within the appendix, due to space constraints.

Sample Selection and Preprocessing

The raw dataset contains thousands of observations and variables (variables will be referred to as "features" for the remainder of this paper, keeping with tradition in machine learning research). However, for this study, I focused on a subset of economic indicators that were easiest to perform clustering analysis on. This meant focusing on features which were naturally ordinal and directly economically relevant. I also chose the features which PSID had already performed imputation on, so that each observation had a value for every feature. This involved selecting the features from the back of the PSID codebook which might be considered "summary" features. The exclusion of categorical features, of which the PSID has many that may be of economic interest, should not be taken lightly, and is addressed below in **Data Limitations**. The initial dataset contained 9569 observations and 5,632 features. After feature selection, clustering was performed on an equal amount of observations, with feature dimensionality reduced to 146. The only necessary preprocessing after obtaining the data was to standardize the features to ensure a comparable scale.

Unit of Analysis

The unit of analysis in this study is the household. Each observation represents a unique household surveyed in the 2019 PSID wave, capturing its financial characteristics at a given point in time.

Data Limitations

While the PSID provides rich longitudinal data, certain limitations must be acknowledged. First, self-reported income, wealth, and consumption measures may introduce reporting biases. Second, the dataset does not account for all financial assets, notably not counting defined-contribution (DC) pensions, potentially underrepresenting household wealth (Cooper et al., 2019). Finally, the imputation method performed by the PSID to fill in all missing data can not be taken to be as reliable as gathered data, introducing some noise into the data.

The feature selection I have performed also introduces biases into the results. The PSID itself has already performed a sort of "feature selection" on the households, by making judgment calls about which questions are worth including and that households are capable of reliably answering. Further selection has been done to only provide data publicly which does not identify the households. My own feature selection also introduces issues, mainly that some useful information has certainly been excluded. For example, it would be easy to argue that geographic (the state of residence, for example) data should have been included. However to include this data would involve expanding the feature-space by roughly 50, as to restore the required ordinality, each area would need its own binary feature. Therefore the inclusion of this data must be weighed against the consequences of increased dimensionality.

Methodology

This section outlines the research methods and computational techniques used to cluster households based on economic characteristics. I describe the tools used, the clustering algorithms applied, and the validation techniques used to assess model performance.

Clustering Methods

To identify natural groupings within the PSID dataset, I applied five unsupervised clustering techniques:

- k-means Clustering: A centroid-based algorithm that partitions data into k clusters by minimizing the within cluster sum of squares (Hartigan and Wong, 1979). The optimal number of clusters (k) was explored using both the "Elbow Method" and the Silhouette score.

- DBSCAN (Density-Based Spatial Clustering of Applications with Noise): A density-based approach that identifies clusters of varying shapes by grouping points with high local density (Ester et al., 1996). The key hyperparameters (ϵ and minPts) were tuned based on distance threshold analysis.

- k-medians/PAM (Partitioning Around Medoids): A robust alternative to k-means that selects actual data points (medoids) as cluster centers, minimizing dissimilarity within clusters (Kaufman, 2005). The number of clusters was chosen using silhouette analysis.

- Agglomerative Clustering: A hierarchical method that builds nested clusters by iteratively merging the closest pair of clusters based on a linkage criterion until k clusters exist (Kaufman, 2005).

- Spectral Clustering: A graph-based algorithm that leverages eigenvalues of a similarity matrix to perform dimensionality reduction before applying clustering (Ng et al., 2001). Similarly to k-means, the data is partitioned into k clusters, however this method is particularly useful for detecting non-linearly separable clusters.

Dimensionality Reduction

High-dimensional data presents challenges for clustering due to the aforementioned "Curse of Dimensionality". In order to be robust against these effects, Principal Component Analysis (PCA) analysis was used to perform dimensionality reduction on the dataset. It was found that the 146 features could be reduced to 110 principal components when selecting for 95% variance explained. The clustering algorithms were then run on the

PCA and non-PCA dataset concurrently to determine if there were meaningful differences between the results. As the reduction in number of dimensions was small, visual inspection yielded no significant difference, and as PCA complicates further analysis, the PCA results were not included in the final manual cluster validation.

Validation Metrics

Since clustering is an unsupervised technique, evaluating its quality requires alternative assessment methods:

- Internal Metrics: The Silhouette score, was computed to assess cluster compactness and separation. Silhouette scores are a common measure of cluster cohesion, assigning high values to points which are similar to their cluster and dissimilar to other clusters. They range between -1 to +1, with clusterings with scores below .50 generally considered weak. This metric is also biased towards spherical clusters.
- Manual/Indirect Validation: Clusters were examined by checking the features with the highest absolute mean difference between each cluster and the largest cluster. As one of my clusters was overwhelmingly larger than all of the others, I labeled that cluster as the "control" cluster and used it to establish a point of comparison. I then analyzed whether clustering outcomes aligned meaningfully with economic theory. Future work may consider the use of the mean of all data points as a more effective comparison point, as it does not rely on the existence of a large enough "control" cluster and allows for analysis of the largest cluster.

Implementation and Computational Tools

The clustering algorithms and validation techniques were implemented in Python using the scikit-learn and umap-learn libraries. Data preprocessing was conducted using pandas, and visualizations were generated using matplotlib and seaborn.

Results

This section presents the clustering results from the PSID dataset. I analyze cluster characteristics, compare algorithm performance, and discuss economic insights.

Clustering

To determine the optimal number of clusters for k-Means, I used the Elbow Method and the Silhouette score. The silhouette scores are displayed in Table 5 within the appendix. As Silhouette scores below .50 are generally considered weak (Rousseeuw, 1987), I concluded that the data may not be appropriately spherical for this method. Figure 1 shows the within-cluster sum of squares (WCSS) plotted against different values of k. A common method for choosing the optimal number of clusters k is to find the point in the graph where there is a sharp "elbow" and select that as your optimal k. The linear nature of this graph suggests as with the Silhouette score that k-means is a suboptimal method for clustering this data.

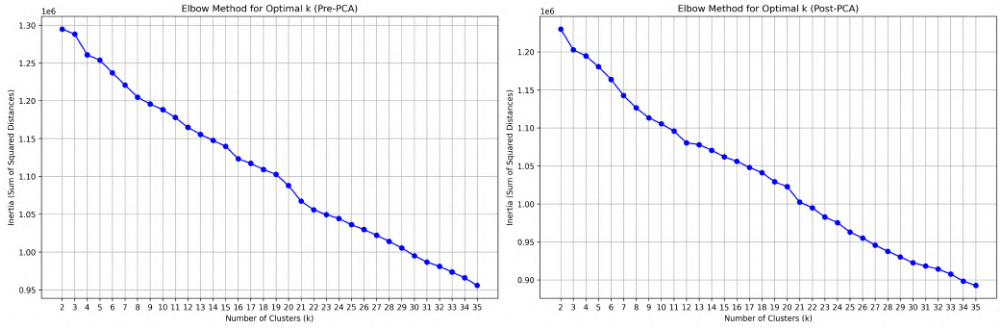

Figure 1. Elbow Method for k-means

For DBSCAN, analysis was first performed on a standard range of ϵ (0.5-5) and minPts (2, 3, 5) values. After this analysis did not provide visually pleasing results, I attempted to follow the heuristics in Sander et al., 1998, leading to a choice of $\epsilon = 10$ and $minPts = 300$. The k-distance graph used is included as Figure 2. As neither of these methods yielded visually validated results, the DBSCAN analysis was abandoned. Future analysis may revisit the choice of hyper-parameters, as there is a great deal of literature on the topic (Schubert et al., 2017).

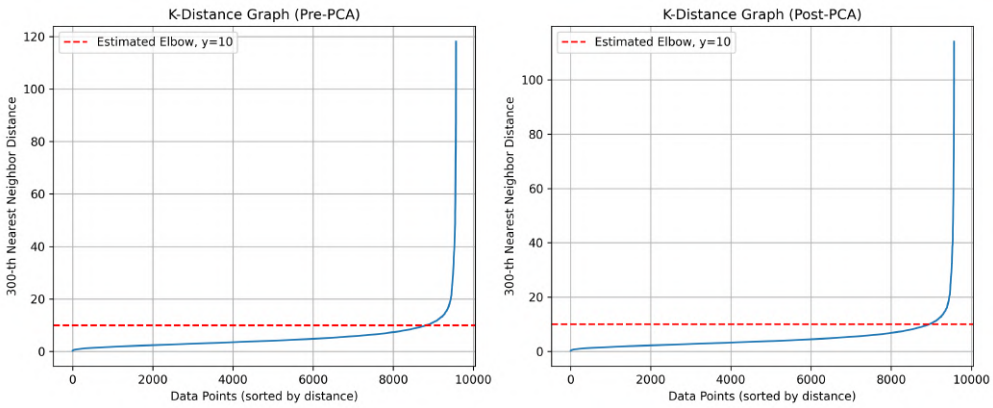

Figure 2. Elbow Method for DBSCAN

Both partitioning around medoids (PAM), otherwise known as k-medians and agglomerative clustering were also tested in this context. As PAM suffers from similar distance and spherical issues as k-Means, it produced similarly unusable results in this context. Agglomerative clustering with ward linkage produced results which were more congruent with the t-SNE visualizations, however this method was also abandoned for final analysis in favor of spectral clustering, due to the preferable visual results of the latter. Future work may revisit this method testing alternative linkage criteria or by combining agglomerative clustering with other methods.

Spectral clustering was performed similarly to k-means, on every k in the range from 2 to 25, with the silhouette score being highest for $k = 2$. As manual selection of k provided much better visual results, I chose to present and analyze the $k = 19$ result, as it was the closest number to my manual read of the t-SNE results. Such a choice is inherently arbitrary, and can only be validated by a much more extensive analysis of results than is time-appropriate for this paper. Spectral clustering was the most visually successful of the results, likely due to its congruence with t-SNE (Ng et al., 2001; Linderman and Steinerberger, 2017).

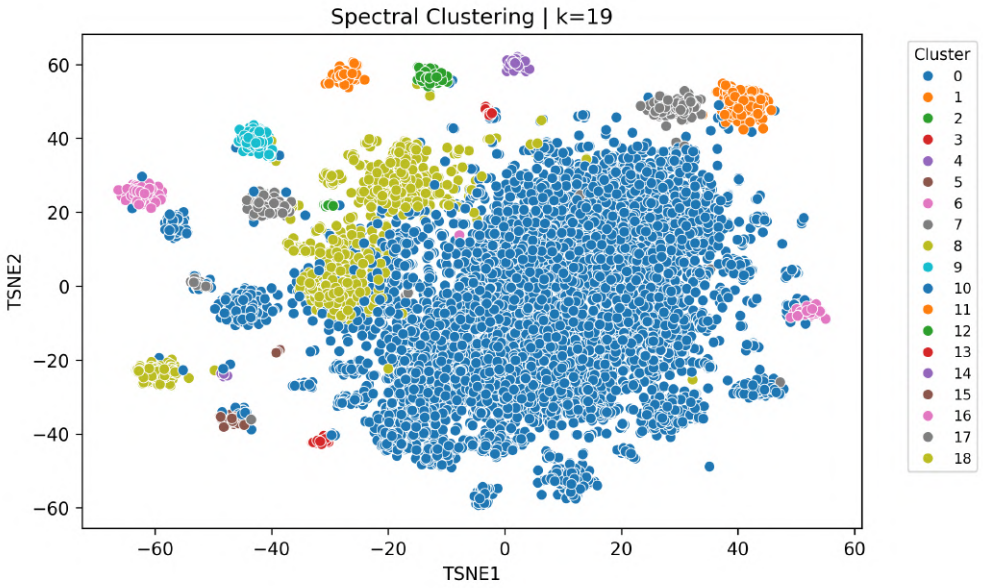

Figure 3. k=19 Spectral Clustering Result

As a final attempt to force the most visually optimal results, DBSCAN was performed on the post-dimensionality reduced t-SNE results. The resulting visualizations provide by far the most cleanly separated clusters. Although some may criticize this method with over-reliance on t-SNE which is not by default a clustering algorithm, recent research suggests that t-SNE performs similarly to spectral clustering and can produce results similar to "ground truth" in a variety of cases (Shaham and Steinerberger, 2017). Future work should compare analysis on this method with the results of spectral clustering itself and determine which method produces the most economically useful clusters.

Due to the immense size of the clustering figures, full results for all clusters are included within the appendix, in Figures 12-16.

Cluster Characteristics

All manual cluster analysis was performed on the $k = 19$ spectral clustering result. The method of analysis was to take the mean of each feature for every cluster and compare it to the mean of the features for the "control" or "mainland" cluster, which was by far the largest cluster in the t-SNE visualization. As mentioned earlier, an alternative method of

analysis might be to compare each cluster to the mean the features for all data points, which would be easier to perform on clusterings which did not have a clear "control" cluster, and would allow for analysis of the largest cluster as well. Future work should weigh the merits of these two methods and choose accordingly.

In my analysis, the feature means were standardized to allow comparability between features, and then the absolute mean difference in features was taken for each cluster compared to the "control". The absolute mean differences were ranked and the top 5 largest absolute mean difference characteristics were graphed for the purpose of selecting the critical features which defined each cluster. Despite, or perhaps due to its simplicity, this method provided a surprising amount of clarity to the clusters, with most clusters being easily identifiable as clear economic groups with only these 5 characteristics. For example, cluster 18 (Figures 4 & 5) might easily be defined as retirees.

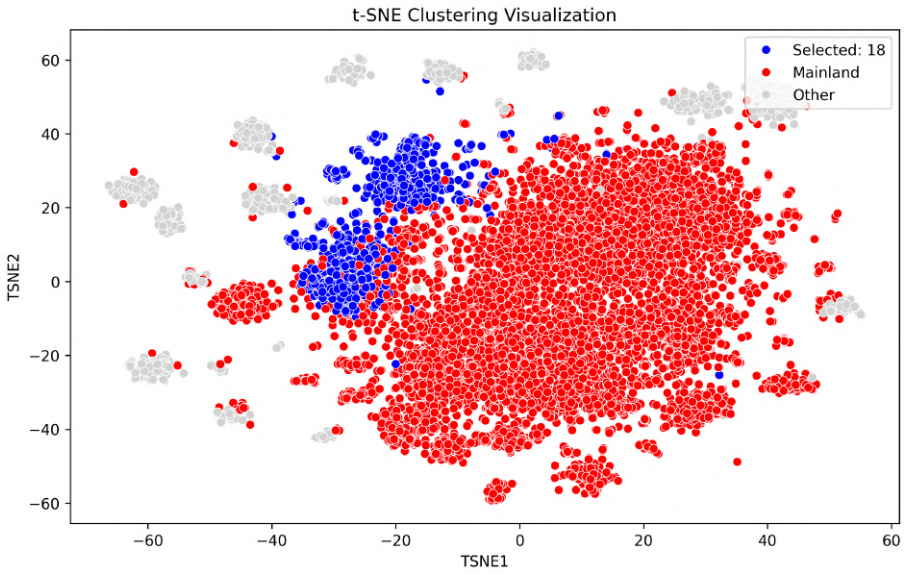

Figure 4. Cluster 18 Visualization

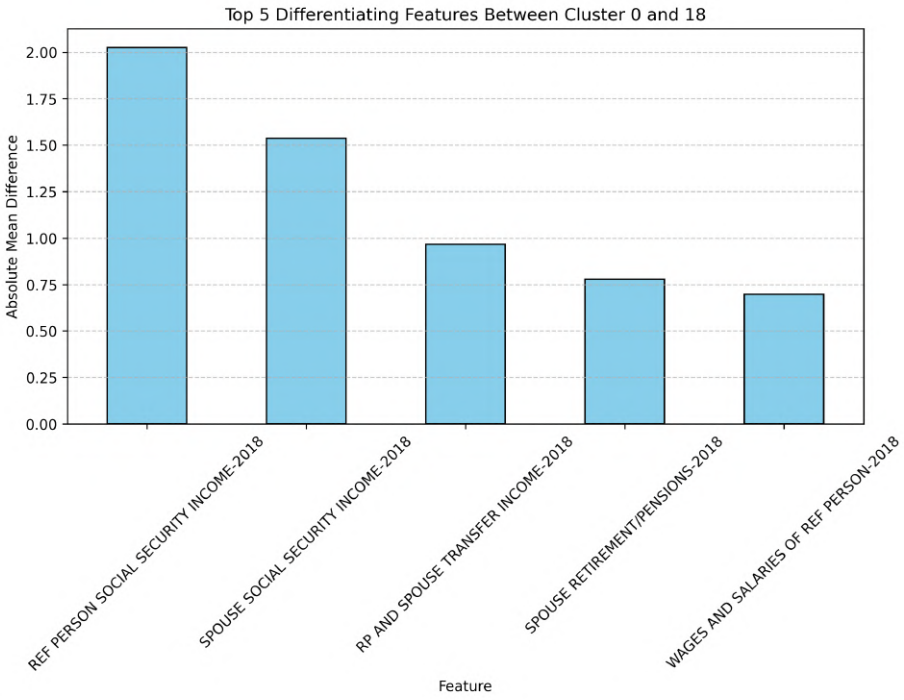

Figure 5. Cluster 18 Characteristics

My analysis stopped short of verifying the cluster identities defined by back-checking the results with the corresponding PSID data. This was necessary due to the time constraint of the study, however a more expanded study could perform this analysis by checking the overlap between cluster-defined (and human labelled) groups and those who self report to have certain characteristics, for example: those who answer in the PSID that they are retired. Such validation methods would greatly increase the robustness of the study.

Key Findings

The clustering analyzed shows a variety of groups of varying economic usefulness. A few of the groups found include:

- Cluster 3 (Figures 6 & 7) shows people with high TANF (Temporary Assistance for Needy Families) income. By examining the other differing characteristics (total consumption, income, wages & salaries) we can see the characteristics that define welfare

recipients and explore areas policy makers can change to move this cluster closer to the others.

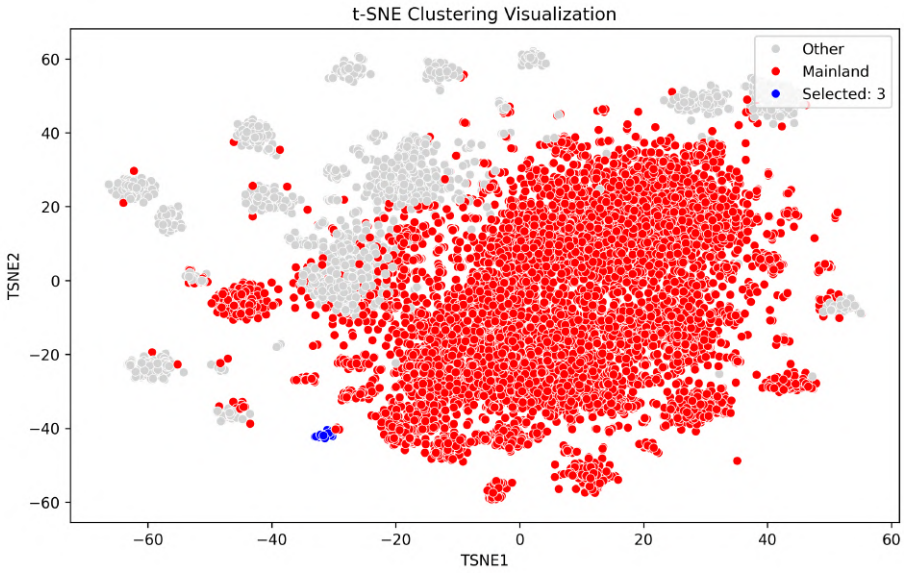

Figure 6. Cluster 3 Visualization

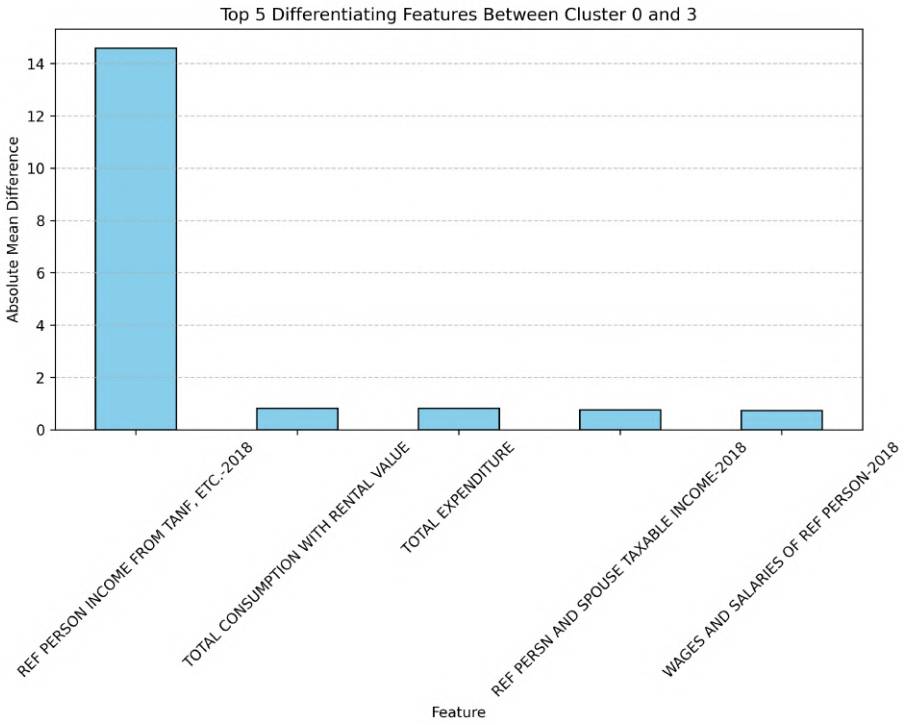

Figure 7. Cluster 3 Characteristics

- Cluster 10 (Figures 8 & 9) is people with high gas and sewer expenditures, perhaps indicating this may be a cluster of rural family units.

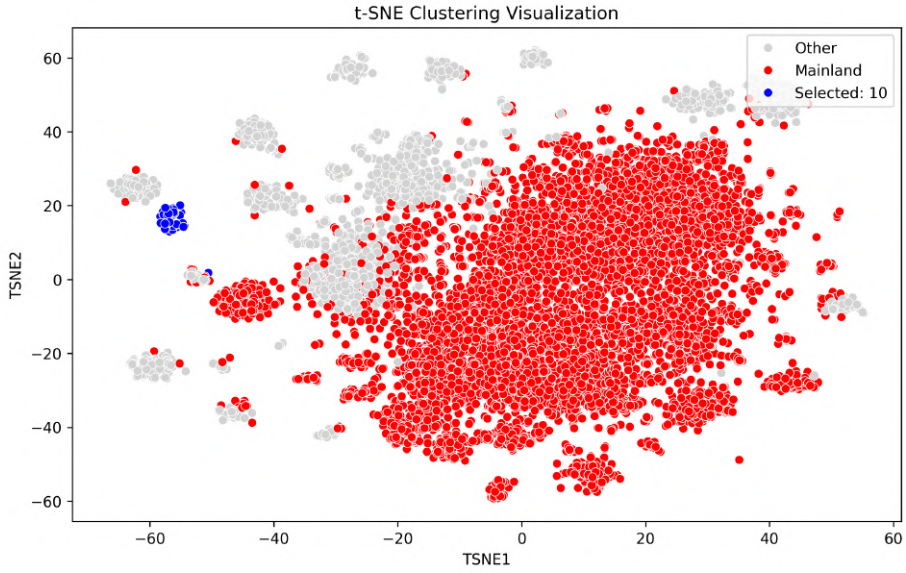

Figure 8. Cluster 10 Visualization

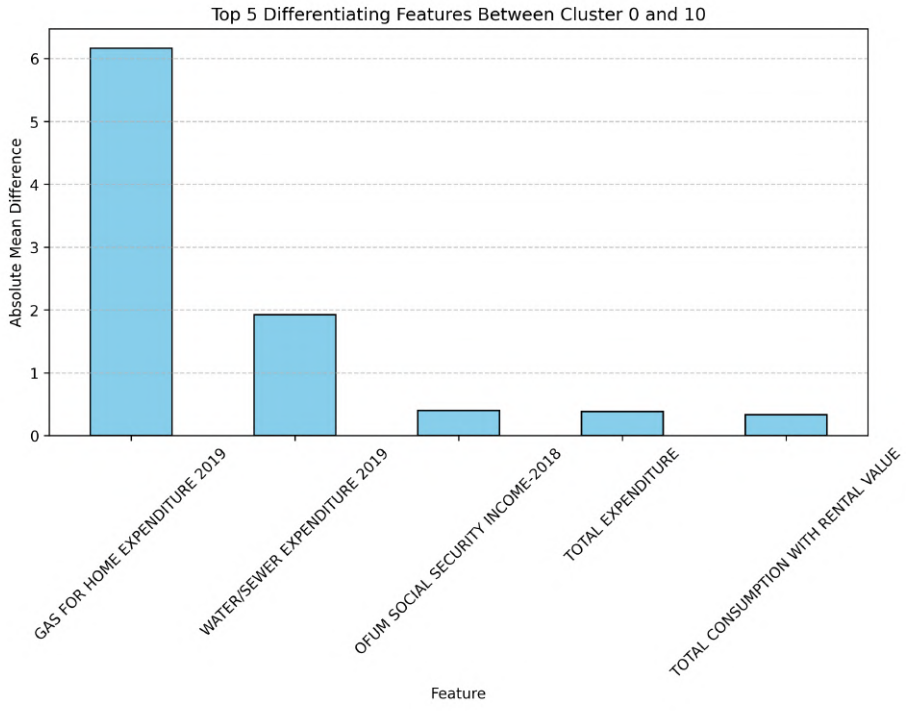

Figure 9. Cluster 10 Characteristics

- Cluster 11 (Figures 10 & 11) shows a cluster of higher educated family units, with defining characteristics in completed education for reference person and spouse, as well as higher wage rates and labour income.

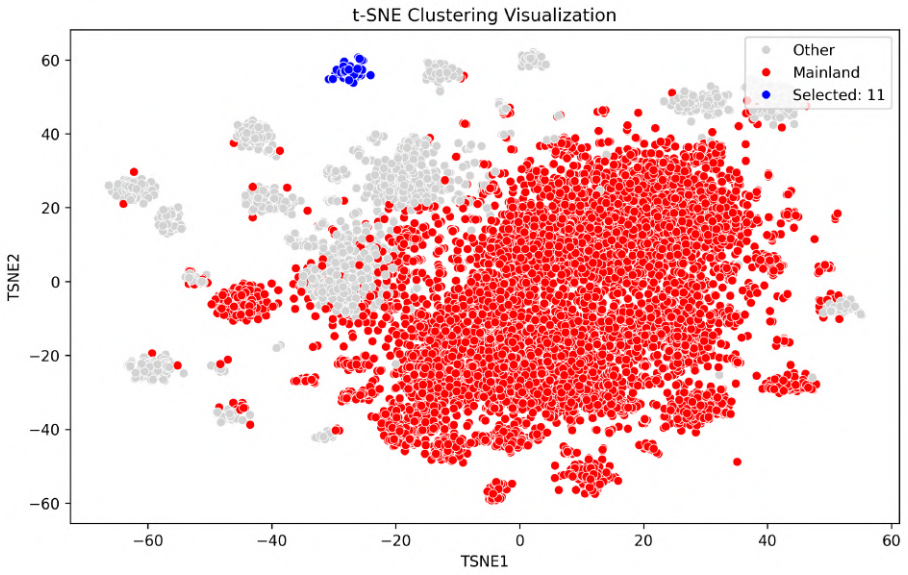

Figure 10. Cluster 11 Visualization

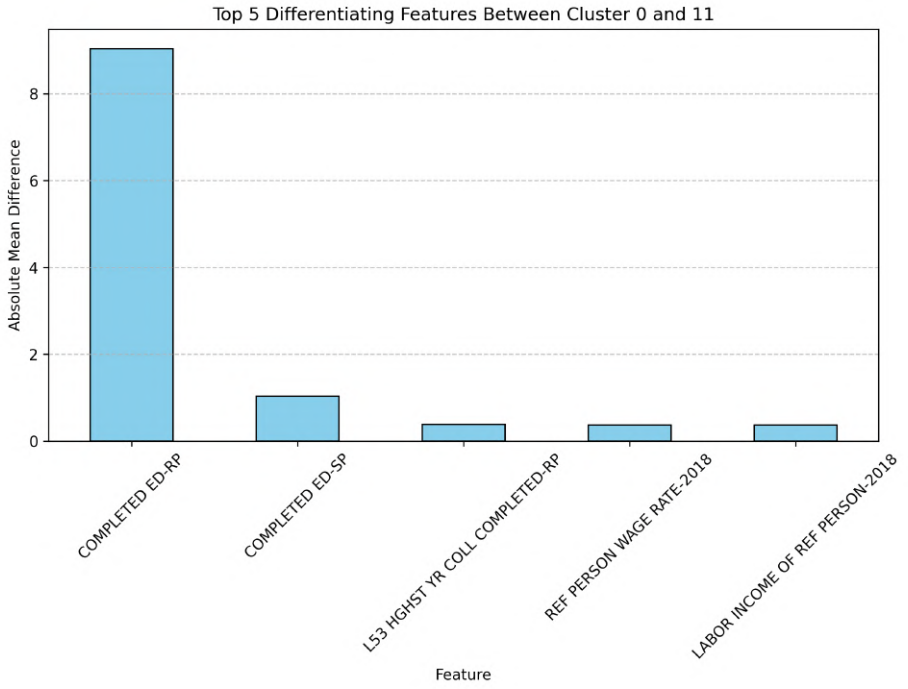

Figure 11. Cluster 11 Characteristics

- Cluster 18 (Figures 4 & 5, p. 13-14) shows people with high reference personal social security income (SSI), high spousal SSI, high pension income, and lower wages and salaries. This cluster is likely retirees.

The number of clusters is too great to fully analyze within this paper. However, a few other results bear consideration when examining the usefulness of this technique. Some other clusters are most largely characterized by features which do not lend themselves to clear interpretation. For example, cluster 2 appears to be largely separated by other utility expenditure, a category most economists would likely not consider to be overly interesting, however further analysis may uncover meaning there. Additionally, clusters 4 & 8 also are characterized with high reference person SSI, and greater analysis would need to be done (perhaps by expanding the number of characteristics examined, or by including other information from the PSID study) in order to determine why these clusters ended up separated from cluster 18 by the algorithm.

Discussion

This section interprets the clustering results and explores their implications for economic research and policy. I also discuss the limitations of this study and potential avenues for future research.

Interpretation of Findings

The clustering results suggest that unsupervised machine learning can effectively identify economically meaningful household groups based on income, wealth, and consumption patterns. These groups have the advantage of being not human defined, but emerging organically from the data.

These findings demonstrate that clustering can capture economic heterogeneity in a way that traditional classification methods may overlook. In practice, this could help policymakers target financial aid, tax benefits, or social programs more effectively by identifying groups that are economically similar rather than relying on broad demographic categories.

Implications for Economic Research

The study highlights the potential of unsupervised learning techniques to enhance economic classification by moving beyond predefined income brackets and enabling data-driven segmentation. This approach can uncover previously unidentified economic subgroups, offering several important implications. Clustering can contribute to more nuanced economic models by grouping households based on actual financial behaviors rather than arbitrary thresholds, thereby improving the precision of both macroeconomic and microeconomic analyses. It enables better policy targeting, allowing governments and organizations to structure social programs around empirically identified economic clusters for more effective resource allocation. This study underscores the growing value of incorporating artificial intelligence and machine learning techniques into traditional economic research, paving the way for more innovative and data-informed approaches to understanding economic behavior.

One particularly interesting possible method for policy makers may be to shield or "mask" certain features from the dataset, perform clustering again, and analyze the resulting clustering. For example, policy makers may hope that those receiving temporary assistance are not meaningfully distinct from the rest of society during the period which they receive this assistance. By masking the Temporary Assistance for Needy Families (TANF) feature and performing clustering again, interested parties would be able to determine if the TANF cluster still exists absent that direct information. This information could then be used to argue for greater or more effective transfer schemes.

Limitations

While the results are promising, several limitations must be acknowledged. One major challenge lies in the high-dimensional and correlated nature of economic data, which can hinder the effectiveness of traditional clustering methods. Although techniques like PCA and t-SNE were employed to reduce dimensionality, this process may have resulted in the loss of important information. Additionally, the absence of ground truth labels in clustering introduces subjectivity into validation; while Silhouette scores and visual inspection offer some guidance, there is no definitive measure of clustering quality in this context. Another limitation stems from feature selection, as the features were drawn from PSID data selectively, potentially overlooking key economic dimensions. Future work should consider incorporating broader indicators and alternative feature selection strategies. Finally, clustering outcomes are sensitive to hyperparameter choices—such as the number of clusters in k-means or the ϵ parameter in DBSCAN—and although parameters were tuned systematically, the results remain contingent on these selections, particularly my choice to perform manual analysis on a single spectral clustering result.

Future Extensions

Future research can build on these findings in several meaningful ways. Exploring alternative clustering methods, such as Gaussian Mixture Models (GMM), Expectation Maximization (EM), or self-organizing maps (SOMs), could yield deeper insights into

economic segmentation. Additionally, incorporating the longitudinal nature of the PSID dataset through time-series clustering would allow researchers to analyze how households transition between clusters over time, providing a dynamic and unique view of economic mobility. Validating clustering outcomes with external benchmarks—such as credit scores, tax brackets, or other alternative indicators—could further assess the robustness and real-world applicability of the methodology. Since many of the clustering algorithms tested rely on Euclidean distance, future studies should also consider alternative distance metrics like Manhattan distance, Mahalanobis distance, or cosine similarity, which may better capture the complexities of economic relationships. Additionally, improving feature selection by including psychological, geographic, or behavioral variables could lead to a more comprehensive and nuanced understanding of economic groupings.

Conclusion

This study applied unsupervised machine learning techniques—specifically clustering—to the Panel Study of Income Dynamics (PSID) dataset to identify distinct household economic groups. My findings suggest that clustering methods can effectively segment households based on income, wealth, consumption, and education, providing a data-driven alternative to traditional economic classifications.

Key Takeaways

- **Economic heterogeneity can be explored through clustering:** The results revealed clusters likely aligned with groups such as retirees, the highly educated, and emergency assistance recipients. These clusters align with economic theory while emerging naturally from the data rather than being predefined.
- **Machine learning may provide alternative economic classifications:** Traditional classifications may oversimplify reality. Clustering offers a more holistic perspective by grouping households based on similar patterns in the data, which can enhance economic theory and policy targeting, particularly in the age of economic

empiricism.

- **Spectral clustering provided the most visually congruent results:** Other clustering methods failed to produce results in this study that were visually aligned with the t-SNE plots. Although there are some criticisms of t-SNE as a clustering approach, recent research suggests that it performs well in some cases, and manual validation suggested that the groups produced here may have utility in economic research.

Implications for Policy and Research

The ability to classify economic groups without relying on arbitrary income cutoffs holds significant potential for both research and policy applications. For policymakers, this approach enables more effective targeting of government assistance programs by focusing on data-driven economic groupings rather than rigid demographic classifications. In the financial sector, banks and lenders could utilize clustering methods to assess household financial stability in a more nuanced manner than traditional credit scores allow. Additionally, researchers can integrate these clustering techniques into predictive models to better understand how economic shocks affect different segments of the population, enhancing the accuracy and relevance of economic forecasting. The longitudinal nature of the PSID may allow for researchers to track the movement and size of these groups over time, allowing for a more nuanced and holistic understanding of society's progression.

Limitations and Future Research

While the study provides a few valuable insights, the nascent nature of the field presents several opportunities for further improvement. Future research could benefit from expanding feature selection to include psychological, behavioral, or geographic factors, which may help refine clustering results. Additionally, exploring alternative clustering techniques—such as Gaussian Mixture Models (GMM), employing different distance metrics like L1 distance, or selecting clustering parameters using a range of internal

validity metrics—could uncover new and insightful patterns. Given that the PSID is a panel dataset, incorporating longitudinal analysis would allow researchers to track how households transition between clusters over time, offering a more dynamic perspective on financial outcomes. Finally, validating cluster classifications against external economic benchmarks or classifications withheld from the clustering, such as credit scores, or government-defined poverty thresholds would enhance the robustness and real-world relevance of the findings.

Final Thoughts

This study demonstrates the potential of unsupervised learning in economic research and that clustering techniques can uncover meaningful economic patterns in household survey data. While clustering provides a powerful tool for grouping family units in the Panel Survey of Income Dynamics, there are challenges in validation and interpretation. The results suggest that machine learning can complement traditional economic analysis by offering data-driven classification methods. With further refinement, these approaches could become a valuable tool for economists, policymakers, and researchers aiming to better understand financial behaviors at scale and as machine learning techniques continue to advance, their integration into economic modeling has the potential to revolutionize how we analyze and classify financial stability, inequality, and policy effectiveness.

References

Athey, S., & Imbens, G. W. (2019). Machine learning methods that economists should know about. *Annual Review of Economics*, *11*(1), 685–725. https://doi.org/10.1146/annurev-economics-080217-053433

Bellman, R., Bellman, R., & Corporation, R. (1957). *Dynamic programming*. Princeton University Press. https://books.google.ca/books?id=rZW4ugAACAAJ

Cooper, D., Dynan, K., & Rhodenhiser, H. (2019, December). *Measuring household wealth in the panel study of income dynamics: The role of retirement assets* (L. Bean, Ed.; Federal Reserve Bank of Boston Research Department Working Papers). Federal Reserve Bank of Boston. https://doi.org/10.29412/res.wp.2019.06

Einav, L., & Levin, J. (2014). Economics in the age of big data. *Science*, *346*(6210), 1243089. https://doi.org/10.1126/science.1243089

Hartigan, J. A., & Wong, M. A. (1979). Algorithm as 136: A k-means clustering algorithm. *Applied Statistics*, *28*(1), 100. https://doi.org/10.2307/2346830

Hastie, T., Tibshirani, R., & Friedman, J. H. (2017). *The elements of statistical learning: Data mining, inference, and prediction* (Second edition). Springer.

Hennig, C. (2015, March). Clustering strategy and method selection [arXiv:1503.02059]. https://doi.org/10.48550/arXiv.1503.02059

Kansal, T., Bahuguna, S., Singh, V., & Choudhury, T. (2018). Customer segmentation using k-means clustering. *2018 International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS)*, 135–139. https://doi.org/10.1109/CTEMS.2018.8769171

Kaufman, L. (2005). *Finding groups in data: An introduction to cluster analysis*. Wiley.

Linderman, G. C., & Steinerberger, S. (2017, June). Clustering with t-SNE, provably [arXiv:1706.02582]. https://doi.org/10.48550/arXiv.1706.02582

Ng, A., Jordan, M., & Weiss, Y. (2001). On spectral clustering: Analysis and an algorithm. In T. Dietterich, S. Becker, & Z. Ghahramani (Eds.), *Advances in Neural*

Information Processing Systems (Vol. 14). MIT Press. https://proceedings.neurips.cc/paper_files/paper/2001/file/801272ee79cfde7fa5960571fee36b9b-Paper.pdf

Panel Study of Income Dynamics. (2025). Panel Study of Income Dynamics, public use dataset. https://psidonline.isr.umich.edu/

Paramita, A. S., & Hariguna, T. (2024). Comparison of k-means and dbscan algorithms for customer segmentation in e-commerce. *Journal of Digital Market and Digital Currency*, *1*(1), 43–62. https://doi.org/10.47738/jdmdc.v1i1.3

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. https://doi.org/10.1016/0377-0427(87)90125-7

Sander, J., Ester, M., Kriegel, H.-P., & Xu, X. (1998). Density-Based Clustering in Spatial Databases: The Algorithm GDBSCAN and Its Applications. *Data Mining and Knowledge Discovery*, *2*(2), 169–194. https://doi.org/10.1023/A:1009745219419

Schubert, E., Sander, J., Ester, M., Kriegel, H. P., & Xu, X. (2017). Dbscan revisited, revisited: Why and how you should (Still) use dbscan. *ACM Transactions on Database Systems*, *42*(3), 1–21. https://doi.org/10.1145/3068335

Shaham, U., & Steinerberger, S. (2017, February). Stochastic Neighbor Embedding separates well-separated clusters [arXiv:1702.02670]. https://doi.org/10.48550/arXiv.1702.02670

Tabianan, K., Velu, S., & Ravi, V. (2022). K-means clustering approach for intelligent customer segmentation using customer purchase behavior data. *Sustainability*, *14*(12), 7243. https://doi.org/10.3390/su14127243

van der Maaten, L., & Hinton, G. (2008). Viualizing data using t-SNE. *Journal of Machine Learning Research*, *9*, 2579–2605.

Appendix

Table 1

Education Variables

Code	Description
ER76922	L53 HGHST YR COLL COMPLETED-RP
ER77599	COMPLETED ED-RP
ER77600	COMPLETED ED-SP

Table 2

Income Variables

Code	Description
ER77291	TOTAL BUSINESS INCOME-2018
ER77293	NUMBER OF BUSINESSES OWNED BY FU IN 2018
ER77294	FARM INCOME OF REF PERSN AND SPOUSE-2018
ER77296	RP LABOR INCOME FROM BUSINESS-2018
ER77297	RP ASSET INCOME FROM BUSINESS-2018
ER77298	NUMBER OF BUSINESSES OWNED BY REF PERSON
ER77299	WAGES AND SALARIES OF REF PERSON-2018
ER77301	BONUS INCOME OF REF PERSON-2018
ER77303	OVERTIME INCOME OF REF PERSON-2018
ER77305	TIPS OF REF PERSON-2018
ER77307	COMMISSION INCOME OF REF PERSON-2018
ER77309	PROFESSIONAL PRACTICE OF REF PERSON-2018
ER77311	REF PERSON ADDITIONAL JOB INCOME-2018
ER77313	MISC LABOR INCOME OF REF PERSON-2018
ER77315	LABOR INCOME OF REF PERSON-2018

Code	Description
ER77316	REF PERSON RENT INCOME-2018
ER77318	REF PERSON DIVIDENDS-2018
ER77320	REF PERSON INTEREST INCOME-2018
ER77322	RP INCOME FROM TRUSTS/ROYALTIES-2018
ER77324	SP LABOR INCOME FROM BUSINESS-2018
ER77325	SP ASSET INCOME FROM BUSINESS-2018
ER77326	NUMBER OF BUSINESSES OWNED BY SPOUSE
ER77327	WAGES AND SALARIES OF SPOUSE-2018
ER77329	BONUS INCOME OF SPOUSE-2018
ER77331	OVERTIME INCOME OF SPOUSE-2018
ER77333	TIPS OF SPOUSE-2018
ER77335	COMMISSION INCOME OF SPOUSE-2018
ER77337	PROFESSIONAL PRACTICE OF SPOUSE-2018
ER77339	SPOUSE ADDITIONAL JOB INCOME-2018
ER77341	MISC LABOR INCOME OF SPOUSE-2018
ER77343	LABOR INCOME OF SPOUSE-2018
ER77344	SPOUSE RENT INCOME-2018
ER77346	SPOUSE DIVIDENDS-2018
ER77348	SPOUSE INTEREST INCOME-2018
ER77350	SPOUSE INCOME FROM TRUSTS/ROYALTIES-2018
ER77352	REF PERSN AND SPOUSE TAXABLE INCOME-2018
ER77353	REF PERSON INCOME FROM TANF, ETC.-2018
ER77355	REF PERSON SSI-2018
ER77357	REF PERSON OTHER WELFARE-2018
ER77363	REF PERSON ANNUITIES-2018
ER77365	REF PERSON IRAS-2018

Code	Description
ER77367	REF PERSON OTHER RETIREMENT-2018
ER77369	REF PERSN UNEMPLOYMENT COMPENSATION-2018
ER77371	REF PERSON WORKERS COMPENSATION-2018
ER77373	CHILD SUPPORT RECEIVED BY REF PERSN-2018
ER77375	REF PERSON INCOME FROM ALIMONY-2018
ER77377	REF PERSON HELP FROM RELATIVES-2018
ER77379	REF PERSON HELP FROM OTHERS-2018
ER77381	REF PERSON MISCELLANEOUS TRANSFERS-2018
ER77383	SPOUSE INCOME FROM TANF, ETC.-2018
ER77385	SPOUSE SSI-2018
ER77387	SPOUSE OTHER WELFARE-2018
ER77389	SPOUSE VA PENSION-2018
ER77391	SPOUSE RETIREMENT/PENSIONS-2018
ER77393	SPOUSE ANNUITIES-2018
ER77395	SPOUSE IRAS-2018
ER77397	SPOUSE OTHER RETIREMENT-2018
ER77399	SPOUSE UNEMPLOYMENT COMPENSATION-2018
ER77401	SPOUSE WORKERS COMPENSATION-2018
ER77403	CHILD SUPPORT RECEIVED BY SPOUSE-2018
ER77405	SPOUSE ALIMONY-2018
ER77407	SPOUSE HELP FROM RELATIVES-2018
ER77409	SPOUSE HELP FROM OTHERS-2018
ER77411	SPOUSE MISCELLANEOUS TRANSFERS-2018
ER77413	RP AND SPOUSE TRANSFER INCOME-2018
ER77414	REF PERSON WAGE RATE-2018
ER77415	SPOUSE WAGE RATE-2018

Code	Description
ER77416	TOTAL LABOR INCOME OF OTR FU MEMBRS-2018
ER77418	TOTAL ASSET INCOME OF OTR FU MEMBRS-2018
ER77420	TAXABLE INCOME OF OTHER FU MEMBERS-2018
ER77423	OTR FU MEMBERS SSI-2018
ER77425	OTR FU MEMBERS OTHER WELFARE-2018
ER77427	OTHER FU MEMBERS VA PENSION-2018
ER77429	OTHER FU MEMBR RETIREMENT/ANNUITIES-2018
ER77431	OFUM UNEMPLOYMENT COMPENSATION-2018
ER77433	OTR FU MEMBERS WORKERS COMPENSATION-2018
ER77435	OFUM INCOME FROM CHILD SUPPORT-2018
ER77437	OTR FU MEMBERS HELP FROM RELATIVES-2018
ER77439	OFUM MISCELLANEOUS TRANSFERS-2018
ER77441	TOTAL TRANSFER INCOME OF OFUMS-2018
ER77442	REF PERSON SOCIAL SECURITY INCOME-2018
ER77444	SPOUSE SOCIAL SECURITY INCOME-2018
ER77446	OFUM SOCIAL SECURITY INCOME-2018
ER77448	TOTAL FAMILY INCOME-2018

Table 3

Wealth Variables

Code	Description
ER77451	IMP VALUE FARM/BUS ASSET (W11A) 2019
ER77453	IMP VALUE FARM/BUS DEBT (W11B) 2019
ER77457	IMP VAL CHECKING/SAVING (W28A) 2019
ER77461	IMP VAL CD/BONDS/TB (W28) 2019
ER77465	IMP VAL OTH REAL ESTATE ASSET (W2A) 2019

Code	Description
ER77467	IMP VAL OTH REAL ESTATE DEBT (W2B) 2019
ER77471	IMP VALUE STOCKS (W16) 2019
ER77473	IMP VALUE VEHICLES (W6) 2019
ER77477	IMP VALUE OTH ASSETS (W34) 2019
ER77481	IMP VALUE ANNUITY/IRA (W22) 2019
ER77485	IMP VAL CREDIT CARD DEBT (W39A) 2019
ER77489	IMP VAL STUDENT LOAN DEBT (W39B1) 2019
ER77493	IMP VAL MEDICAL DEBT (W39B2) 2019
ER77497	IMP VAL LEGAL DEBT (W39B3) 2019
ER77501	IMP VAL FAMILY LOAN DEBT (W39B4) 2019
ER77505	IMP VAL OTHER DEBT (W38B7) 2019
ER77507	IMP VALUE HOME EQUITY 2019

Table 4

Consumption Variables

Code	Description
ER77514	FOOD AT HOME EXPENDITURE 2019
ER77516	FOOD AWAY FROM HOME EXPENDITURE 2019
ER77518	FOOD DELIVERED EXPENDITURE 2019
ER77520	HOUSING EXPENDITURE 2019
ER77521	MORTGAGE EXPENDITURE 2019
ER77523	VALUE OF HOME IF RENTED 2019
ER77525	RENT EXPENDITURE 2019
ER77527	PROPERTY TAX EXPENDITURE 2019
ER77529	HOME INSURANCE EXPENDITURE 2019
ER77531	UTILITY EXPENDITURE 2019

Code	Description
ER77533	GAS FOR HOME EXPENDITURE 2019
ER77534	ELECTRICITY EXPENDITURE 2019
ER77535	WATER/SEWER EXPENDITURE 2019
ER77536	OTHER UTILITY EXPENDITURE 2019
ER77537	TELEPHONE/INTERNET EXPENDITURE 2019
ER77539	TRANSPORTATION EXPENDITURE 2019
ER77540	VEHICLE LOAN PAYMENT EXPENDITURE 2019
ER77542	VEHICLE DOWN PAYMENT EXPENDITURE 2019
ER77544	VEHICLE LEASE PAYMENT EXPENDITURE 2019
ER77546	AUTO INSURANCE EXPENDITURE 2019
ER77548	ADDITIONAL VEHICLE EXPENDITURE 2019
ER77550	VEHICLE REPAIR EXPENDITURE 2019
ER77552	GASOLINE EXPENDITURE 2019
ER77554	PARKING EXPENDITURE 2019
ER77556	BUS/TRAIN EXPENDITURE 2019
ER77558	TAXICAB EXPENDITURE 2019
ER77560	OTHER TRANSPORTATION EXPENDITURE 2019
ER77562	EDUCATION EXPENDITURE 2018
ER77564	CHILDCARE EXPENDITURE 2018
ER77566	HEALTH CARE EXPENDITURE 2019
ER77567	HOSPITAL/NURSING HOME EXPENDITURE 2018
ER77569	DOCTOR EXPENDITURE 2018
ER77571	PRESCRIPTIONS/OTHER EXPENDITURE 2018
ER77573	HEALTH INSURANCE EXPENDITURE 2019
ER77575	COMPUTING EXPENDITURE 2018
ER77577	HOUSEHOLD REPAIRS EXPENDITURE 2018

Code	Description
ER77579	HOUSEHOLD FURNISHING EXPENDITURE 2018
ER77581	CLOTHING EXPENDITURE 2018
ER77583	TRIPS EXPENDITURE 2018
ER77585	OTHER RECREATION EXPENDITURE 2018
ER77587	TOTAL EXPENDITURE
ER77588	TOTAL CONSUMPTION WITH RENTAL VALUE

Table 5

K-Means Silhouette Scores

Code	Description
2	0.2761
3	0.2764
4	0.1737
5	0.1690
6	0.0856
7	0.0666
8	0.0726
9	0.0637
10	0.0639
11	0.0606
12	0.0608
13	0.0610
14	0.0620
15	0.0635
16	0.0653
17	0.0653

Code	Description
18	0.0676
19	0.0618
20	0.0678
21	0.0431
22	0.0444
23	0.0195
24	0.0195
25	0.0176
26	0.0179
27	0.0186
28	-0.0067
29	-0.0054
30	-0.0118
31	-0.0086
32	-0.0131
33	-0.0062
34	-0.0169
35	-0.0154

Figure 12. All k-means clustering results, visualized

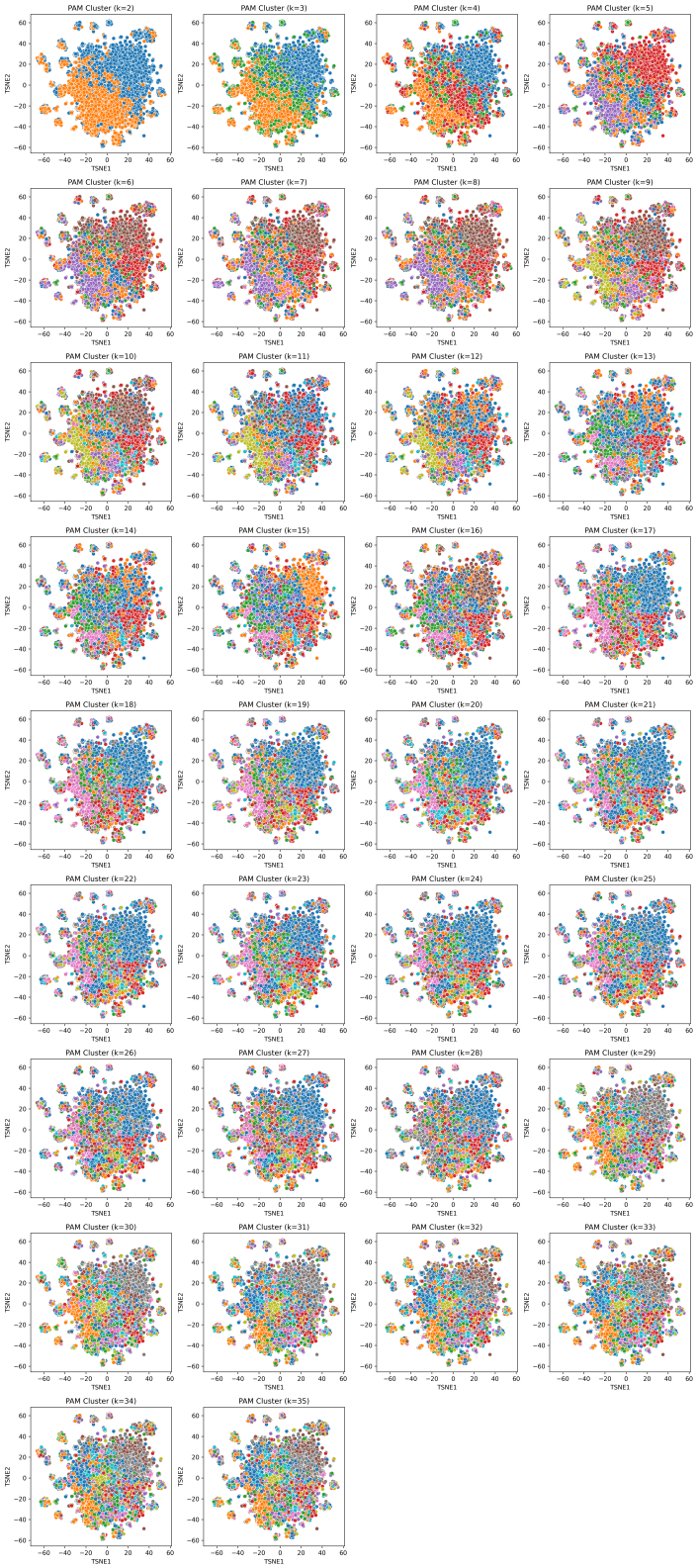

Figure 13. All PAM/k-medians clustering results, visualized

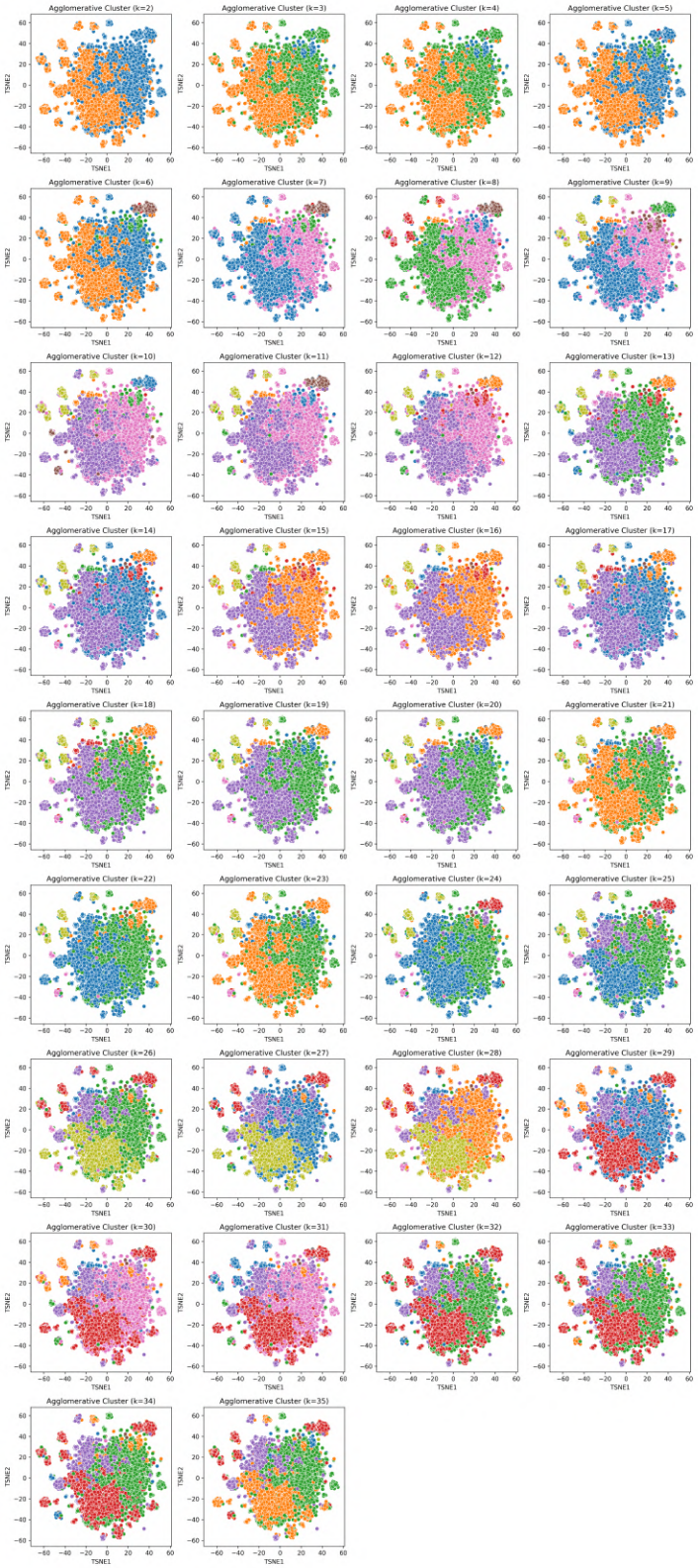

Figure 14. All agglomerative clustering results, visualized

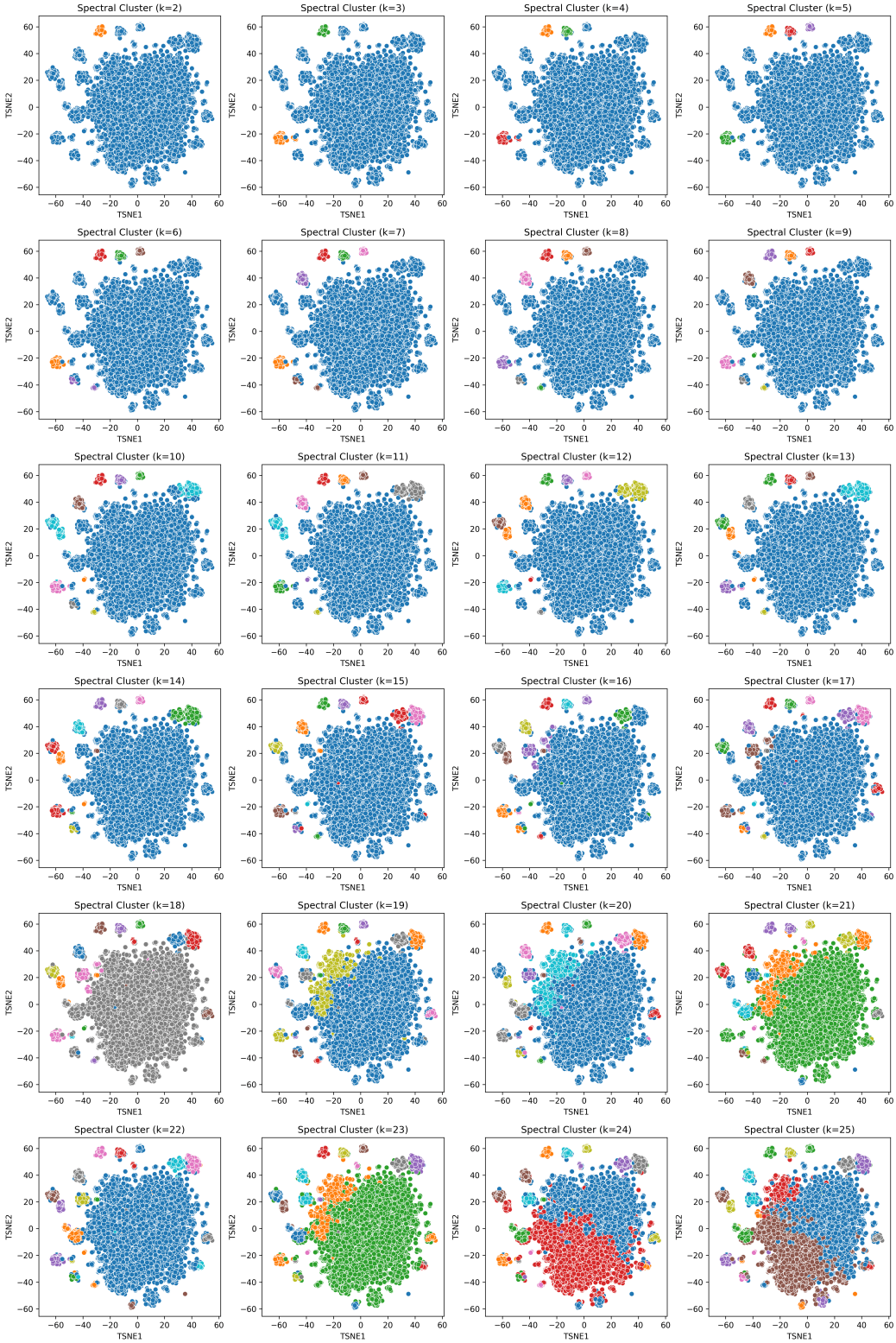

Figure 15. All spectral clustering results, visualized

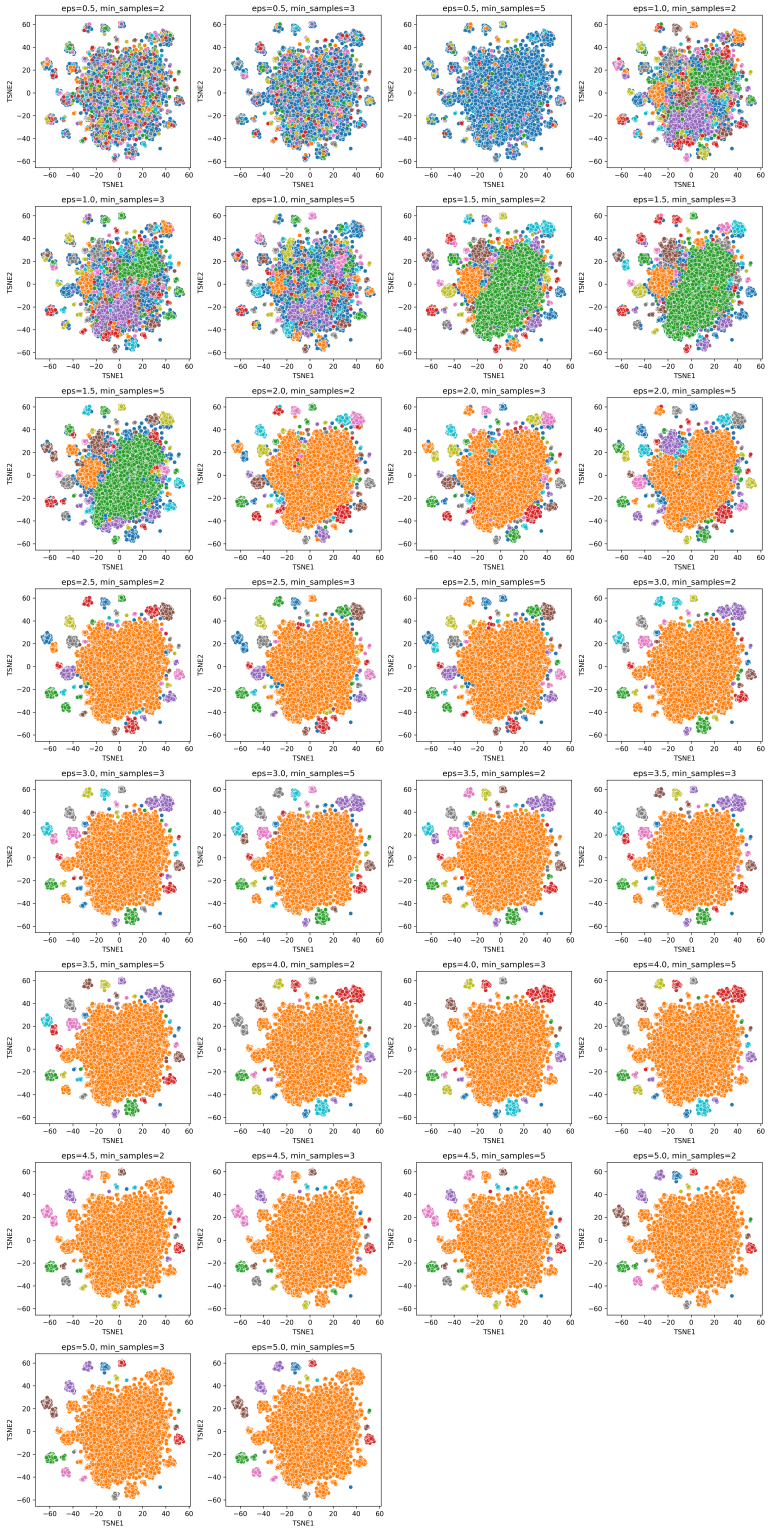

Figure 16. All DBSCAN clustering results, performed post-t-SNE, visualized