Fuzzy Clustering with Hierarchical Agglomerative Clustering for Applications in Sustainability

Arsh Siddiqui
Department of Computer Science
Virginia Tech
Blacksburg, VA 24060
arshsiddiqui@vt.edu

Prayash Joshi
Department of Computer Science
Virginia Tech
Blacksburg, VA 24060
prayash@vt.edu

Charles Bruner
Department of Computer Science
Virginia Tech
Blacksburg, VA 24060
charlesb32@vt.edu

Abstract

To establish better policy regarding sustainability and sustainable based practices in the United States, they must be developed and applied at a more granular level. However, neglecting similarities as well can lead to an over-creation of certain policies that may better be applied to multiple standards (or states). To establish a level of similarity between states on different (sometimes conflicting) metrics, a hierarchical agglomerative clustering approach is proposed that employs a fuzzy clustering based algorithm, involving an entropy based weighting metric. The DP-SIR framework was employed as part of the data collection process, for the goal of attaining an efficient spread of sustainability based and related data. It was found that the classical metrics (centroid, single, and complete link) lacked balanced structures of dendrograms, whereas the fuzzy metrics (fuzzy membership matrix, naive weight, entropy weight measure) had greater structural balance between the generalized groups, which made them easier to analyze and understand to one degree. However, the fuzzy membership matrix and naive weight metrics both had inversion issues, making the dendrogram harder to comprehend and, thus, analyse. The entropy weight measure metric was qualitatively quite balanced and had no inversions, at least for this particular case.

1 Introduction

The United States has set forth various sustainability goals with a target set for the year 2050. These goals, although admirable, have been critiqued for their general nature (including questionable attainability and nationwide generalizations), and many critics argue that they may not be fully achievable by the stipulated deadline. To address this challenge, our project aims to design a model by which we can classify the states of the United States based upon their similarity. In this way, one can more minutely classify and cluster sustainability goals to prevent generality, without forming hyper niche policies with little translational power. This alternative recognizes that each state possesses different energy capacities and capabilities, and thus our model intends to resolve generality issues.

There have been many different machine learning papers and research covering and relating to the issue of sustainability. One directly covered the issue of machine learning for attaining sustainability, particularly covering the classification of certain generations of energy (through sustainable sources, like solar for instance) in attaining sustainable development goals ([1]). This paper used a Regression Kriging model, and found that it helped to achieve a number of environmental objectives. [2] utilized fuzzy clustering and supervised machine learning techniques to assess the sustainability of certain countries. These researchers discovered that methods involving clustering approaches were more robust in assessing sustainability goals than those without. In particular, they employed fuzzy clustering (specifically the fuzzy c-means algorithm), where each cluster is associated with a membership function, allowing them to belong to more than one cluster with a degree of membership. [3] attempted to assess and model the sustainability of food consumption. They employed a five part framework, involving economic input/output analysis, feature scaling (non-dimensional normalization), determination of a sustainability index value, a k-means clustering conditional to the various attributes of food industries, and a logistic regression model based upon the cluster indices retrieved from the previous step. The results of their study itself were quite rich and the model itself had a high accuracy. [4] attempted to forecast the levels of sustainability of locations at a micro-territorial level (as opposed to such things as states and countries). They did this by identifying a set of sustainability indicators which they employed in decision trees, artificial neural networks, and support vector machines to evaluate the sustainability levels of these micro-territories. They found that using statistical and machine learning models for identifying behavioral patterns of influence upon micro-territories has merit, with primary limitations coming from a lack of information and data at such a granular level.

2 Problem Description

The question of determining and understanding sustainability is an inherently difficult question. It is difficult to come up with a model, firstly, for something with very loose definitions that can be somewhat conflicting or tremendously vague. With respect to our specific considerations and applications, the sustainability of the fifty states of the United States is an, at first, difficult question to answer. It implies that a state is or isn't sustainable, which in turn, establishes sustainability as a binary question. This is not a tremendously useful metric. Sustainability might be better considered as a dynamic or sliding scale.

In considering the relationship between different states of the United States, each have different performance metrics which overlap and differ from other states in non-obvious ways. In effect, a state can be both similar and different to another state depending on the particular sustainability-based metric that is used to compare them.

The question of this problem demands a unique approach to determine object similarity across certain metrics and attributes. The objects themselves being so indistinct and distinct at the same time would need to be considered as well.

In considering the levels of similarity and dissimilarity that is present among states, a hierarchical clustering model could be considered. It would allow for similarities and differences to be gleamed at a general and structured level, permitting one to notice differences from a most minute to general level.

Employing fuzzy clustering based metrics to this hierarchical model could then further the juxtaposition between similarity and difference previously outlined. This would additionally allow for the addition of knowledge of how states are similar on certain attributes and differ on others (Virginia and Arizona may have somewhat similar population sizes, but their geographical climates are quite different, leading to a presumed difference in energy consumption across different sources of energy). Additionally, the importance of certain sustainability metrics are themselves quite varied, ranging from such attributes as carbon output to pesticide exposure, thus, a weight matrix should be applied to the aforementioned fuzzy clustering model to permit a knowledge of difference in attribute importance to be present in the model.

3 Literature Review

3.1 The DPSIR Framework

The DPSIR (Driving forces-Pressures-State-Impacts-Responses) Framework was established by the European Environment Agency and the Organization of Economic Cooperation and Development for the purposes of efficient and effective management of social economic systems ([5]). It has been employed in a variety of different sustainability based research studies. One study from [5], for instance, was employed in determining the efficacy of environmental protection measures taken by the Chinese government, and another, from [6], was used to determine how effective certain European member states were able to promote and employ sustainable development. [5] notes that the DPSIR model is primarily used to evaluate and identify areas of improvement in sustainable activity. However, we plan on utilizing it for to a similar nature as [6], particularly with respect to data collection and segmentation. The model, fundamentally, is able to determine and model interactions and relationships between certain management procedures with their environmental and sustainable impact ([7]). Based upon these core tendencies and abilities provided by the DPSIR model, this paper seeks to develop and utilize a classified collection of sustainability evaluation indicators to use as a method of segmenting our data, but then employing it all for our model.

3.2 Fuzzy Clustering

Fuzzy clustering is a particularly useful evaluation metric under circumstances in which one needs to evaluate the state of some particular object without having any particular evaluation criteria ([6]). In this manner, it provides "an uncertainty of belonging", when it comes to class membership, which can range in value in the unit interval $[0, 1]$ for each cluster for each data point (where 0 would be points not in the cluster and 1 would be points in the center of the cluster). In this way, data points can have non-zero "degrees of belonging" to multiple clusters, but the summation of its "degree of belonging" to each cluster should be 1 ([8], [9]). The structures of clusters established by this method can be considered to be a representation of the relationships among, not just the data within the clusters (i.e. the internal cluster structures), but additionally amongst the clusters themselves ([9]). This can be considered, thus, a more generalized model of clustering, which, for our use case, takes into account the considerations and juxtapositions of similarities and differences between our data points in question (being individual states).

3.2.1 Fuzzy Clustering Iteration Model

Fuzzy Clustering Iteration, as a particular development of general Fuzzy Clustering, was put forth by [10], and it allows one to consider the weight of features being implemented and applied to the model in question ([11]). The model additionally, in effect, iteratively determines a fuzzy membership matrix and fuzzy cluster centroid matrix. The model has been used extensively in natural weather applications. For instance, [11] applied a fuzzy clustering iteration model to flood classification problems, and found it to provide a better fit and be more comprehensive, while also having potential in areas which involve classification without evaluation criteria. Another example would be in precipitation, where [12] used it to study the distribution of precipitation over time with relation to human activity to separate and analyze human-caused activity and its impact or relation onto the precipitation itself. While the iteration model outlined here in its particularities would not appear to be directly necessary to our application (particularly with respect to the hierarchical aspects), the idea of iteration with fuzzy clustering, and especially the idea of classification without evaluation criteria, are particularly useful ideas to keep in mind.

3.3 Agglomerative Hierarchical Clustering

Agglomerative Hierarchical Clustering is a method by which, at each algorithmic step, a pair of clusters are combined into one, until we end with a singular cluster. This is a particular form of hierarchical clustering, which contrasts with divisive hierarchical clustering (starting from one initial node that clusters all data points, and slowly becoming more specific). Consider each cluster as a partition of our dataset, where at each step in the algorithm, two of the current partitions are merged together on some condition. This is repeated until we have a single partition, which would represent the full dataset clustered together. A tree can be constructed from this series of merge operations, from the bottom up, which can then be referred to as a dendrogram ([13]).
[14] employed a method that utilized fuzzy clustering and the concepts therein within a hierarchical clustering model. It had allowed the researchers to be more flexible in their hierarchical clustering approach. [15] described their methods behind the hierarchical fuzzy clustering, which followed a technique by which they employed fuzzy membership functions onto each cluster centroid, which are then combined based upon a particular distance metric. Each of the data elements themselves are fuzzily assigned to certain centroids based on the fuzzy partitioning.

4 Methods and Data Sources

4.1 Sustainability Classification and Indication System

To determine and find a selection of data that we can employ for our model, we must determine a collection of classified data that can quantify the sectors of our DPSIR model. In this sense, we would be following the method of [6], but adapting it to make the sustainability model based upon American metrics, as opposed to European metrics. To assuage any concerns over the necessity of indicators for our model, we consider their necessity in the ability to communicate certain needs to the public and officials. Indicators are inherent to being able to share and express ideas among the public and, thus, to support policy development [16].

4.1.1 DPSIR Framework Classification and Indication System

Through the five layers of the DPSIR Framework, we can establish and determine a set of indicators of various ranges (economic, population, etc.) by which we can quantify each layer at a more specific level that would additionally be relevant to the United States. This will allow us to cover each of the layers of the model and thus model interactions and relationships between our development (through broad terms) and the environment. Thus, we put forth the following table used to classify and determine our necessary indicators of economic and environmental conditions.

DPSIR Criteria	**Indicator of Criteria**	**Data Type**
Driving Forces	Population; Population of Income Determination; Population above 25; Households; Housing Units; Unemployment Base; Low Income Quantity; Unemployed Quantity; Less than High School Education Quantity; Demographics Index; Land Area in Square Meters; Water Area in Square Meters; ...	Numeric Data; Percentages
Pressures	Total Nonrenewable Energy; Energy consumption; Consumption per capita; CO2 emissions; Coal Use; Natural Gas Use; Crude Oil Use; Biofuels Use; Wood and Waste; ...	Numeric Data; Percentages
States	Toxic Releases to Air; Superfund Proximity; RMP Facility Proximity; Hazardous Waste Proximity; Wastewater Discharge; Particulate Matter; Ozone; Air Toxics Cancer Risk; Air Toxics Respiratory HI; ...	Numeric Data
Impacts	Total energy price; Total energy expenditure; Energy expenditure per capita; energy expenditures as a percent of GDP; Average Retail Price of Electricity to Residential Sector; Gasoline Prices Dollars; Gasoline Expenditures; Petroleum Prices; Petroleum Expenditures; Natural Gas Prices; Natural Gas Expenditures; ...	Numeric Data (incl. prices)
Resources	Total Renewable Energy; Natural Electric Power Use; Underground Storage Tanks; ...	Numeric Data; Percentages

4.1.2 Sources of Data

Data has been retrieved per state in the USA. They come from a myriad of locations, mostly from American governmental institutions. A large amount of the data comes from the EPA (Environmental Protection Agency), but there is a lot of other data coming from agencies like the Energy Information Administration, the US Census Bureau, and the Centers for Disease Control. The data itself is all from the most recent iteration of data collection, and the data is all numeric, but of various units (in terms of money, population size, British thermal units, percentages, etc.). This data does get normalized, however, for determining feature weighting, under the metrics in which it is relevant, the non-normalized data is employed to prevent the weight methods from being influenced by a normalization.

4.2 Hierarchical Agglomerative Clustering

This clustering method takes a bottom up approach to creating and combining clusters. It starts with all data points as initial clusters of themselves (in this case, this refers to our fifty states each being their own cluster), and at each iteration, two clusters (or partitions) are merged together based upon some similarity or distance metric.

We developed and worked with a number of different methods that would determine this similarity or distance metric, and we endeavored to compare them all. To set a baseline, the centroid link, single link, and complete link metrics were utilized. However, due to the overlapping nature of sustainability among states (where certain states overlap on certain metrics and differ on others), incorporating fuzzy clustering based techniques and knowledge into our own metrics appeared to us to potentially provide aid and a better understanding of the necessary juxtaposition of our data. This provides ample levels of information to compare linkage and similarity/distance upon.

4.2.1 Fuzzy Membership Matrix Based Metric

The first metric we established was done through the use of membership matrices that determine the likelihood of the relation of some cluster to another cluster. We begin with a matrix of size $n * k$ (R^{n*k}), where $n =$ the current number of data points (our fifty states to begin with) and $k =$ the current number of clusters. At each iteration, the membership matrix will contain data on the distances/belonging of each cluster in the matrix to each other cluster. The two clusters with the highest belonging to each other are chosen and merged into one cluster. This process repeats until we have one full cluster containing all of the initial data points.

4.2.2 Naive Weighted Fuzzy Membership Matrix Based Metric

The second metric follows much the same pattern as the previous, but accounts for weightedness to a naive degree. [17] employed a method of weighting distances to greater express the relevancy of certain features of their data. The weighting measure was fairly simplistic in nature, appearing as follows:

$$w_k = \frac{v_k}{\sum_{k=1}^{m} v_k}$$

where

$$v_k = \frac{\sigma_k}{\overline{x}_k}$$

where k is the current attribute, σ_k is the standard deviation of attribute k, and $\overline{x}_k$ is the mean of attribute k.

The weight matrix that this method results in would be applied to the fuzzy membership matrix that was developed in 4.2.1, thus incorporating some aspect of knowledge, albeit naive, into the initial concept.

4.2.3 Entropy Weight Based Fuzzy Membership Matrix Based Metric

The third metric follows much the same pattern as the second metric. The primary difference is in the measure used for determining the weights, which was done through the Entropy Weight Method. This method effectively measures the dispersion of the data itself, measuring, thus, how much information can be derived from the data (this particularly comes from the calculation of the entropy E_i, where a larger value means more information can be derived for that attribute). It has shortcomings primarily in relation to distortion, both from if there are too many 0's in the dataset and the ignoring of rank discrimination, but these are not extremely relevant factors for our use case or with our data ([18]). The calculation for this measure follows a series of steps adapted from [18]:

We begin by standardizing the data itself. n is the number of data samples, m is the number of attributes, the value of the ith attribute of the jth sample is denoted as x_{ij}. In the following equation p_{ij} refers to the standardized data at each data point for each attribute.

$$p_{ij} = \frac{x_{ij}}{\sum_{j=1}^{n} x_{ij}}$$

The entropy for each data point is then calculated below. Note that if $p_{ij} = 0$, we can consider the entropy E_i at that point to be 0 to reduce computational difficulty (this goes back to the potential issue of distortion that arises from using the entropy weight method).

$$E_i = -\frac{\sum_{j=1}^{n} p_{ij} * \ln p_{ij}}{\ln n}$$

Thus, we can then calculate the weight itself as follows:

$$w_i = \frac{1 - E_i}{\sum_{i=1}^{m} (1 - E_i)}$$

Similarly to 4.2.2, the weight matrix that results from the entropy weight method is incorporated into the fuzzy membership matrix that was developed like in 4.2.1, thus incorporating a greater amount of knowledge with particular respect to the relevancy of the attributes of our data back into the initial concept.

4.3 Fuzzy Clustering

Fuzzy clustering has been employed for each of the models that we have designed. There are a number of methods by which one can implement fuzzy clustering and there are a number of ways that we had integrated them into the hierarchical agglomerative clustering models. The concepts that we had employed to establish our fuzzy clustering models were established through our literature review of the subject. It employs ideas from fuzzy clustering iteration ([10]), general fuzzy clustering ([6]), and through former fuzzy clustering based hierarchical agglomerative clustering models ([15]).
Additionally, different weight metrics were employed on top of these models based upon other literature research relating to the implications of our data itself, in particular, we had employed a model without weights, one with a basic weight metric, and one with an entropy weight based metric (these were as outlined in 4.2.1, 4.2.2, and 4.2.3). Here, though, we outline the general idea of fuzzy clustering, and additionally add an explanation as to how the weight matrix is employed within the model (for the two cases, 4.2.2 and 4.2.3, in which we have employed weight matrices).
In effect, much of the fuzzy clustering based knowledge that is employed as part of our hierarchical agglomerative metrics are centered around the idea of the fuzzy membership matrix. A fuzzy membership matrix showcases the likelihood of one cluster being clustered to another cluster, for all currently available clusters. The matrix is, at first, initialized with a set of random numbers, where the total probability per each cluster's belonging into any other cluster would have the following constraint:

$$\sum_{i=1}^{N} x_{ji} = 1$$

Where N is the number of clusters, x_{ji} is the current belonging, and j is the index of the current cluster.
After being initialized, the membership values of the matrix themselves are updated. Here, at each belonging, the distance between the two clusters is calculated (if a weight metric is used, it is employed first upon the data within these clusters based upon the correlating attributes therein), then a density metric is calculated based upon the distances. This density metric is just used to correlate all the distances between one cluster and all other clusters in the data based upon a fuzzy indicator that the user chooses. The membership index is then updated with the inverse of this density.
These two concepts, the initialization of the matrix and the updating of the membership values within, is implemented iteratively in our final model. At each state, because two clusters are being merged into one, we will have one less overall cluster in the next iterations, thus, a new matrix of a new size is necessary, along with an updating of the membership values within.

5 Results

As comparisons by which to understand our findings, we had used three of the classical metrics for hierarchical agglomerative clustering: centroid link, single link, and complete link. The dendrograms produced on our data with these metrics are as follows.

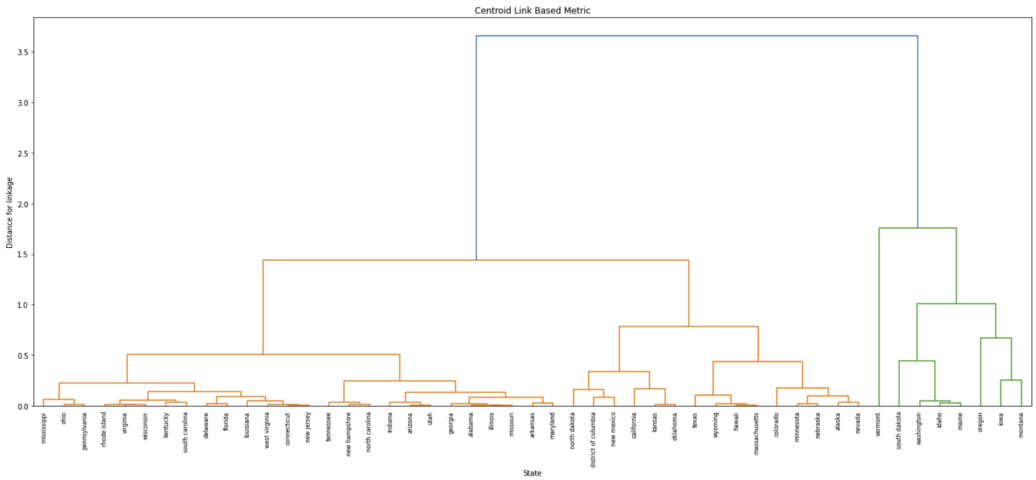

Figure 1: Centroid Link Metric

Note that the coloring in this diagram, and for those that follow, is the basic one imposed by scipy, being $0.7 * max(Z[:\ 2])$, where Z is a linkage matrix, which refers to a matrix of how the clusters/partitions are linked together over the iterations. In this case, two primary groupings are produced in terms of falling below the outlined distance threshold, where most of the states are skewed to the orange grouping, making this structure somewhat imbalanced.

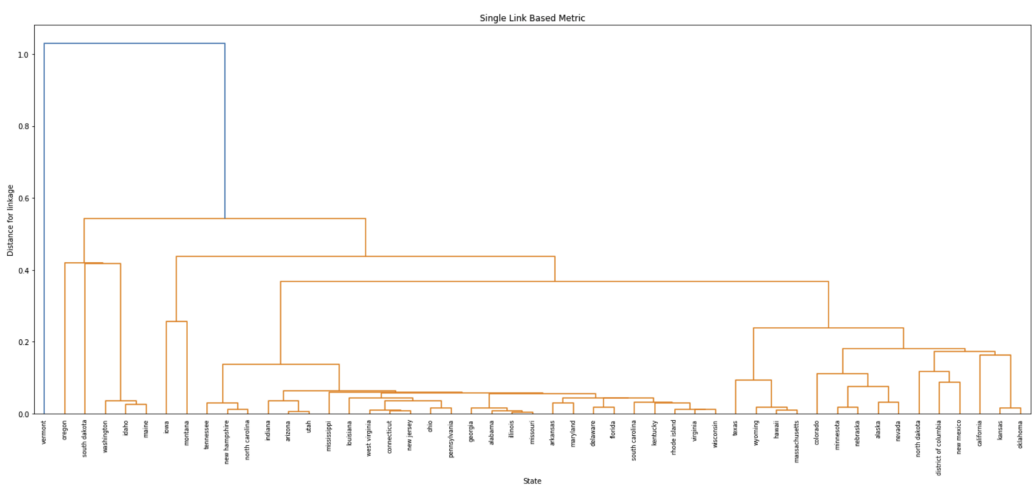

Figure 2: Single Link Metric

In this case, the single link metric is very skewed towards the orange grouping, where there is only one state outside of it, being Vermont. This could arise from the fact that the single link metric ignores the overall structure of the clusters, making it sensitive to outliers (and noise).

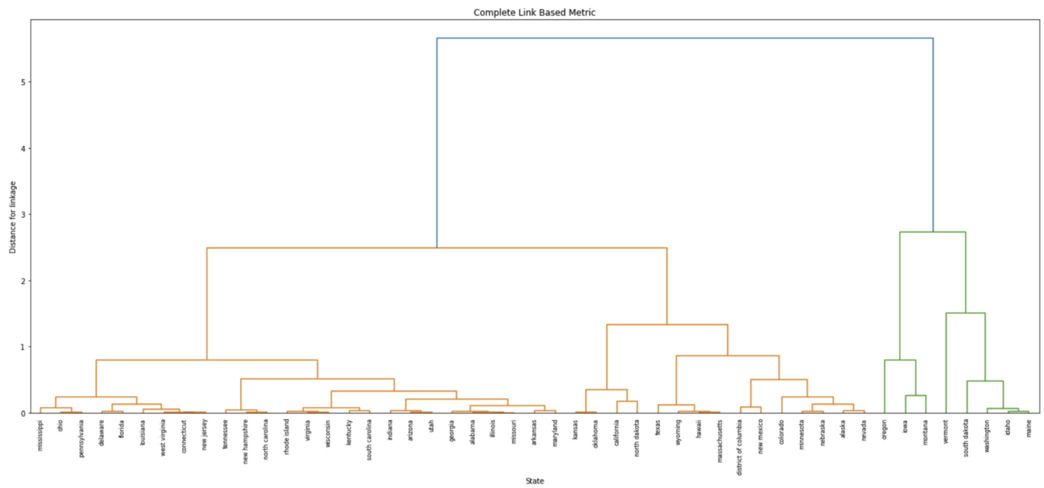

Figure 3: Complete Link Metric

The complete link metric, ignoring the leaves, has a structure that is reminiscent of the centroid link metric. We, additionally, have two primnary groupings that fall under the distance threshold. This metric, similar to the single link metric, is sensitive to noise.

The next three dendrograms that will be displayed are built upon the metrics that was previously outlined in 4.2: fuzzy membership matrix based metric, naive weighted fuzzy membership matrix based metric, and entropy weight based fuzzy membership matrix based metric.

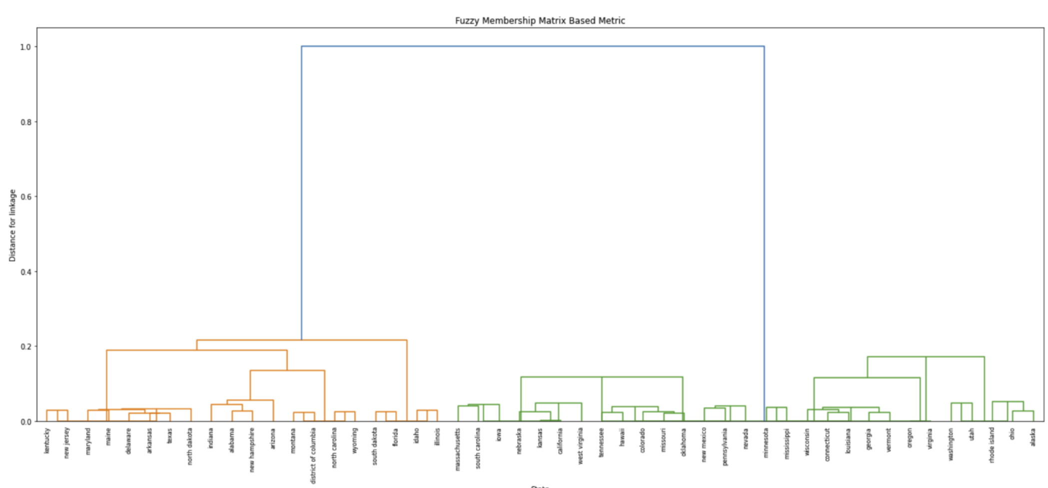

Figure 4: Fuzzy Membership Matrix Based Metric

This metric employs no weighting, and can be seen to have a somewhat distinct grouping below the distance threshold. In this case, the orange and green groupings appear to be reasonably dissimilar, leading to a large distance between them at the level of two clusters. Note that this metric clearly has the issue of clusters/partitions being merged at a height below both clusters, this makes the dendrogram harder to understand and makes the visualization less clear. This issue is present in the centroid metric as well, though not on the one we demonstrated with our data ([19]).

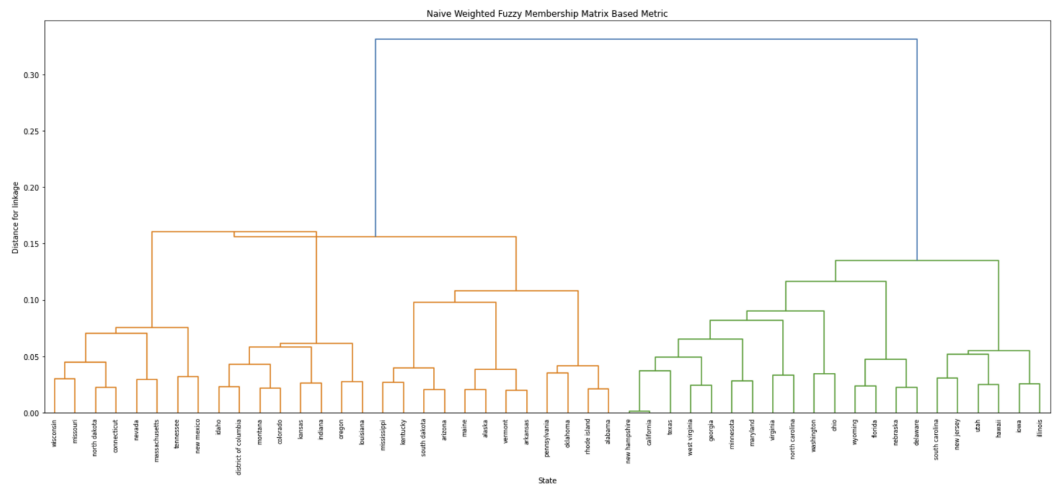

Figure 5: Naive Weight Based Fuzzy Membership Matrix Based Metric

This metric employs the naive weighting outlined in 4.2.2, and has a much more balanced structure between the two primary groupings and within. This distance between the two is lesser than that of the fuzzy membership matrix based metric, and each individual state can be seen to be of a more evenly based distance to their nearest states or partitions. This metric also has the inversion merging issue as seen in the former metric, though to a much lesser degree.

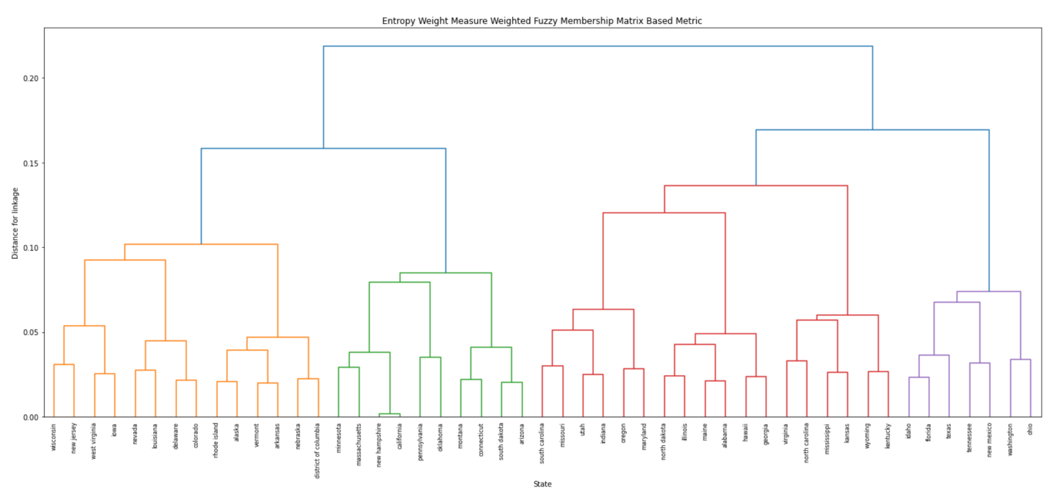

Figure 6: Entropy Weight Measure Based Fuzzy Membership Matrix Based Metric

This metric employs the entropy weight measure outlined in 4.2.3, and has a qualitatively very different structure from any of the previous metrics. Firstly, it is indeed a fairly balanced structure between the four primary groupings. Secondly, this metric found a larger number of primary groupings of clusters over time. Additionally, at least in this particular example, there appears to be no inversion merging issues that plagued the basic fuzzy membership matrix metric and was present in the naive weighting method. The cluster distances themseleves appear to be more evenly balanced and spread out from each other, making this graph somewhat easier to follow than some previous metrics.

6 Conclusions

A more granular approach to the development of sustainability policy is key to achieving sustainability goals at a more accurate level. Given the nature of certain groups having achieved a certain level in a certain metric yet disagreeing in their similarities on other metrics, this paper has employed a metric for hierarchically clustering groups with a weighted fuzzy clustering based metric to determine similarity.

It is critical to note that there is no ground truth to the similarity of states based upon sustainability metrics. In this way, most conclusions drawn from the paper will inherently be qualitative, based upon the appearance, or the structure, of the dendrograms generated by their respective similarity measures.

The classical metrics used in determining similarity (centroid link, single link, and complete link), were found to have generalized groupings that were not very balanced, with a vast majority of the clusters falling under one of two potential general groups. The single link metric most predominately shows this, with only one state falling outside of the primarily generalized grouping.

The metrics design in this paper for determining similarity (fuzzy membership based metric, naive weight based fuzzy membership based metric, and entropy weight measure based fuzzy membership based metric) had a mixed level of balance between the generalized groupings of states. The fuzzy membership based metric is somewhat balanced structurally, but has a substantial distance split between the two generalized groupings. The naive weight based metric did not have this issue, and was much more structurally balanced. The same is true for the entropy weight measure based metric, which appears to be more structurally balanced metric than any of the former ones, and also has a larger number of generalized groupings, making it a bit easier to dissect and follow. One new problem that arose, however, was very evident in the first two metrics, which had inversion problems with the merging of certain centroids. This makes those dendrograms harder to understand, thus making it harder to find discoveries.

As mentioned, it is difficult to determine the accuracy of any of these modelings of the data and the clustering within, due to the lack of a ground truth. In this manner, all we can consider is the visual appearance and qualitative nature of these dendrograms. This finds that the naive weighted method and the entropy weighted methods suggest a possibility of producing more balanced dendrograms that make following the similarities and differences of certain members simpler than other metrics, and allowing for more flexibility in the generation of hierarchical models for cases in which the observations in the data has juxtapositions of agreement and disagreement within their respective attributes, because of the lack of a strong border, and the inclusion of a fuzzy membership matrix into the similarity metric.

6.1 Future Research

For any future research in this subject, it would be preferred to determine a method by which we can more accurately and quantitatively compare the dendrograms produced by certain metrics. One major pitfall of this research is the lack of that possibility, where most determinations of ”model success” was derived from qualitatively comparing the models we had produced with those produced by the classical metrics, thus leading to a majority of the discussion on the models being based around the balance in the structure of the dendrogram and the amount of inversions in merging certain clusters or partitions.

With respect to the metrics themselves, we would like, in the future, to have a weight matrix based on expert knowledge. Due to the nature of sustainable metrics, it can be inherently difficult to determine the respective importance of some attribute (for instance, how should we compare the weight of airborne pesticide pollution versus the weight of out-factory water pollution). This may allow for a more rich knowledge base to be generated by a metric that takes into account former knowledge and research attained in sustainability.

Another future point of research we would like to work towards would be incorporating a greater corpus of data. Currently, it is quite difficult to parse through and gain access to certain sustainability-related data, simply due to the somewhat tedious ways that the data it stored. It makes mass access to the data fairly difficult. This becomes more-so true when considering geographic data. In this sense, we would like to expand our DPSIR framework to include and account for a richer level of indicators. This would allow us a greater set of metrics to select upon.

References

[1] P. Asha, K. Mannepalli, R. Khilar, *et al.*, "Role of machine learning in attaining environmental sustainability," *Energy Reports*, 2022.

[2] M. Nilashi, P. F. Rupani, M. M. Rupani, *et al.*, "Measuring sustainability through ecological sustainability and human sustainability: A machine learning approach," *Journal of Cleaner Production*, 2019.

[3] G. M. Abdella, M. Kucukvar, N. C. Onat, H. M. Al-Yafay, and M. E. Bulak, "Sustainability assessment and modeling based on supervised machine learning techniques: The case for food consumption," *Journal of Cleaner Production*, 2020.

[4] N. I. Molina-Gómez, K. Rodríguez-Rojas, D. Calderón-Rivera, J. L. Díaz-Arévalo, and P. A. López-Jiménez, "Using machine learning tools to classify sustainability levels in the development of urban ecosystems," *Sustainability*, 2020.

[5] H. Chen, J. Xu, K. Zhang, *et al.*, "New insights into the dpsir model: Revealing the dynamic feedback mechanism and efficiency of ecological civilization construction in china," in *Journal of Cleaner Production*, 2022.

[6] X. Y. Qiang Wang, "Investigating the sustainability of renewable energy – an empirical analysis of european union countries using a hybrid of projection pursuit fuzzy clustering model and accelerated genetic algorithm based on real coding," in *Journal of Cleaner Production*, 2020.

[7] S. Gebremedhin, A. Getahun, W. Anteneh, S. Bruneel, and P. Goethals, "A drivers-pressure-state-impact-responses framework to support the sustainability of fish and fisheries in lake tana, ethiopia," in *Sustainability*, 2018.

[8] M.-S. Yang, "A survey of fuzzy clustering," in *Mathematical and Computer Modelling*, 1993.

[9] E. H. Ruspini, J. C. Bezdek, and J. M. Keller, "Fuzzy clustering: A historical perspective," *IEEE Computational Intelligence Magazine*, 2019.

[10] C. Shouyu, "Theory and application engineering fuzzy sets," in *National Defence Industry Press*, 1998.

[11] Q. Zou, L. Liao, Y. Ding, and H. Qin, "Flood classification based on a fuzzy clustering iteration model with combined weight and an immune grey wolf optimizer algorithm," in *Water*, 2019.

[12] L. Hui-yun, W. Guo-li, W. Ben-de, and Z. Chi, "A fuzzy clustering iterative model on temporal distribution of precipitation and its application," in *Sixth International Conference on Fuzzy Systems and Knowledge Discover*, 2009.

[13] F. Murtagh and P. Contreras, "Algorithms for hierarchical clustering: An overview," *WIREs: Data Mining and Knowledge Discovery*, 2012.

[14] A. Ciaramella, D. Nardone, and A. Staiano, "Data integration by fuzzy similarity-based hierarchical clustering," *BMC Bioinformatics*, 2020.

[15] V. Torra, "Fuzzy c-means for fuzzy hierarchical clustering," in *The 14th IEEE International Conference on Fuzzy Systems*, 2005.

[16] E. Smeets and R. Weterings, "Environmental indicators: Typology and overview," European Environment Agency, Tech. Rep., 1999.

[17] H. Bei, Y. Mao, W. Wang, and X. Zhang, "Fuzzy clustering method based on improved weighted distance," *Mathematical Problems in Engineering*, 2021.

[18] Y. Zhu, D. Tian, and F. Yan, "Effectiveness of entropy weight method in decision-making," *Mathematical Problems in Engineering*, 2020.

[19] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning with Applications in R Second Edition*. Springer, 2021.