CS 4824

Fuzzy Clustering Based Hierarchical Clustering

Applications in Sustainability

Machine Learning

Problem & Motivation

- **Initial Problem:** Model to generate the best set of sustainability investments by State to reach Net-Carbon
  - **Issues**: Long-Term Energy Generation and Consumption Forecasting
  - **Observations:** Plenty of energy and socio-economic data for states, need of proper clustering models for US states

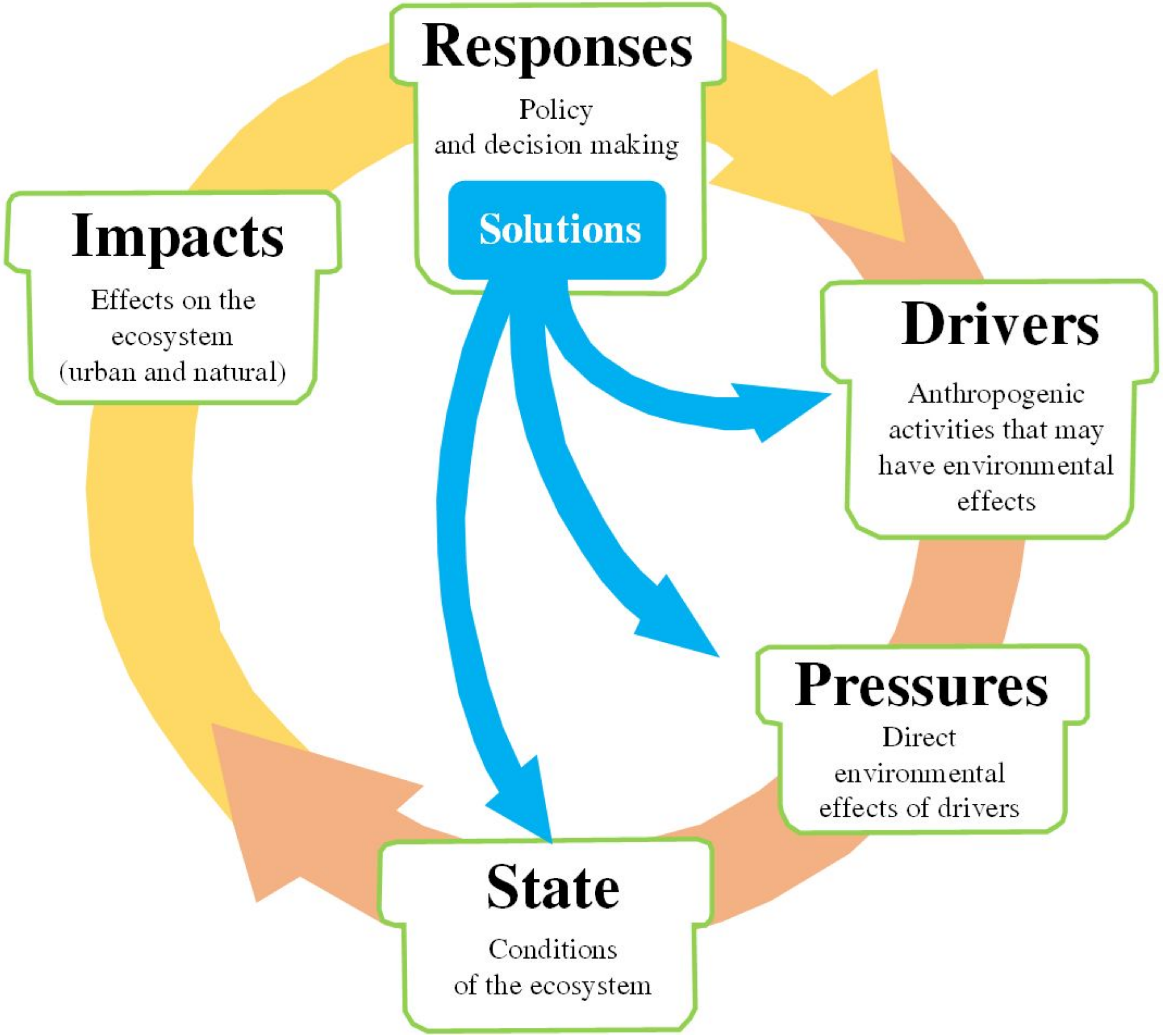
- **Revised Problem:** Group US State Territories by their sustainability measure
  - **Approach**: Fuzzy C-Means, Entropy-Weighted Hierarchical Fuzzy C-Mean Clustering

- **Motivation**: Federal Government can create Legislation accordingly by Cluster

Data Collection

DPSIR Model for Data

- 4 Datasets
  - Socio_Economic_By_State
    - 48*70
  - Production_Threashold_By_State
    - 48*15
  - Consumption_Emmisions_By_State
    - 48*11
  - Price_Expenditure_By_State
    - 48*10



Methodology

Fuzzy Clustering

General Fuzzy Clustering

“Uncertainty of belonging” for class membership

Observations can “belong” to multiple clusters

Structures of clusters can reveal relationships between clusters

Fuzzy Clustering Iteration

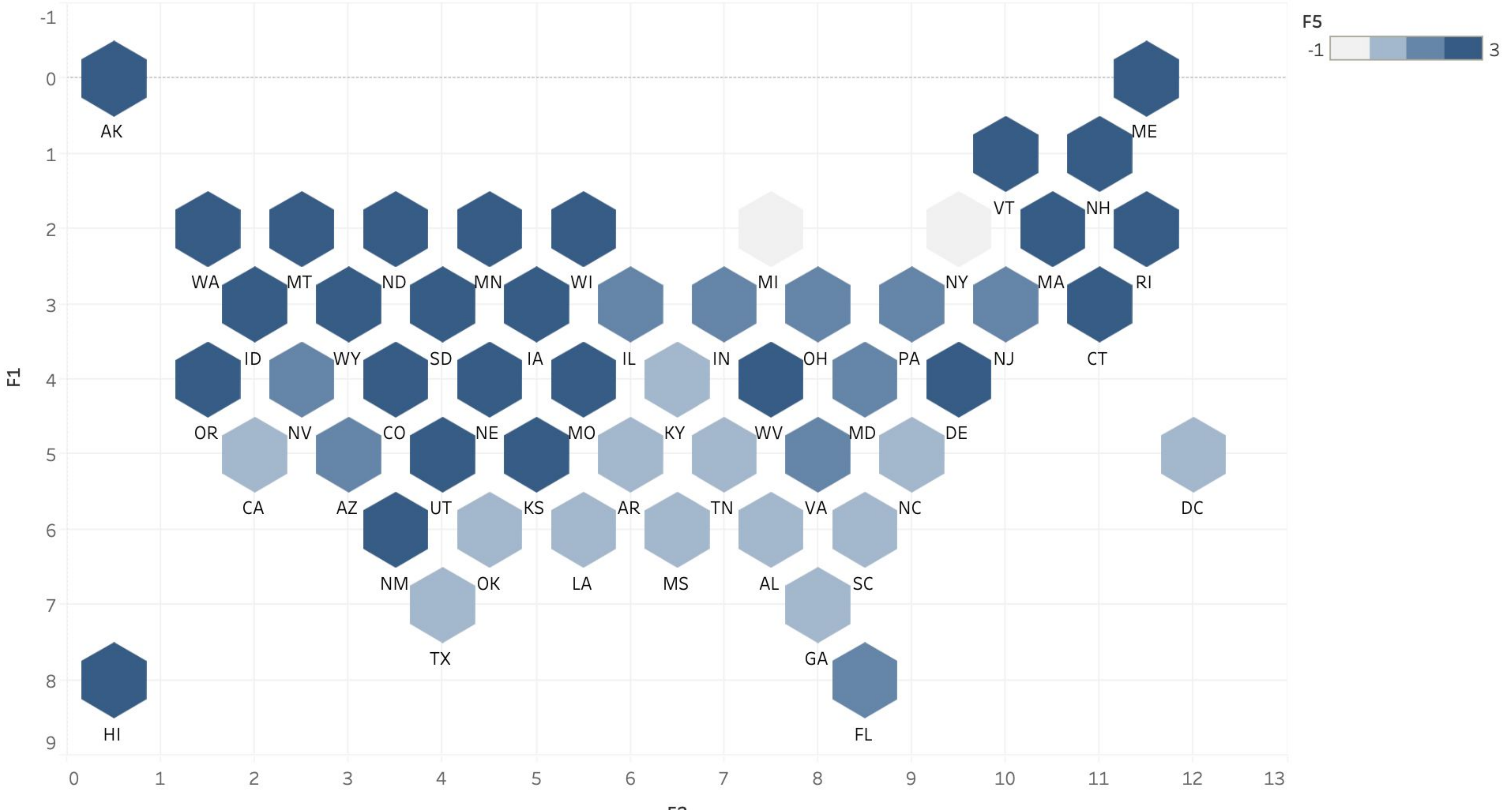
Introduced by Shouyu, 1998

Implements weighted relevancy of features

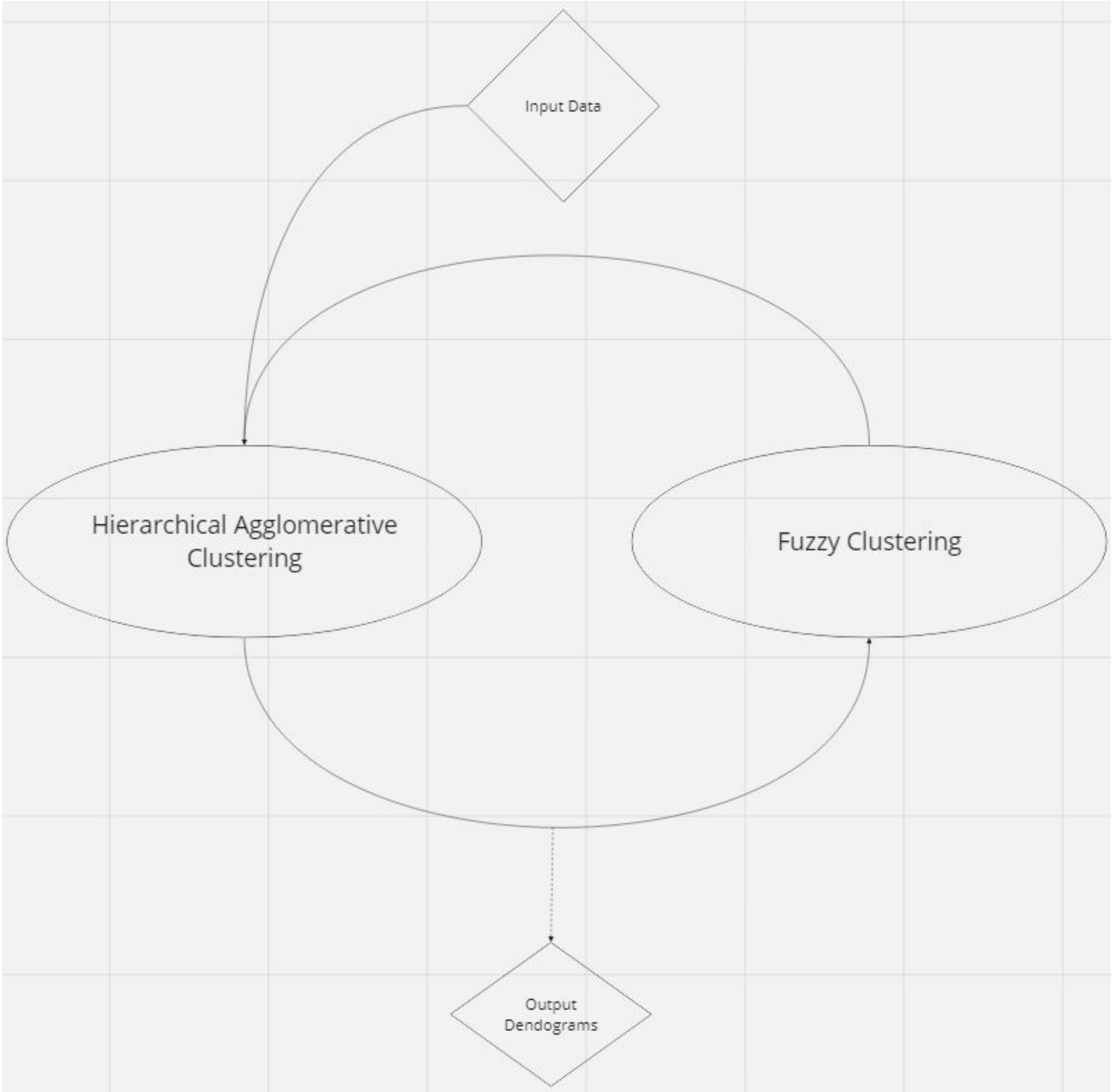
Iteratively determining a fuzzy membership matrix and the fuzzy cluster centroids

Has potential in areas without evaluation criteria

Fuzzy C-Means Clustering of States



Project Flow



-Input data reflects certain sustainability criteria per state (i.e. energy consumption, CO2 emissions, energy expenditure)

-Merge clusters at each iteration, in a bottom up approach, using fuzzy clustering membership matrix values as merging criteria

-Output is a dendrogram that shows which states are most related to each other at each iteration

Weight

None	Naive Weighting	Entropy Weight Measure
Using no weight matrix All attributes are of equal importance	Simplistic weighting based on feature relevance	Measure dispersion of the data How much information can be derived

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

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

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

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

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

Hierarchical Agglomerative Clustering

General Model

At each step, merge two partitions together based on some similarity metric

Start with all observations and merge till all fall under one cluster

Construct a tree (dendrogram) based on the combinations

Fuzzy Clustering Incorporation

Use fuzzy clustering as a method of determining similarity between partitions

Allows for flexibility in hierarchical approach

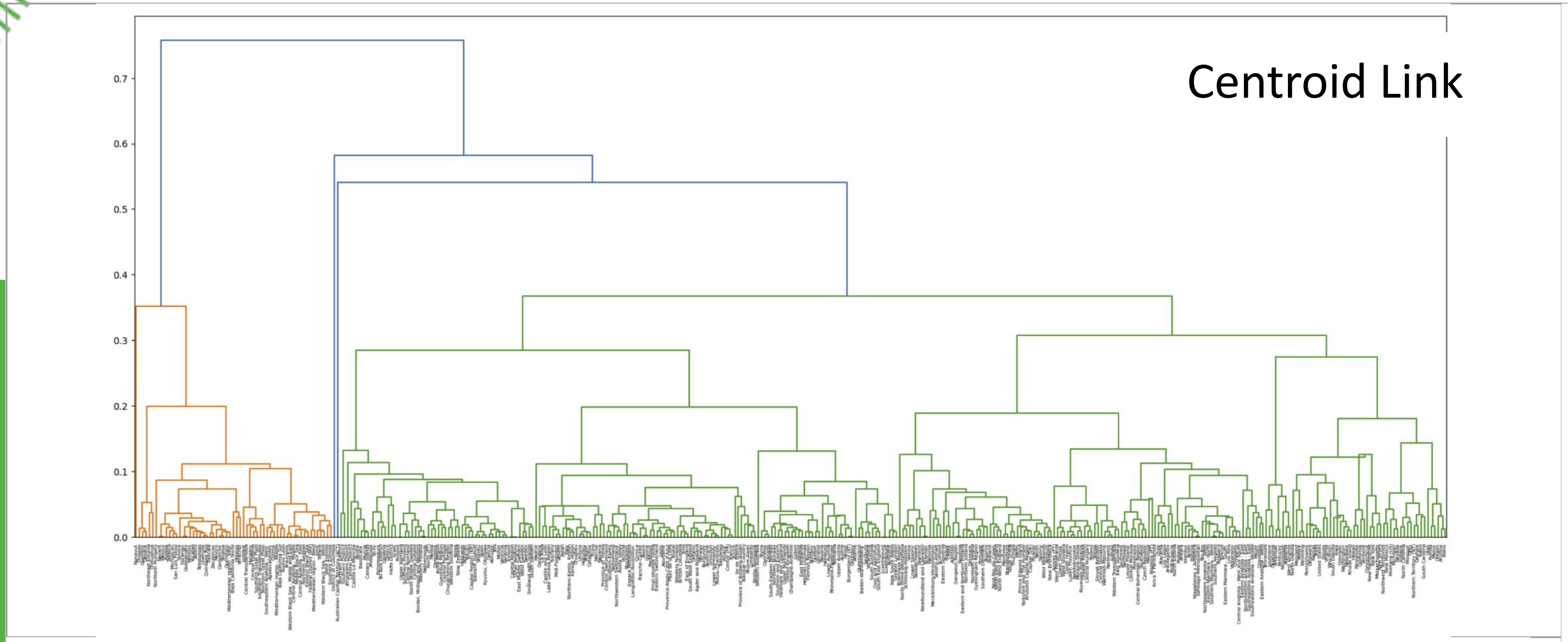
Employ membership functions to determine a distance metric

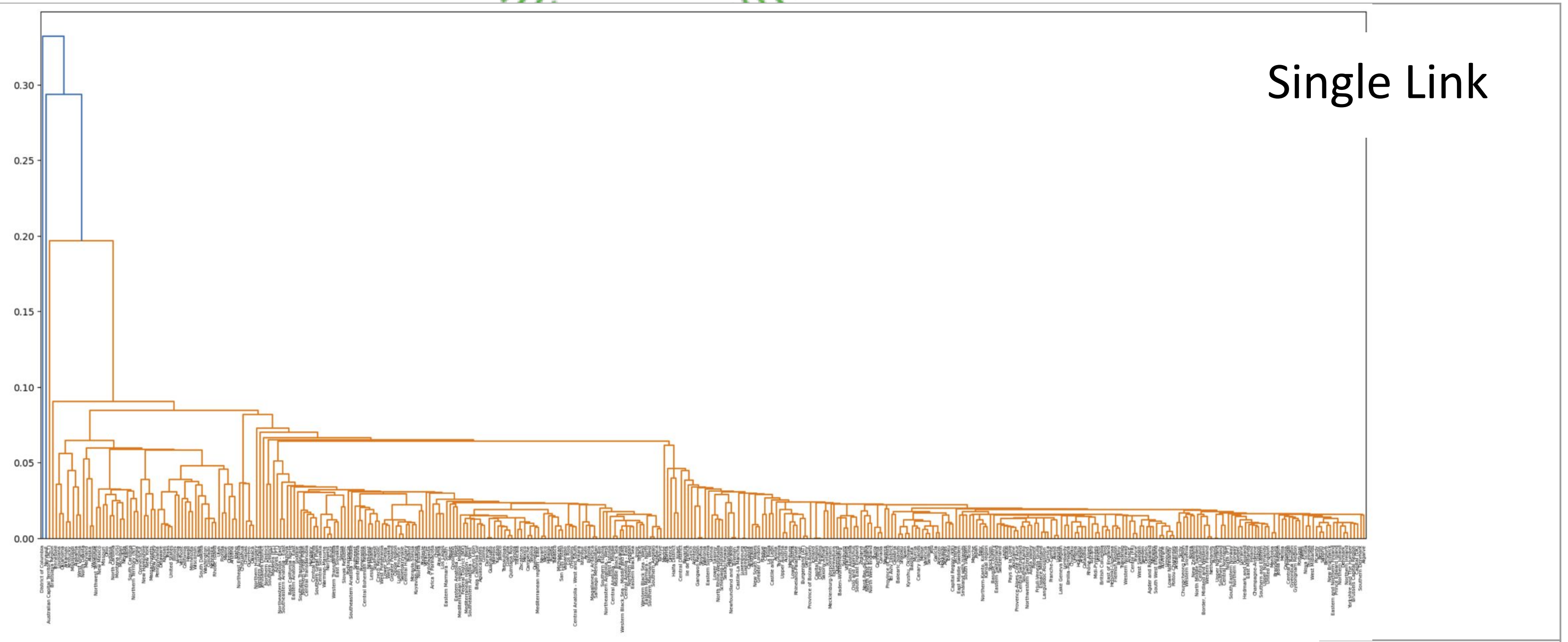
Similarity Metrics

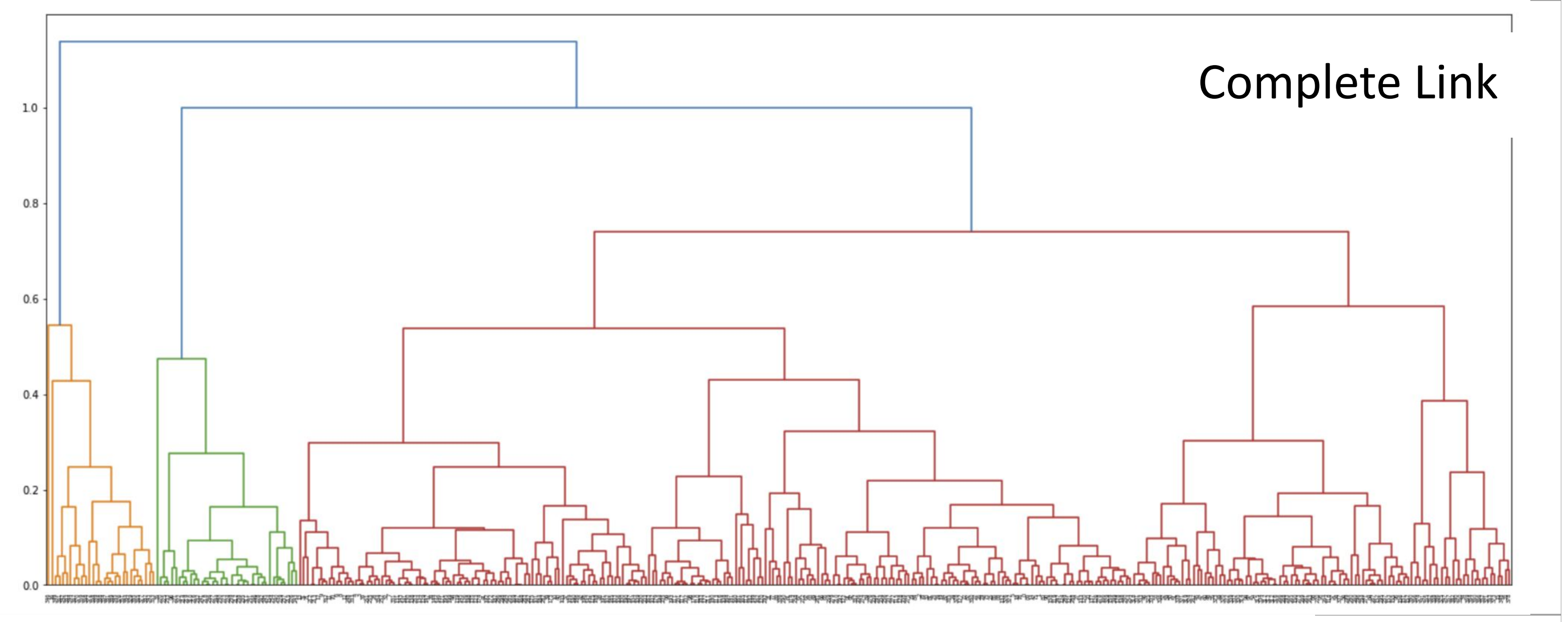
We implemented all of these metrics manually

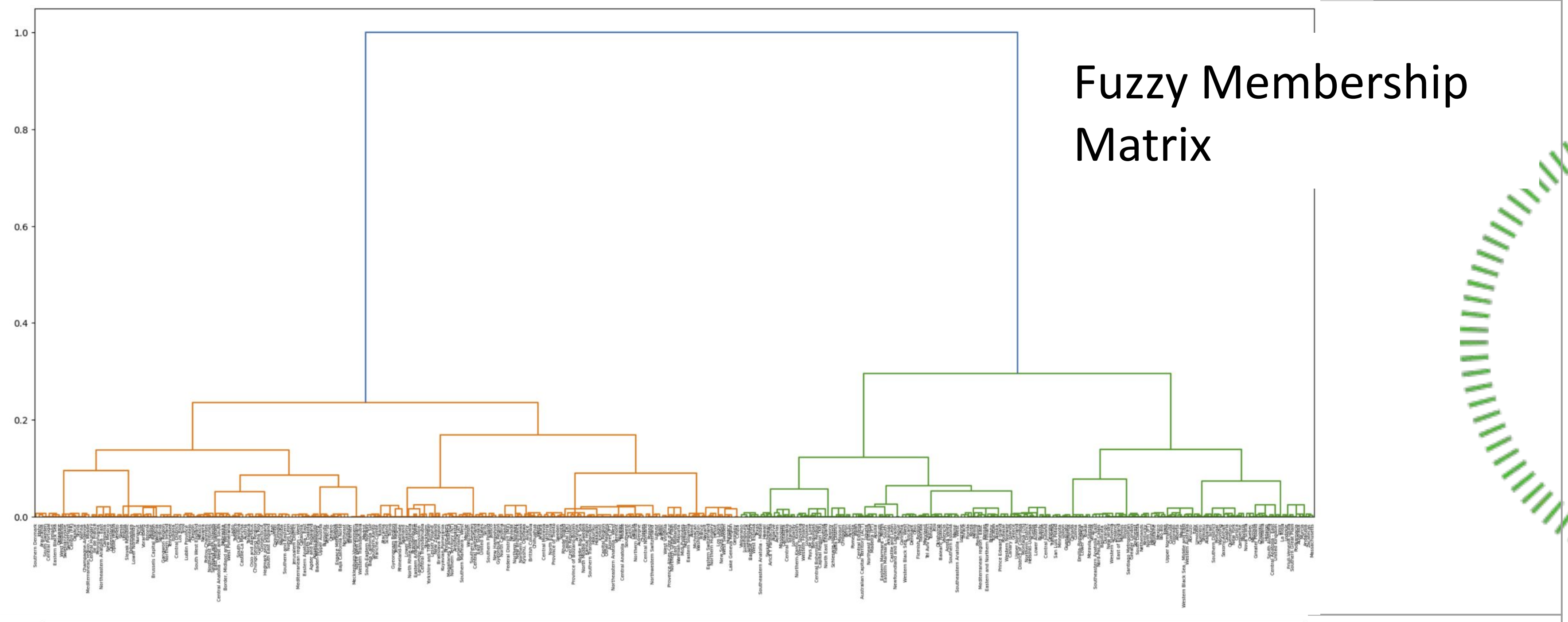
Centroid Link	**Single Link**	**Complete Link**
Distance between centroids	Distance between most similar members	Distance between most dissimilar members
Fuzzy Membership Matrix	**Naive Weighted Fuzzy Membership Matrix**	**EWM Weighted Fuzzy Membership Matrix**
Highest similarity between clusters from entire matrix	Implementing the naive weighting onto the fuzzy membership matrix	Implementing the entropy weight based weighting onto the fuzzy membership matrix

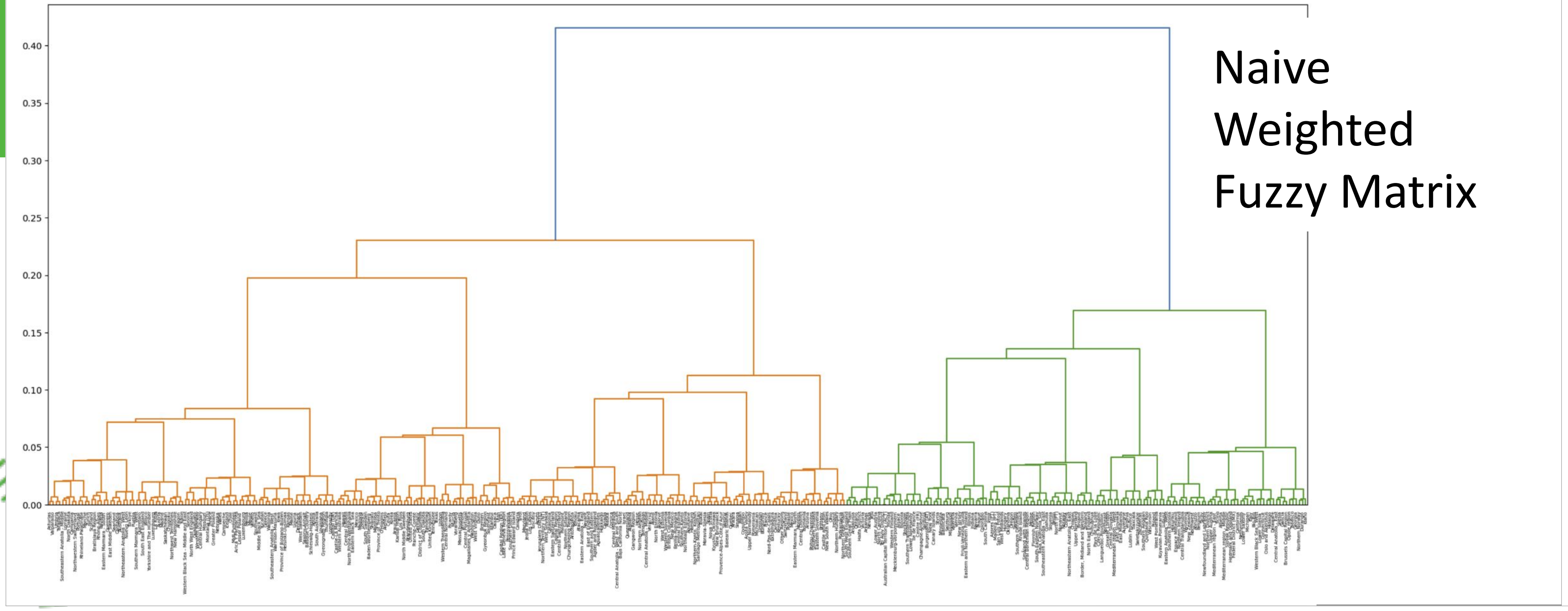
SIMPLE DATA

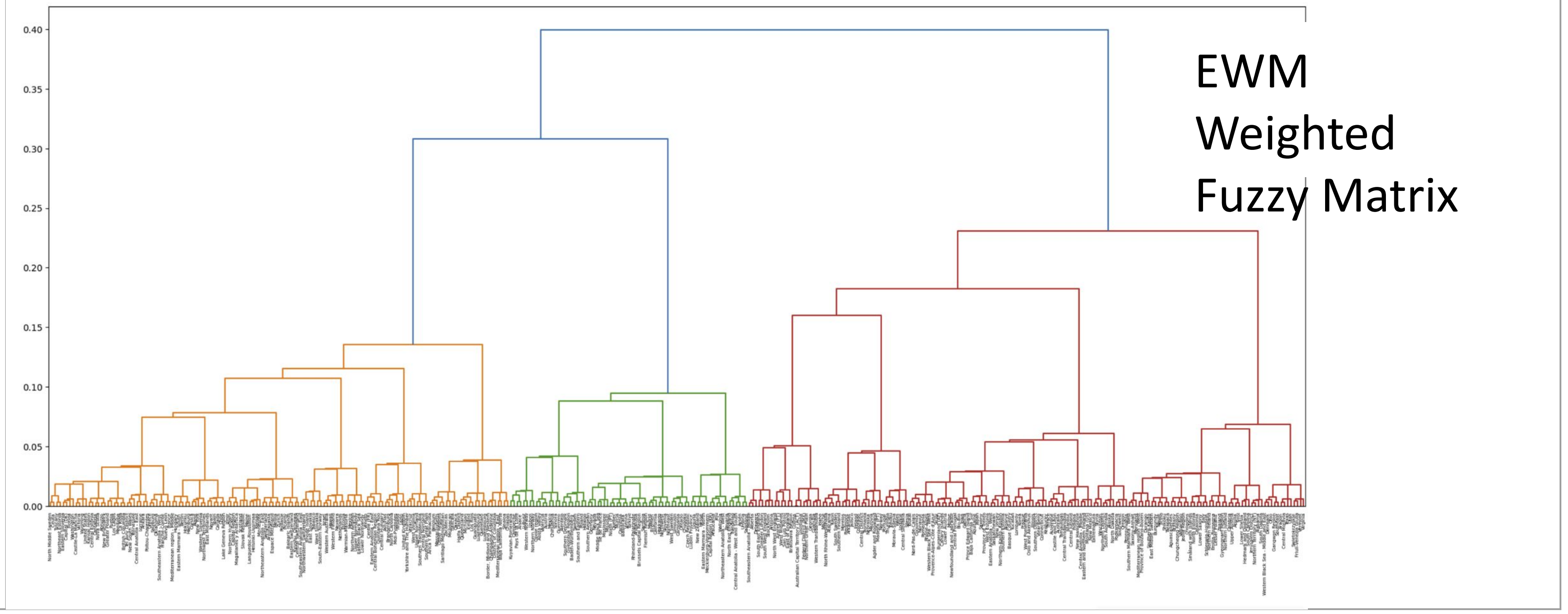

Centroid Link


Single Link


Complete Link

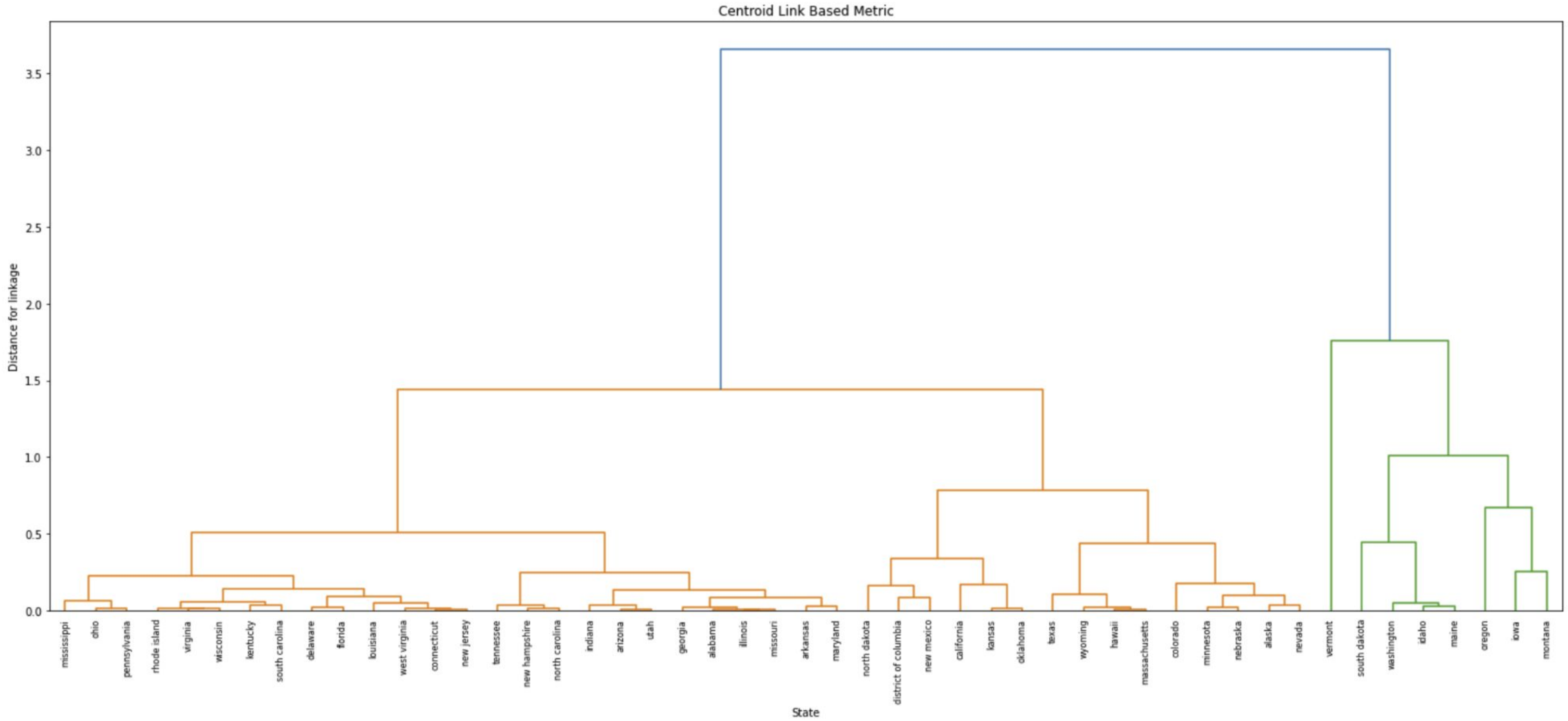

Fuzzy Membership Matrix


Naive Weighted Fuzzy Matrix

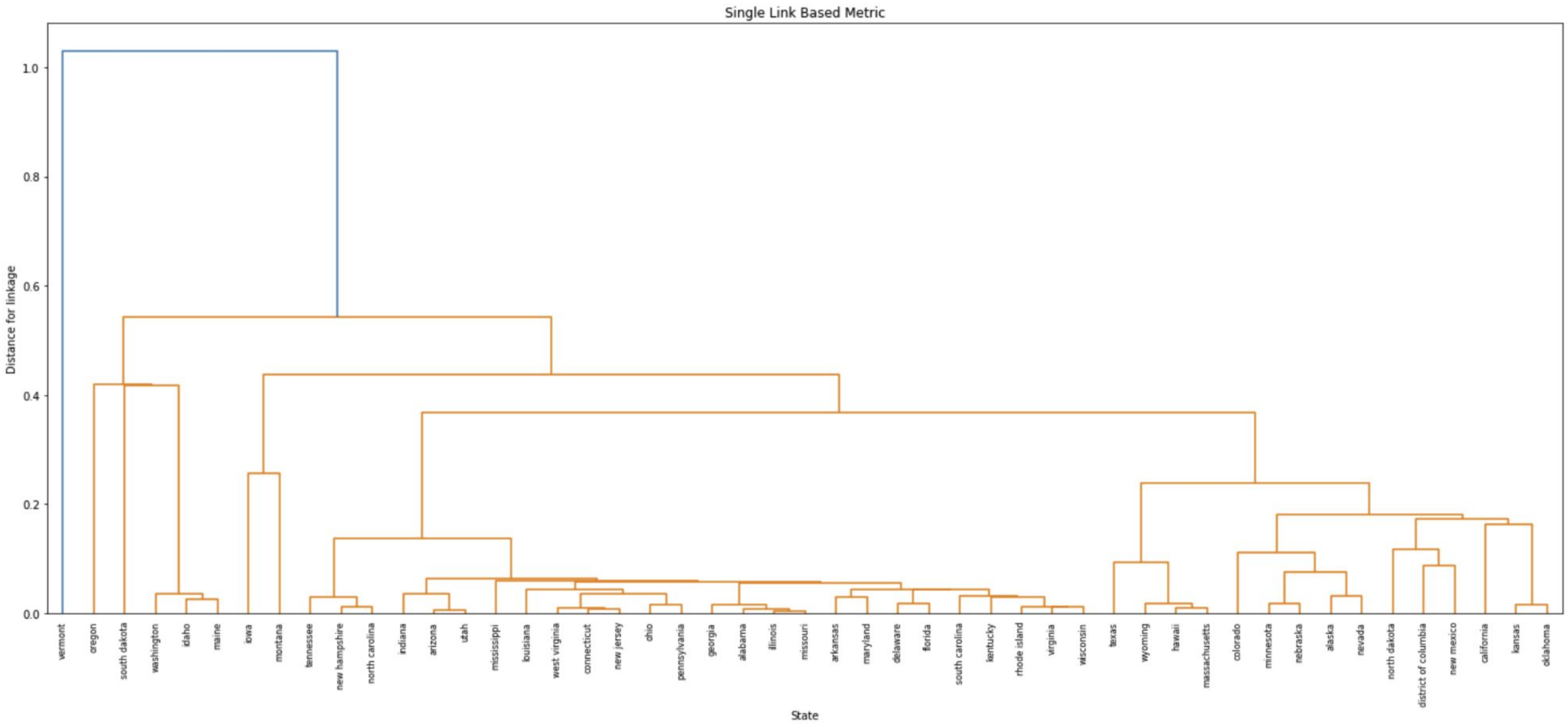

EWM Weighted Fuzzy Matrix

Findings

Centroid Link

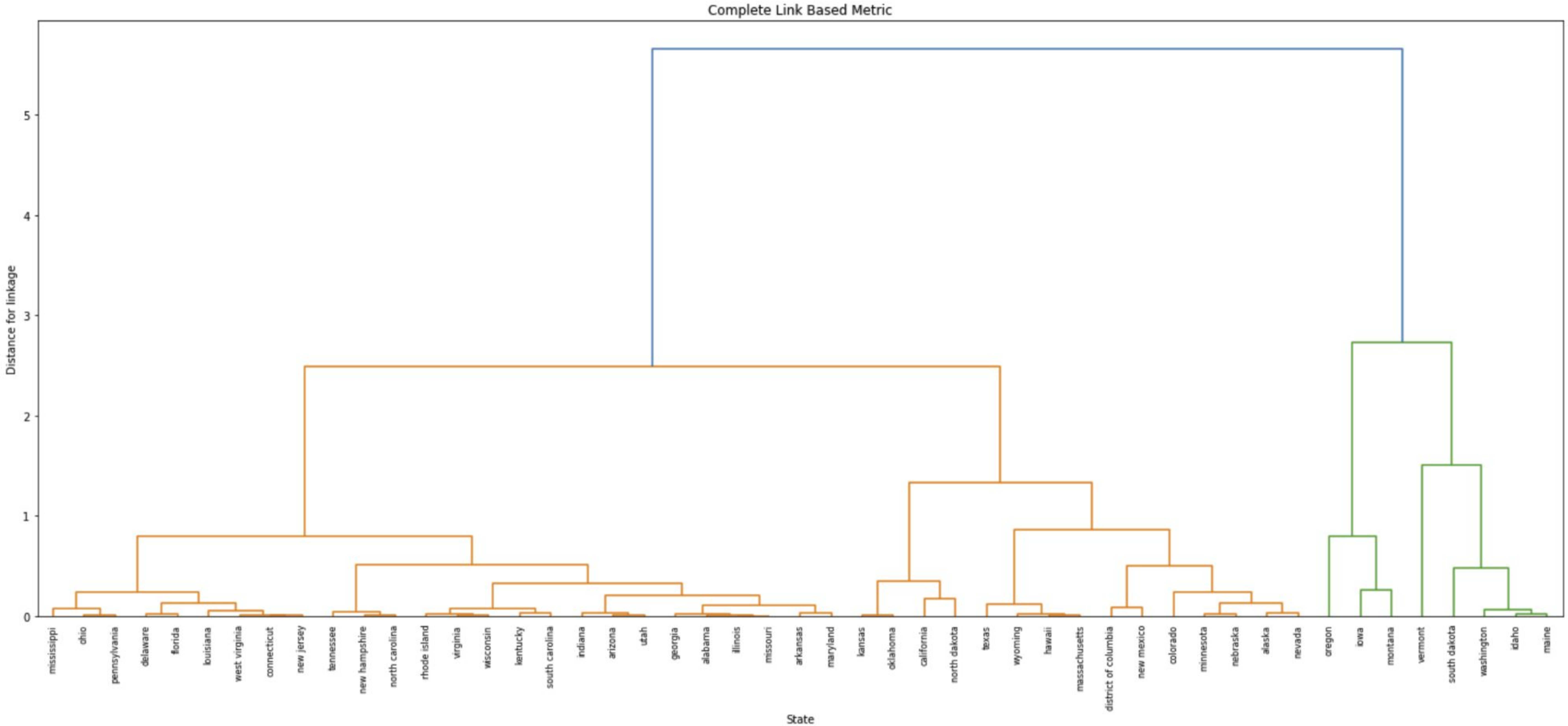


Single Link

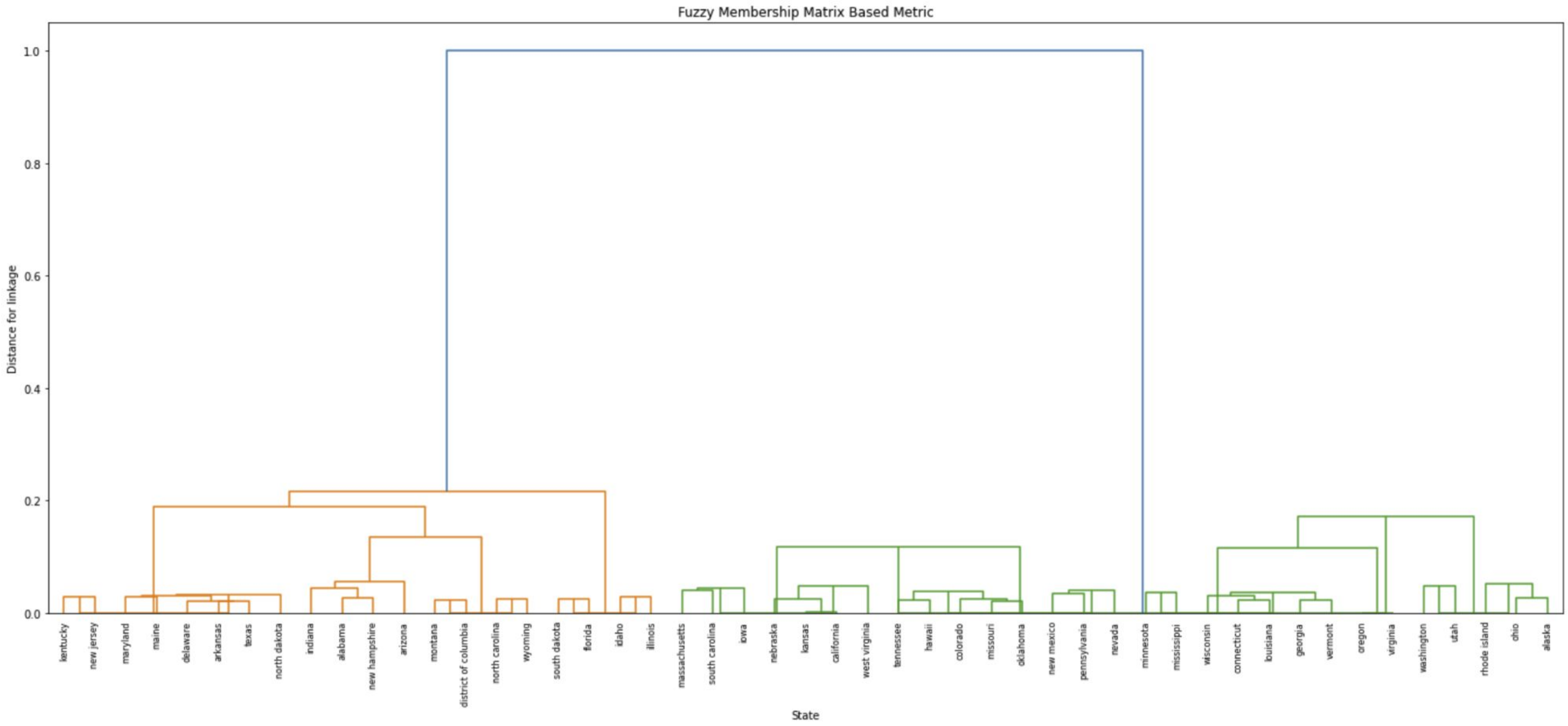


Single Link Based Metric
Distance for linkage
1.0
0.8
0.6
0.4
0.2
0.0
vermont
oregon
south dakota
washington
idaho
maine
iowa
montana
tennessee
new hampshire
north carolina
indiana
arizona
utah
mississippi
louisiana
west virginia
connecticut
new jersey
ohio
pennsylvania
georgia
alabama
illinois
missouri
arkansas
maryland
delaware
florida
south carolina
kentucky
rhode island
virginia
wisconsin
texas
wyoming
hawaii
massachusetts
colorado
minnesota
nebraska
alaska
nevada
north dakota
district of columbia
new mexico
california
kansas
oklahoma
State

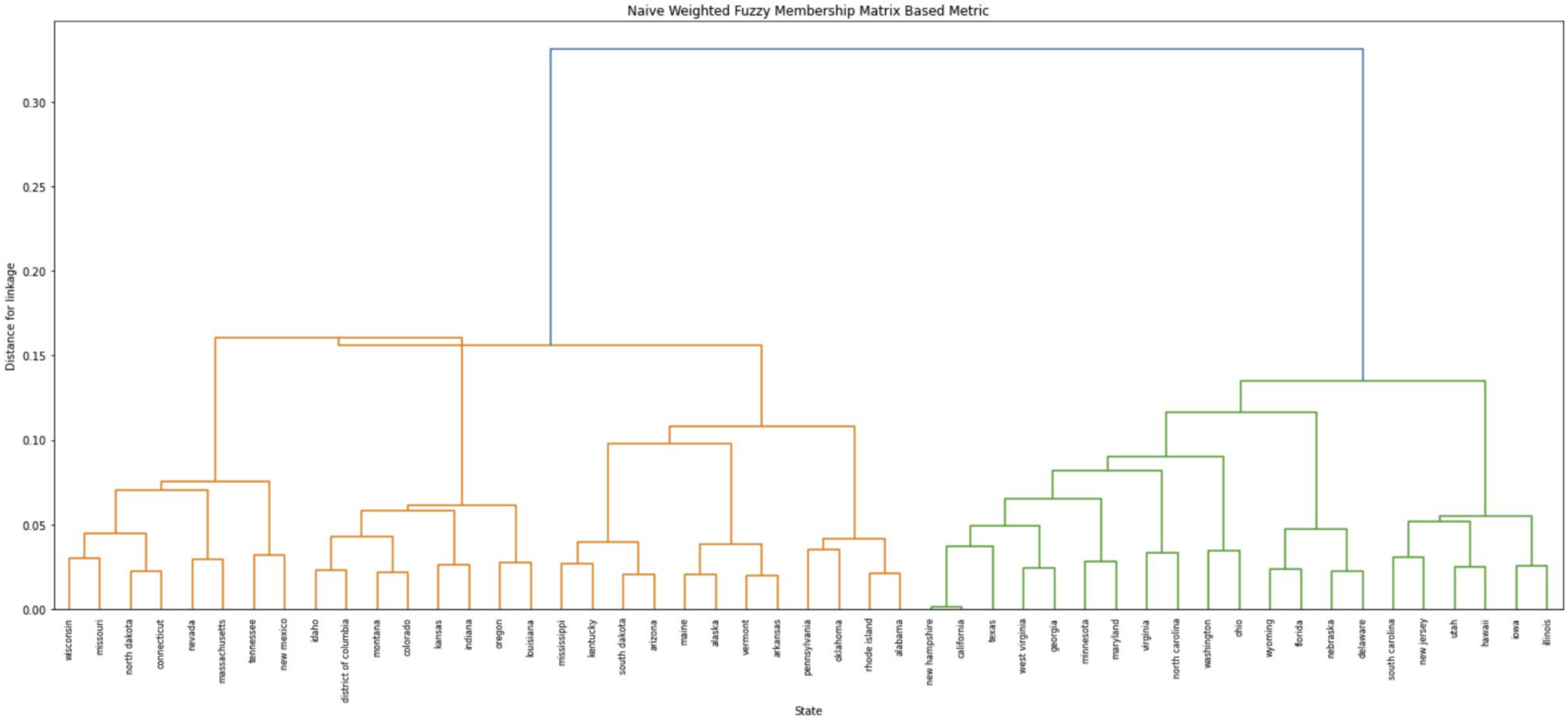
Complete Link



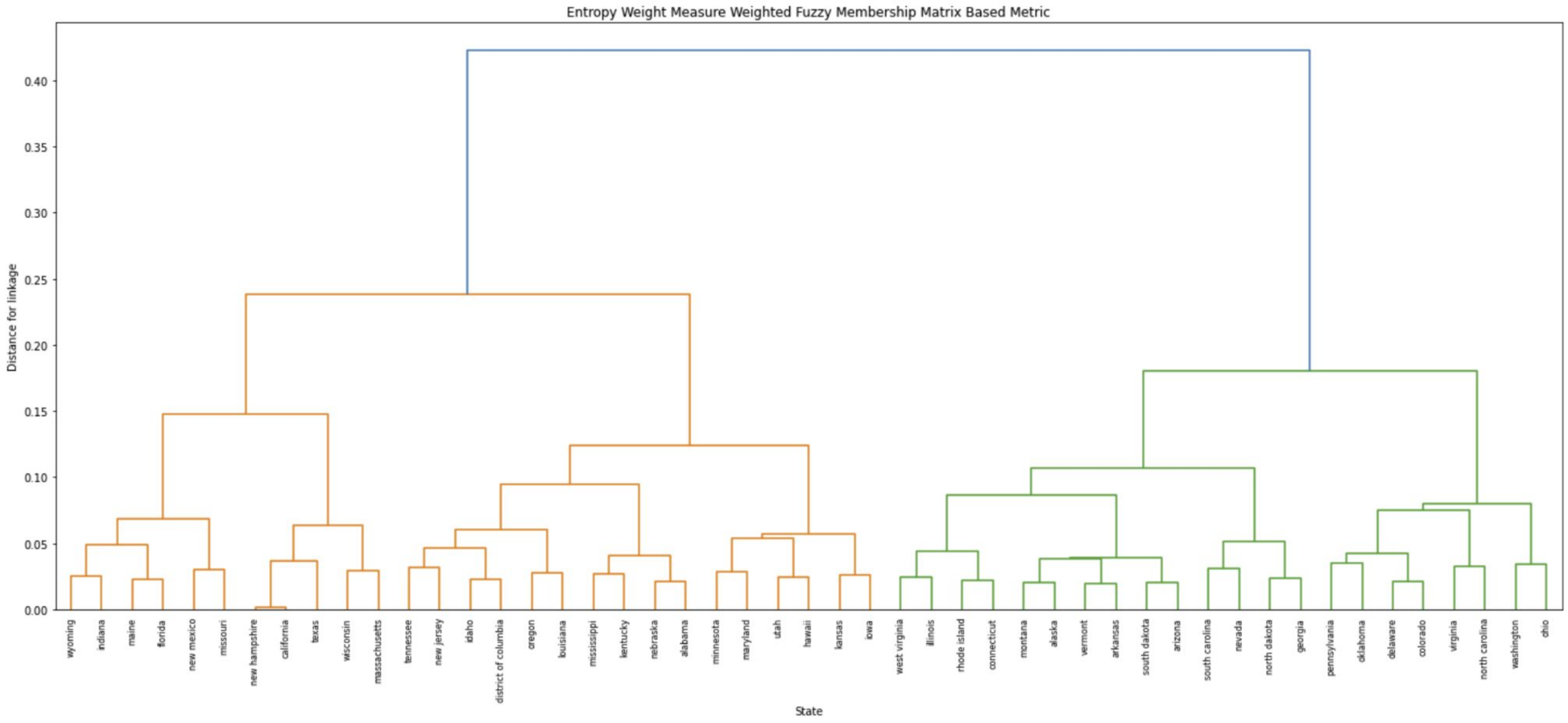
Fuzzy Membership Matrix



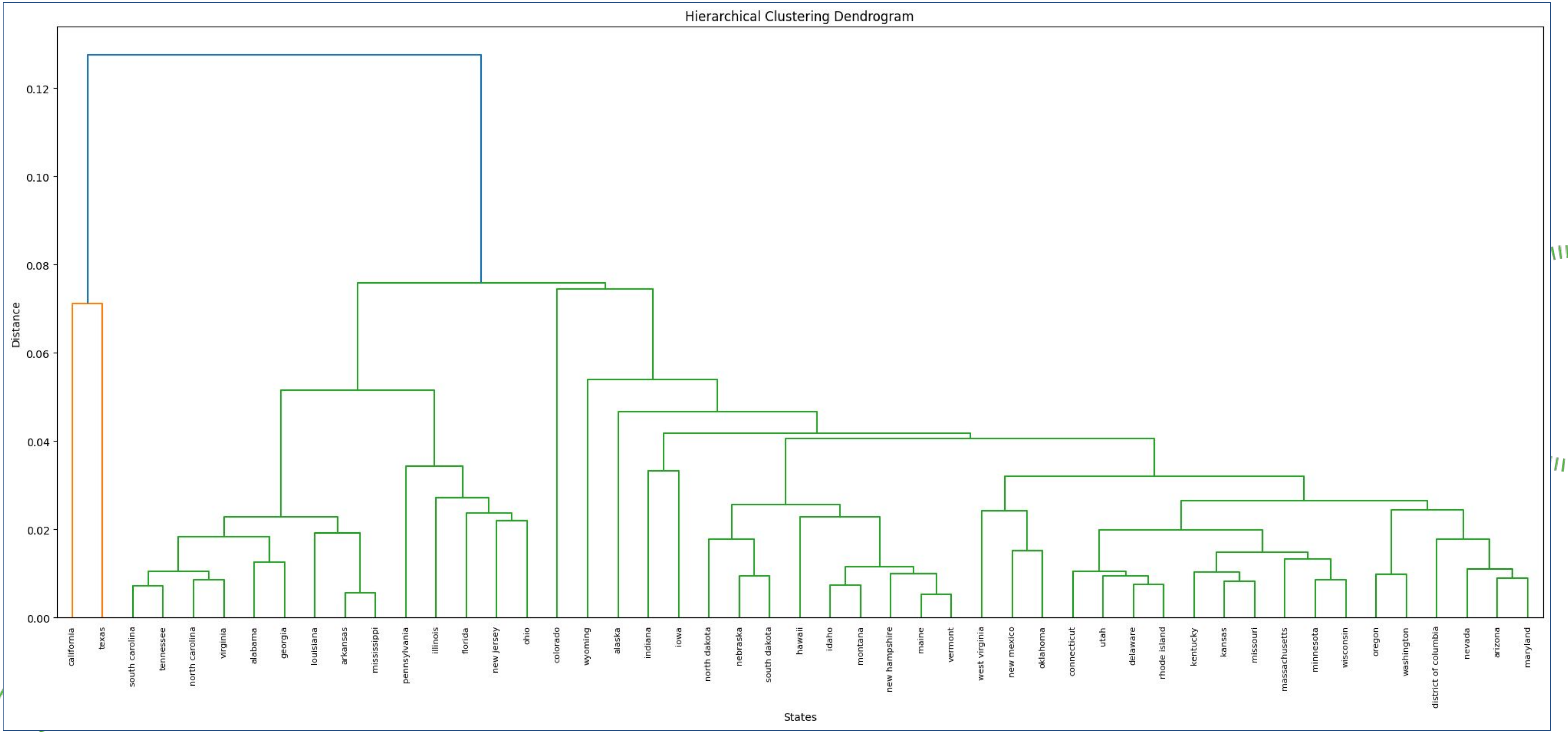
Naive Weighted Fuzzy Matrix



EWM Weighted Fuzzy Matrix



EWM Weighted Matrix with Ward's Criterion



Conclusions

Conclusion

01 Classical metrics lack balanced structures

02 Fuzzy based metrics had greater balance

03 Fuzzy membership and naive weight metrics have inversion issues

04 EWM metric is qualitatively balanced and has no inversions for this particular case

05 No ground truth exists to compare by

Future Research

Quantitative validity

Determine a method to compare the dendrograms quantitatively.

Expert based weighting

Allow for rich knowledge base from prior sustainability research.

Greater corpus of data

Expand the DPSIR framework for a richer level of indicators to select upon.

Citations

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-G´omez, K. Rodr´ıguez-Rojas, D. Calder´on-Rivera, J. L. D´ıaz-Ar´evalo, and P. A. L´opez Jim´enez, “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.

Citations

[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.

Questions?