A Study of Hierarchical Correlation Clustering for Scientific Volume Data

– Yi Gu and Chaoli Wing, Michigan Technological University

CS677

Topics in Large Data Visualisation

Group -6

Adarsh Mudgal

K. Prajwal Subudhi

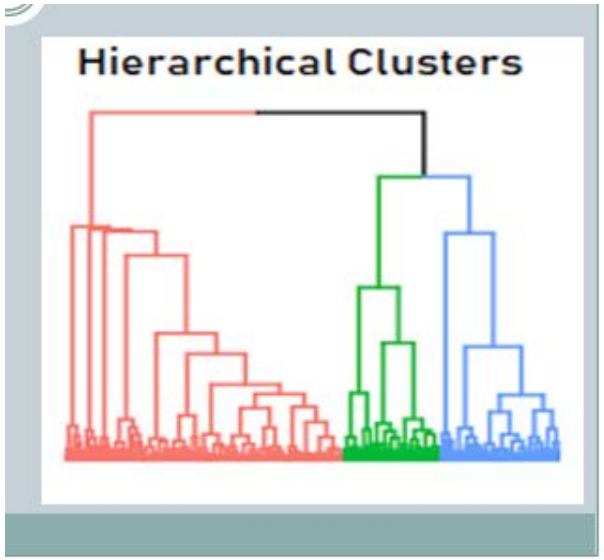
Ruchita Rani

Sevak Shekokar

Indian Institute of Technology Kanpur

Objective:

- Visualization and Analysis of **time-varying multivariate volume data** (climate data)
- Explore various hierarchical clustering methods for classifying volumetric samples based on the similarity
- Comparative analysis of clustering methods in terms of **quality** and **performance**
- Integrate Clustering Results with Volume Rendering and Parallel Coordinates for Comprehensive Visualization



http://www.sthda.com/english/wiki/beautiful-dendrogram-visualizations-in-r-5-must-known-methods-unsupervised-machine-learning

Introduction:

- Identifying connections within time-varying multivariate data is crucial in various scientific fields so some solutions are mentioned
- One effective solution is to **cluster voxels** based on correlation similarity
- Many research efforts adopted the **standard correlation coefficients** to study the **linear correlation** between variables, yet little work is done to build a hierarchy for coarse-to-fine exploration of data correlation
- Hierarchical clustering can show **cluster within clusters** and much as in multiresolution visualization, it provides us a flexible means to adaptively examine the data
- The evaluation includes side-by-side qualitative comparison of clustering results and quantitative comparison using **silhouette plot**

Past Techniques:

Researchers have applied various techniques in their analysis.

1. **Standard Pointwise Correlation**
2. **New user interfaces** were also developed to visualize multivariate data relationships
3. **Parallel Coordinates**

Standard Pointwise Correlation:

- Basic statistical method used to measure the linear relationship between two variables at specific points in time or space
- In the context of **multivariate data analysis**, it is used to analyze relationships between pairs of variables at specific instances,
- **Problem -** It does not account for the potential relationships among multiple variables simultaneously

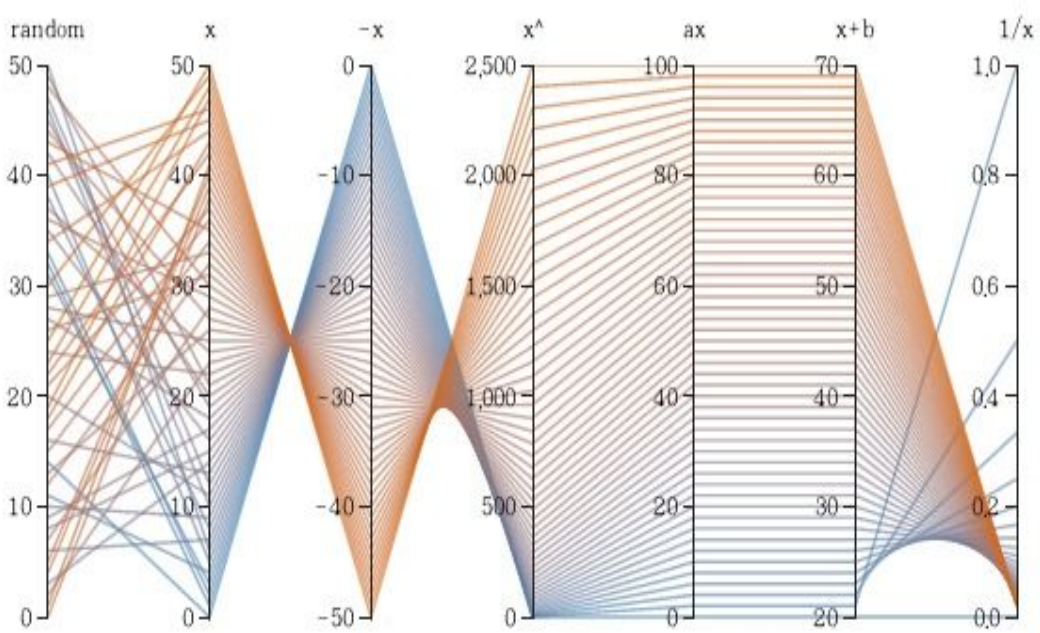
New User Interfaces:

- To allow users to interact with and explore complex datasets providing visual representation.

Past Techniques: Conti....

Parallel Coordinates:

- Popular technique for visualizing relationships among a large collection of variables
- **Issue** : For parallel coordinates is to order dimensions to reveal multivariate data patterns

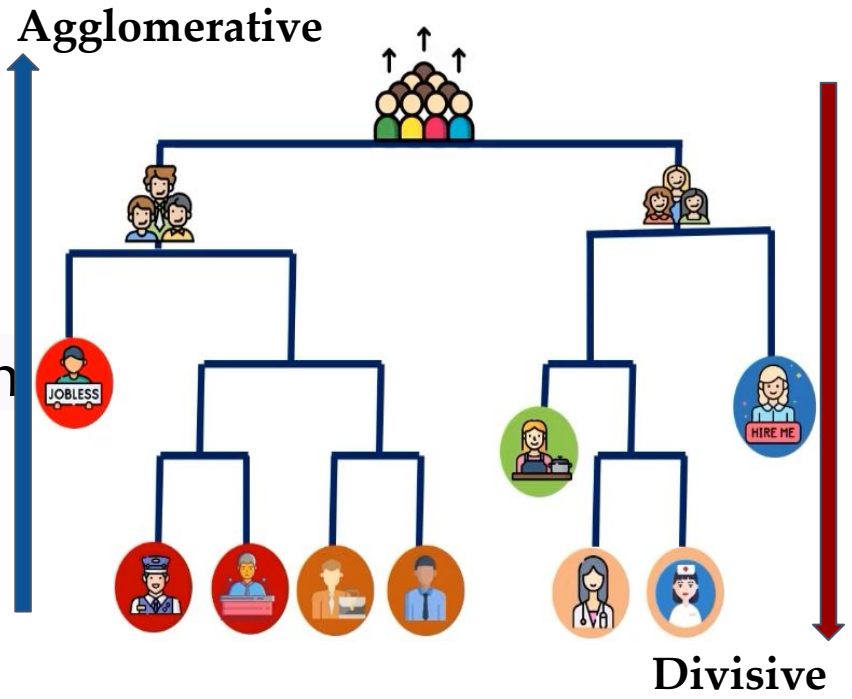


https://en.m.wikipedia.org/wiki/File:Parallel_coordinates-sample.png

- **Possible Modifications:**
  - One way to achieve this is based on the evaluation of similarity between dimensions
  - Yang et al. built a hierarchical dimension structure and allowed dimension reordering and filtering
  - Use parallel coordinates with two correlation-based distance measures to visualize quantitative correlation information for hierarchically clustered volume samples

Hierarchical Clustering:

- A technique used in data analysis and machine learning to group similar objects into clusters based on their characteristics
- Hierarchical clustering can show cluster within clusters and much as in multiresolution visualization
- Creates a tree-like structure called a dendrogram, which illustrates the arrangement of the clusters formed at various levels of similarity
- Its ability to create a hierarchy of clusters allows for a complex qualities understanding of the relationships within the data

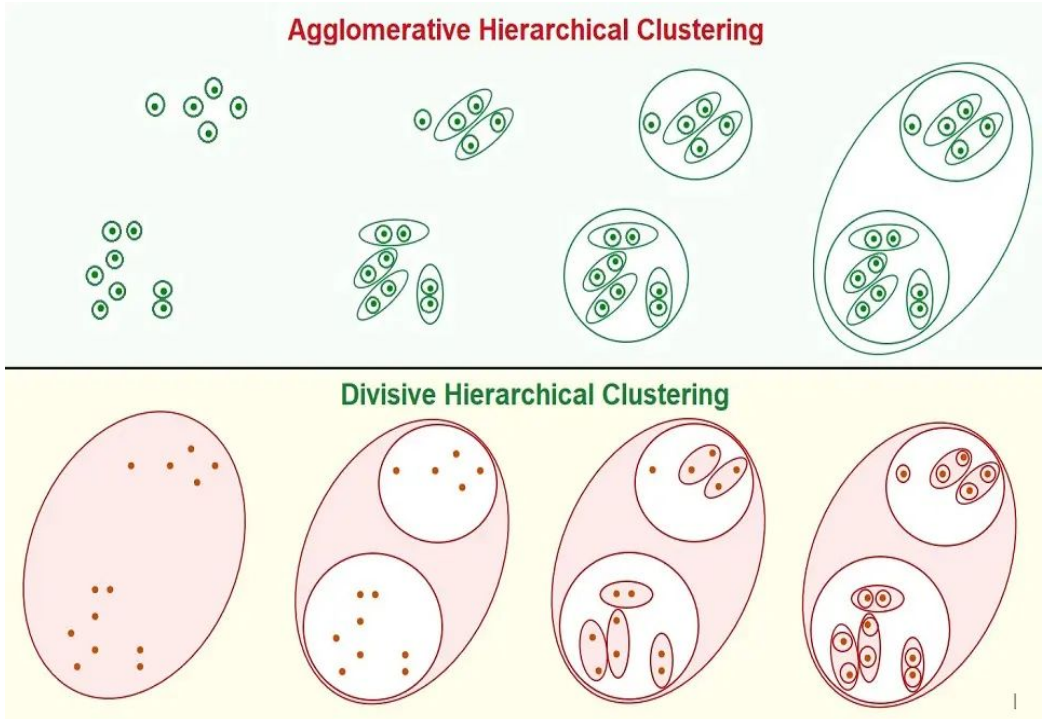


https://www.google.com/url?sa=i&url=https%3A%2F%2Fmedium.com%2F%40sachinsoni600517%2Fmastering-hierarchical-clustering-from-basic-to-advanced-5e770260bf93&psig=AOvVaw1y4RFO29lbVpUBuUbmwFg0&ust=1730448167547000&source=images&cd=vfe&opi=89978449&ved=0CBQQjRxqFwoTCPDS1YGUuIkDFQAAAAAdAAAAABAp

Hierarchical Clustering: Conti....

- Two types:
  1. **Agglomerative (Bottom-up)** -
     Start with individual samples, merge clusters
  2. **Divisive (Top-down)** -
     Start with all samples in one cluster, then split clusters

- **Key Advantages**:
  1. Visualizes "**clusters within clusters**" for deeper insights
  2. Allows for flexible, **multilevel exploration** of data
  3. Allows user to observe the cluster according to DFS/BFS transversal order



Illustrates the arrangement of the clusters produced by the corresponding analyses.

https://www.ejable.com/tech-corner/ai-machine-learning-and-deep-learning/hierarchical-clustering-in-machine-learning/

Hierarchical Clustering: Top-Down vs Bottom-Up

Top-Down (Divisive) Approach:

- Starts with one large cluster containing all samples
- Successively splits the cluster into smaller clusters
- Continues splitting until each sample is in own cluster or certain criteria are met
- **Example**: Hierarchical k-Means (used in the paper)

Bottom-Up (Agglomerative) Approach:

- Starts with each sample as its own cluster
- Merges clusters based on similarity until only one cluster remains
- Builds clusters from finer to coarser levels
- **Example**: Hierarchical Quality Threshold, Random Walks

The correlation matrix and distance measure can be used to cluster the samples in a hierarchical manner

Correlation:

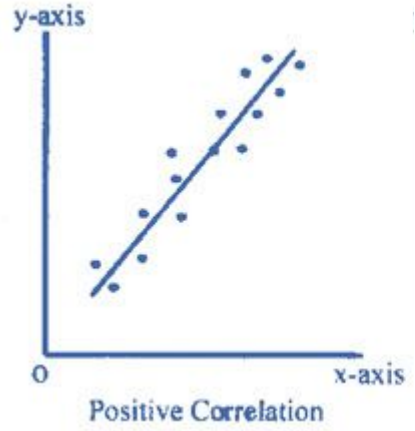
- **Pearson correlation coefficient** measures the linear correlation between two time series

$$\rho_{XY} = \frac{1}{T}\sum_{t=1}^{T}\left(\frac{X_t - \mu_X}{\sigma_X}\right)\left(\frac{Y_t - \mu_Y}{\sigma_Y}\right)$$

where, ρ_{XY} - Correlation value between X and Y

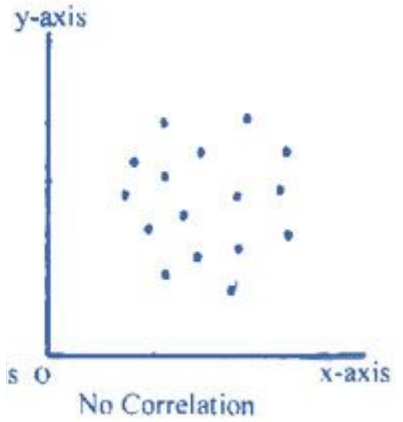
T - The number of time stamps

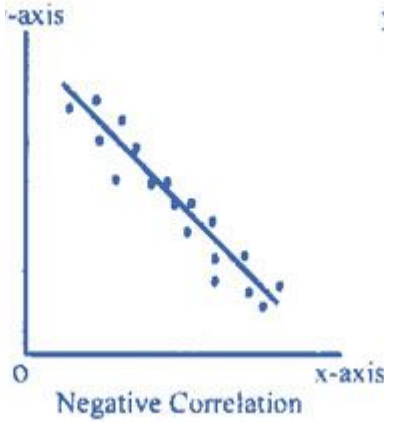
σ_X, μ_X - Standard deviation and mean of variable X



https://www.scribbr.com/statistics/pearson-correlation-coefficient/

- Used to assess the relationship between two variables, like when ρ_{XY} is
  - **1**: Perfect positive linear relationship
  - **-1**: Perfect negative linear relationship
  - **0**: No linear relationship





- *M* is known as correlation matrix where ***M(i,j)*** is $\rho_{X_i X_j}$

Distance Measure:

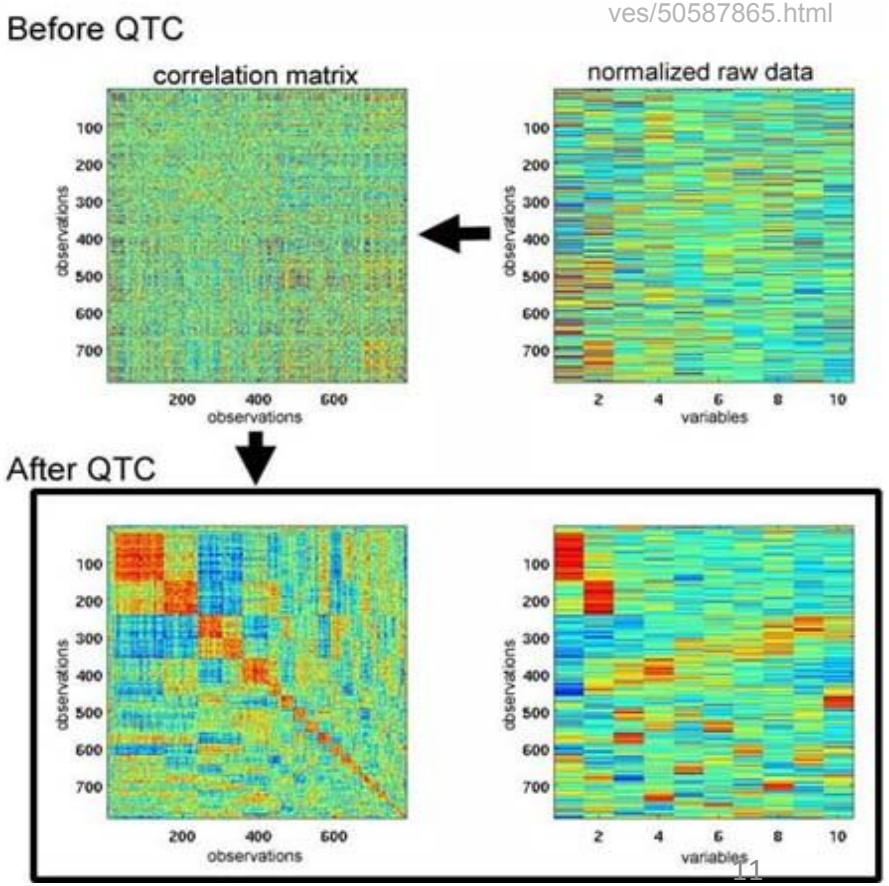
- **First distance measure (d_s):**
  - $d_s(X_i, X_j) = 1 - |\mathbf{M}_{i,j}|$
  - This distance indicates the strength of correlation between two variables
  - Ranges from 0 to 1:
    i. Closer to **0** means high correlation
    ii. Closer to **1** means low correlation

- **Second distance measure (d_v) :**
  - $$d_v(X_i, X_j) = \sqrt{\sum_{k=1}^{N} \left(\mathbf{M}_{i,k} - \mathbf{M}_{j,k}\right)^2}$$
  - Considers correlations between one sample and all other samples, where N is number of samples, it is normalised from 0 to 1 for use
  - **More accurate** for clustering than simple distance

Hierarchical Quality Threshold (HQT) Clustering:

- **Type:** Bottom-up clustering
- **Method:** Utilizes distance thresholds to form clusters at various levels. Clusters are formed by iteratively merging samples that are closer than the specified threshold

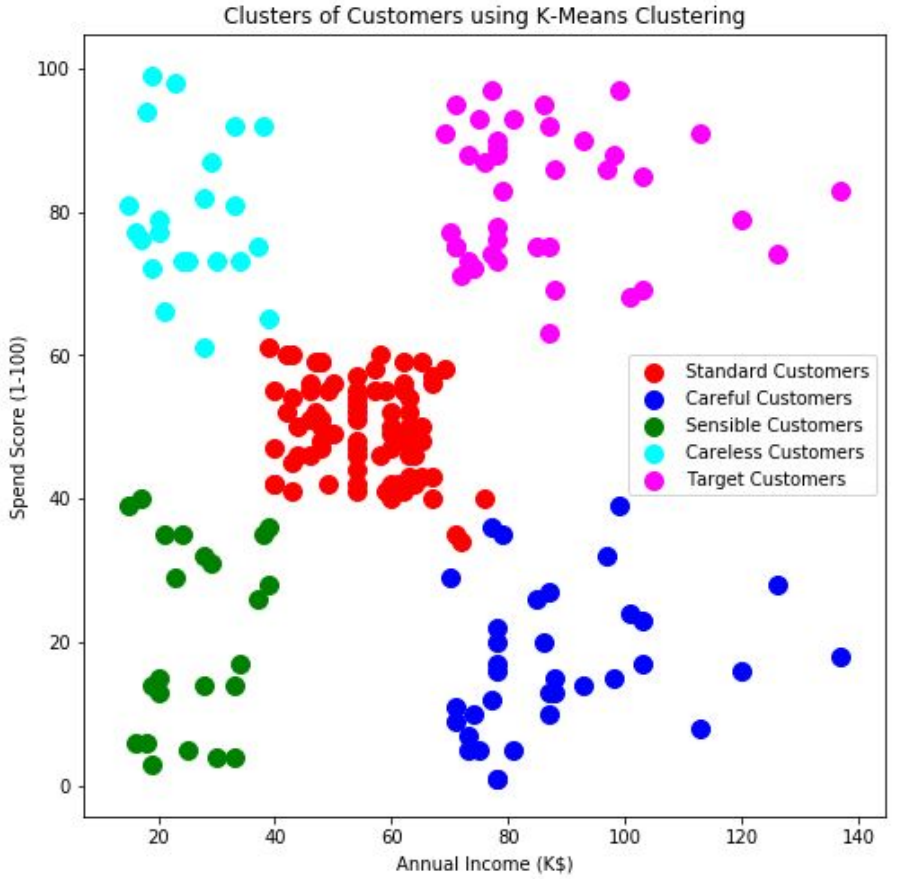
https://blog.livedoor.jp/brain_network/archives/50587865.html



- Define thresholds $\{\delta_0, ..., \delta_l\}$ with $\delta_0 = 0$, $\delta_l = 1$, and $\delta_i < \delta_j$ for $i < j$
- Form candidate clusters for each sample by including samples with distances smaller than the current threshold, and select the largest cluster
- Remove classified samples and repeat until all samples are classified or one cluster remains

Hierarchical K-Means Clustering:

- **Type:** Top-down clustering
- **Method:**
  - Start with **k** clusters (k < number of points) and assign a centroid for each and points are re-assigned to the nearest centroid, forming new clusters, and centroids are recalculated
  - Repeat reassignment and recalculation until convergence
  - Use the previous iterations' clusters as input for the next k-means iteration, building a hierarchy, and stop when a set number of levels is reached or average distortion falls below a threshold



https://www.researchgate.net/figure/Ejemplo-del-algoritmo-k-means-con-cinco-clusters_fig3_366370864

Hierarchical Random Walks Clustering:

- **Type:** Bottom-up clustering
- **Method:**
  - This algorithm considers the N ×N correlation matrix **M** as a fully connected graph where we treat $|\mathbf{M}_{ij}|$ as the weight for edge e_{ij} .
  - At each step, a walker moves from a vertex v_i to an adjacent vertex v_j, with the transition probability based on the weight $|\mathbf{M}_{ij}|$ of the edge connecting them
  - The probability that an adjacent vertex vj is chosen is defined as $\boldsymbol{P_{ij}} = |M_{ij}|/d_i$, where

$$d_i = \sum_{j=1}^{N} |\mathbf{M}_{ij}|$$

  - A random walk probability matrix $\mathbf{P^t}$ is created to track the probability of reaching v_j from v_i in t steps. With $\mathbf{P^t}$, we define the distance between v_i and v_j as:

$$r_{ij}^t = \sqrt{\sum_{k=1}^{N} \frac{\left(\mathbf{P}_{ik}^t - \mathbf{P}_{jk}^t\right)^2}{d_k}},$$

Hierarchical Random Walks Clustering: Conti....

- **Method:**
  - It start with every vertex in its own cluster. Then the algorithm iteratively merges two clusters with the minimum mean distance into a new cluster, and updates all the distances between clusters.
  - This process continues until we only have one single cluster left. The probability of going from a cluster C to vj in t steps is defined as-

$$\mathbf{P}_{Cj}^{t} = \frac{1}{|C|} \sum_{i \in C} \mathbf{P}_{ij}^{t},$$

  - The distance between two clusters C and D is defined as-

$$r_{CD}^{t} = \sqrt{\sum_{k=1}^{N} \frac{\left(\mathbf{P}_{Ck}^{t} - \mathbf{P}_{Dk}^{t}\right)^{2}}{d_k}}.$$

Data Set Overview: Tropical Oceanic Climate Data

Source - Tropical oceanic data simulated by NOAA's Geophysical Fluid Dynamics Laboratory (GFDL) using the CM2.1 global circulation model.

Purpose - Provides comprehensive climate measurements focused on the equatorial upper-ocean for in-depth correlation studies.

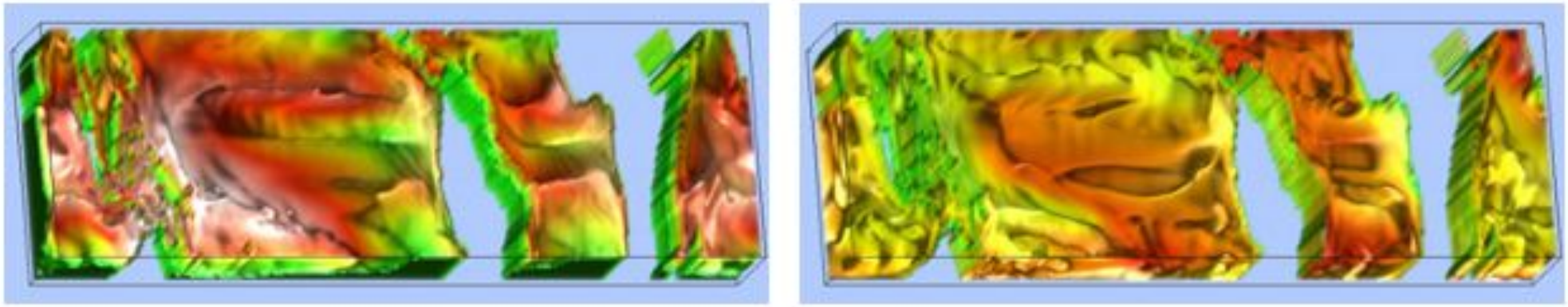
Temporal Coverage - 100 years, capturing long-term trends.

Temporal Resolution - 1,200 monthly time steps (100 years x 12 months)

Spatial Resolution - Organized in a 360 (longitude) x 66 (latitude) x 27 (depth) grid, covering:

- **Longitude**: Complete range of tropical ocean, represented by x-axis
- **Latitude**: From 20° South to 20° North (66 grid points), represented by y-axis
- **Depth**: From 0 to 300 meters (27 layers), represented by z-axis

Snapshot of Temporal and Salinity Fields in the First Time Step



(a) temperature field (b) salinity field

Fig. 1. Snapshots of the temperature and salinity fields at the first time step. Green, yellow, and red are for low, medium, and high scalar values, respectively.

- Temperature Field (a) : Displays the spatial distribution of temperature in the tropical ocean
- Salinity Field (b) : Shows the spatial distribution of salinity
- Green represents low values
- Yellow represents medium values
- Red represents high values

Spatial Sampling Strategy: Focusing on Key Oceanic Regions

Domain-Driven Sampling: Following guidance from NOAA scientists, we tailored our spatial sampling to focus on high-impact oceanic regions.

Sampling Criteria:

- **Exclude Continental Voxels**: Only oceanic regions are considered, as the focus is on oceanic climate data
- **Equatorial Emphasis**: A denser sampling grid is used near the equator, where climate interactions are particularly dynamic
- **Surface Prioritization**: Sampling density increases near the sea surface, which is crucial for capturing surface-level climate interactions

Spatial Sampling Strategy: Focusing on Key Oceanic Regions

Sampling criteria:

- **Non-Uniform Sampling Design**:
  a. **Gaussian Function for Latitude**: Applied along the y-axis to prioritize sampling near the equator
  b. **Exponential Function for Depth**: Applied along the z-axis, focusing sampling density closer to the sea surface
  c. **Rational**: This approach ensures high-resolution data in regions of interest while maintaining manageable data volume
- **Sample Sizes**: Two sample sets of 500 and 2000 voxels were selected for comparison in correlation clustering

Temporal Sampling Strategy: Reducing Volume, Retaining Trends

To reduce computational load, we applied a systematic temporal sampling approach, selecting a subset of time steps

Temporal Stroding:

- **Consistent Seasonal Representation**: Starting with the first time step, we selected every 12th step (i.e., the same month each year) to capture seasonal patterns across 100 years
- **Outcome**: This selection yielded **100 time steps**, balancing detail with computational efficiency

Advantages of Stroding:

- **Seasonal and Inter-Annual Variation**: By selecting the same month each year, we retain essential seasonal and annual patterns while reducing redundant data
- **Efficient Correlation Matrix Calculation**: The reduced set of 100 time steps is sufficient for robust correlation analysis without overwhelming data volume

Evaluation Approach

- **Objective** - To fairly assess the effectiveness of different hierarchical clustering algorithms
- **Consistency in Clustering** - For a fair comparison, each method generates a similar number of clusters
- **Direct Comparison**:
  - **Side-by-Side Volume Comparison**: Clusters are compared visually in the volume space
  - **Limitation**: Visual comparison is subjective and may lack precision
- **Complementary Quantitative Analysis**:
  - **Silhouette Plot**: Provides an objective measure of clustering quality
  - **Purpose**: Offers insights beyond visual comparisons, quantifying how well each data point is clustered

Silhouette Plot Analysis

Silhouette Score: Measures each point's fit within its cluster

- **Score Calculation**:
  - Calculate each point's **average similarity** a_i within its cluster (C)
  - Then for any other cluster C_j , we calculate p_i 's avg similarity d_{ij} with all count in C_j
  - **Minimum** b_i of all d_{ij} for p and corresponding cluster C_k (second best cluster for p_i)
  - **Silhouette Value** $s_i = \frac{b_i - a_i}{\max(a_i, b_i)}$, ranging from -1 to 1
- **Interpretation**:
  - $s_i \approx 1$: Strongly clustered; $s_i \approx -1$: Poorly clustered; $s_i \approx 0$: Boundary case

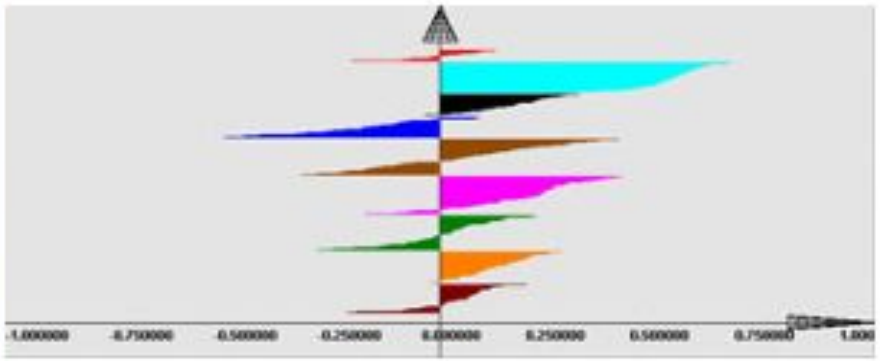
Evaluating Overall Clustering:

- **High Mean Silhouette Values**: Indicate effective clustering
- **Threshold**: If most scores are below 0.25, clustering quality may be low due to data overlap or algorithm limitations

Visual Insight: Silhouette plots offer a clear, quantitative view of each algorithm's performance

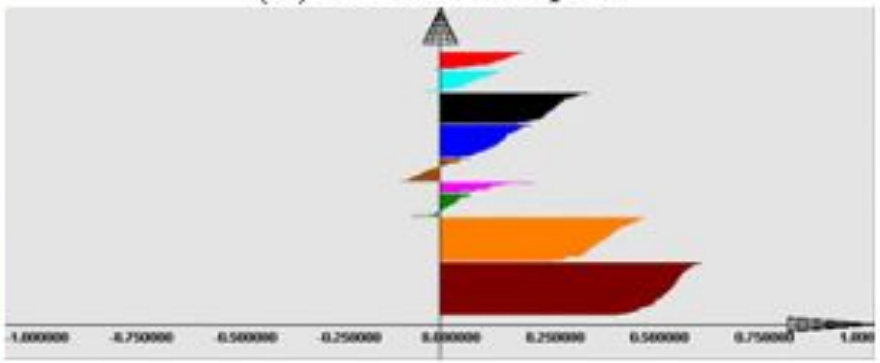
Distance Measure Comparison in Clustering Performance

(a) clustering with d_s



(b) silhouette plot

(c) clustering with d_v



(d) silhouette plot

Caption: Comparison of distance measures d_s and d_v with random walks. Both methods used 500 samples, creating nine clusters (shown in different colors). Silhouette plots in (b) and (d) indicate that d_v achieves better clustering quality than d_s .

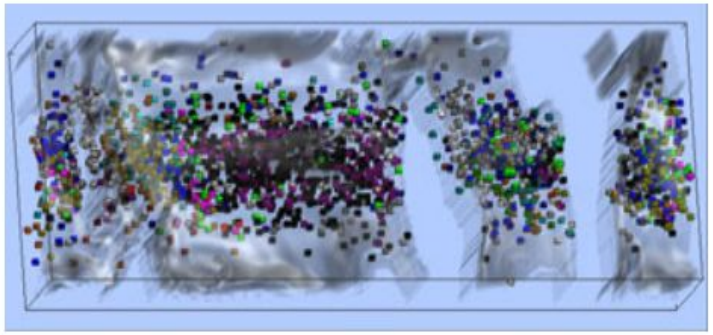
Silhouette Plot Analysis and Results

- **Silhouette Plot**: Provides a quantitative measure of clustering quality by indicating how well each sample is clustered.
- **Results**:
  - **With d_v**: 45.4% of samples have a silhouette value above 0.25, compared to only 22.4% with ds.
  - **Poorly Clustered Samples**: Only 8.8% of samples have a silhouette value below 0.0 with d_v, while 26.8% are poorly clustered with d_s.
- **Conclusion**: The silhouette plot clearly shows d_v yields better clustering quality.

Reason for dv Superiority:

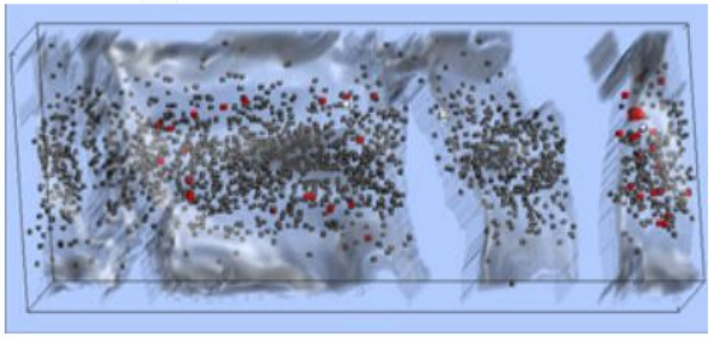
- **Broader Context**: Unlike d_s, d_v considers correlations with all samples, capturing a more comprehensive similarity structure, which results in higher clustering quality.
- **Application**: Based on these results, d_v was chosen as the distance measure in analyses. uent

Level of Detail Exploration in Hierarchical Clustering

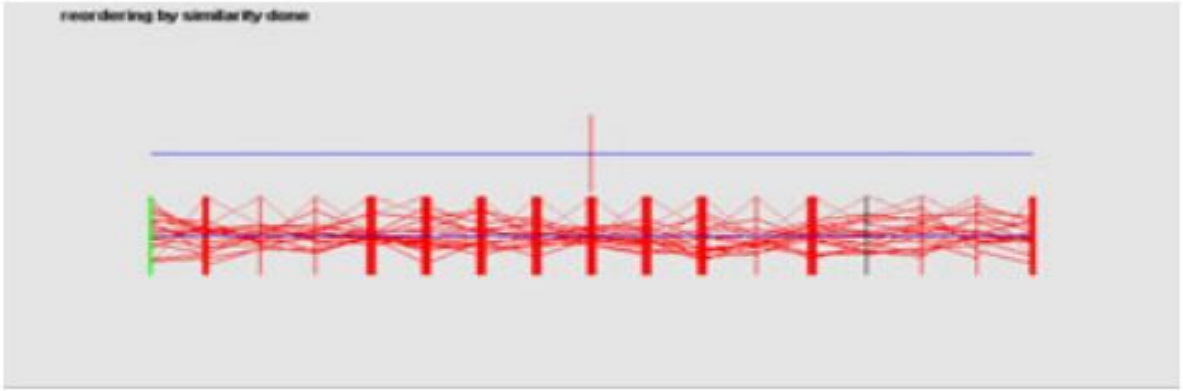


(a) coarse level-of-detail

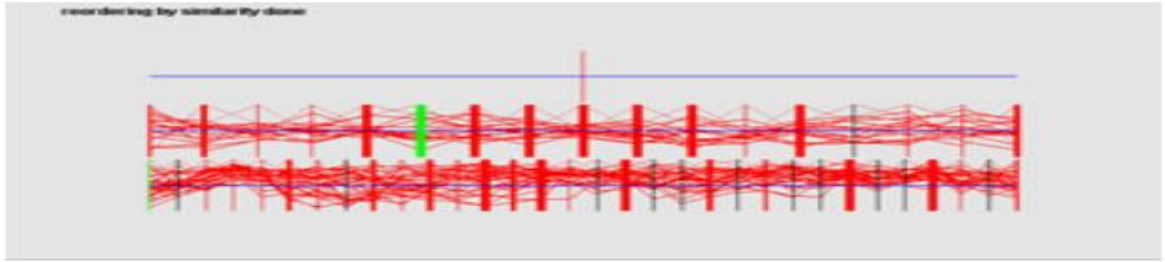
(c) medium level-of-detail



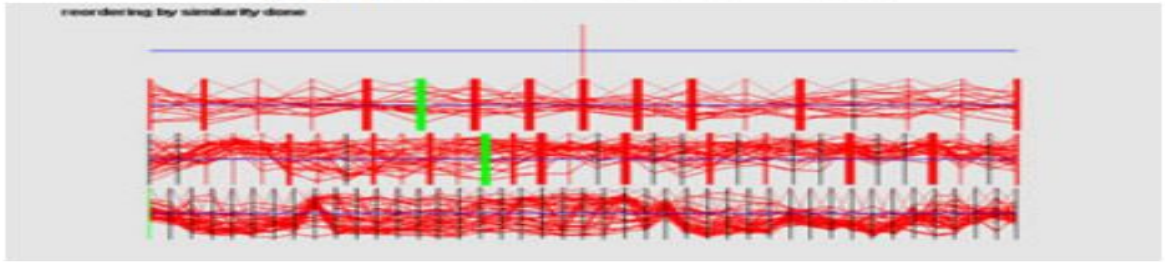
(e) fine level-of-detail



(b) parallel coordinates



(d) parallel coordinates



(f) parallel coordinates

Caption: Level-of-detail exploration of 2000 samples **using hierarchical quality threshold.**

(a) shows all samples at a coarse level, (c) highlights only the current level, and (e) de-emphasizes other samples in gray.

Parallel coordinates display correlation relationships, with axes reordered by similarity, each representing a sample.

Role of Parallel Coordinates in Correlation Exploration

Parallel Coordinates: A tool for visualizing quantitative relationships within clusters.

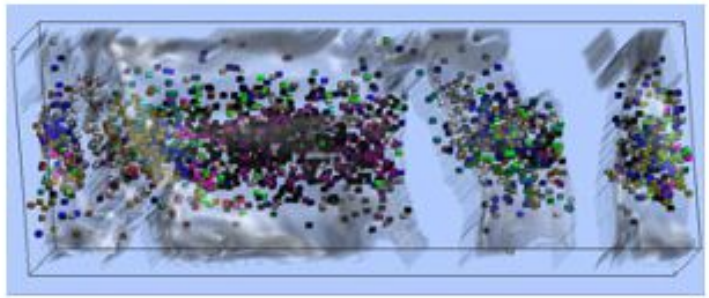
Working-

- **Axes Representation**: Each axis corresponds to a sample in the current level of detail, with the number of axes equaling the number of samples.
- **Axis Thickness**: Reflects the number of samples in the next hierarchical level, providing visual hints for user interaction.
- **Sorted by Similarity**: Axes are reordered by correlation similarity, enhancing visibility of patterns within clusters.

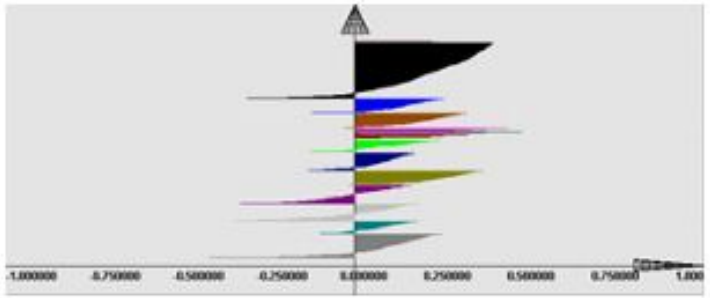
Linked Views:

- **Volume and Parallel Coordinates**: Samples along the path from root to current level are highlighted in white (volume view) and green (parallel coordinates), offering an intuitive, coordinated exploration.
- **Interactive Exploration**: Linking both views gives users control over the depth and perspective of their analysis.

Clustering Algorithm Comparison

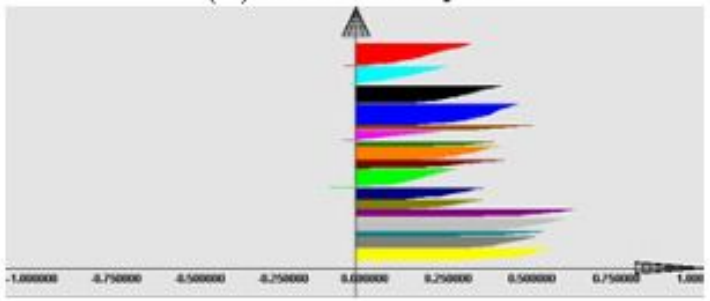


(a) hierarchical quality threshold



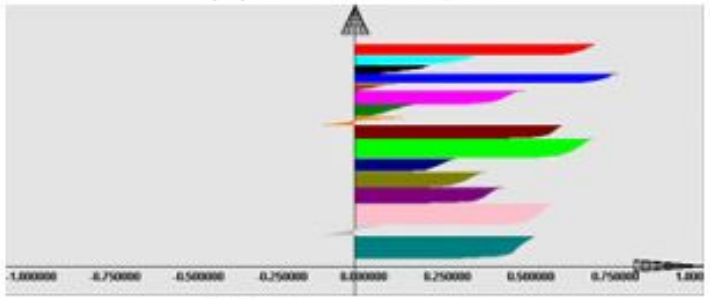
(b) silhouette plot

(c) hierarchical k-means



(d) silhouette plot

(e) random walks



(f) silhouette plot

Silhouette Plot Analysis:

- **HQT**: Performs the worst; many samples have low silhouette scores, indicating poor clustering quality.
- **K-means**: Achieves moderate clustering quality but requires longer computation time.
- **Random Walks**: Outperforms the others with more samples having high silhouette values (>0.5), providing better-defined clusters and faster computation.

Clustering Algorithm Analysis

	HQT	K-means	Random Walks
Advantages	• Offers flexibility in clustering, allowing for different levels of granularity based on threshold adjustments	• Allows for efficient clustering by leveraging the k-means framework within a hierarchical structure.	• Fast and parameter-free, making it suitable for large datasets without the need for extensive parameter tuning.
Disadvantages	• Sensitive to parameter selection, particularly the choice of thresholds, which can affect the final clustering results.	• Requires the number of clusters to be defined in advance, which can lead to suboptimal clustering if k is not chosen appropriately. It can also be computationally intensive.	• The interpretation of the graph structure may be less intuitive, and the effectiveness can depend on the quality of the correlation measures used.

Clustering Algorithm Comparison

Hierarchical Quality Threshold (HQT):

- Performs worst in clustering quality.
- High sensitivity to parameters (levels and thresholds) leads to unstable results.
- Limited appeal due to inconsistency in output.

Hierarchical K-means:

- Achieves fairly good clustering quality.
- Requires significantly more computation time than Random Walks.
- Needs pre-defined parameters (number of clusters), making it less flexible.

Random Walks:

- Best trade-off between quality and computational efficiency.
- Higher percentage of well-clustered samples (silhouette values > 0.5).
- No parameter requirements (e.g., thresholds), offering simplicity and flexibility.
- Consistently stable and fast, making it the preferred choice.

Clustering Algorithm Comparison

	quality threshold	k-means	random walks
strategy	agglomerative	divisive	agglomerative
parameters	# levels threshold for each level	# initial clusters termination threshold	none
randomness	no	yes	yes
tree style	general	general	binary
speed	unstable (184.4s, 22.2s)	slow (673.9s, 30.1s)	fast (188.4s, 4.5s)
quality	bad (22.0%, 67.6%)	good (65.1%, 60.8%)	good (72.3%, 45.4%)
stability	unstable	stable	stable

Caption - Comparison of three hierarchical clustering algorithms. The two timings (percentages) in the speed (quality) entry are for the clustering time in second on an AMD Athlon dual-core 1.05 GHz laptop CPU (samples with silhouette value larger than 0.25) with 2000 and 500 samples, respectively.

Conclusion

Study Overview:

This study examined hierarchical correlation clustering for analyzing time-varying, multivariate climate data.

Data Selection:

Climate samples were selected based on domain knowledge, prioritizing important spatial and temporal regions.

Analytical Tools:

- **Parallel Coordinates**:

  Used to visualize quantitative correlation information across clusters.
- **Silhouette Plots**:

  Employed to assess and compare clustering quality objectively.

Conclusion

Algorithm Comparison:

- **Finding**:

  Among the three clustering methods tested (Hierarchical Quality Threshold, K-means, and Random Walks), **Random Walks** emerged as the best in terms of both clustering quality and computational efficiency.

Key Insight:

Our methodology offers a structured approach for understanding correlations in complex, multivariate climate datasets, highlighting Random Walks as an optimal choice for clustering

Future work:

- Evaluating the approach and results with the domain scientists
- Investigating the uncertainty/error introduced in the sampling in terms of the clustering accuracy

Thank You