Coarse-Grain Cluster Analysis of Tensors
With Application to Climate Biome Identification

DeSantis, Derek - CNLS, T3 - ddesantis@lanl.gov
Wolfram, Phil - T3 - pwolfram@lanl.gov
Alexandrov, Boian - T1 - boian@lanl.gov

Los Alamos NATIONAL LABORATORY EST.1943

Background

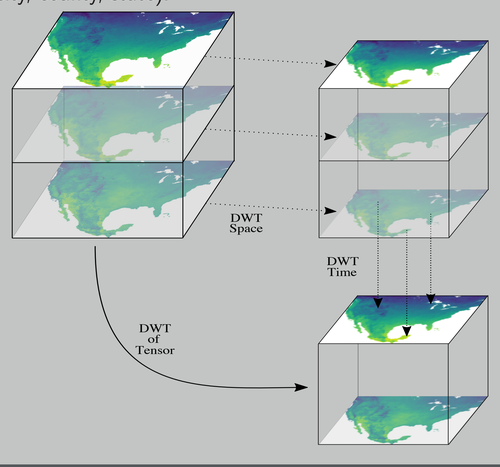
- Many data structures come in the form of tensors. Frequently, one is interested in the relationship between particular slices of the tensor.
- Clustering data along a chosen slice provides a method to parse the complexity-complete data into digestible bits of information.
- It is well known that the results of clustering methods are heavily dependent on the algorithm of choice, as well as the chosen hyperparameters of the algorithm. What isn't understood is the further dependence on "hidden parameters" such as the scale of the data.

Application - Climate Classification

- Classifying the surface of the Earth into climate biomes is a way to parse the complex micro and macroscopic interdependencies down to provide meaningful diagnostics that more readily relate to physical and biological systems.
- It is important to analyze both short and long term trends, as well as small and large scale features.
- Classification of the climate is sensitive to spatial/temporal coarse-graining.

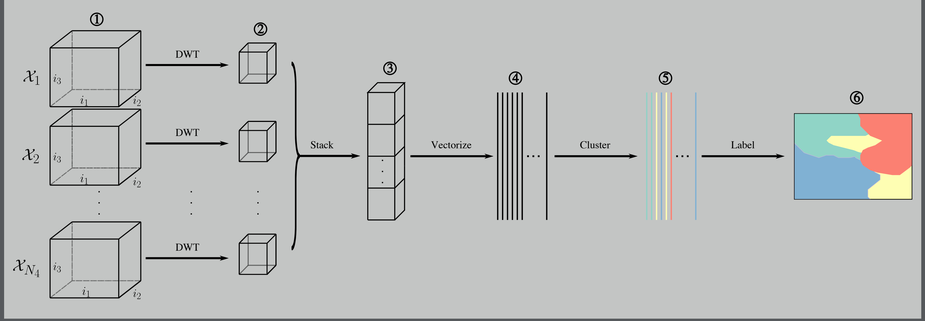
Tool - Discrete Wavelet Transform (DWT)

- The DWT splits a signal into high and low frequency.
- Low temporal signal captures climatology (seasons, years, decades), while low spatial signal captures regional features (city, county, state).
- Each application of DWT reduces size of data by factor of 2 (per dimension).
- DWT for tensors is computed via repeated 1D DWT on the low frequency signal along each axis.



Algorithm - Coarse Grain Clustering

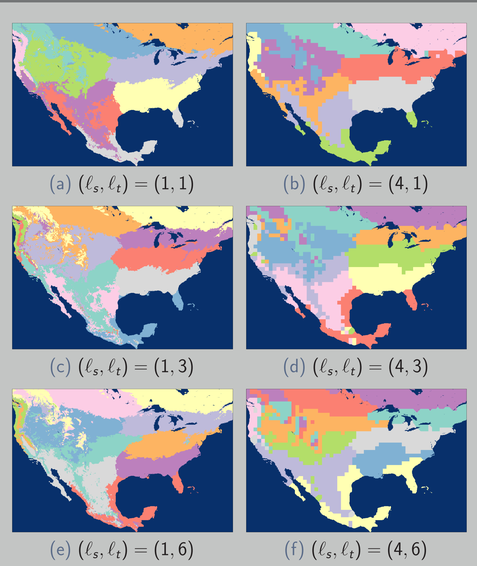
1. Separate the tensor into 3-way tensors of a single variable (e.g. temperature data).
2. Pick integers l_s, l_t, and wavelets w_s for space and w_t for time and for each tensor of data, take l_s DWT in space, and l_t DWT in time.
3. Stack the new tensors for classification.
4. Vectorize the stack of wavelet coefficients.
5. Cluster the vectors using a chosen clustering algorithm (e.g. K-means).
6. Map the clustering on the DWT back to the initial latitude-longitude points.



Proof of Concept - LOCA Statistical Downscaled climate data

- Gridded climate data set of North America.
- Grid cell is monthly data from 1950-2013, six kilometers across.
- Available variables used: precipitation, maximum temperature, minimum temperature.
- Haar wavelets for space, db2 in time.
- Max spatial scale $l_s = 4$ ($\sim$ 100 km) , max temporal scale $l_t = 6$ ($\sim$ 5 years).
- Clustering algorithm used: K-means.

Results - Effect of Coarse-Graining



(a) $(\ell_s, \ell_t) = (1, 1)$ (b) $(\ell_s, \ell_t) = (4, 1)$
(c) $(\ell_s, \ell_t) = (1, 3)$ (d) $(\ell_s, \ell_t) = (4, 3)$
(e) $(\ell_s, \ell_t) = (1, 6)$ (f) $(\ell_s, \ell_t) = (4, 6)$

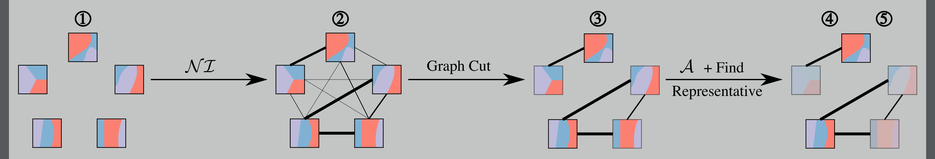
Figure: Clusterings obtained for $k = 10$ at resolution (ℓ_s, ℓ_t).

Tool - Mutual Information

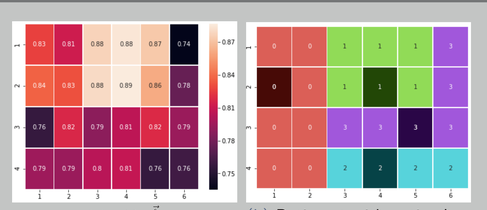
Given partitions of data $U = \{U_j\}_{j=1}^{k}$, $V = \{V_j\}_{j=1}^{l}$, the **Mutual Information** $\mathcal{NI}(U, V)$ measures how knowledge of one clustering reduces our uncertainty of the other.

Algorithm - Mutual Information Ensemble Reduce

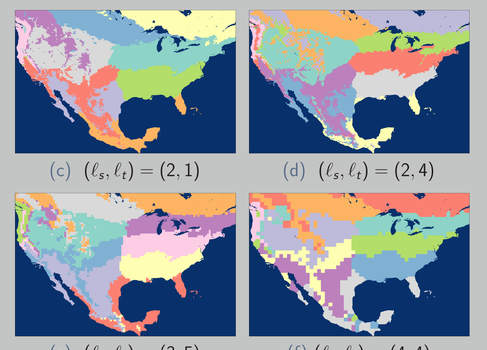
1. Cluster across range of resolutions.
2. Compute pairwise mutual information - creating the mutual information graph.
3. Perform a graph cut to find the most similar connected components.
4. Compute average mutual information between each cluster and all other memembers of its component.
5. In each component, select cluster with highest expected mutual information.



Results - Key Resolutions



(a) $\mathcal{A}(U^{\vec{\ell}})$ (b) Ratio-cut with optimal clusters highlighted



(c) $(\ell_s, \ell_t) = (2, 1)$ (d) $(\ell_s, \ell_t) = (2, 4)$
(e) $(\ell_s, \ell_t) = (3, 5)$ (f) $(\ell_s, \ell_t) = (4, 4)$

Figure: Output of the MIER algorithm for $k = 10$. a) $\mathcal{A}(U^{\vec{\ell}})$ for each resolution $\vec{\ell} \in \mathcal{L}$, b) results of the ratio cut algorithm and key resolutions, c-f) plots of key resolutions.

Discussion and Future Work

- Clustering the DWT illustrates additional structure within the data.
- Hence, an **optimal clustering should be an ensemble of the wavelet clusterings to best characterize the impact of spatio-temporal variability in the climate data.**

LA-UR-20-21692