Discovery of Climate Indices using Clustering

Antonios Makrymallis
Papanikolaou Amalia

University of Edinburgh
School of Informatics
Data Mining and Exploration

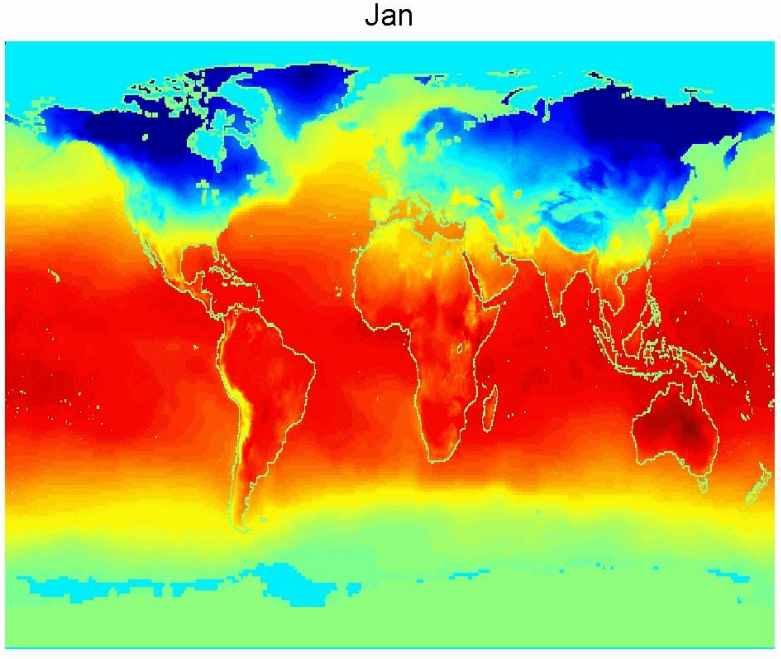
20 Feb 2009

Climate Teleconnections

Research goal

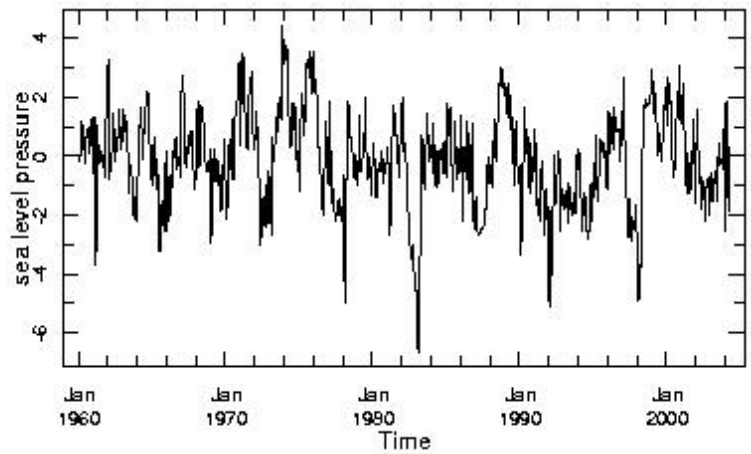
- Analysis of the effect of the oceans and atmosphere on land climate.
- Interest in explaning climate phenomena finding patterns.
- Use of Climate Indices.

Average Monthly Temperature



Climate Indices

- Time series based on sea level pressure (SLP) and sea surface temperature (SST) in ocean regions.
- They distill climate variability at a regional or global scale into a single time series.



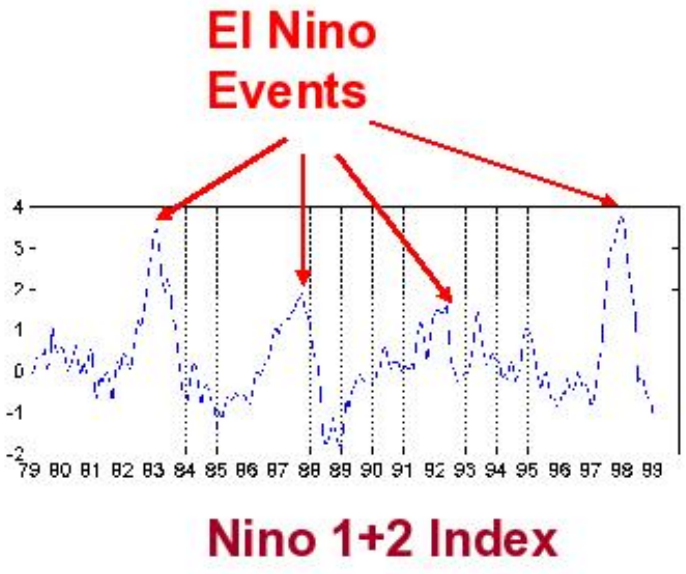
Example: NINO 1+2 index

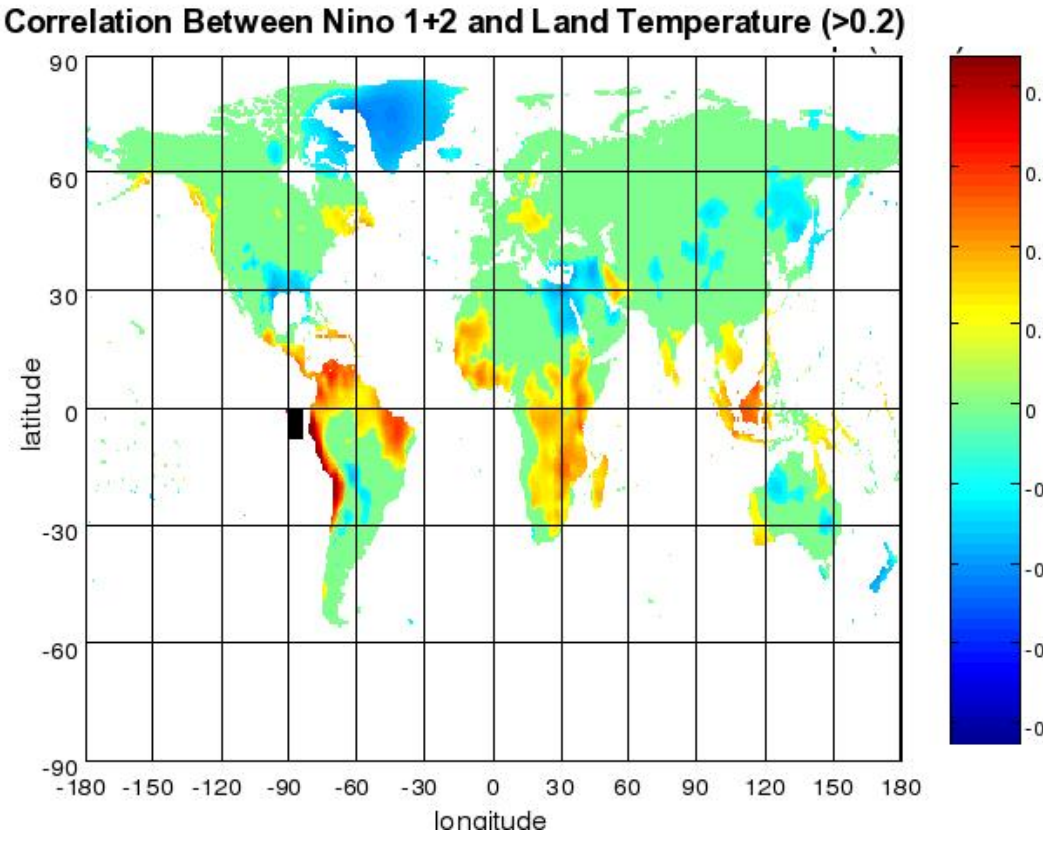
El Niño-Southern Oscillation

An important fluctuation in surface waters of the tropical Eastern Pacific ocean.
It is associated with floods, droughts and heavy rainfall in a range of locations around the world such as Australia and South America.

NINO 1+2: climate index that is related with El Niño phenomenon.

NINO 1+2 index correlation with land temperature

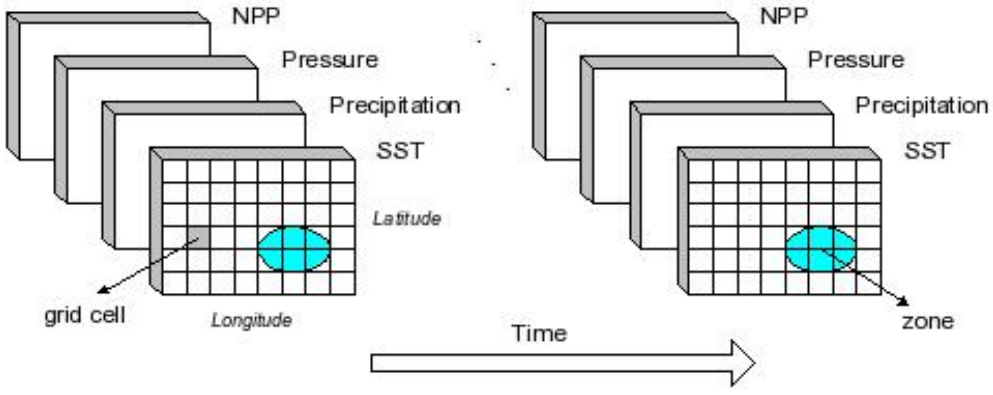




Unexpected correlations.

Data preprocessing

- Global snapshots of measurement values for a number of variables.
- Time series data could be noisy, have cycles and regularity or contain long term trends.



Monthly Z score transformation

- Removes seasonality and temporal autocorrelation.
- For a given month substract off the mean and divide by standard deviation.

Methods for discovering climate indices

Direct observation

El Niño was first noticed by a Peruvian fisherman!

Eigenvalue analysis techniques of data produced by satellite observations

Principal components analysis (PCA) / Singular value decomposition (SVD).

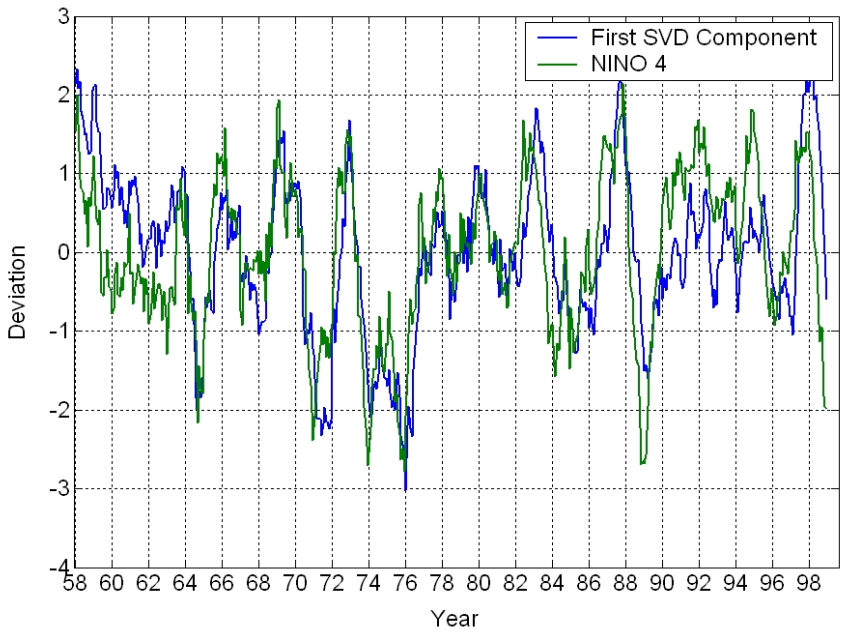
Clustering

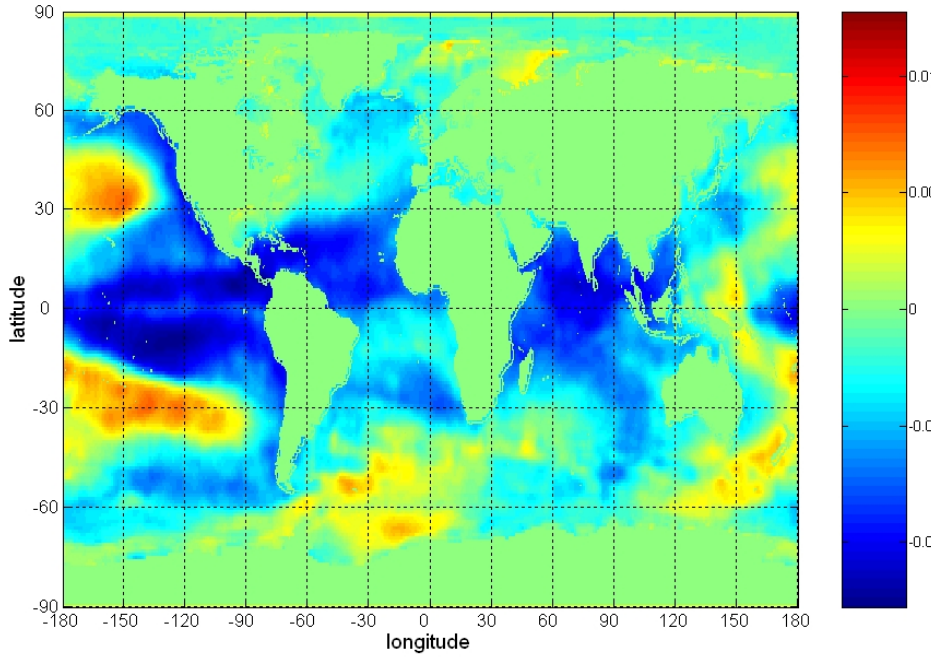
Cluster centroids.

SVD Analysis

- We have a data matrix whose rows consist of time series from various points on the globe.
- Singular Value Decomposition(SVD) or Empirical Orthogonal Functions(EOF) decomposes the matrix into:
  - a set of spatial patterns,
  - a set of temporal patterns.
- Singular values: the strongest patterns are associated with the largest singular values.

Application to the global Sea Surface Temperature(SST)





The strongest temporal pattern is highly related to the El Niño phenomenon.

Limitations of SVD based approaches

- Weaker signal may be masked by stronger signals.
  - Only the first few vectors are typically regarded as trustworthy.
- The discovered patterns are constrained to be orthogonal to each other.
- SVD finds patterns if they fall into independent subspaces.

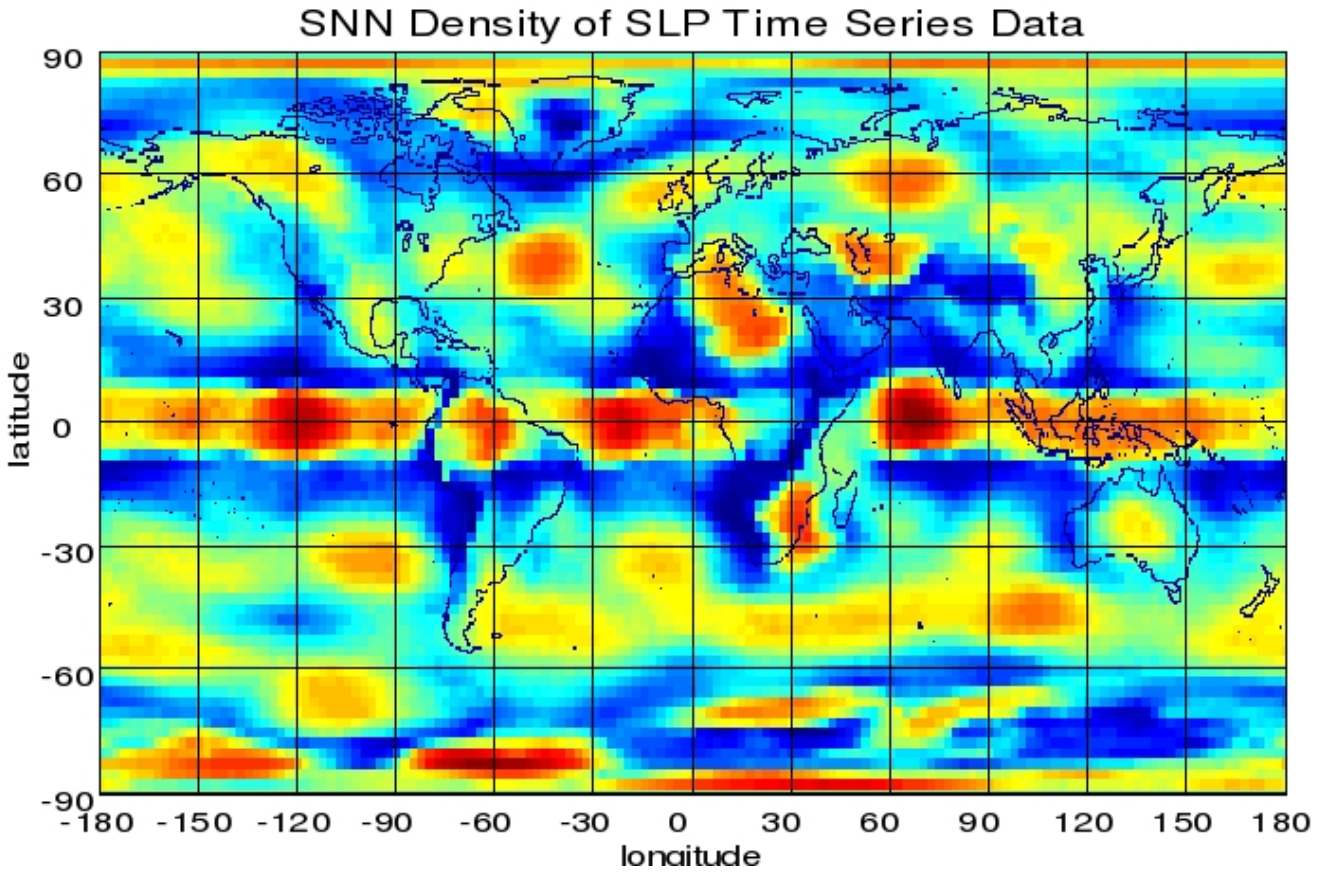
Clustering Based Methodology for the discovery of Indices

- Clustering provides an alternative approach for finding candidate indices.
- Clusters represent ocean regions with relatively homogeneous behavior.
- The centroids of these clusters are time series that summarize the behavior of these ocean areas, and thus, represent potential climate indices.
- Need to evaluate the influence of potential indices on land points.

Shared Nearest Neighbour(SNN) Clustering

- List the k nearest neighbours for every point of the data set.
- Define similarity between 2 points to the extend that their neighbour list is similar.
- Define Density for every point.
  - The value of SNN similarity of the kth nearest neighbour of the point.
  - The sum of the SNN similarities of a point's k nearest neighours.
- Perform the clustering using the density.
  - Identify and eliminate noise and outliers, which are points with low density.
  - Identify core points, which are time series with high density.
  - Build clusters around the core points.

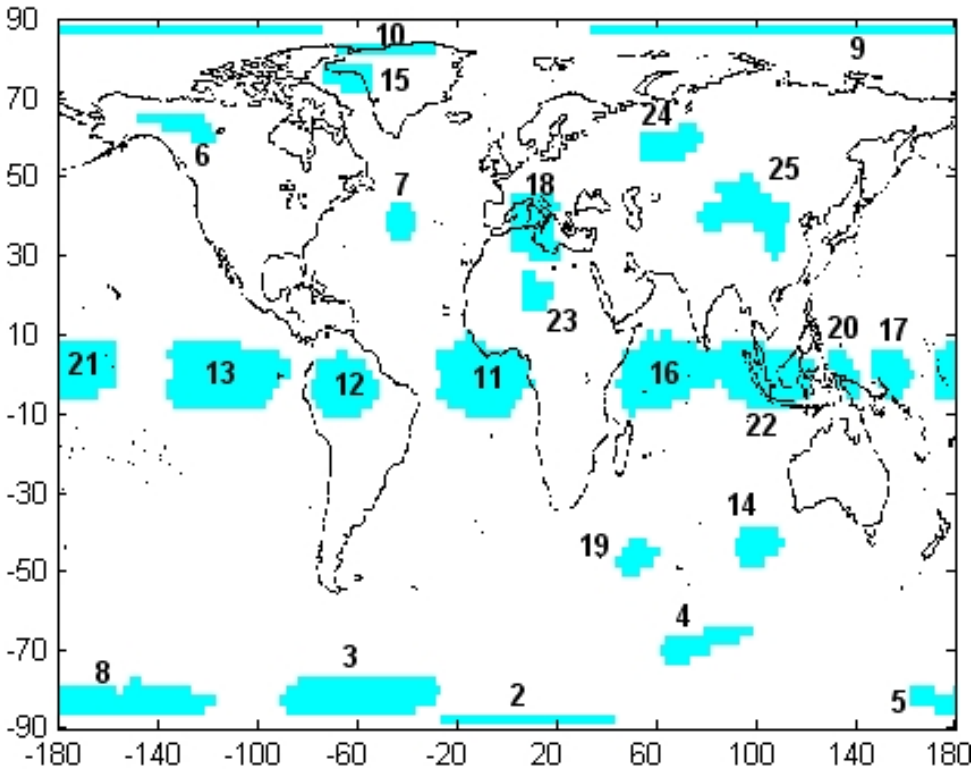
SNN Density of SLP Time Series



Redder areas are high density, i.e., high homogeneity.

SLP Clusters

25 SLP Clusters



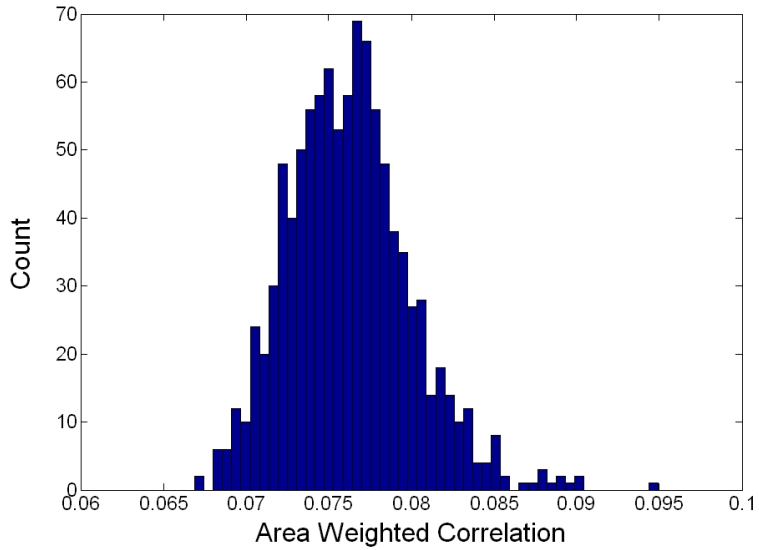
Influence of Climate Indices on Land

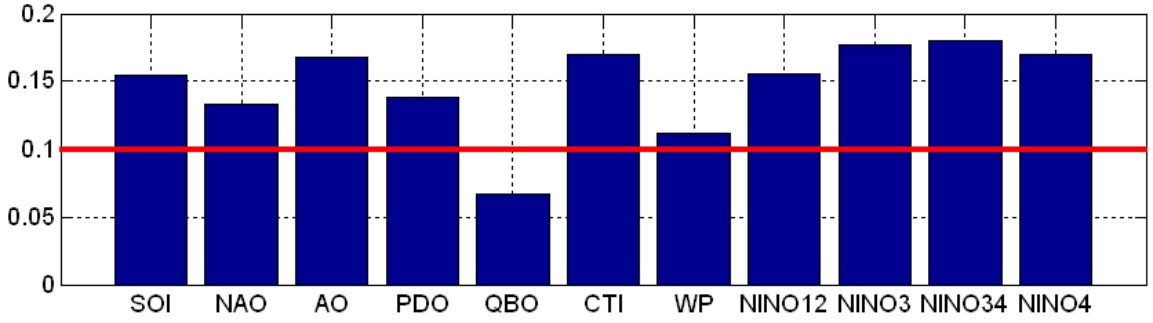
Area Weighted Correlation

- For each grid point, compute the correlation of the candidate climate index with a time series representing the temperature at that point.
- Then compute the weighted average of the absolute correlations of each land point.
- The weights are the areas of the grid points.

Baseline for Area Weighted Correlation

- Need to establish what level of area weighted correlation is significant.
  - Baseline based on correlation of random time series to land temperature.
  - Typical values of current indices.





Overview

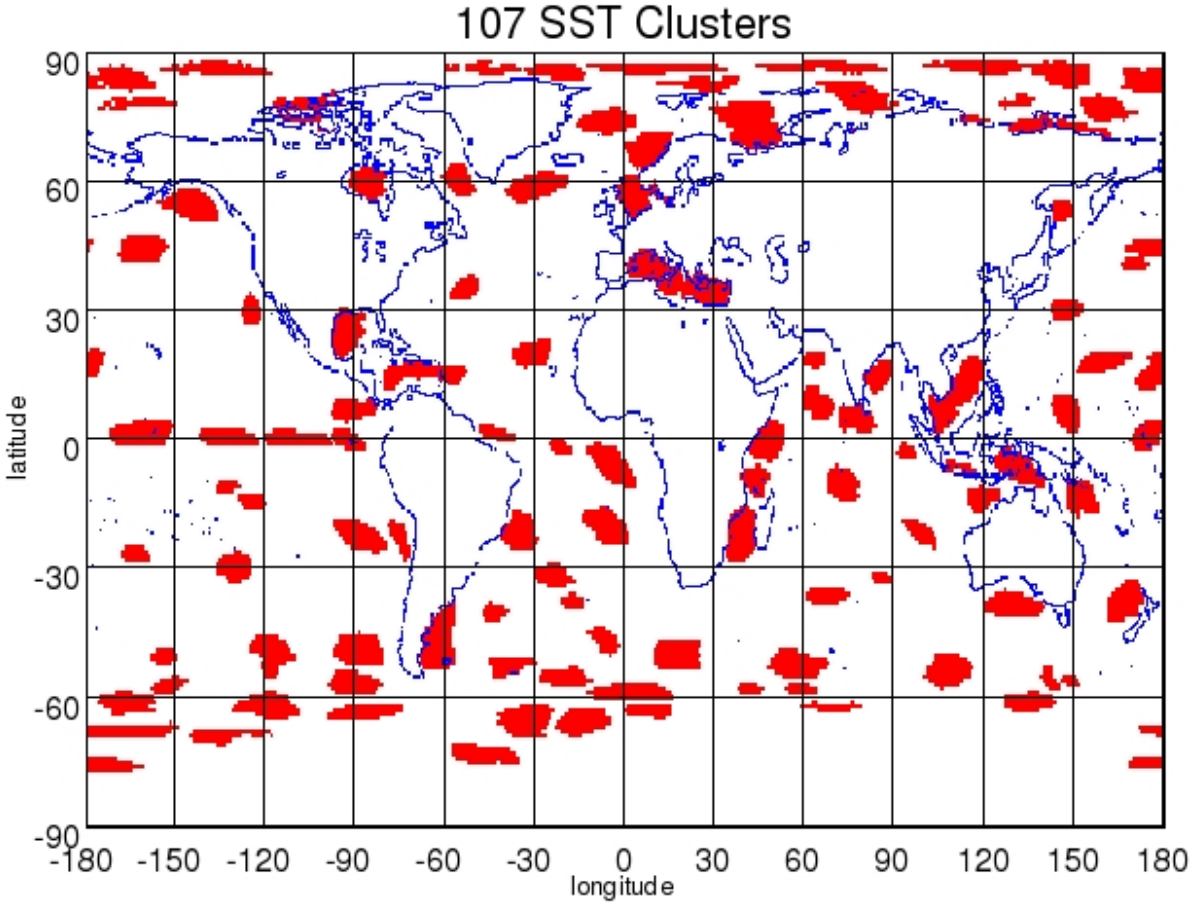
We have seen:

- Singular Value Decomposition and Clustering methods to Earth science data.

Next...

- Results
- Application of SSN to SST data.

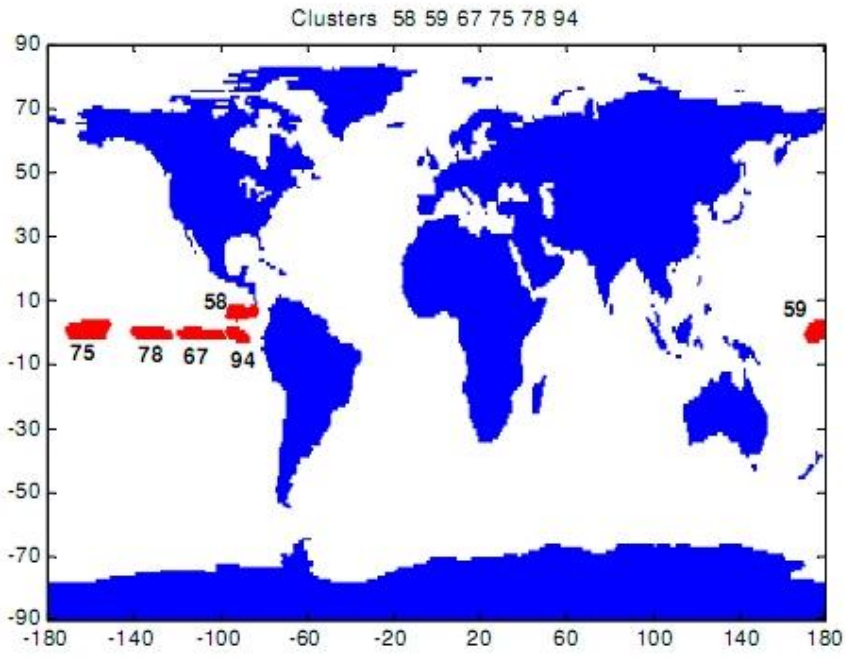
SST-based Candidate indices



- Elimination of all clusters with poor area-weighted correlation.
- The cluster centroids represent potential indices.

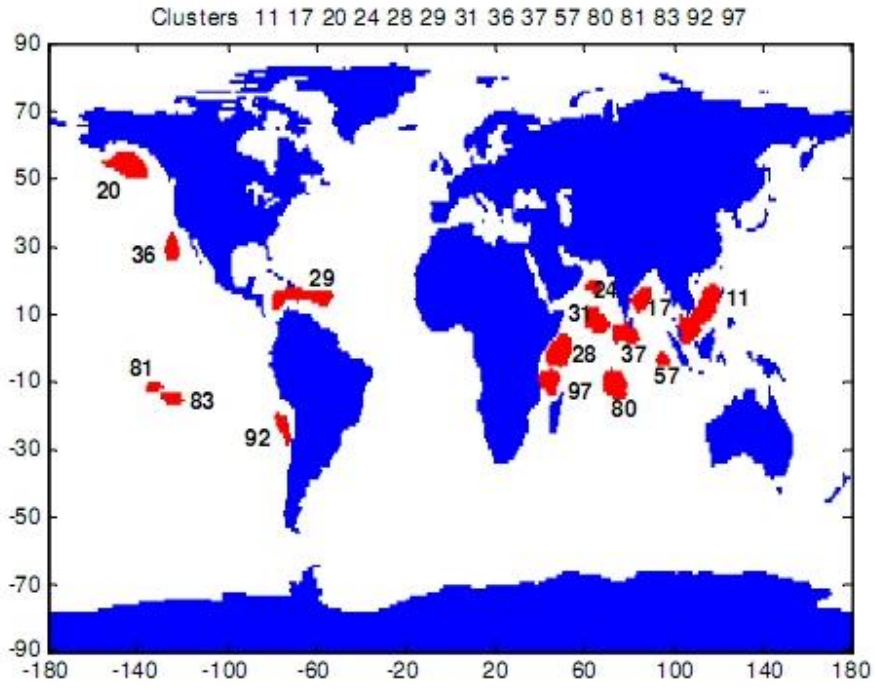
Clustering results

- Group G0: Cluster centroids highly correlated to known indices.
  Method validation.
- Group G1: Variants of known indices.
- Groups G2,G3: May represent new indices.



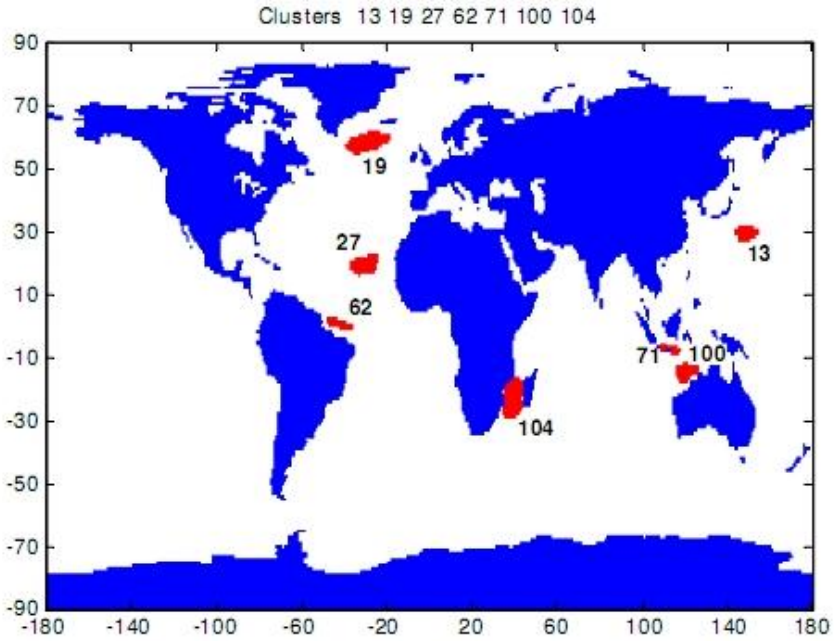
Clustering results

- Group G0: Cluster centroids highly correlated to known indices.
  Method validation.
- Group G1: Variants of known indices.
- Groups G2,G3: May represent new indices.



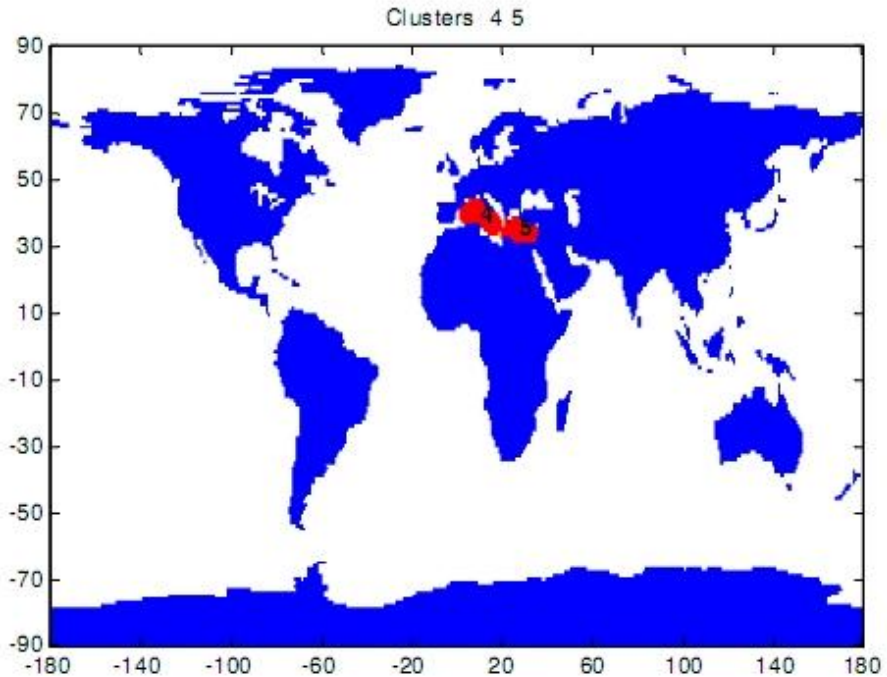
Clustering results

- Group G0: Cluster centroids highly correlated to known indices.
  Method validation.
- Group G1: Variants of known indices.
- Groups G2,G3: May represent new indices.



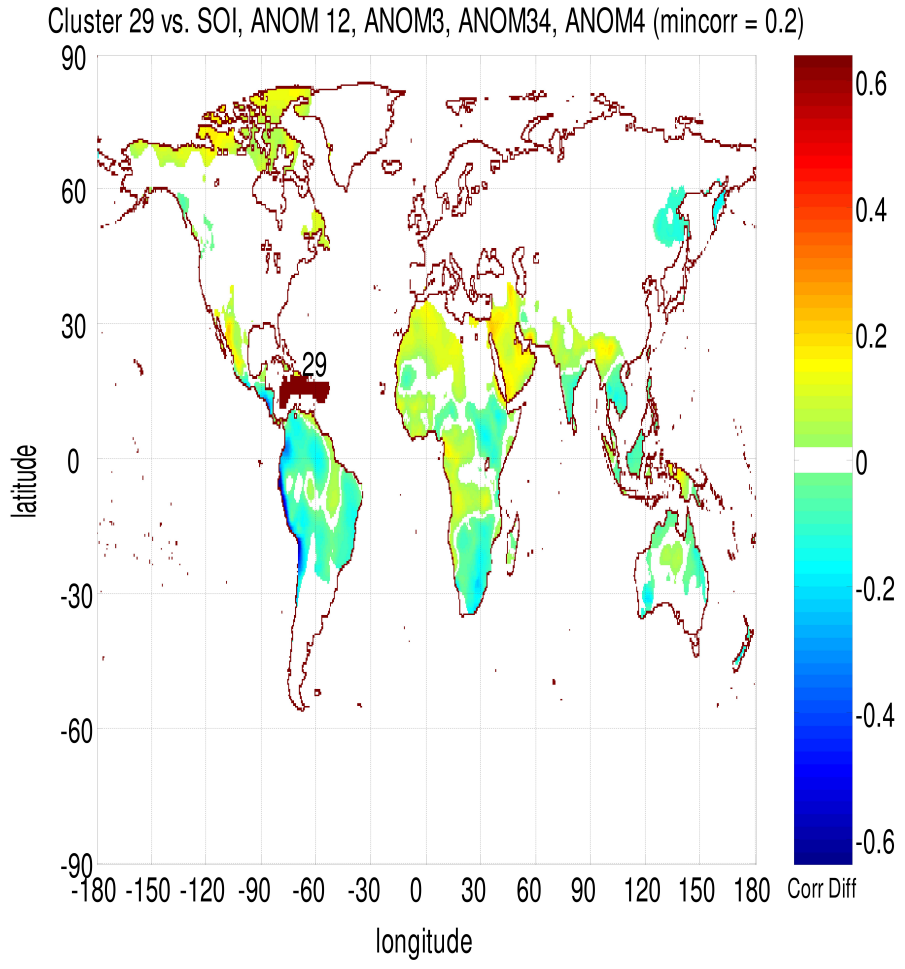
Clustering results

- Group G0: Cluster centroids highly correlated to known indices.
  Method validation.
- Group G1: Variants of known indices.
- Groups G2,G3: May represent new indices.



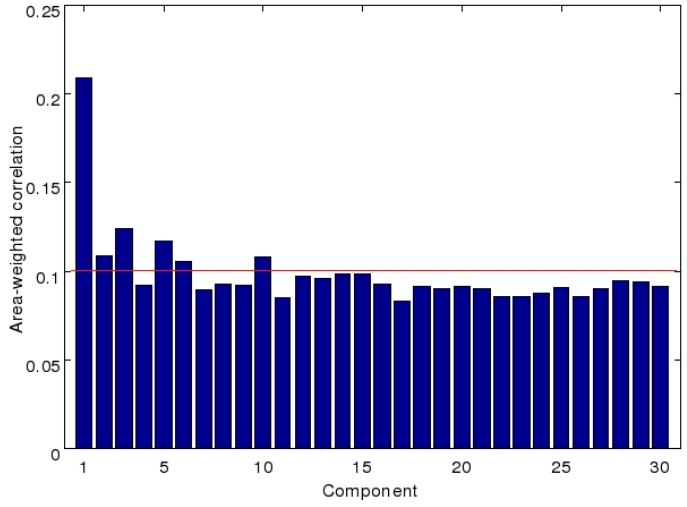
Clusters representing variants of known indices

- Some cluster centroids may provide better coverage for some areas of land.
- Color indicates the difference in correlation.
- There are areas where cluster outperforms the known indices.

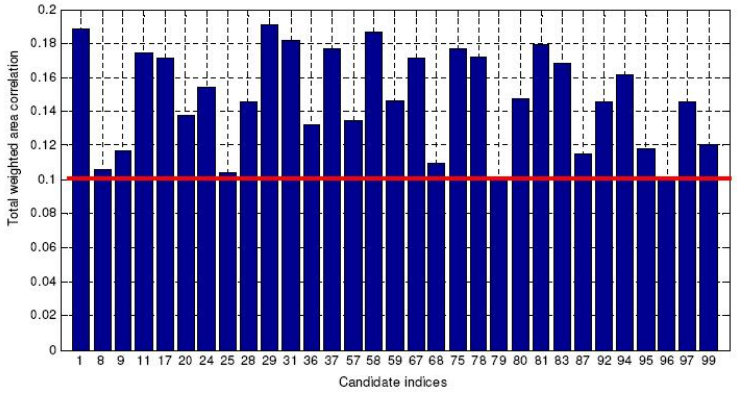


Results: SVD Vs SNN Approach

- Found the top 30 SVD components for SST



- Highest Cluster Centroids for SST



SVD Vs SNN Approach

Correlation of Known Indices with SST Cluster Centroids and SVD Components

Known Indices	Best Matching SST Cluster Centroid	Best Matching SVD Component
SOI	0.7006	0.5427
NAO	0.2973	0.1774
AO	0.2383	0.2301
PDO	0.5172	0.4684
QBO	0.2675	0.3187
CTI	0.9147	0.6316
WP	0.2590	0.1904
NINO1+2	0.9225	0.5419
NINO3	0.9462	0.6449
NINO3.4	0.9196	0.6844
NINO4	0.9165	0.6894

Red indicates higher magnitude of correlation.

SVD Vs SNN Approach

Area-weighted Correlation for known Indices with SST Cluster Centroids and SVD Components

Known Indices	Area Weighted Correlation for		
	Index	Best Centroid	Best SVD Component
SOI	0.1550	0.1768	0.1240
NAO	0.1328	0.1387	0.0929
AO	0.1682	0.1912	0.0929
PDO	0.1378	0.1377	0.0891
QBO	0.0671	0.1377	0.0850
CTI	0.1702	0.1708	0.1240
WP	0.1117	0.1714	0.1240
NINO1+2	0.1558	0.1608	0.2091
NINO3	0.1774	0.1708	0.2091
NINO 3.4	0.1800	0.1714	0.2091
NINO 4	0.1696	0.1768	0.2091

Red indicates higher correlation.

Conclusions

- Clustering is a viable alternative to eigenvalue based approaches for discovering climate indices.
- Centroids of clusters built using SNN replicate and others outperform many well-known climate indices.
- Some indices may represent new Earth Science phenomena.
- Cluster-derived Climate Indices have higher area weighted correlation than SVD-derived Indices, in most cases.
- No need to pre-select the area to analyze, automatically identify areas of interest

Future Work

- Investigation of candidate indices by Earth Scientists.
- Investigate whether there are climate indices that cannot be represented by clusters.
- Investigate several approaches to modeling dynamic clusters (*e.g. Given sensor readings for SLP at different points in the ocean, how to identify clusters of low/high pressure points that may move with space and time*)
- Noise elimination.
- Aggregation.
- An efficient high dimensional cluster method and its application in global climate sets, Ke Li, Fan Lin, Kunqing Xie, 2007

Bibliography

- Discovery of Climate Indices using Clustering by Michael Steinbach and Steven Klooster and Christopher Potter. In KDD 2003.
- Discovery of Patterns in the Global Climate System using Data Mining, Vipin Kumar, University of Minnesota

Questions

Thank you.

Makrymallis Antonios
Papanikolaou Amalia

University of Edinburgh
School of Informatics
Data Mining and Exploration.