Investigation of rotated PCA from the perspective of network communities applied to climate data

David Hartman Jaroslav Hlinka Martin Vejmelka Milan Paluš

Institute of Computer Science, Academy of Sciences, Prague, Czech Republic

This study was supported by the Czech Science Foundation project No. P103/11/J068.

Role of RPCA

Rotated principal component analysis (RPCA) or Empirical Orthogonal Functions (EOF) with varimax rotation have a long history in climatology [1, 2]. It can play a role within complex network analysis and namely community detection that both appeared in climate research relatively recently [7, 3]. Determination of optimal community structure is well known hard problem and there are several methods excelling in specific situations [4] and several ways of measuring quality of resulting community structure such as modularity [5]. A question under study is how RPCA results can be used either solely as community or as dimensionality reduction preprocessing for further community analysis. We use data from National Centers for Environmental Prediction–National Center for Atmospheric Research (NCEP-NCAR) Reanalysis [8], more specifically SAT and SLP.

Community structure

Roughly speaking, distribution of edges has tendency to concentrate within group of nodes. Let's have a graph $G = (V, E)$ and subset of vertices $C \subset V$. For C to be a community

$$\frac{|E(C)|}{|C|} \gg \frac{|E(V)|}{|V|} \quad \text{and} \quad \frac{|E^{-}(C)|}{|C|} \ll \frac{|E(V)|}{|V|}$$

$E(X)$ denotes edges induced by vertex set $X \subseteq V$ and $E^{-}(C)$ number of outgoing edges from vertices of C.

Several algorithms were tested, algorithm `walktrap` [9] is shown.

All algorithms are from package *python-igraph* accessible at `http://igraph.sourceforge.net/`

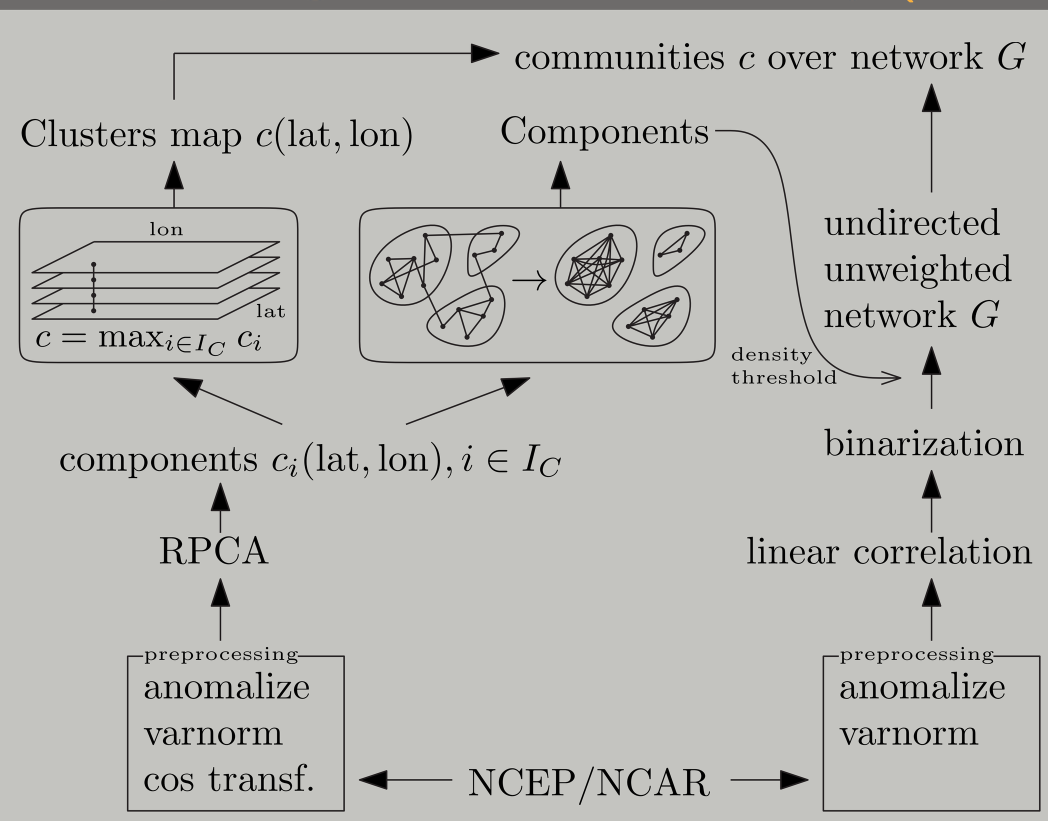
Used measure of quality of community partition

Quality of partition by *modularity* [5]. For graph $G = (V, E), |E| = m$, set of communities $\{C_i\}_{i=1}^{\ell}$ define [4] $\delta(C_i, C_j)$ indicator that vertices i and j have same community.

$$Q = \frac{1}{2m} \sum_{ij} \left(A_{ij} - \frac{k_i k_j}{2m} \right) \delta(C_i, C_j)$$

k_i denotes *degree of vertex i* (number of edges adjacent to vertex i), and A_{ij} elements of adjacency matrix.

Interpreting RPCA as community (scheme)

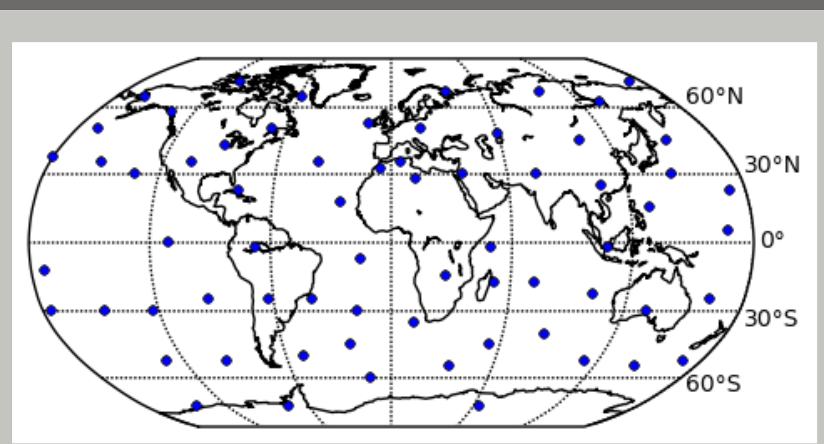


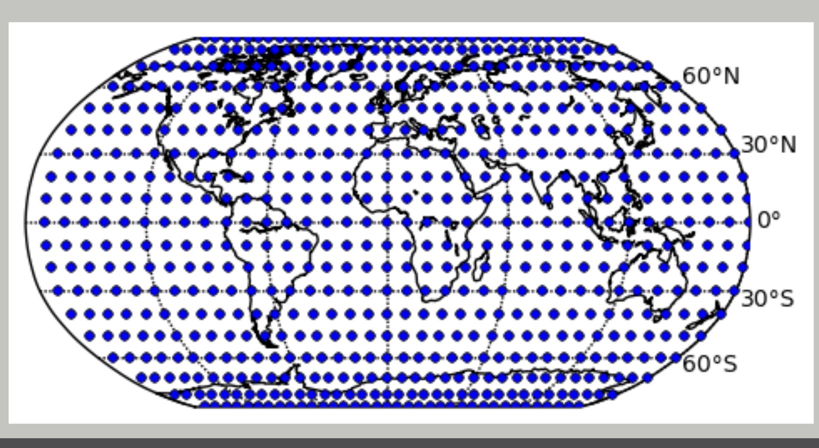
We proceeded as follows:

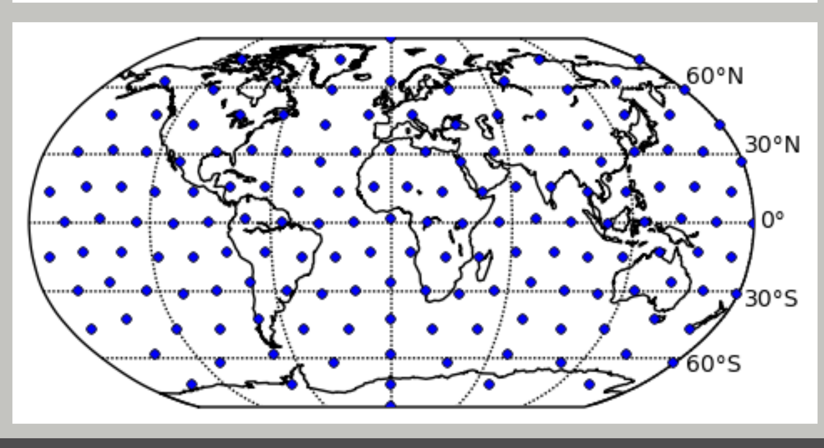
1. Construct RPCA components using preprocessed input data
2. Construct weighted network using the same input without cos. tranform
3. Compute threshold for binarization from density of network constructured as union of complete graphs each corresponding to particular component.
4. Construct unweighted undirected network using computed density
5. Construct community structure as nonoverlapping clusters derived from RPCA components using maximization criterion (i.e. cluster is given by maximal membership in component)

Dimensionality reductions

1. **RPCA components**
   Whole globe 58 SLP, 67 SAT →
2. **GEOGRID Spherical Geodesic Grid [10]**
   162 or 642 grid points, 162 ↘
3. **LATLON subsampling of lat-lon grid [3]**
   Original 73x144 is subsampled to 10° lat x 10° lon that gives 684 points ↓







RPCA as community

RPCA results Q_{RPCA} and algorithms fastgreedy Q_{FG} and walktrap Q_{WT}

SAT: $Q_{RPCA} = 0.4233$ while $Q_{FG} = 0.6549$, $Q_{WT} = 0.7381$

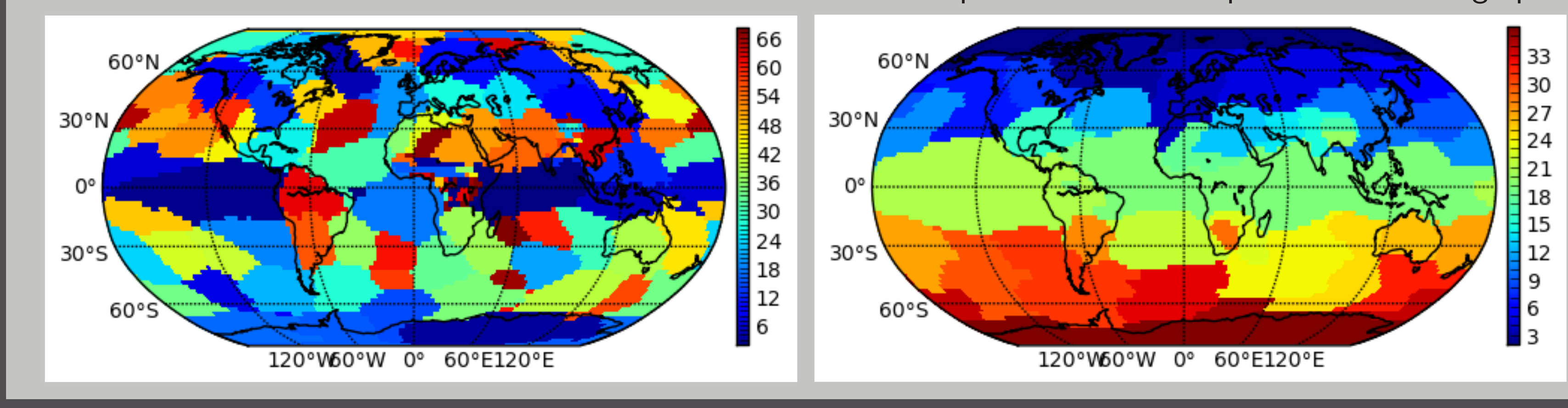
SLP: $Q_{RPCA} = 0.4674$ while $Q_{FG} = 0.5418$, $Q_{WT} = 0.6500$

Discussion (further results in second column)

RPCA seems to preserve some of community structure information – GEOGRID with slightly larger dimension (number of nodes) gives comparable results in sense of modularity. GEOGRID and LATLON with much larger dimension gives larger modularity. Further research with equal graph sizes and data preprocessing is required.

RPCA components and graph community

67 RPCA binarized clusters | 35 waltrap communities computed over whole graph



Comparison of various dimensionality reductions

For every dimensionality reduction (RPCA, GEOGRID, LATLON) proceed as follows.

Whole year

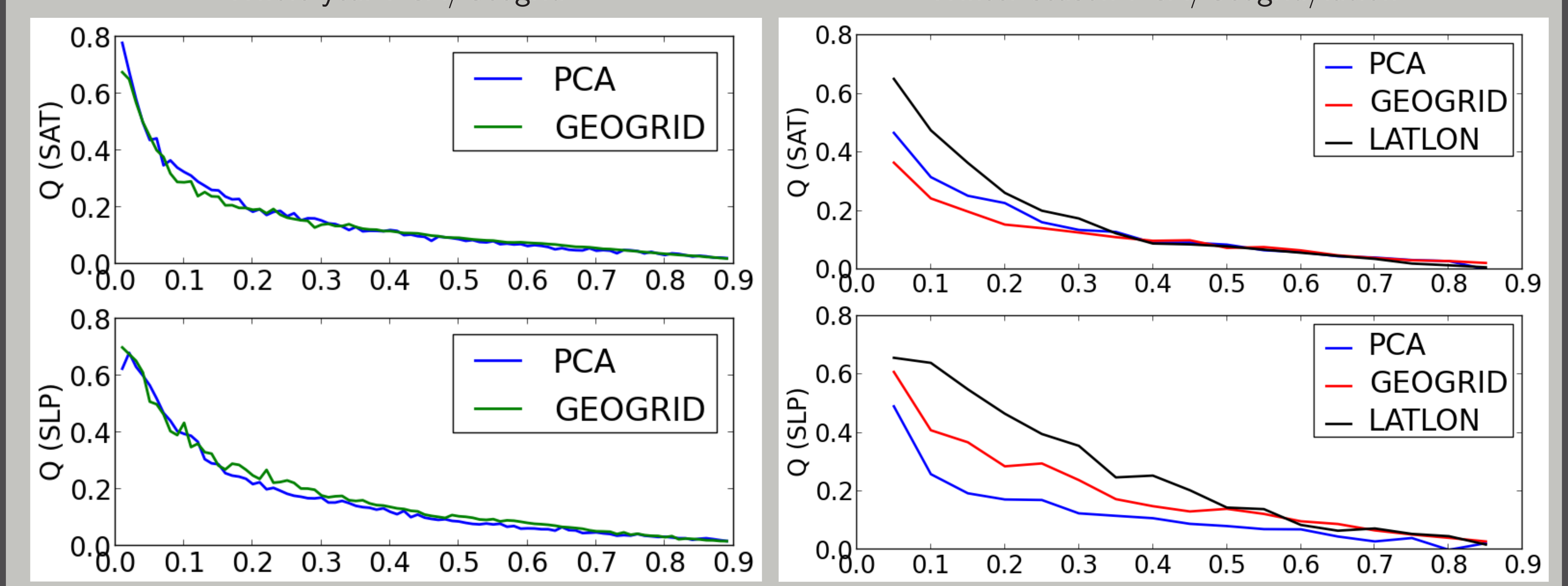
1. Extract whole year data
   Monthly means for all months
   67 RPCA, 162 GEOGRID
2. Preprocess and correlation matrix C
3. For densities in interval
   3.1 Threshold τ binarizes C into A_τ
   3.1 Costruct a network from adjacency A
   3.3 Community structure with Q_τ

Winter season

1. Extract Winter season [3]
   Monthly means for December to February
   67 RPCA, 642 GEOGRID, 684 LATLON
2. Preprocess and correlation matrix C
3. For densities in interval
   3.1 Threshold τ binarizes C into A_τ
   3.1 Costruct a network from adjacency A
   3.3 Community structure with Q_τ

Modularity for RPCA, GEOGRID, LATLON as function of density

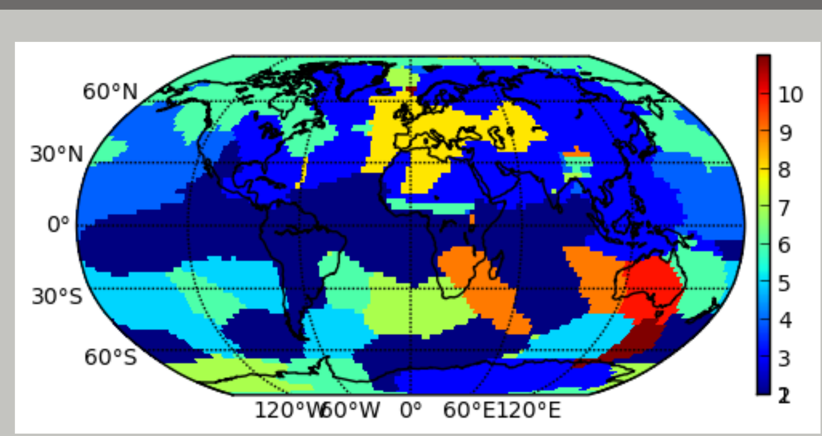
Whole year PCA/Geogrid | Winter season PCA/Geogrid/latlon

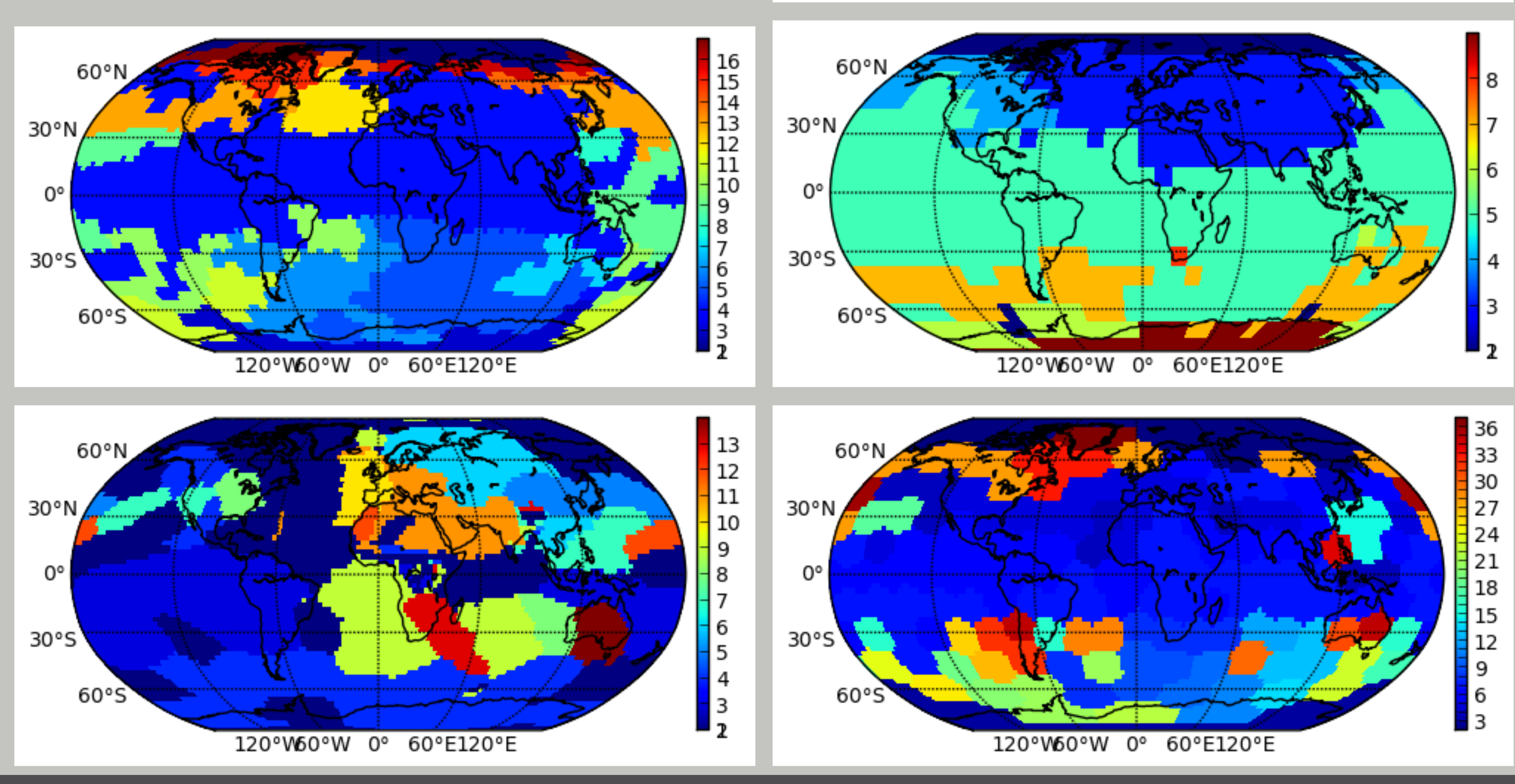


Examples of communities

Resulting SAT communities in chosen density 0.1

- RPCA winter season →
- GEOGRID winter season ↓
- LATLON winter season ↘
- RPCA whole year ↓↓
- GEOGRID whole year ↘↘





References

[1] **Barnston, A.G. and Livezey R.E.**. Classification, seasonality, and persistence of low-frequency atmospheric circulation patterns. *Monthly Weather Review*, 115:1083–1126, 1987.

[2] **Feldstein, S.B.**. The timescale, power spectra, and climate noise properties of teleconnection patterns. *Journal of Climate*, 13(24):4430–4440, 2000.

[3] **Tsonis, AA; Wang, G; Swanson, KL ; Rodrigues, FA; Costa, LD**. Community structure and dynamics in climate networks. *Climate Dynamics*, 37(5–6):933–940, 2011.

[4] **Fortunato, S.**. Community detection in graphs. *Physics Report-Review Section of Physics Letters*, 486(3–5):75–174, 2010.

[5] **Newman, M.E.J. and Girvan N.**. Finding and evaluating community structure in networks. *Physical Review E*, 69(2):026113, 2004.

[6] **Girvan N. and Newman, M.E.J.**. Community structure in social and biological networks. *PNAS*, 99(12):7821–7826, 2002.

[7] **Tsonis, AA and Swanson, KL**. On the origins of decadal climate variability: a network perspective. *Nonlinear Processes in Geophysics*, 19(5):559–568, 2012.

[8] **R. Kistler et al**. The NCEP-NCAR 50-year reanalysis: Monthly means CD-ROM and documentation. *Bulletin of the American Meteorological Society*, 82:247–268, 2001.

[9] **Pons P. and Latapy M.**. Computing communities in large networks using random walks. *LCNS*, 3733:284–393, 2005.

[10] **Jones P.W.**. First- and Second-Order Conservative Remapping Schemes for Grids in Spherical Coordinates. *Monthly Weather Review*, 127:2204–2210, 1998.

http://ndw.cs.cas.cz/ {hartman,hlinka,vejmelka,palus}@cs.cas.cz