Application of sparse PCA, cluster, SSA, wavelets analysis to the GVA data analyzed in sectoral-regional EU context

Saulius Jokubaitis, Dmitrij Celov

Vilnius University, Euro 4 Europe

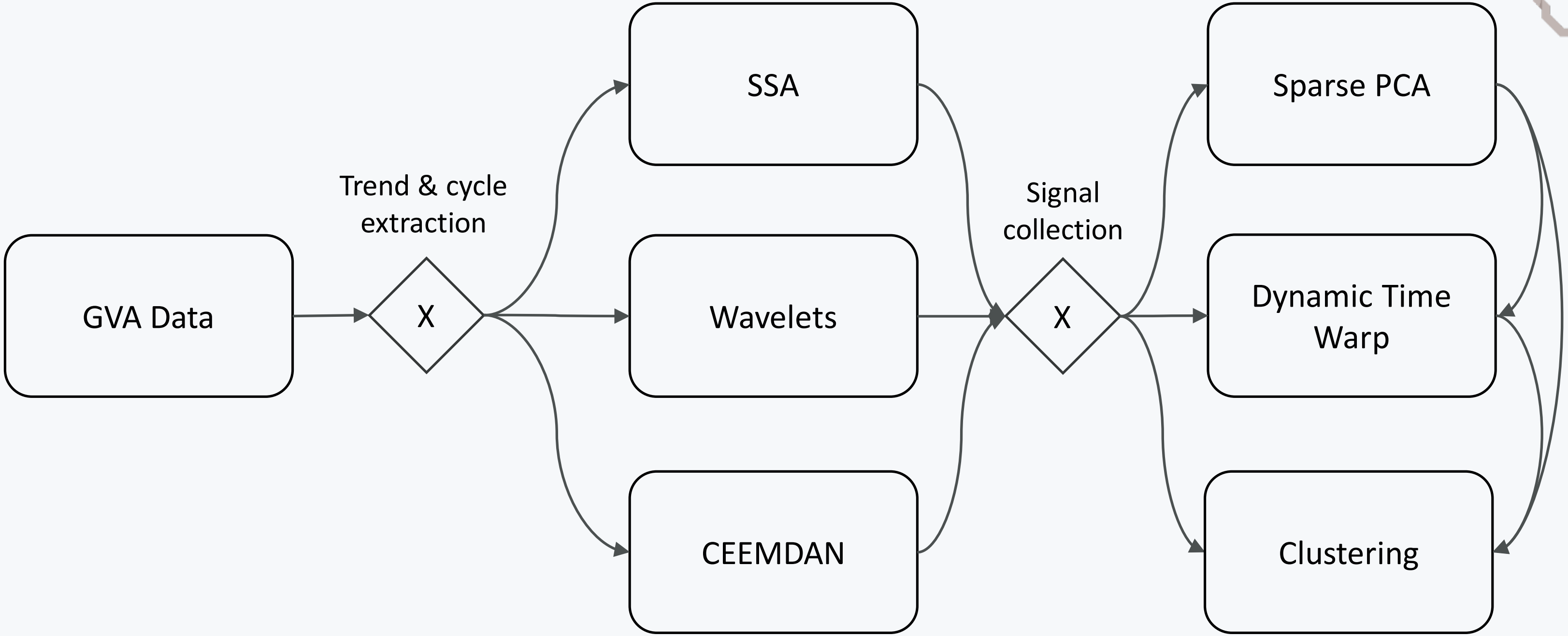
INTRODUCTION

- **Goal**: Evidence for business cycle synchronization – crucial for policy makers with contrasting results in the literature
- **Approach**: through Gross Value Added (GVA) cycles
- **Wavelets**: for trend extraction and cycle decomposition
- **Main scope**: A*10 Sectoral Breakdown of GVA for EU-28 countries
- **Secondary goals**: Wavelet family & methods for deeper exploration

SECTORAL-REGIONAL INVESTIGATION

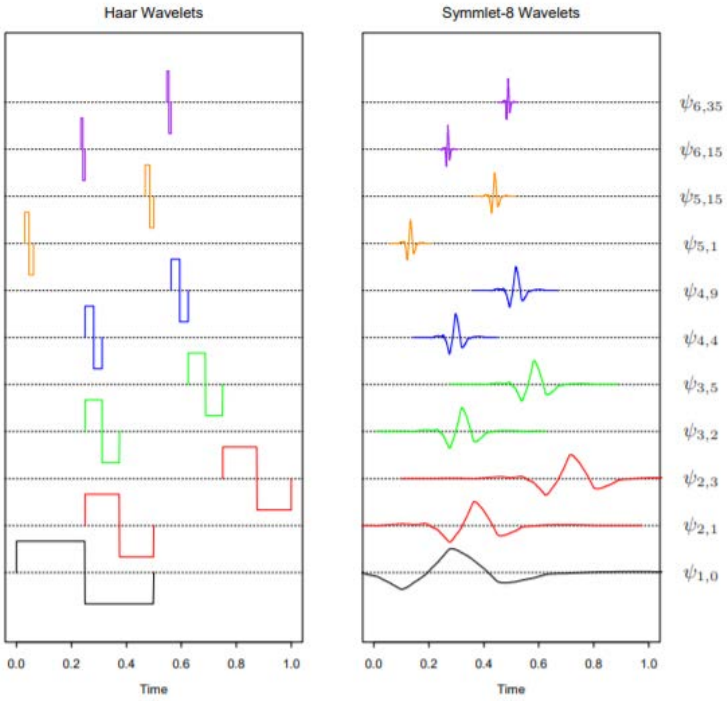
- **Data:** Gross Value Added (Real; Seasonally and Calendar adjusted), Source: *Eurostat*
- **Sample:** 2000 Q1 – 2019 Q1
- **Regions:** EU-28 countries
- **Sectors:** A*10 industry breakdown (NACE rev. 2), 11 combinations

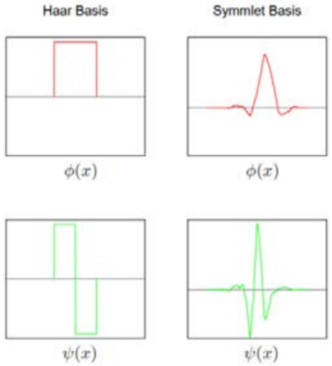
IDEA MAP

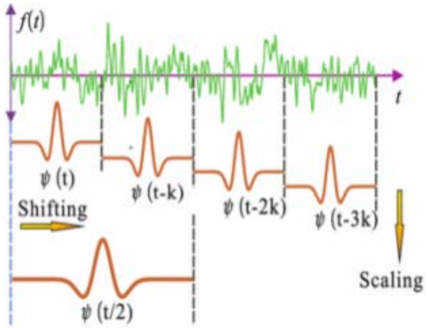

GVA Data
Trend & cycle extraction
X
SSA
Wavelets
CEEMDAN
Signal collection
X
Sparse PCA
Dynamic Time Warp
Clustering

WAVELET TRANSFORMATION

Time & frequency domain decomposition







Where:

$$\phi_{j,k} = 2^{j/2}\phi(2^j x - k)$$

$$\psi(x) = \phi(2x) - \phi(2x-1)$$

$$\psi_{j,k} = 2^{j/2}\psi(2^j x - k)$$

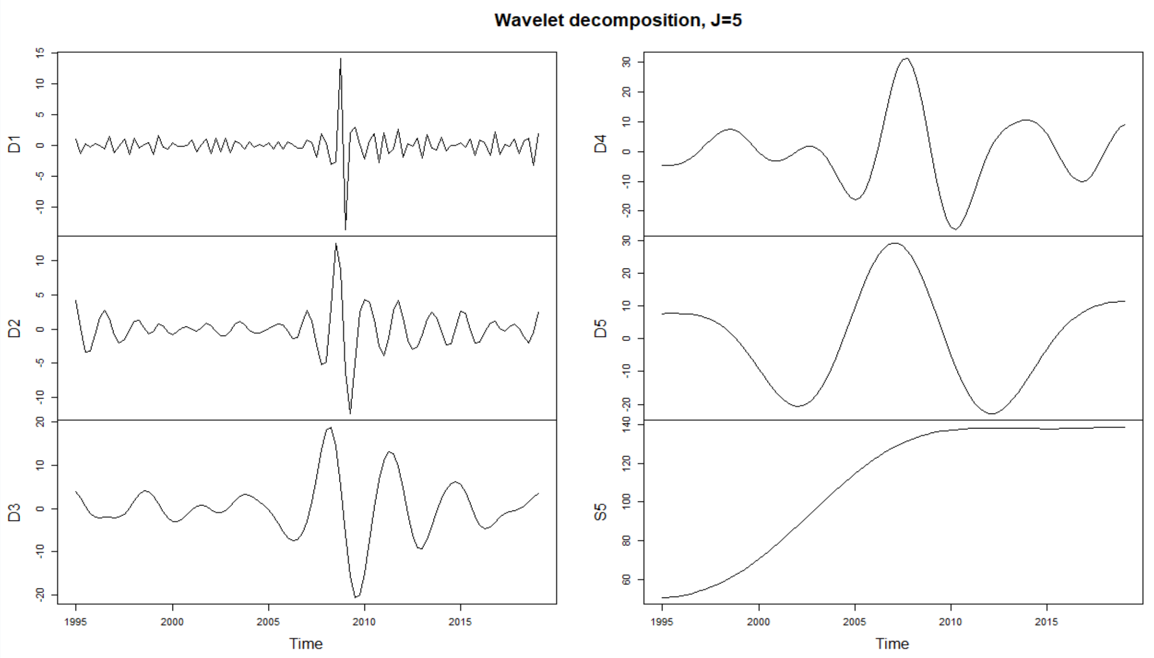
MAXIMUM OVERLAP DISCRETE TRANSFORMATION

- Transformations are very similar to the MODWT have been studied in the literature under the following names:
  - undecimated DWT (or nondecimated DWT)
  - stationary DWT
  - translation invariant DWT
  - time invariant DWT
  - redundant DWT
- Basic idea: use values removed from DWT by downsampling
- Unlike the DWT, MODWT is not orthonormal (in fact MODWT is highly redundant)
- Unlike the DWT, MODWT is defined naturally for all sample sizes (i.e., N need not be a multiple of a power of two)
- Further possible research: MODWPT (packet transform); transformations in Hilbert space

EXAMPLE CASE: CONSTRUCTION; GEO = LITHUANIA

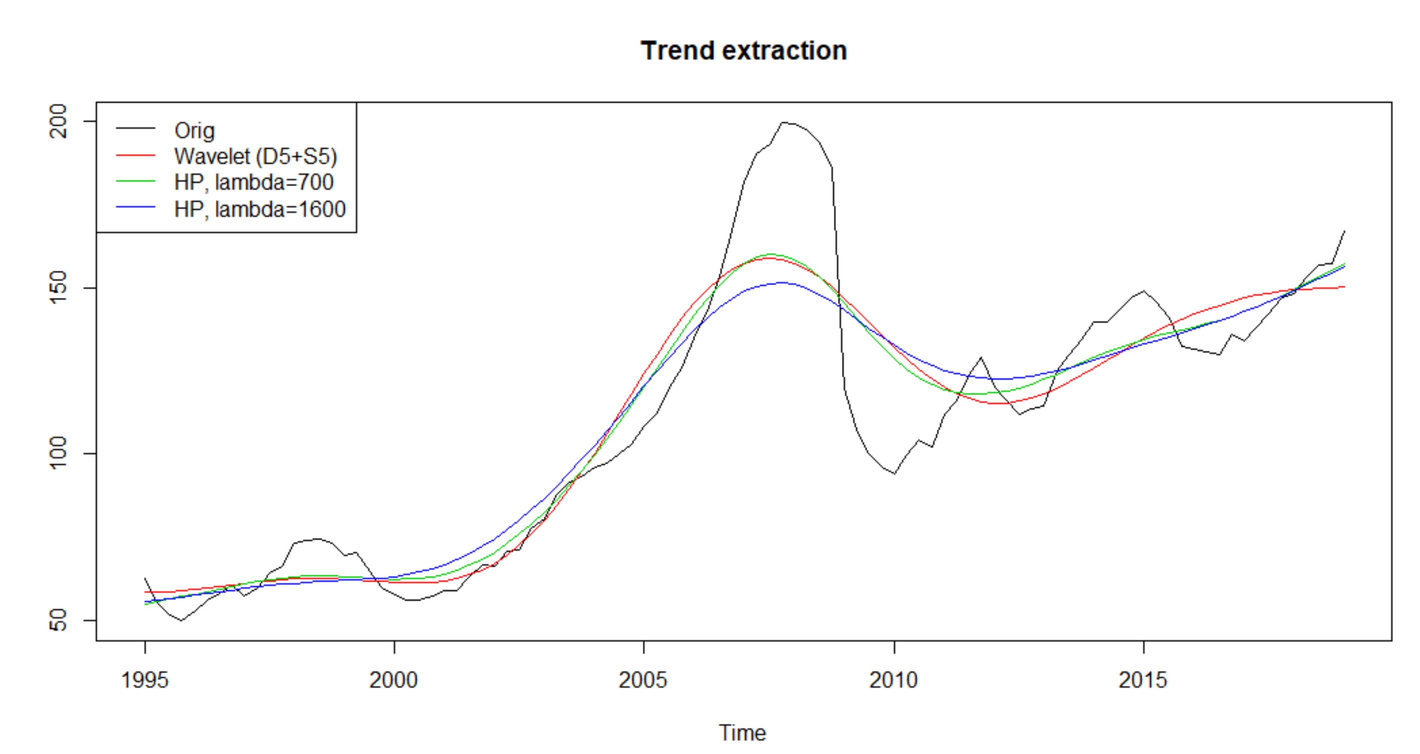
Table 1. Frequency interpretation; *Crowley and Mayes (2008, p. 70)*

Detail	Length	Type
D1	2-4Q	Noise
D2	1-2Y	Noise
D3	2-4Y	BC
D4	4-8Y	BC
D5	8-16Y	CC
S5	16Y+	Trend



EXAMPLE CASE: CONSTRUCTION; GEO = LITHUANIA

Wavelet trend extraction comparison: D5+S5 for trend



SPARSE PCA

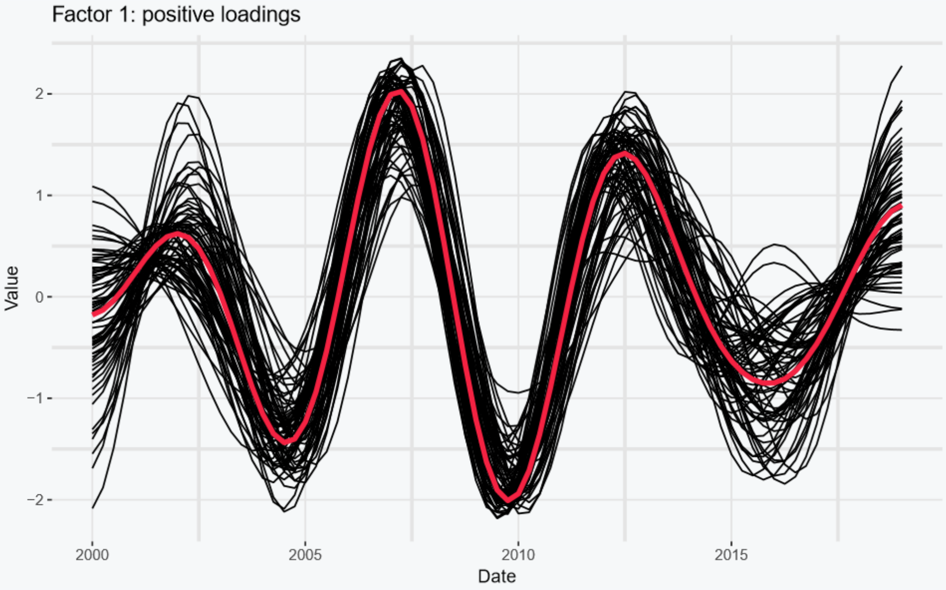
- PCA extension with imposed sparsity restriction on the loading matrix.
- Cardinality of the matrix is reduced with setting the weaker signals to 0.
- This essentially breaks the orthogonality of the signals while possibly improving the signal to noise ratio.

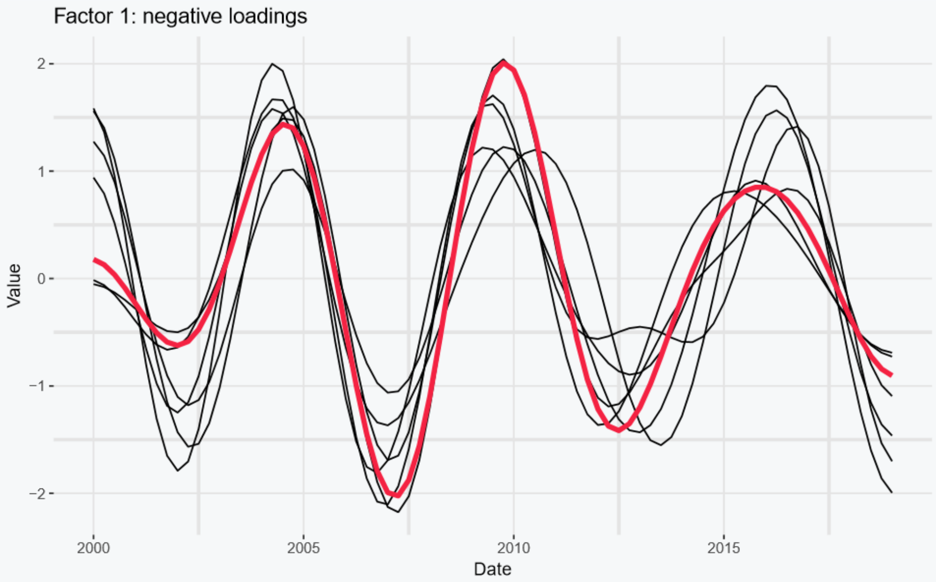
Rationale in the current context:

1) Identify and extract the main signals from the sectoral-regional data
2) The orthogonality of the signals is not a priority: there may be converging cycles
3) Sign invariant: possibility of capturing counter-cyclic signals

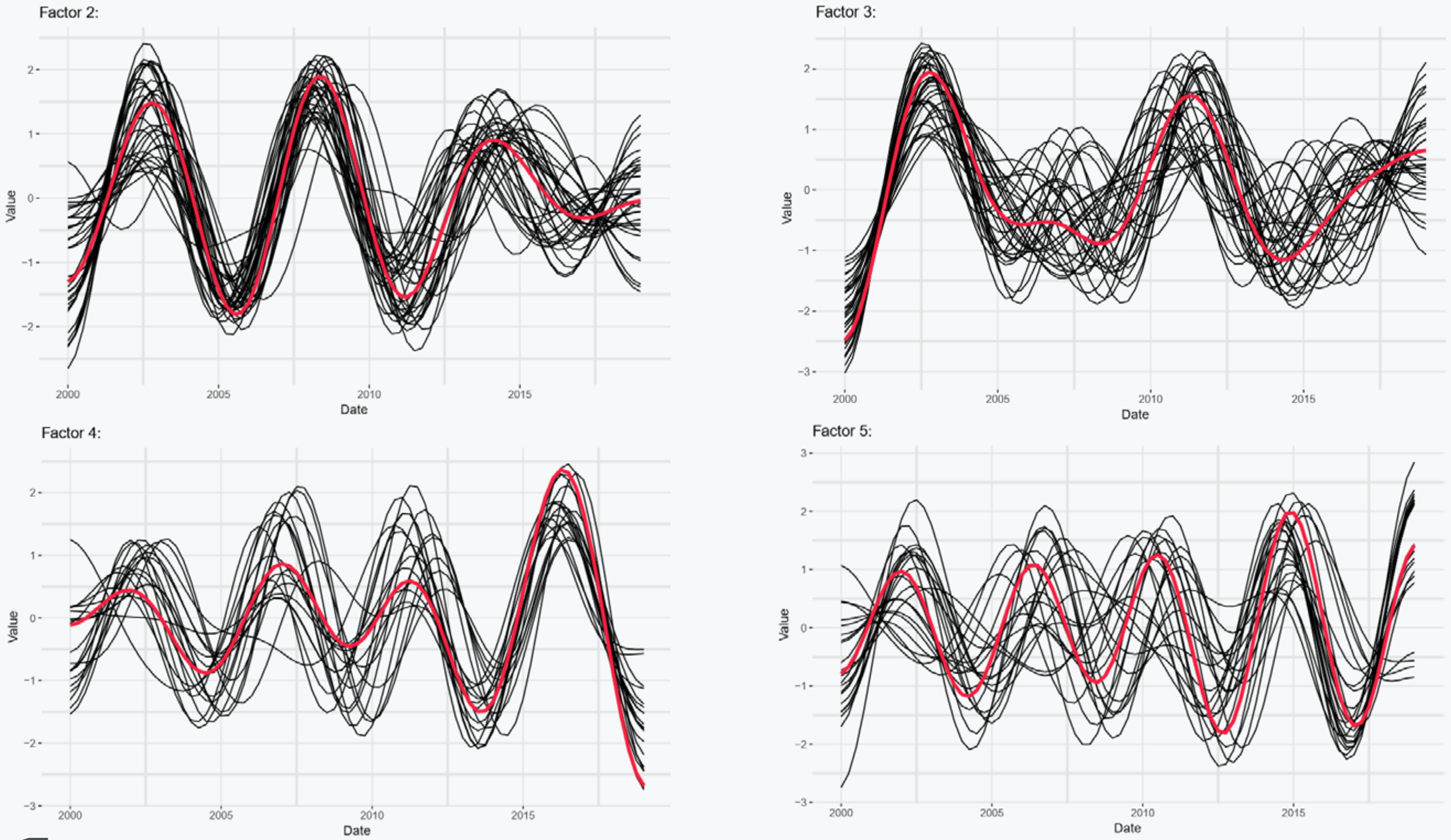
SPCA: D4 WAVES (4-8Y)

- 55% variance explained by 2 factors
- 70% variance explained by 3 factors





Example of sign invariance

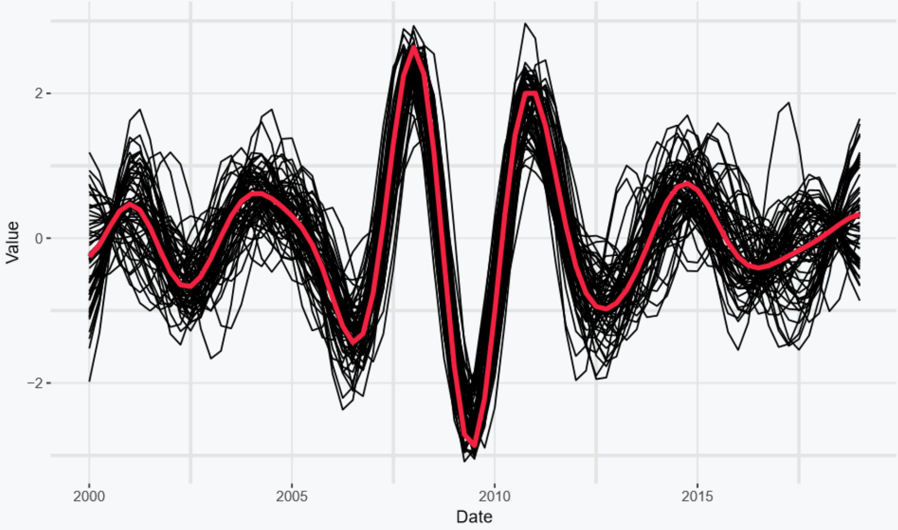


Factor 2:
Value
Date
2000
2005
2010
2015
Factor 3:
Value
Date
Factor 4:
Value
Date
Factor 5:
Value
Date

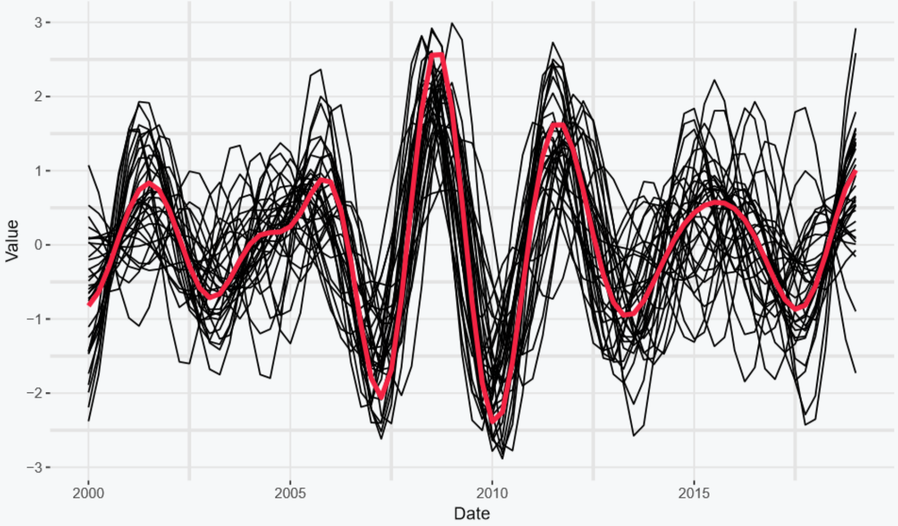
SPCA: D3 WAVES (2-4Y)

- 45% variance explained by 3 factors
- 70% variance explained by 5 factors

Factor 1:



Factor 2:



PCA for cycle Synchronization was used by, among many others,

- Andrle, Brůha & Solmaz (2017), Caporale (1993) and Quah (2013) uses PCA to evaluate if the eurozone is an optimal currency area
- while Selover (1999) and Sethapramote & Thepmongkol (2018) uses PCA to find a common business cycle in the ASEAN,
- finally Kose, Otrok & Whiteman (2003), Kose, Otrok & Prasad (2012) and Ductor & Leiva-Leon (2016) uses it to investigate a global business cycle.
- Nielsen (2018) used it along with cycles, extracted with wavelets

However, in our case, immediate drawbacks are evident by examining the data:

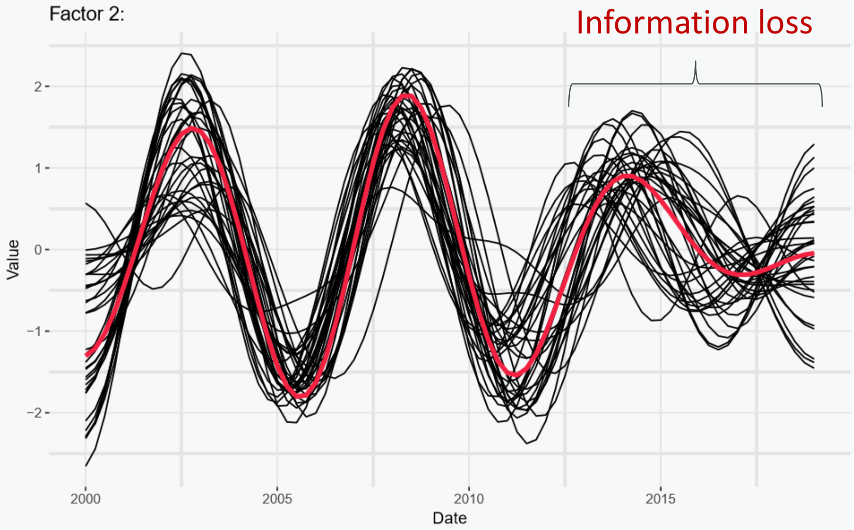
1) Even the strongest signals requires some cleaning, hence the use of Sparse PCA vs PCA.
2) Percentage of variance explained is sensitive to series dynamics

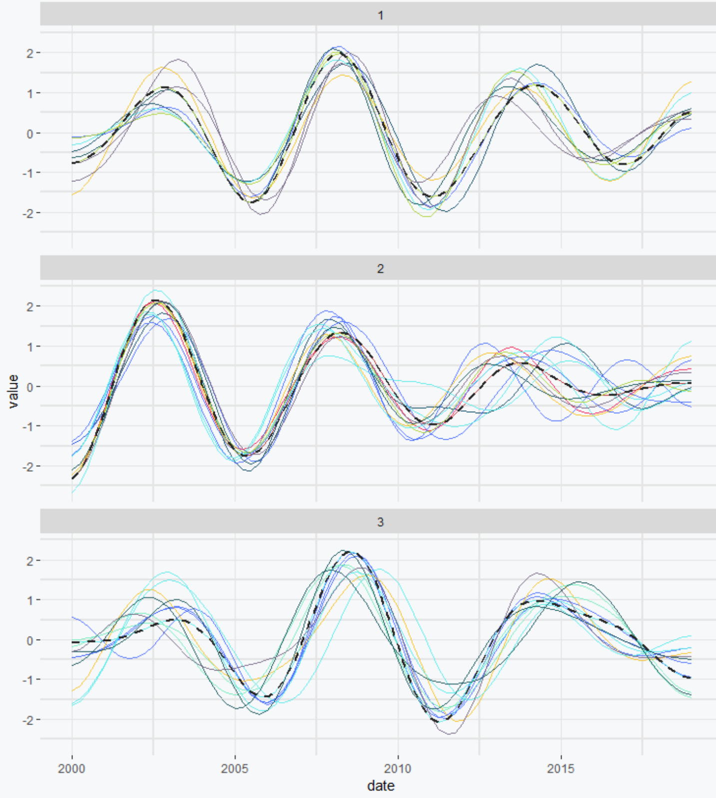
   $\Rightarrow$ weaker signals tend to lose information
3) Percentage of variance explained *can* lead to spuriously grouped signals.

Main takeaway: use SPCA to start the analysis, not end with it.

CLEANING SIGNALS THROUGH CLUSTERING

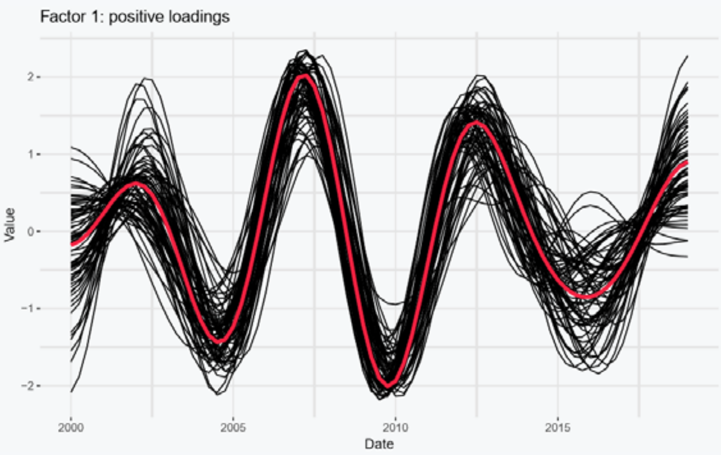
Factor 2:

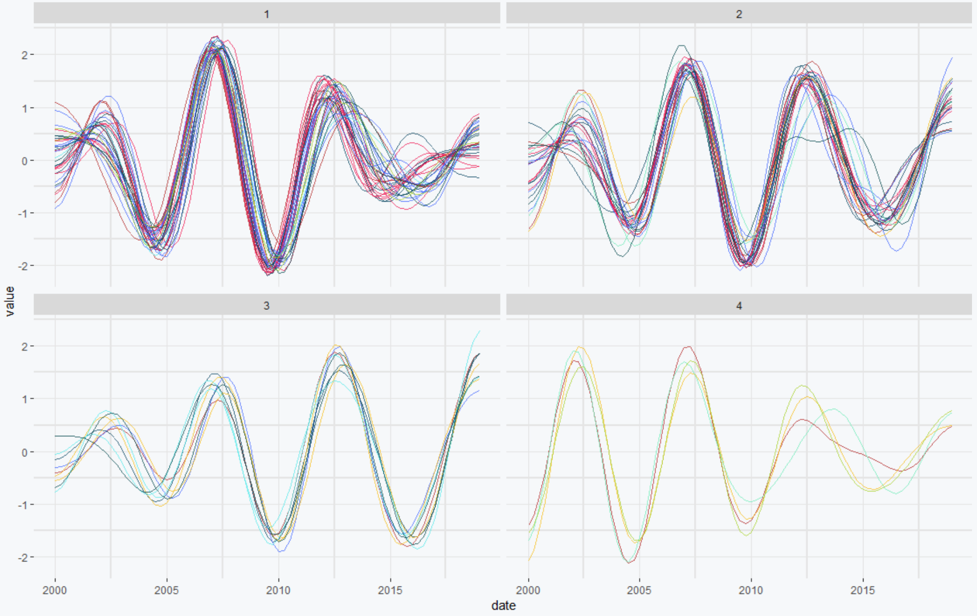




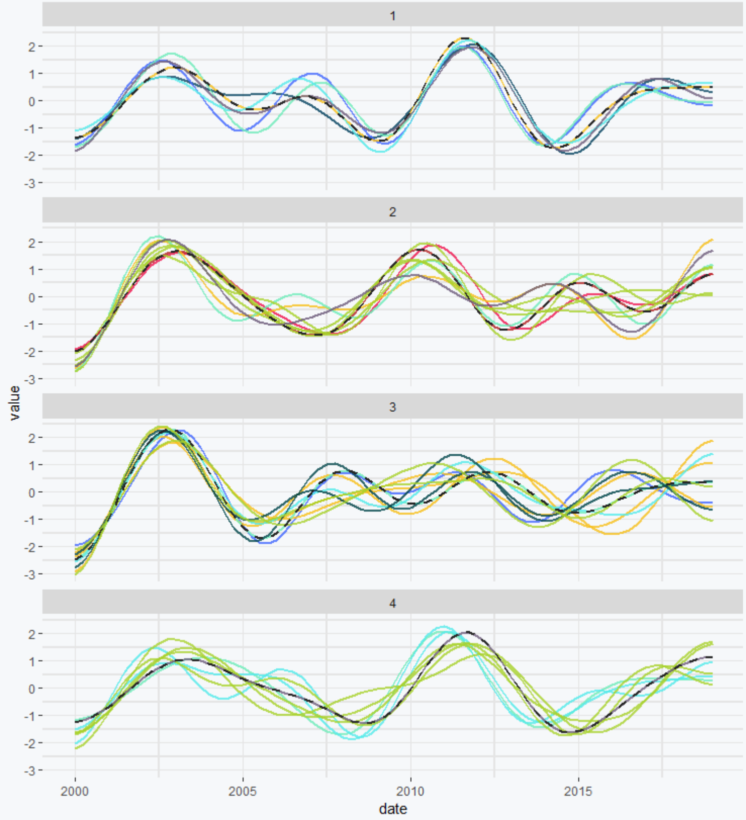
Hierarchical clustering with supremum metric.

- Minimizing the maximum difference separates out potential phase and mode differences (1. vs 2.)





- Cluster 1: strongest synchronization during the shock
- Cluster 2: stronger synchronization after the shock than before
- Cluster 3: strong overall synchronization
- Cluster 4: strong overall synchronization, different modes



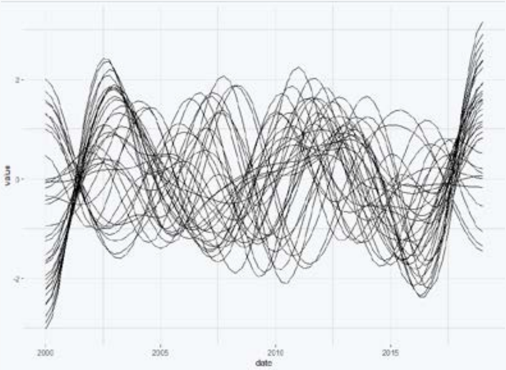
- Weaker signals seem to require more clusters for cleaner separation – how much noise can we allow?
- Cluster 3. appears to separate out series of higher frequency

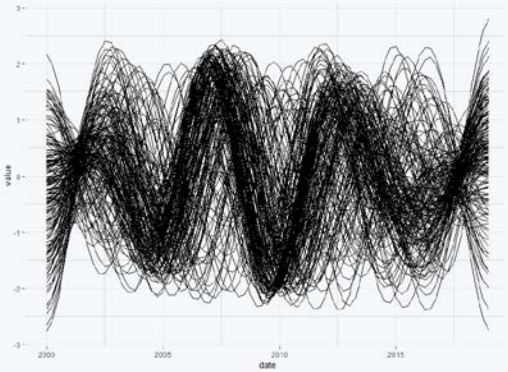
FREQUENCY FILTER

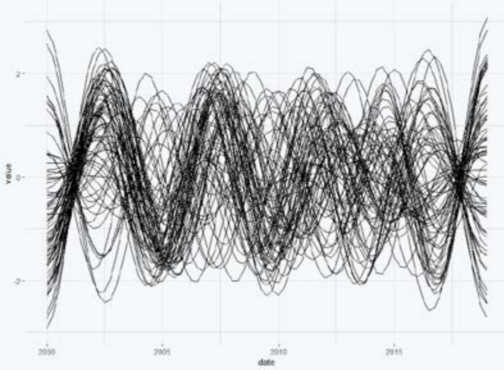
D4 waves should cover signals within the range of 4-8Y (2^4-2^5 Q).
Higher frequency signals, mixed with lower, may add noise to PCA due to interference.

Based on spectral decomposition of the signals, we identify:

1) 10% of signals with unidentified frequency (varying?)
2) 60% of signals with 1/4.625 frequency
3) 25% of signals with 1/3.7 frequency

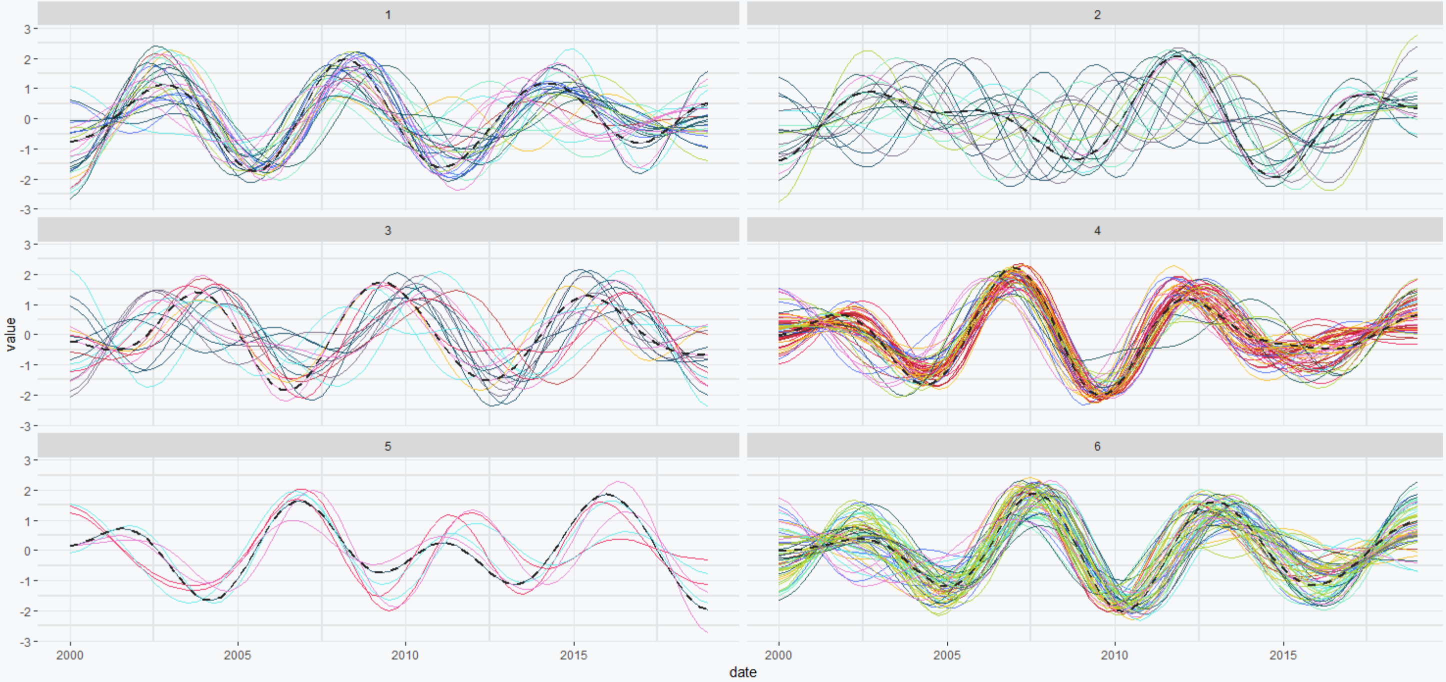






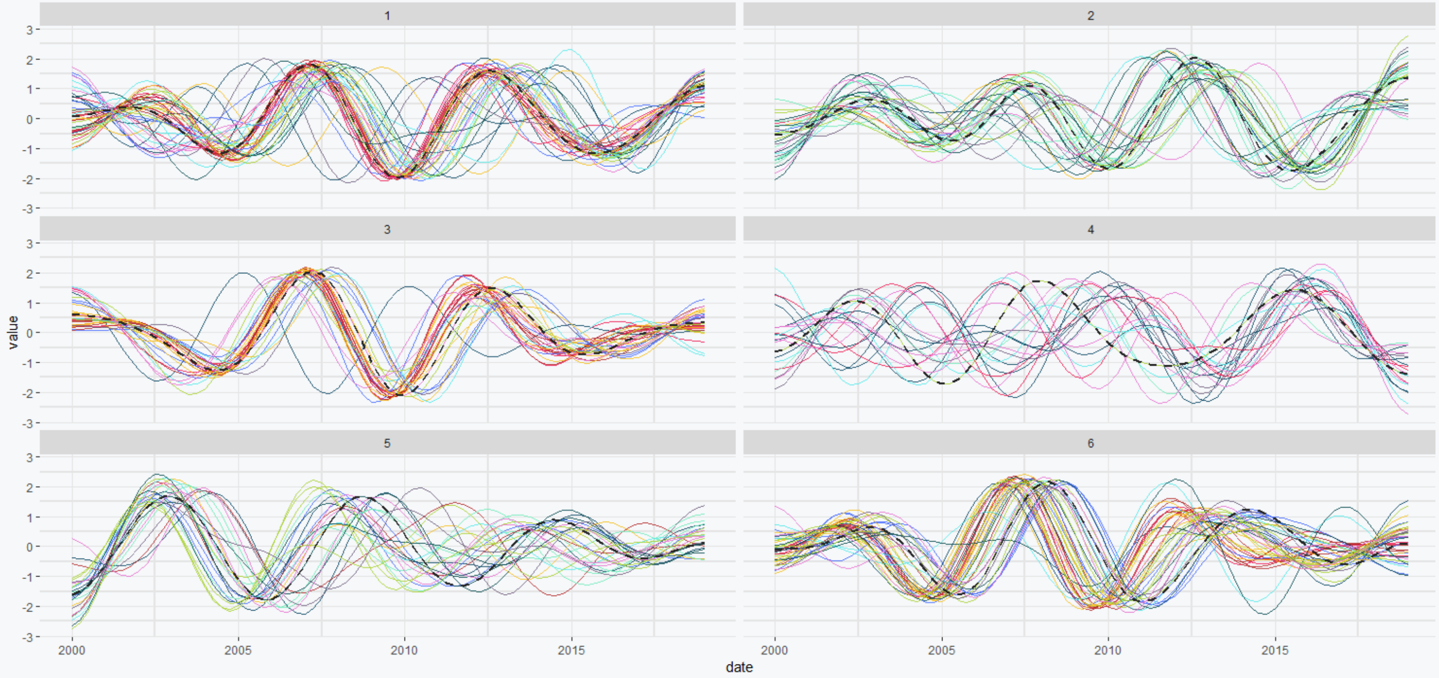
FREQUENCY FILTER + SUPREMUM HCLUST

This doesn't fix the SPCA results, but helps as a first level of hierarchy for hierarchical clustering of the series. We focus on 60% of the series with 1/4.625 frequency.



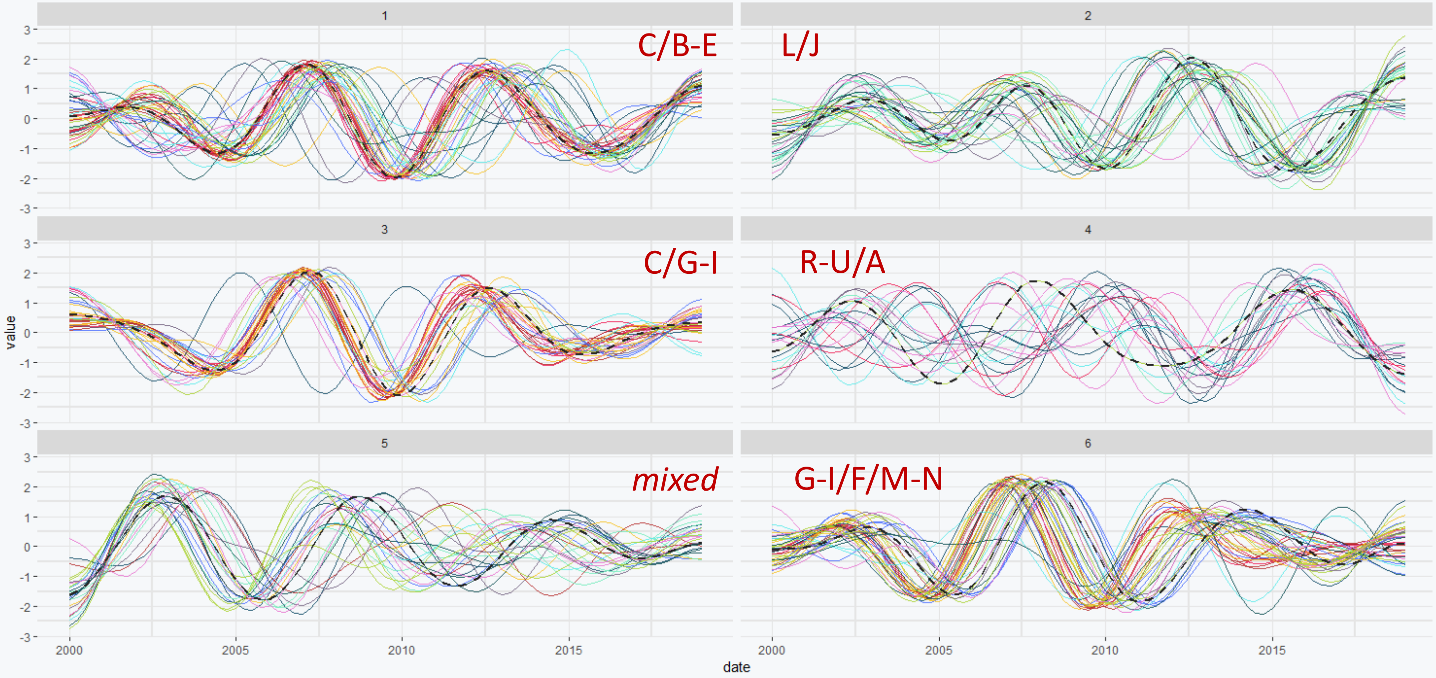
FREQUENCY FILTER + DTW HCLUST

Same series, clusters adjusted by Dynamic Time Warp



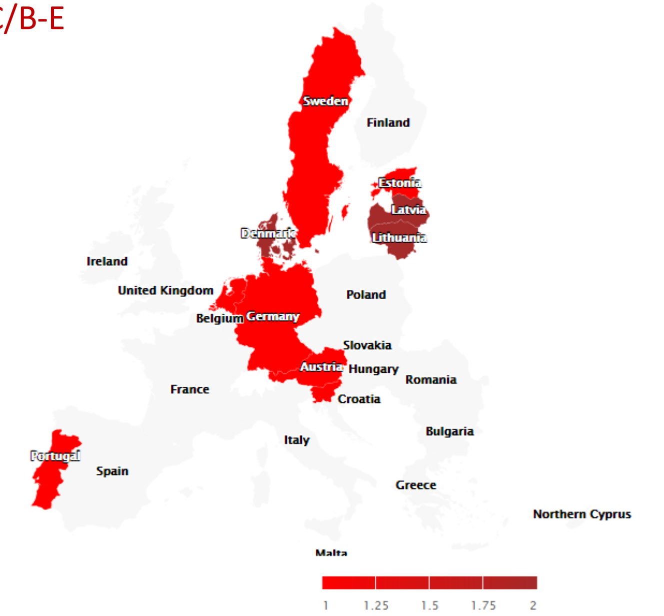
FREQUENCY FILTER + DTW HCLUST

Same series, clusters adjusted by Dynamic Time Warp

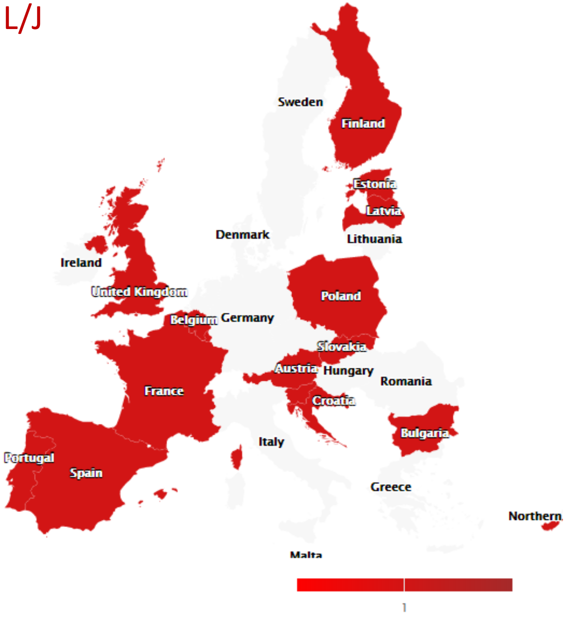


DTW CLUSTERS BY COUNTRY

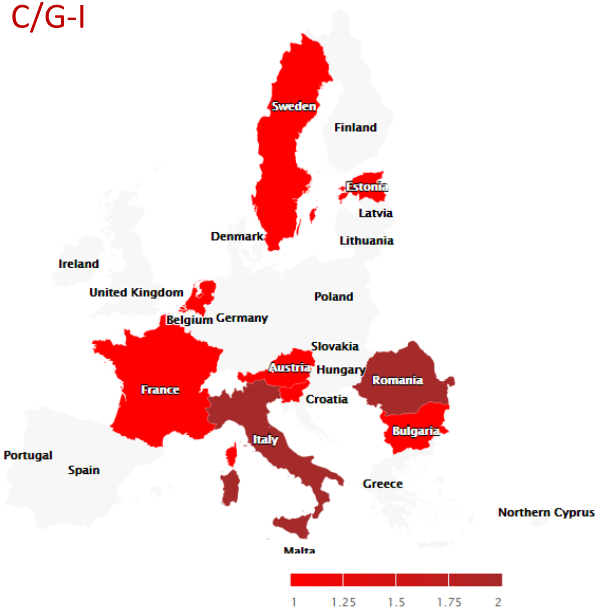
C/B-E



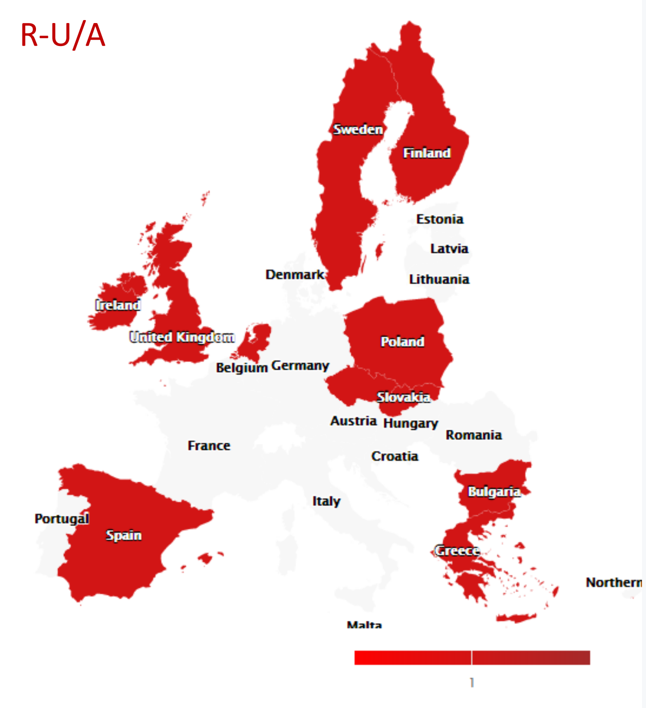
L/J

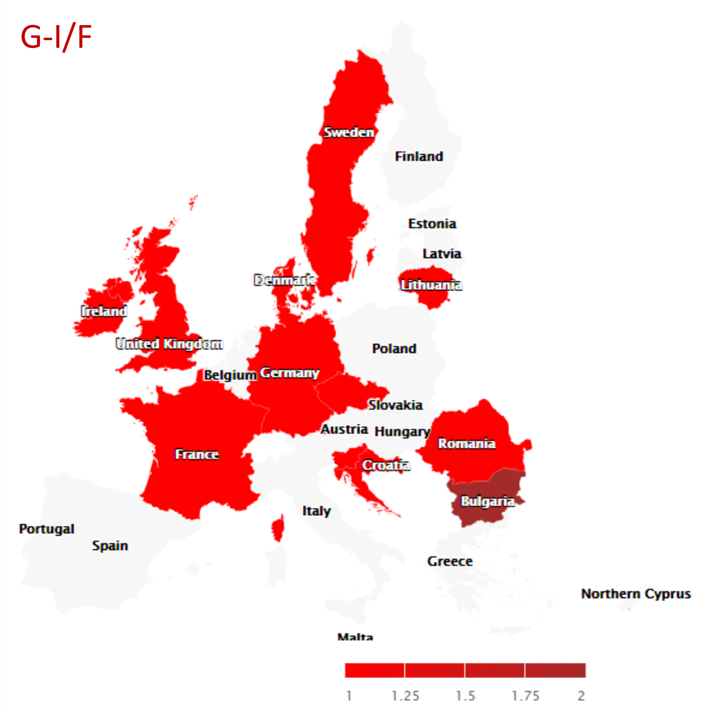


C/G-I



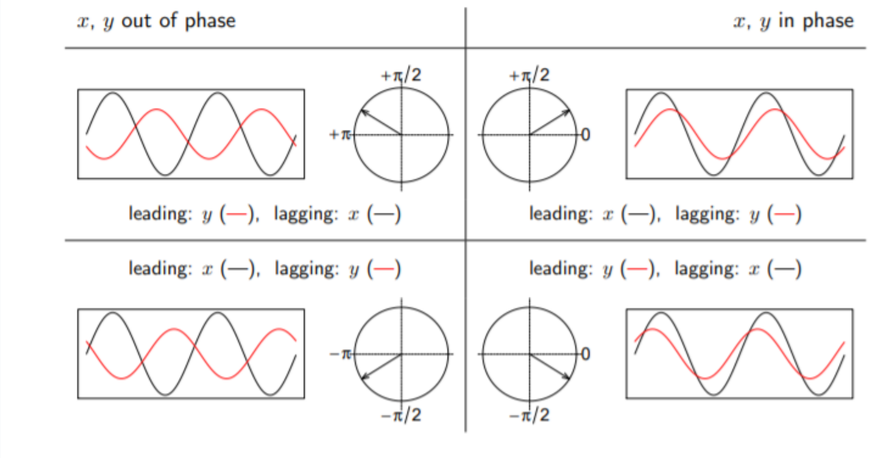
DTW CLUSTERS BY COUNTRY (2)

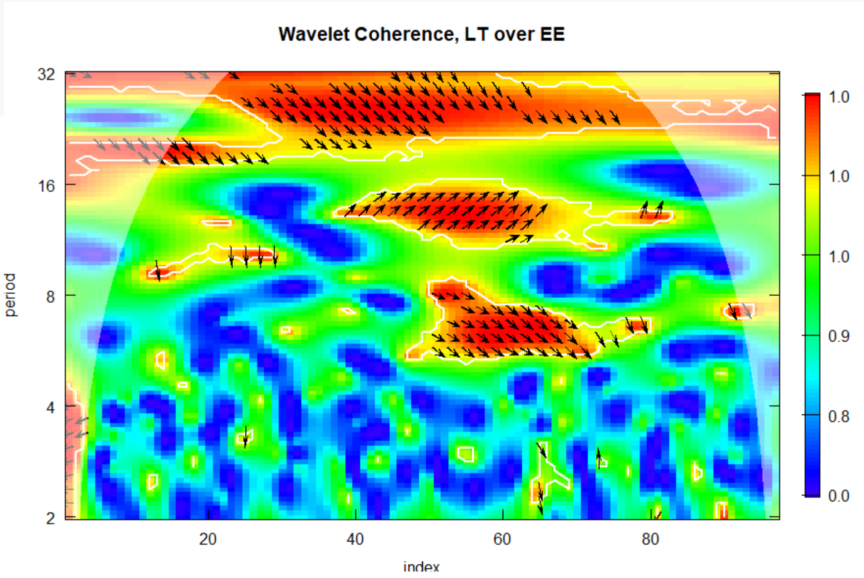


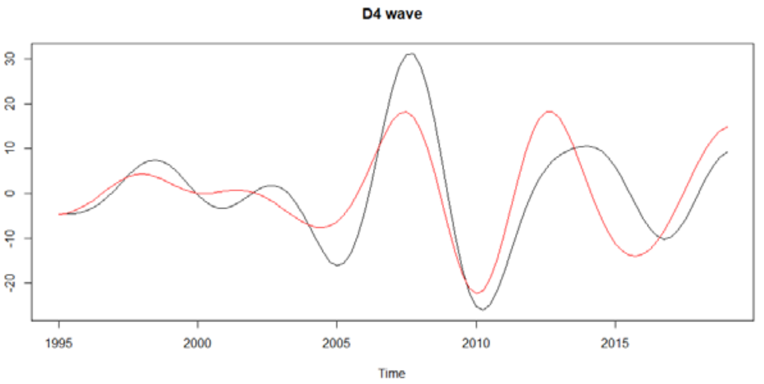


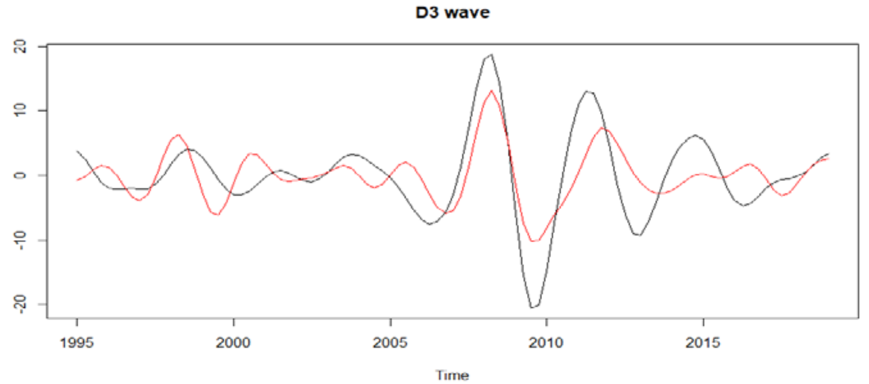
FUTURE RESEARCH

WAVELET COHERENCE



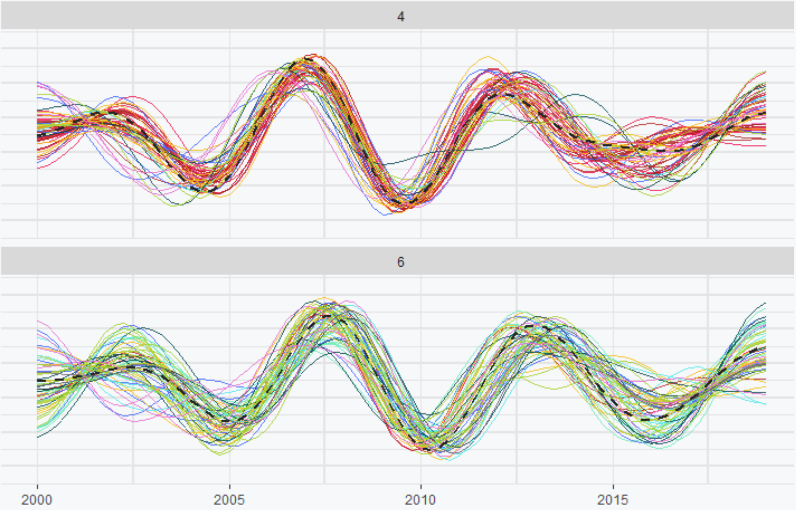






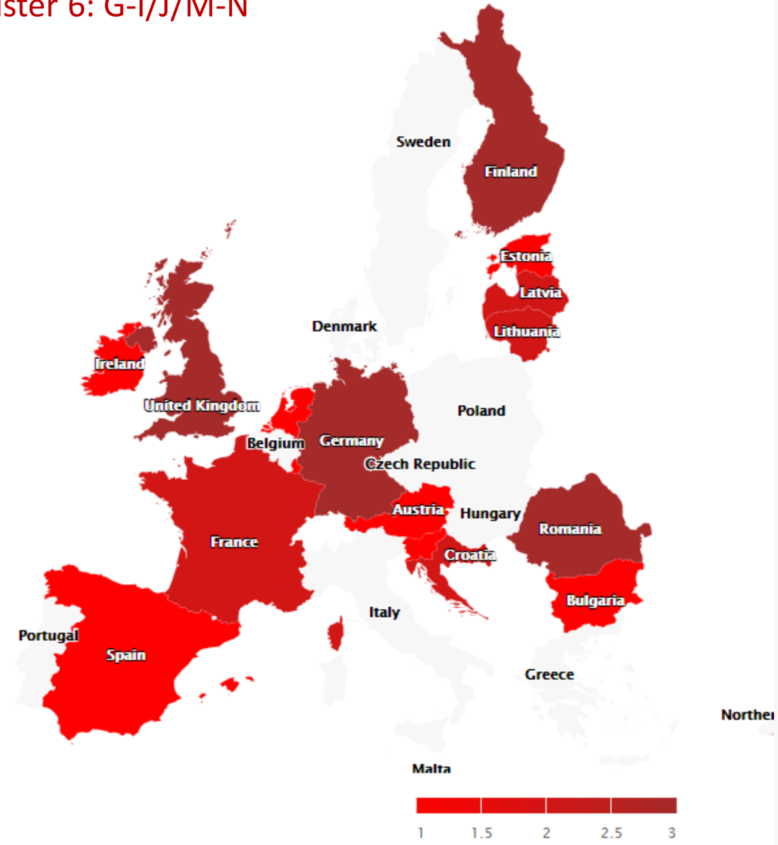
CLUSTER COHERENCE

- Fixing the cluster frequency allows for easy phase shift calculation
- We can notice that cluster 4. is dominated by C/B-E/G-I
- Cluster 6. seems to be dominated by G-I/J/M-N.
- Separated by supremum norm -> likely a shift in time
- Instead of 1 by 1 cluster coherence, can we generalize lead/lag relationship in a cluster level?
    - -> ~80% series from cluster 6 lead cluster 4.



Cluster 4: C/B-E/G-I

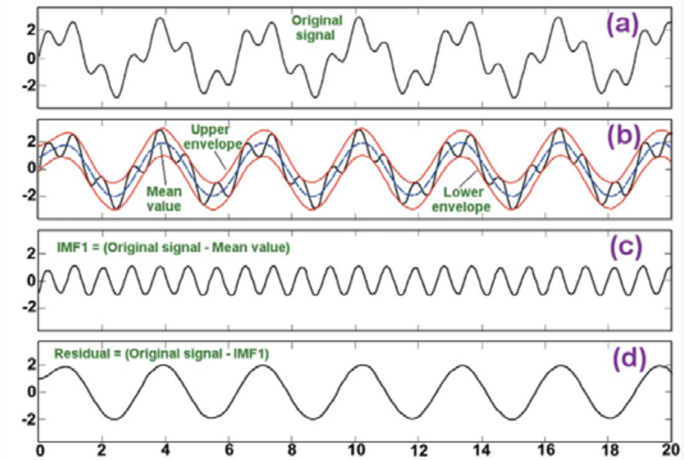
Cluster 6: G-I/J/M-N



CEEMDAN

One possible alternative to Wavelets:
Complete ensemble empirical mode decomposition with adaptive noise.

1) Empirical Mode Decomposition (EMD)
   - decomposes data through Intrinsic Mode Functions (modes)
2) Averaging EMD modes with added Gaussian noise (EEMD)
   - solves mode mixing
3) Smartly chosen Adaptive Noise (EEMD-AN) solves remaining theoretical shortcomings
4) The resulting decomposition is Complete, with negligible residual remaining (CEEMDAN, Torres et. al. (2011))



CEEMDAN

Few hyper-parameters needed to be tuned:

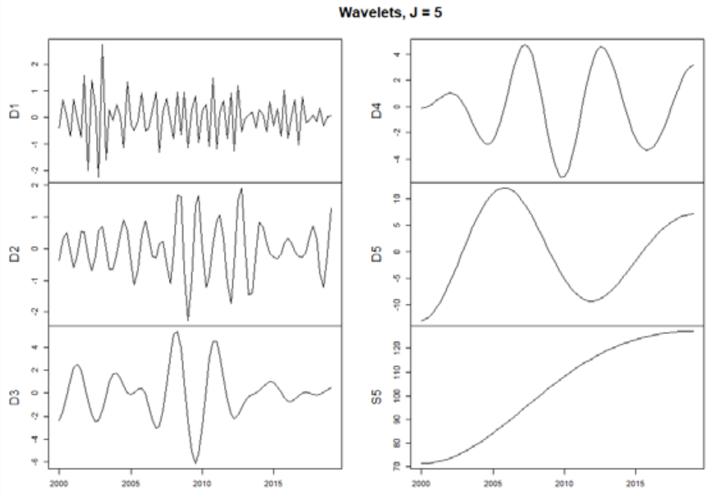
1) Size of the ensemble
2) Noise type and strength
3) Number of zero-crossings for the IMF’s
4) Number of siftings (as for an algorithm stopping criterion)

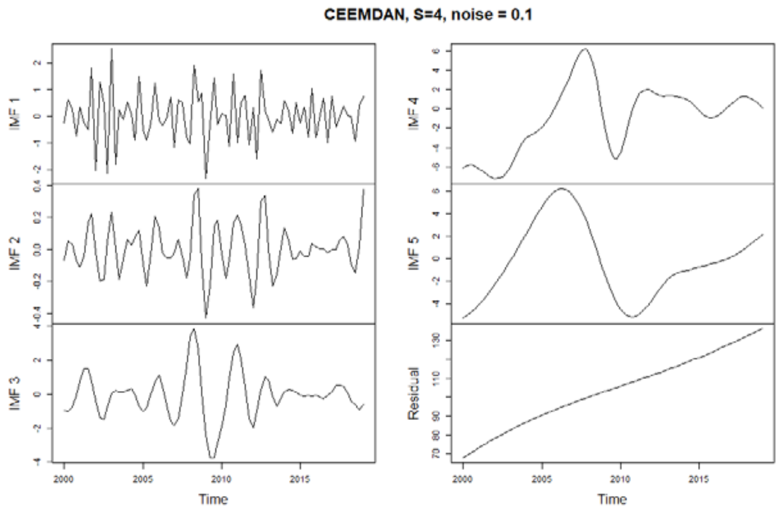
All of which impacts the smoothness or the edge behavior.

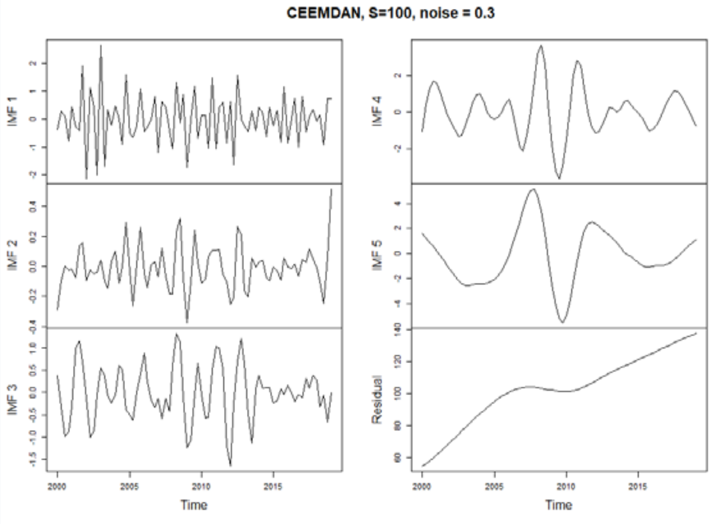
Our observations so far:
1) Less stability over all series
2) Harder to predict the frequency, since IMF’s are not tied to frequency range
3) More variance in the extracted signals
4) Signals are more correlated than wavelets (DTW orthogonal, MODWT less so)

CEEMDAN: EXAMPLE CASE

Example series: B-E sector in Lithuania





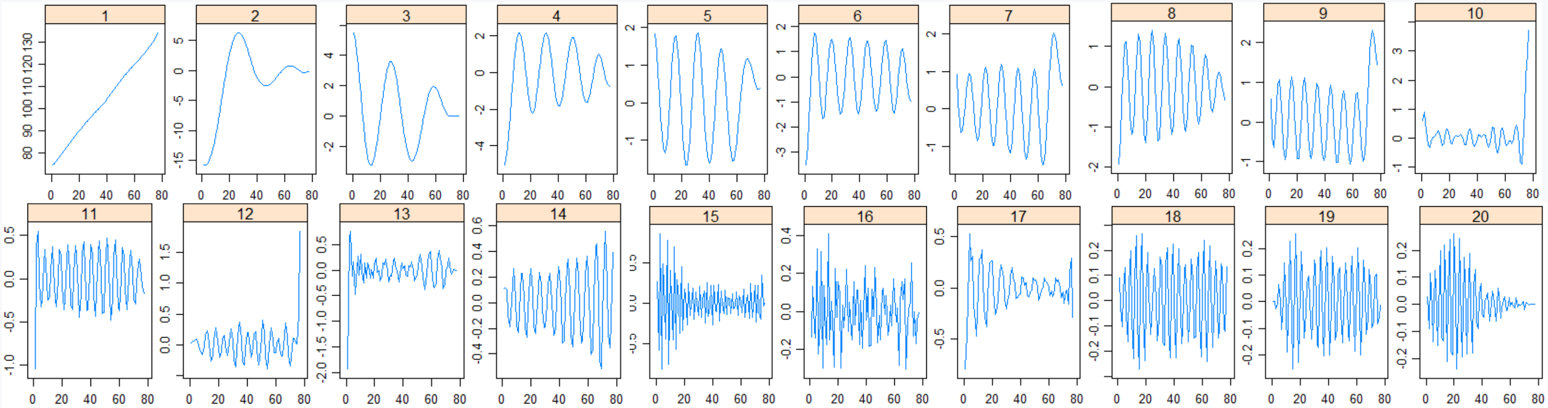


SINGULAR SPECTRUM ANALYSIS

SSA algorithm steps:

1) Embedding of the series
2) Singular Value Decomposition (SVD)
3) Grouping & diagonal averaging

Decomposition of series, used in CEEMDAN example:



THANK YOU!

APPENDIX: COLOR CODES