Singular Spectrum Analysis with Rssa

Maurizio Sanarico

Chief Data Scientist

SDG Consulting

Milano R – June 4, 2014

Motivation

- Discover structural components in complex time series
- Working hypothesis: a signal is composed by trend (possibly multiple ones), oscillatory components (often more than one) and noise (white noise or colored noise)
- It works combining concepts and tools from classical time series analysis, multivariate statistics, multivariate geometry, dynamical systems and signal processing + statistics as a tools for inference

The basics: univariate case

- $X_N = (x_1, \ldots, x_N)$: ***time series***
- L $(1 < L < N)$: **window length**
- construct L-lagged vectors $X_i = (x_i, \ldots, x_{i+L-1})^T$, $i = 1, 2, \ldots, K$, where

$K = N - L + 1$

- Form the **trajectory matrix** **X** from these vectors
- The **eigen-analysis** of the matrix $\mathbf{XX}^T$ (equivalently, the **SVD** of the matrix **X**) yields a collection of L eigenvalues and eigenvectors
- A **combination** r of these **eigenvectors** determine an r-dimensional subspace L^r in R^L , $r < L$. The L-dimensional data $\{X_1, \ldots, X_K\}$ is then projected onto the subspace L^r
- **Averaging** over the diagonals yields some Hankel matrix **X***
- Time series $(x^*_1, \ldots, x^*_N)$, in one-to-one correspondence with **X*** provides an **approximation** either the whole series X_N or a particular component of X_N .

Features of SSA

- Non-parametric and model-free
- Main assumption behind Basic SSA:

  - time series can be represented as a sum of different components such as trend (which we define as any slowly varying series), modulated periodicities, and noise
  - Interpretable components can be approximated by low-rank time series and described via Linear Recurrence Relations
  - Obtaining such components helps interpretation and improve reliability of the analysis

Parameters in Basic SSA

- Window length: L
- Group indices: r

  - Choice of such parameters can be derived by analysis of the results
  - Automatic selection can also be done but it depends on the situations

Applications of SSA

- Forecasting
- Missing value imputation (gap-filling methods)
- Change-point detection
- Density estimation
- Multivariate time series analysis
- Image processing

Where has been applied

- Climatic, meteorological and geophysical time series
- Engineering
- Image processing
- Medicine
- Actuarial sciences
- Predictive maintenance
- Financial and econometric data

The Algorithm

1. Embedding: using the N-L+1 lagged vectors to form the trajectory matrix $\mathbf{XX}^{\mathbf{T}}$
2. Singular value decomposition of the matrix obtained
3. Grouping the components (eigentriples)
4. Reconstruct the series by diagonal averaging (the better the closer components are to independence)

Properties

- Discover automatically:
  - Trends (one o multiple)
  - Oscillatory components (multi-periodic or modulated periodicities)
  - Residual
  - Support natural forecasting methods: Linear recurrent or vector
  - May be used as a pre-processing step for many other analyses:
    - Forecasting
    - Spectral density estimation of specific components
    - Gap filling in time series

Some issues: Separability

- Separability of two time series X(1) and X(2) means the possibility of extracting X(1) from the observed sum X(1) + X(2)
- Time series components can be identified on the basis of the following principle: the form of an eigenvector replicates the form of the time series component that produces this eigenvector
- Graphs of eigenvectors can help in the process of identification
- A sinusoid generates, exactly or approximately, two sine wave components with the same frequency and the phase shift =2
  - Therefore, the scatterplot of a **pair of eigenvectors**, which produces a more or less regular **T-vertex polygon**, can help to identify a sinusoid of period T

Some issues: Separability

- Very helpful information for separation is contained in the so-called w-correlation matrix

  - This is the matrix consisting of weighted correlations between the reconstructed time series components. The weights reflects the number of entries of the time series terms into its trajectory matrix
  - Well-separated components have small correlation whereas badly separated components have large correlation
  - Therefore, looking at the w-correlation matrix one can find groups of correlated elementary reconstructed series and use this information for the consequent grouping. One of the rules is not to include into different groups the correlated components.

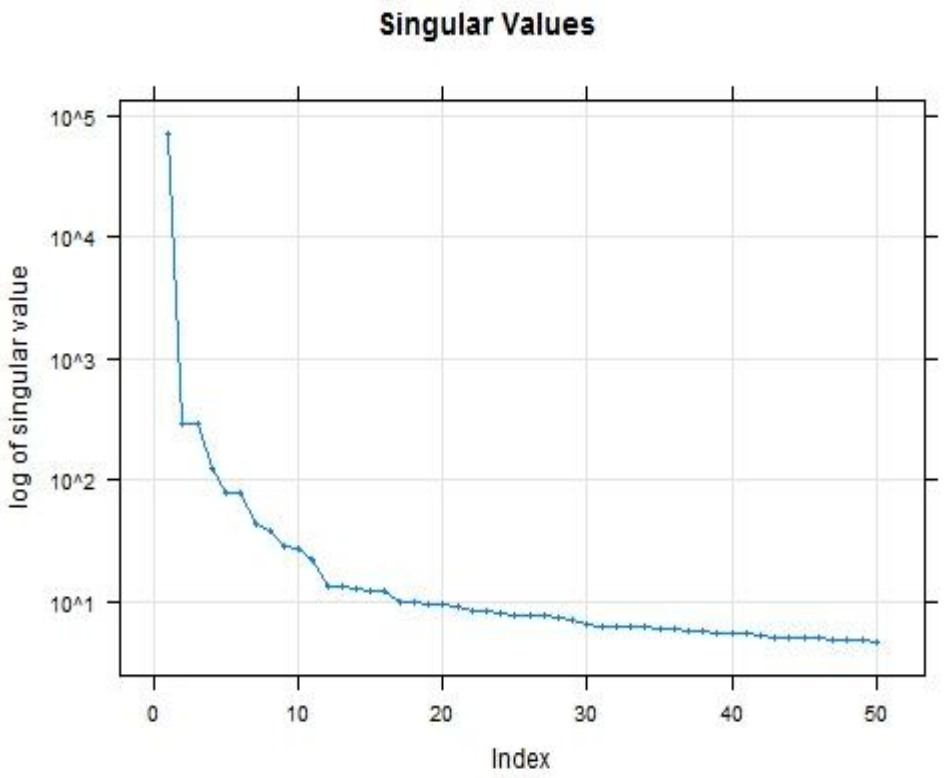
Some issues: Choice of the Embedding

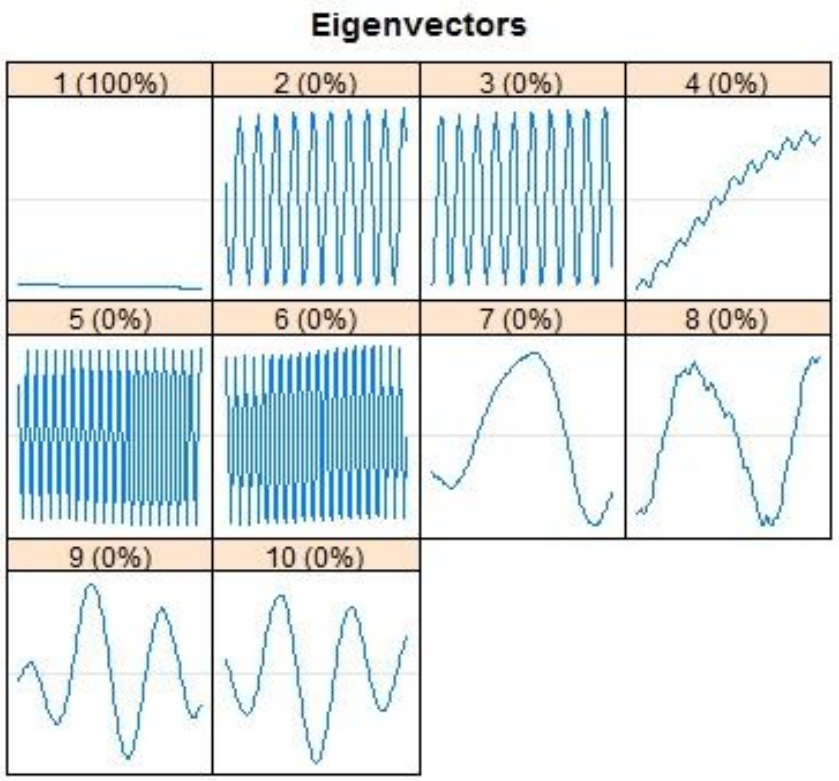
- The embedding L should be large enough (es., L = N/2) and if we want to extract a periodic component with known period, then the window lengths which are divisible by the period provide better separability
- If the time series has a complex structure, the so-called Sequential SSA is recommended. Sequential SSA consists of two stages, at the first stage the trend is extracted with small window length and then periodic components are detected and extracted from the residual with L=N/2

Rssa, SSA in R: first example

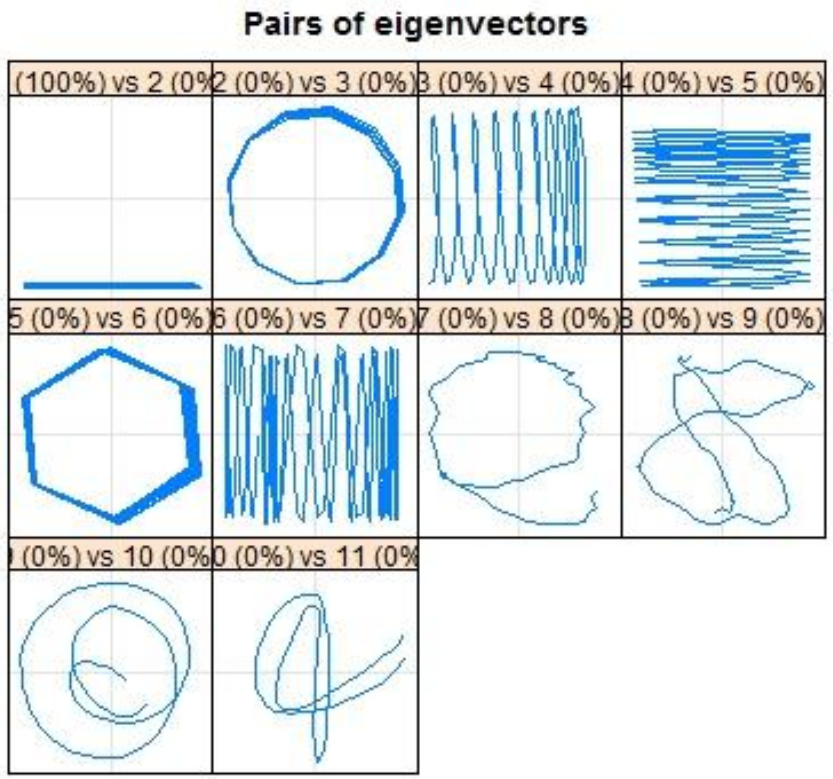
```
library(Rssa)
# Decomposition stage
s <- ssa(co2, L = 120)
plot(s)
plot(s,type="vector")
plot(s,type="paired")
plot(s,type="wcor")
# Reconstruction stage
# Grouping from looking at the W Cor matrix
# The results are the reconstructed series r$F1, r$F2, and r$F3
recon <- reconstruct(s, groups = list(c(1,4), c(2, 3), c(5, 6)))
# Calculate the residuals
res <- residuals(recon)
plot(recon, type = "cumsum")
plot(wcor(s, groups = list(c(1,4), c(2,3), c(5, 6))))
plot(recon)
```

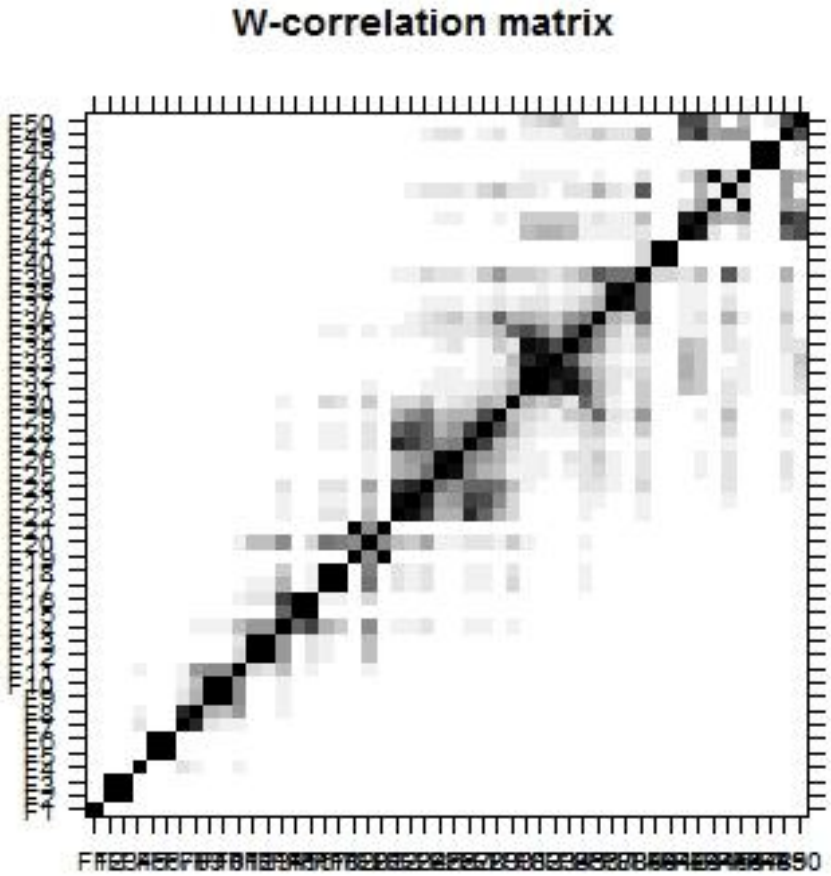
Results


Singular Values
log of singular value
10^5
10^4
10^3
10^2
10^1
0
10
20
30
40
50
Index

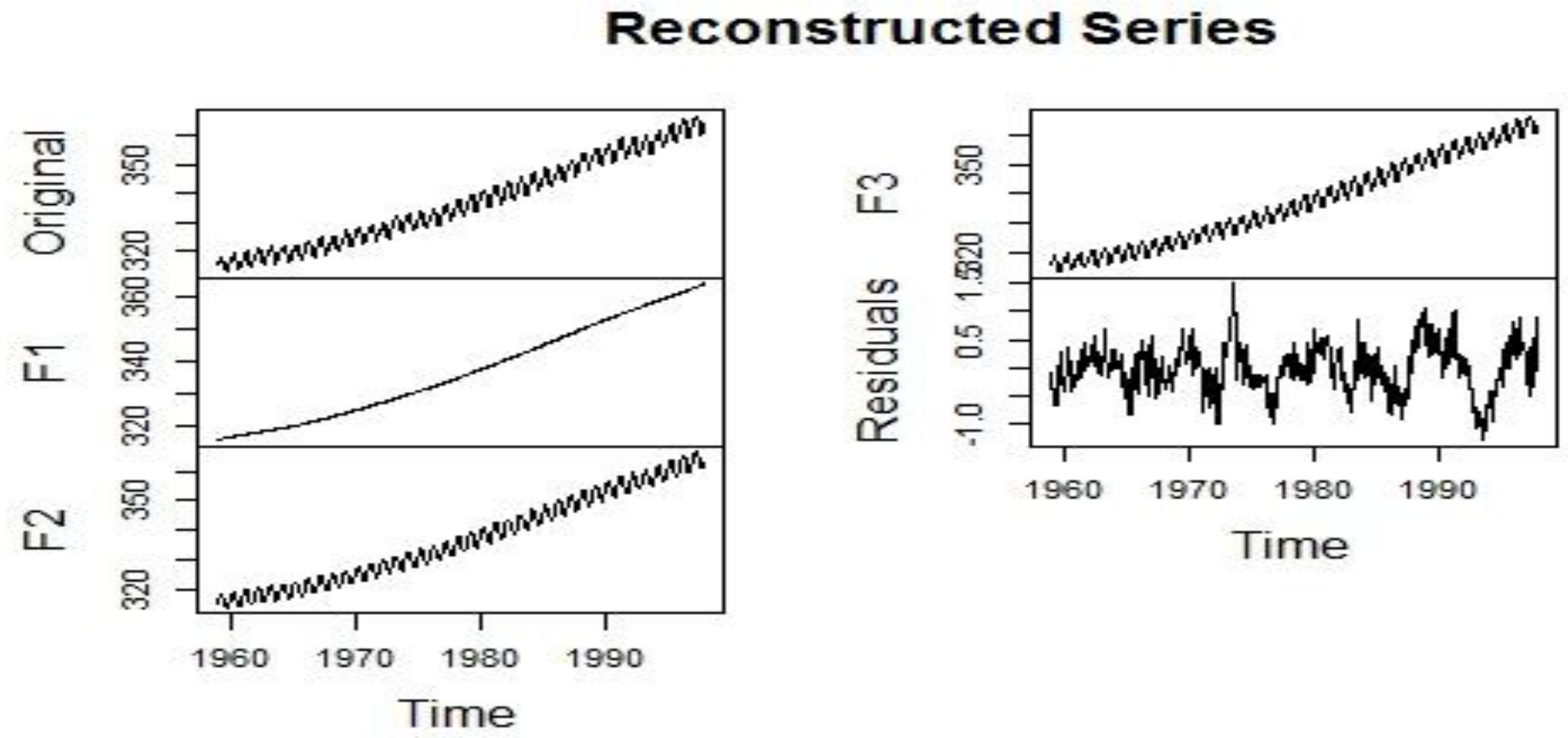

Eigenvectors
1 (100%)
2 (0%)
3 (0%)
4 (0%)
5 (0%)
6 (0%)
7 (0%)
8 (0%)
9 (0%)
10 (0%)

Results

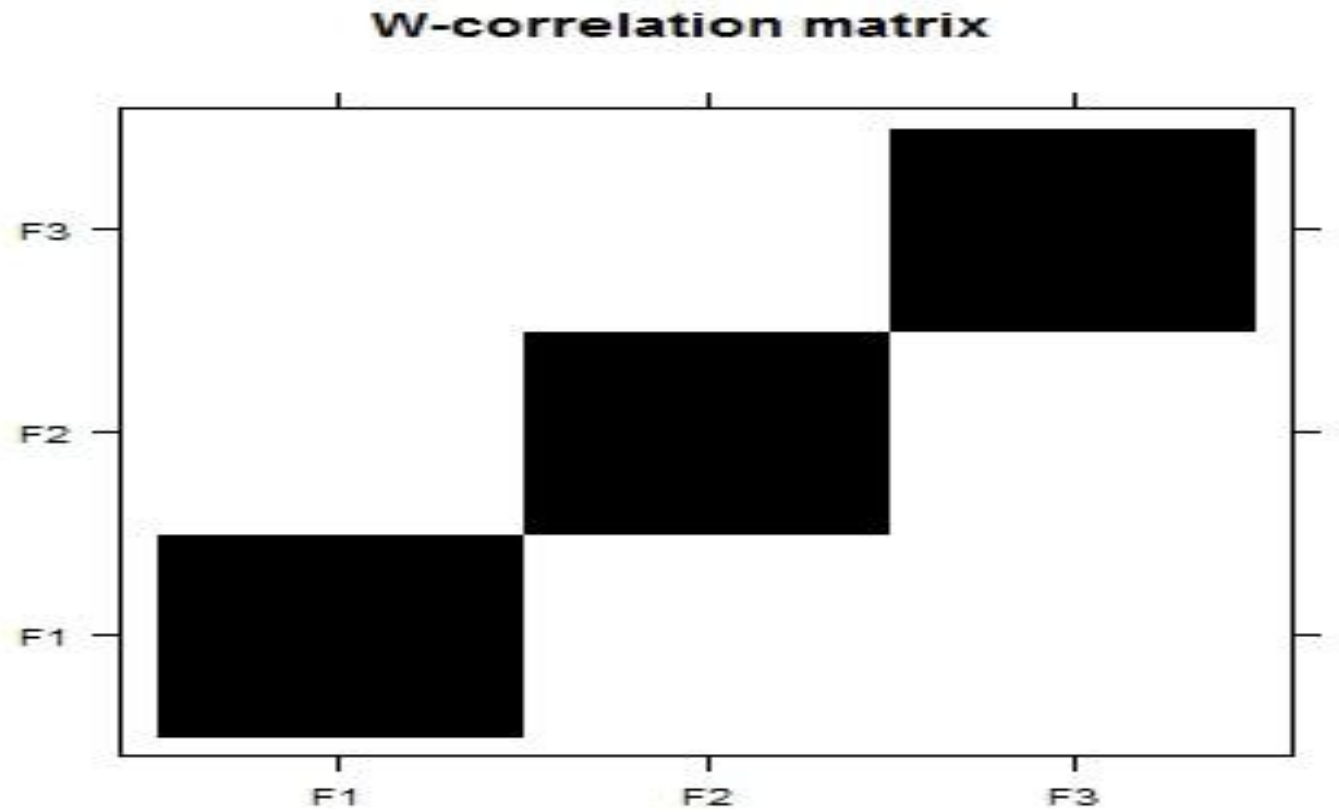

Pairs of eigenvectors
(100%) vs 2 (0%)
2 (0%) vs 3 (0%)
3 (0%) vs 4 (0%)
4 (0%) vs 5 (0%)
5 (0%) vs 6 (0%)
6 (0%) vs 7 (0%)
7 (0%) vs 8 (0%)
8 (0%) vs 9 (0%)
(0%) vs 10 (0%)
0 (0%) vs 11 (0%


W-correlation matrix

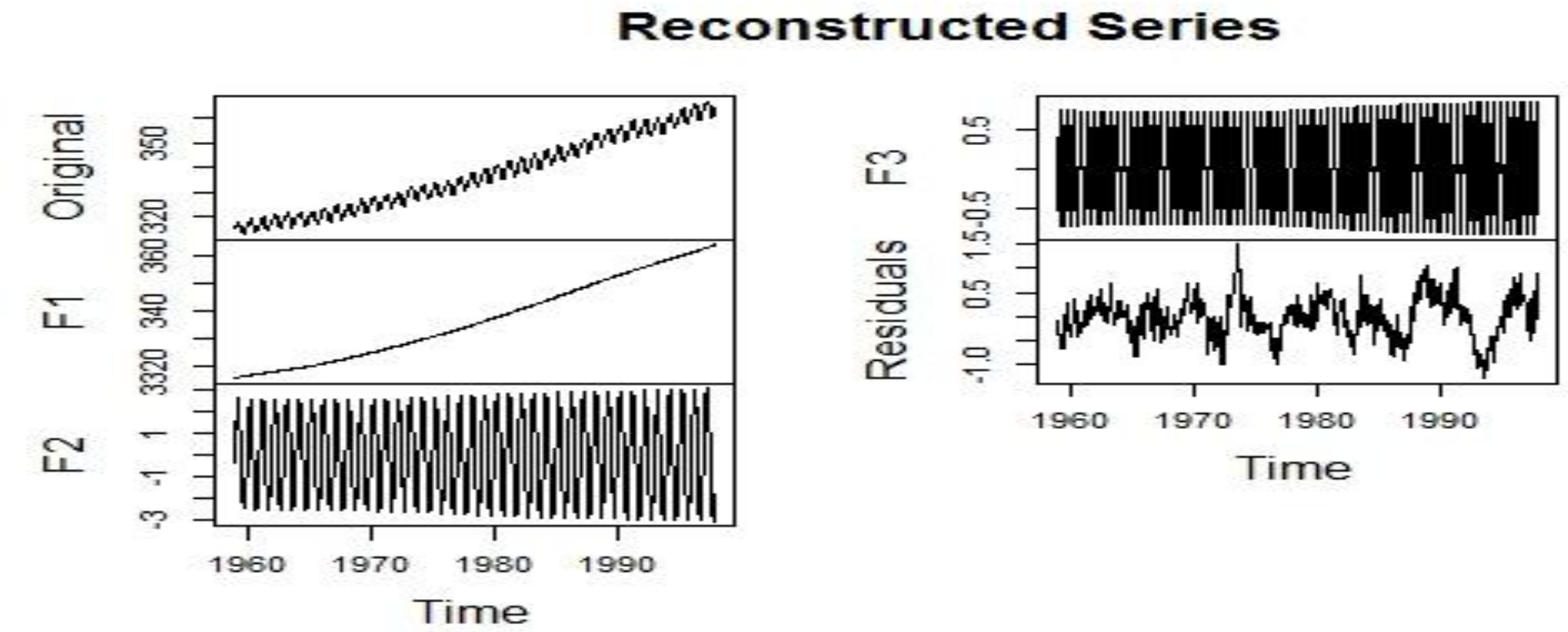
Results: reconstructed series, cumulative view



Results: Wcor reconstructed series



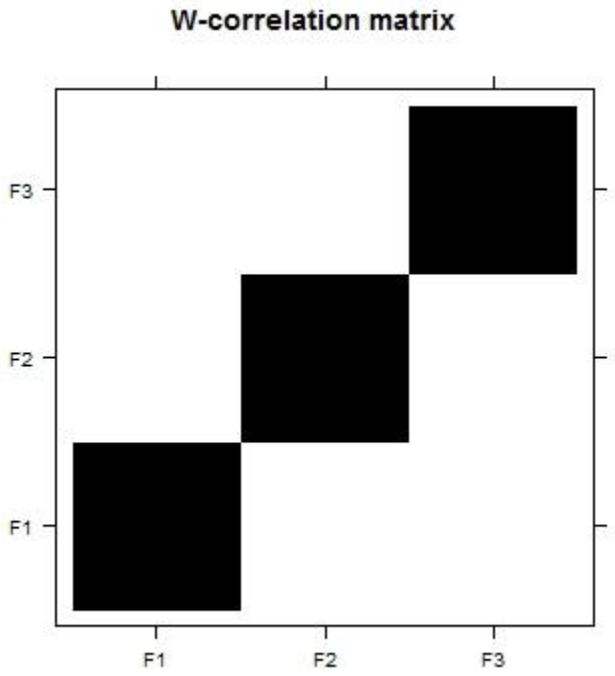
Results: reconstructed series, components

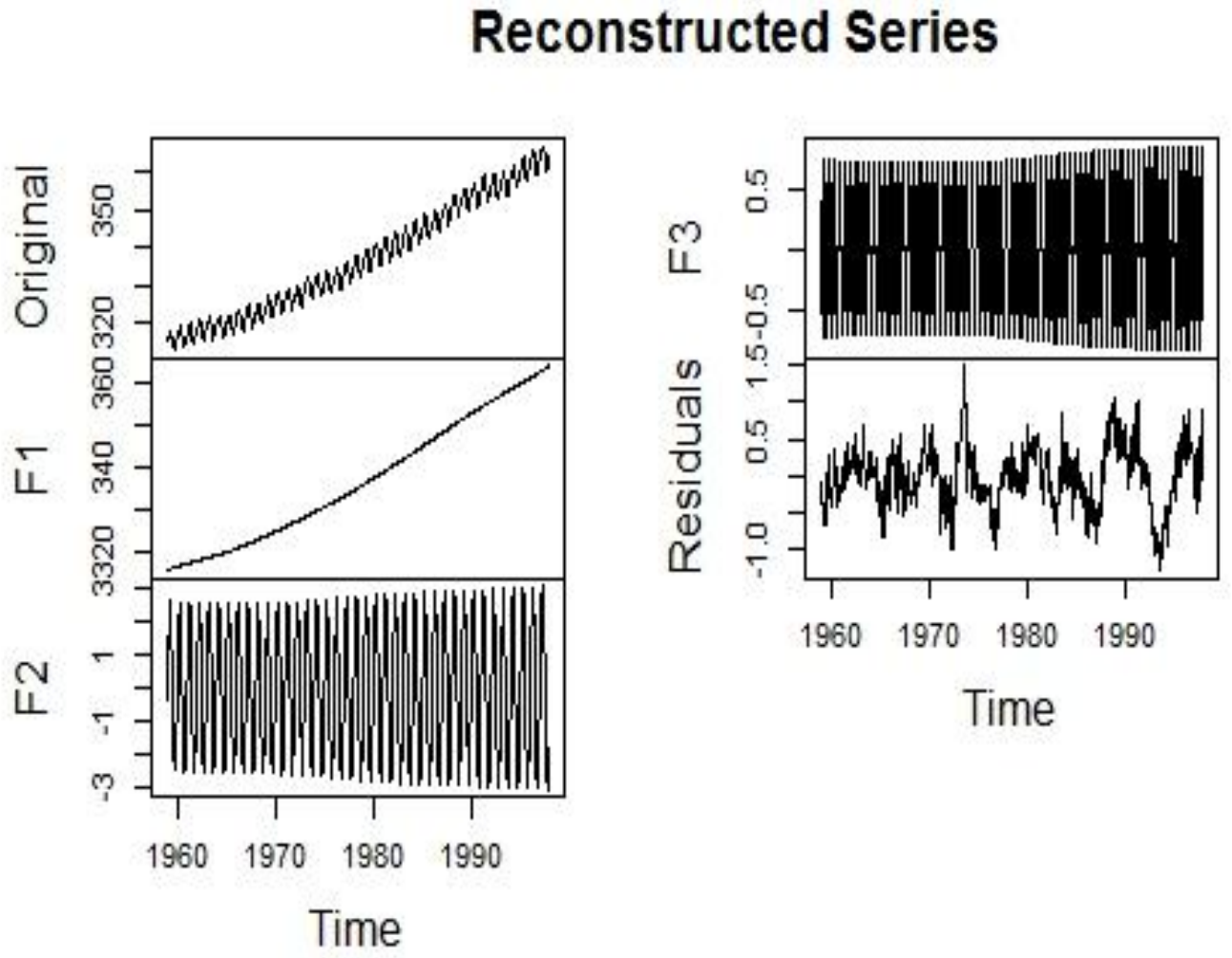


Another grouping method

```
# Alternative grouping, using cluster analysis on the eigen triples
lst<-clusterify(s,group = 1:6, nclust=3)
lst
r <- reconstruct(s, groups = list(c(1), c(2, 3, 4), c(5, 6)))
plot(wcor(s, groups = list(c(1), c(2,3, 4), c(5, 6))))
plot(r)
```

Results


W-correlation matrix
F3
F2
F1
F1
F2
F3

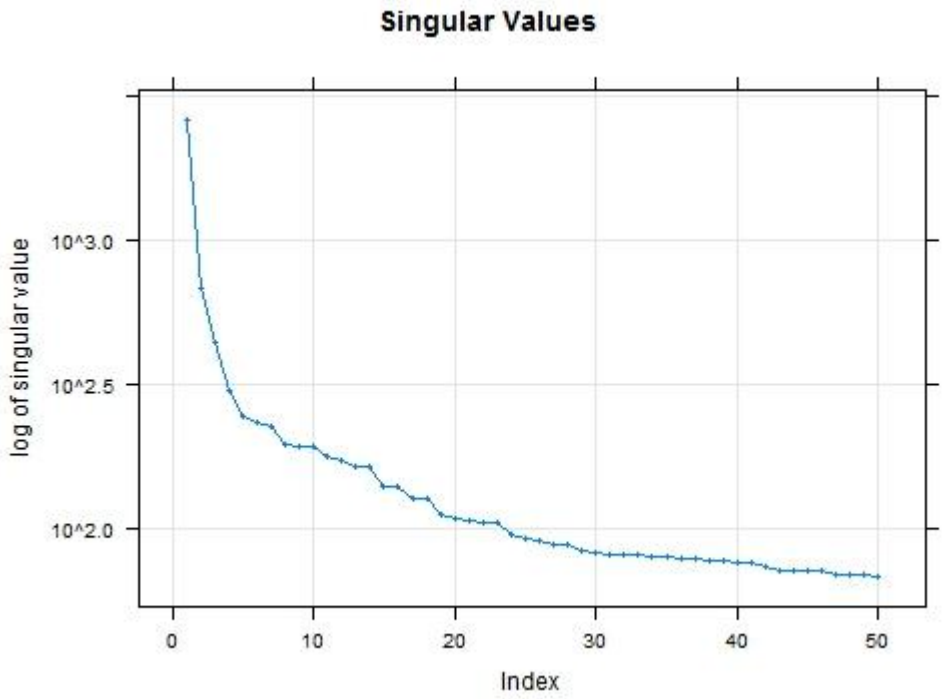

Reconstructed Series
Original
F1
F2
350
320
360
340
3320
1
-1
-3
1960
1970
1980
1990
Time
F3
Residuals
0.5
-0.5
1.5
0.5
-1.0
1960
1970
1980
1990
Time

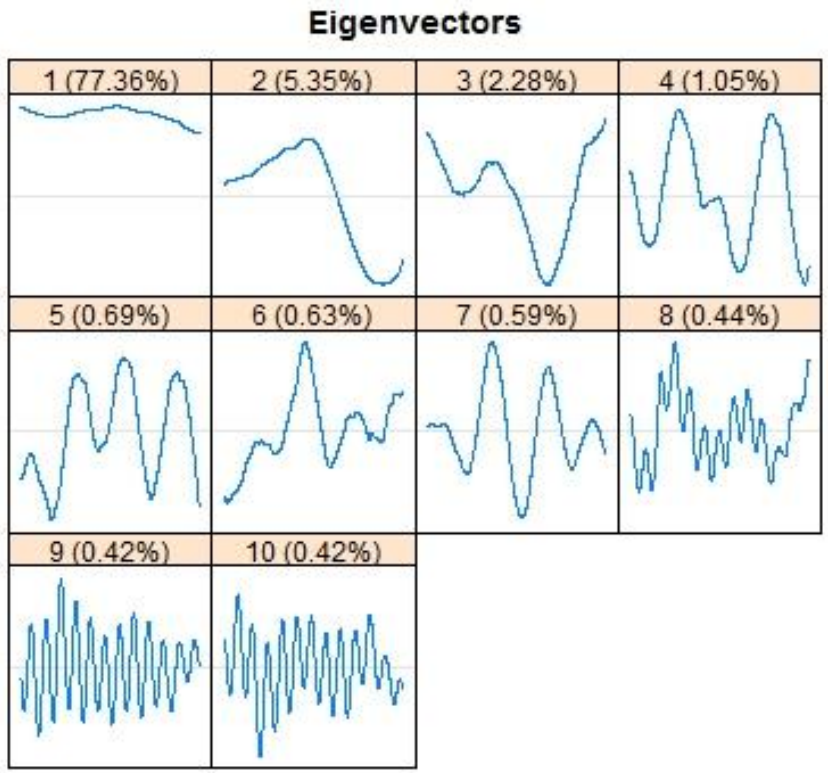
Comparison with wavelets

- Wavelets
- PROS
  - Naturally multiscale
- CONS
  - Choice of the basis
  - Choice of boundary conditions (reflexive / periodic)
  - Choice of the levels in the decomposition
  - Depth of the wavelet decomposition
  - Length of the testing window

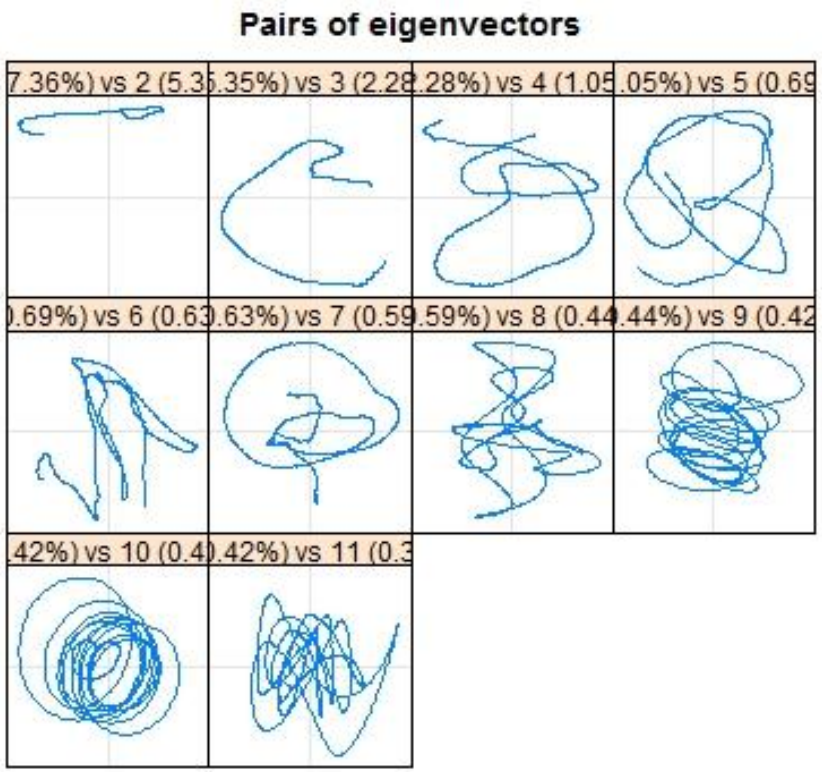
- SSA
- PROS
  - Single parameter (window length)
  - Adaptive / data-driven basis (EOF=empirical orthogonal functions)
- CONS
  - Not naturally multiscale (but can be adapted)

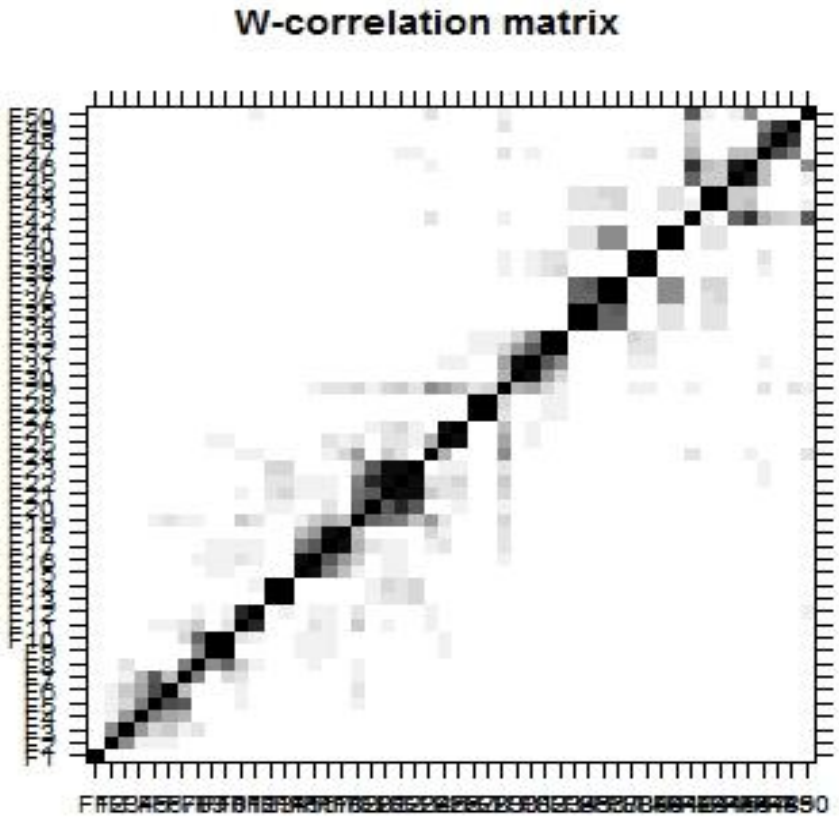
Another example: Power consumption data



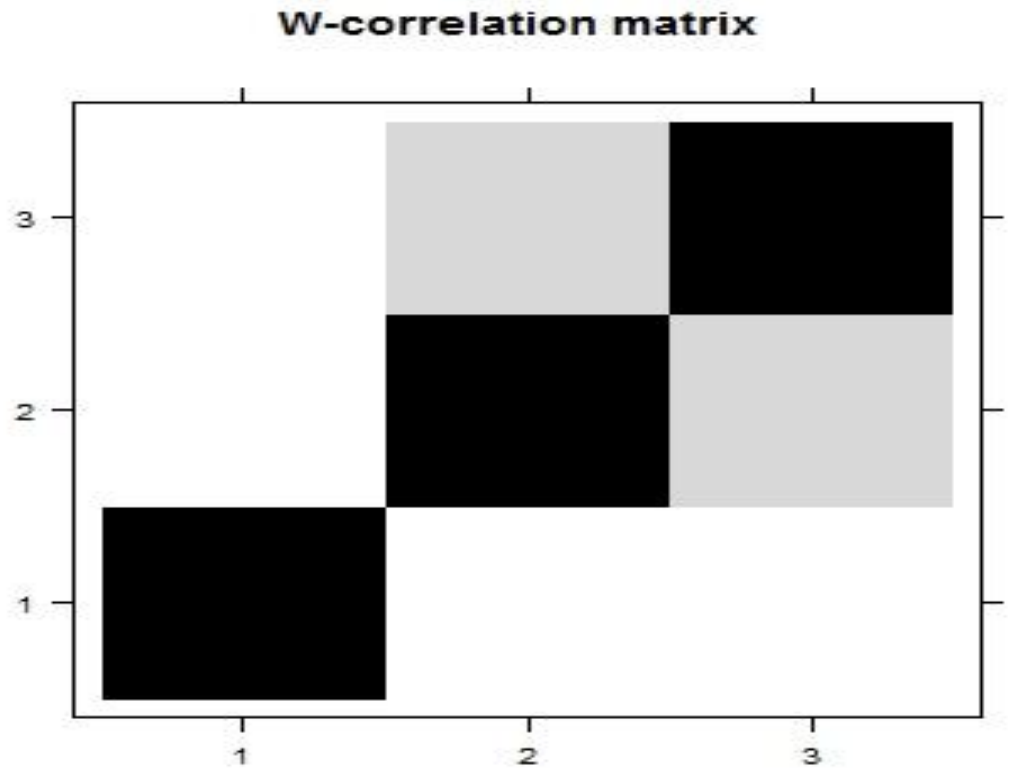


Another example: ctd.

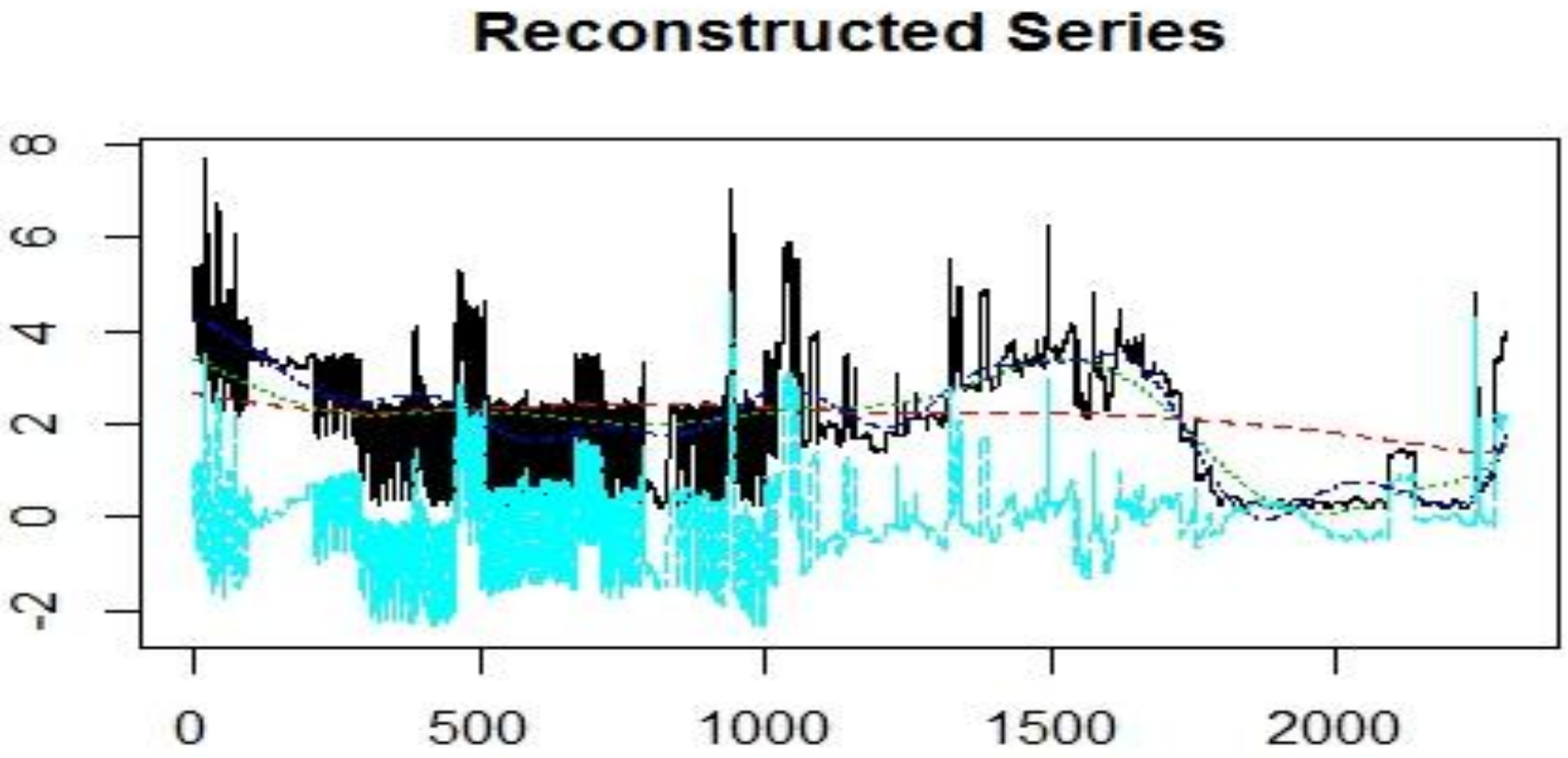




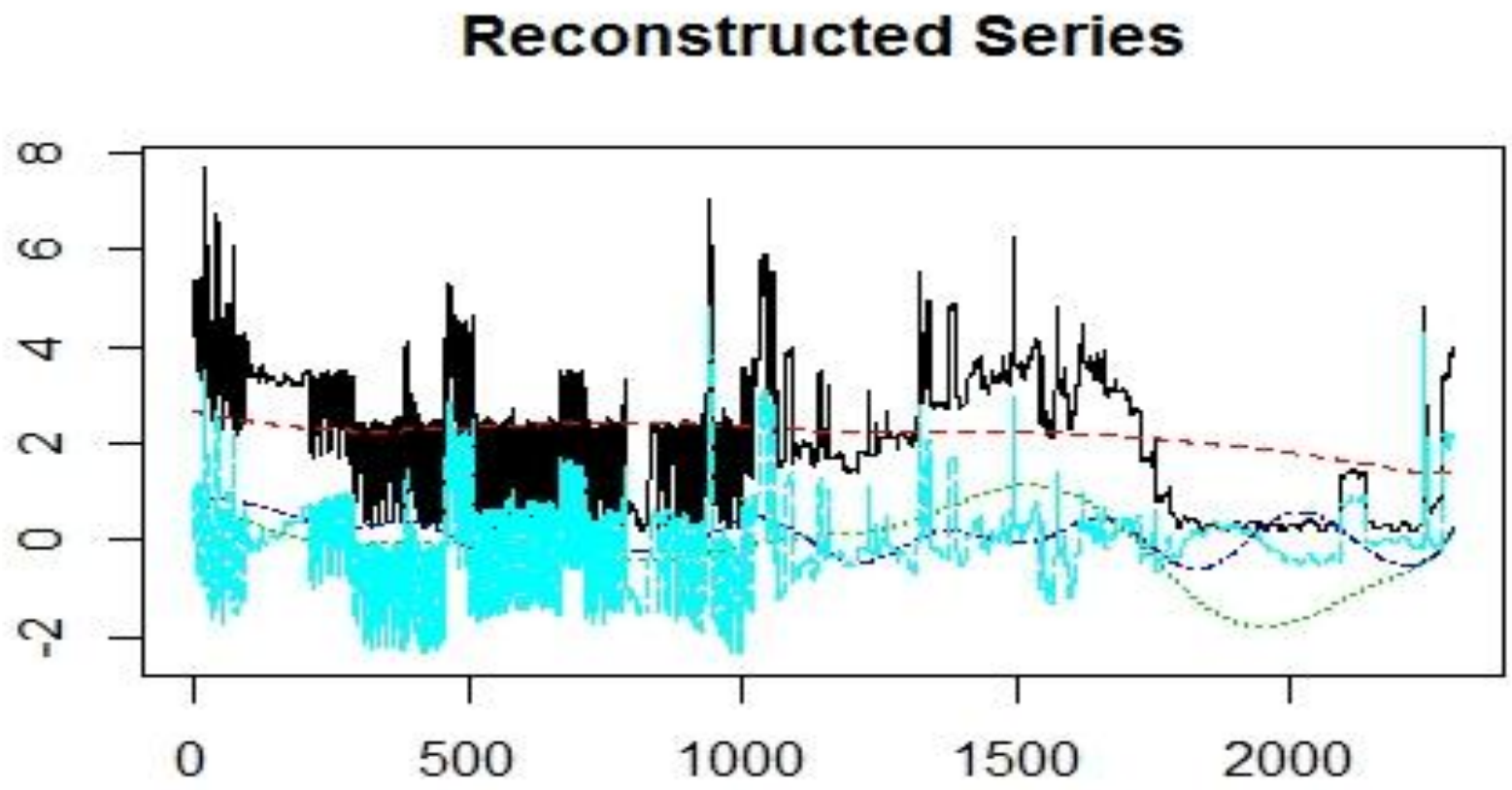
Another example: ctd.



Another example: ctd., cumulative view



Another example: ctd., component view



Forecasting: Co2 series

```
s <- ssa(co2)

f <- forecast(s, groups = list(1:6), method = "bootstrap-recurrent", len = 24, R = 10)

# Plot the result including the last 24 points of the series

plot(f, include = 24, shadecols = "green", type = "l")
```

Forecasting: the CO2 series