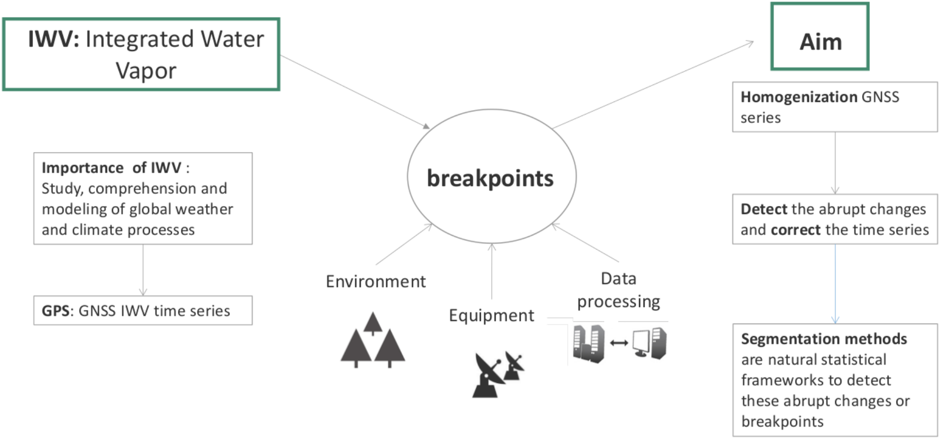
A new homogenization method for the atmospheric GNSS data

Annarosa Quarello,[1,2]
Olivier Bock[1],
Emilie Lebarbier[2]
IPGP, IGN, ENSG, Universite Paris Diderot, Sorbonne Paris,[1]
AgroParisTech[2]

April 18, 2019

Outline

GNSS Time Series and Objective



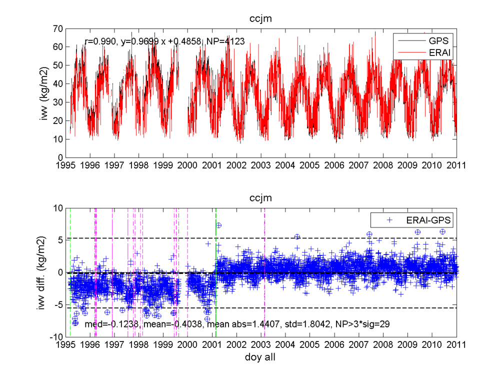
Study of IWV

1. Detection problem: finding break-points in the IWV time series

2. Adjustment problem: correcting the time series

3. Climate analysis on the clean IWV series

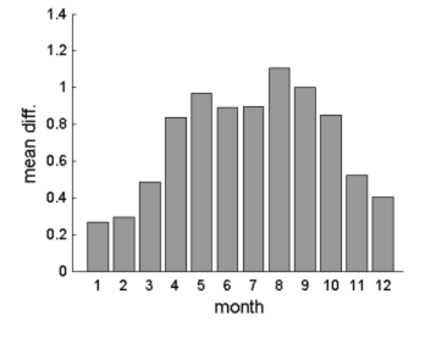
Example of Such Data

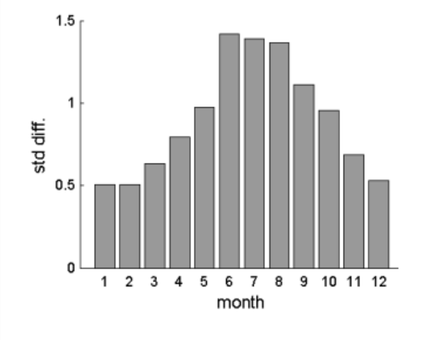


- Natural seasonal tendency
- ERAI : meteorological reanalysis
- $\Delta IWV = IWV_{ERAI} - IWV_{GPS}$

Features of the Data

Monthly mean and variance for the station CCJM

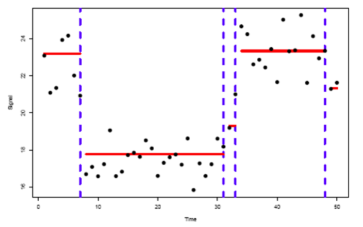




Features of the data.

1. a periodic signal remains
2. a non-stationary of the variability is observed

Breakpoint detection in the mean model with homogeneous variance



Example of segmentation with homoscedastic model

Model. The signal $Y = (Y_1 \ldots, Y_n)$ is such that $Y_t = \mu_k + E_t$,

$$\forall t \in I_k = [t_{k-1} + 1; t_k] \text{ where}$$

$$\{E_t\}_t \quad are \quad i.i.d. \sim \mathcal{N}(0, \sigma^2).$$

Parameters. The $K - 1$ breakpoints $T = \{t_1, \ldots, t_{K-1}\}$, the means and the variance $\theta = \{\mu_1, \ldots, \mu_k, \sigma^2\}$ and the number of segments K.

Classical Inference Strategy: two steps

1. K being fixed, estimation of T and θ by maximum likelihood
   - ⋆ There are $\binom{n-1}{K-1}$ possible segmentations of n points into K segments
   - ⋆ Dynamic Programming (DP) algorithm gives the exact solution in a reasonable complexity time $O(Kn^2)$
   - ⋆ DP can be applied if and only if the quantity to be optimized is additive with respect to the segments

2. Choice of K. Model selection issue: Penalized log-Likelihood.

A New Segmentation Model

$$Y_t = \mu_k + f_t + E_t,$$

$$\forall t \in I_k = [t_{k-1}+1; t_k] \cap I_{\text{month}} = \{t; \text{date}(t) \in \text{month}\}$$

where

- ⋆ f_t will be approximated using a Fourier decomposition of order $4 \rightarrow$ to take into account for the periodic signal

$$f_t = \sum_{i=1}^{4} \left(a_i \cos(i2\pi t/L) + b_i \sin(i2\pi t/L)\right) \quad \text{with } L \text{ the length of a year}$$

- ⋆ $\{E_t\}_t \quad i.i.d. \sim \mathcal{N}(0, \sigma^2_{month}) \rightarrow$ to take into account for the monthly variability

Proposed Inference Strategy $\rightarrow$ to allow the use DP

1. Estimation of the variances σ^2_{month} using a robust approach (robust to the breakpoints) $\rightarrow \hat{\sigma}^2_{month}$

2. Segment the signal with 'known variances'

   ⋆ K being fixed, an iterative procedure, at iteration $[h+1]$:

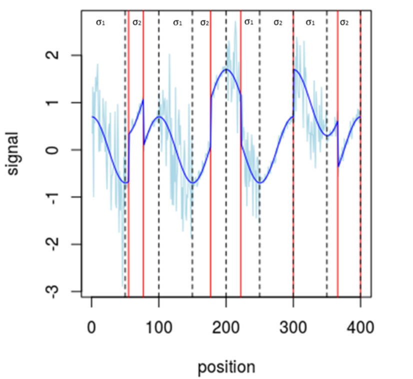
     – Estimation of f on $\tilde{y}_t = y_t - {\mu_k}^{[h]}$ using a weighted least square regression with weights $1/\hat{\sigma}^2_{month}$

     – Estimation of t_k, μ_k on $\tilde{y}_t = y_t - f_t^{[h+1]}$ using DP

   ⋆ Choice of K by adapting model selection criteria (BM , Lav and mBIC)

Simulation Design

- ⋆ $n = 400$
- ⋆ $\sigma_1 = (0.1, 0.5, 0.9)$
- ⋆ σ_2 from 0.1 to 1.5 (by step 0.2)
- ⋆ $f_t = 0.7\cos(\frac{2\pi t}{100})$
- ⋆ $K = 7$
- ⋆ $T = [55, 77, 177, 222, 300, 366]$
- ⋆ $\mu = [0, 1, 0, 1, 0, 1, 0]$



Example of simulated data

Quality Criteria

- ⋆ The difference between the predicted number of segments and the true number one: $\hat{K} - K_{True}$

- ⋆ The root mean squared error between the function f and its estimate :

$$\text{RMSE}(f) = \left[\frac{1}{n} \sum_{t=1}^{n} \left\{ f_t - \hat{f}_t \right\}^2 \right]^{1/2}$$

- ⋆ The root mean squared error between the mean μ and its estimate : $\text{RMSE}(\mu)$

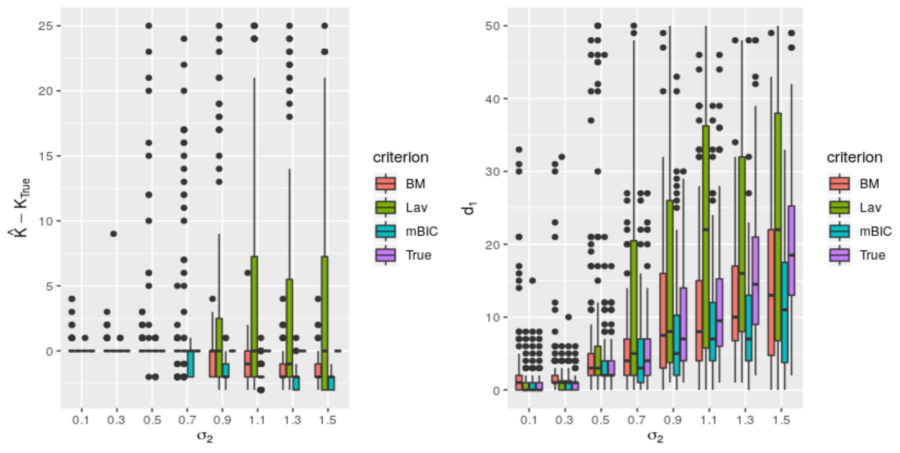
- ⋆ Two components of Hausdorff distance: $d_1(\hat{T}, T_{True})$ and $d_2(\hat{T}, T_{True})$, where

$$d_1(a, b) = \max_b \min_a |a - b|, d_2(a, b) = d_1(b, a)$$

$\rightarrow$ A perfect segmentation results in $d_1 = d_2 = 0$

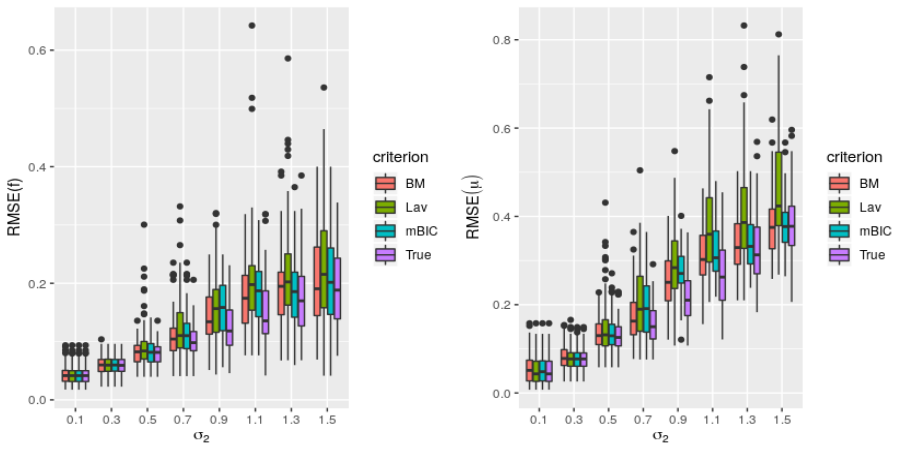
Simulation Results

Results for $\sigma_1 = 0.5$ and $\sigma_2 = (0.1, 0.3, \ldots, 1.5)$



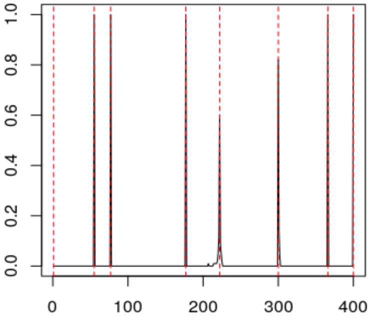
Simulation Results

Results for $\sigma_1 = 0.5$ and $\sigma_2 = (0.1, 0.3, \ldots, 1.5)$



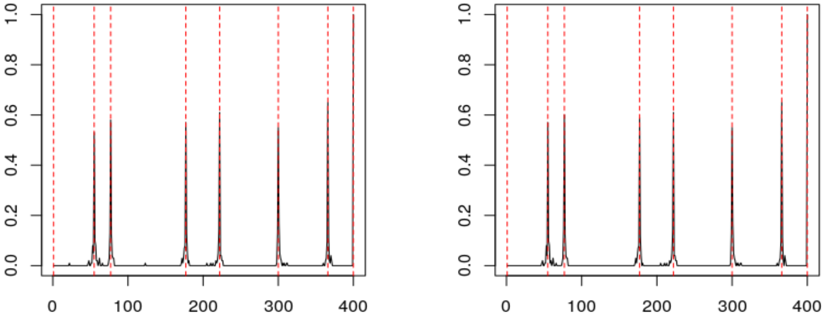
Simulation Results

Breakpoints positioning. Frequencies of all possible breakpoints estimator for $\sigma_1 = 0.5$ and $\sigma_2 = 0.1$ for K_{mBIC} and K_{True}



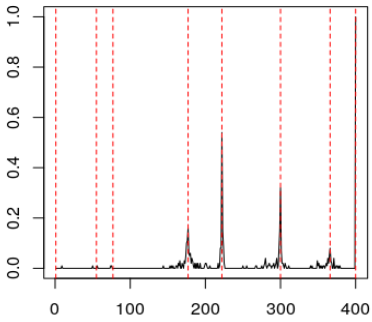
Simulation Results

Breakpoints positioning. Frequencies of all possible breakpoints estimator for $\sigma_1 = 0.5$ and $\sigma_2 = 0.5$ for K_{mBIC} and K_{True}



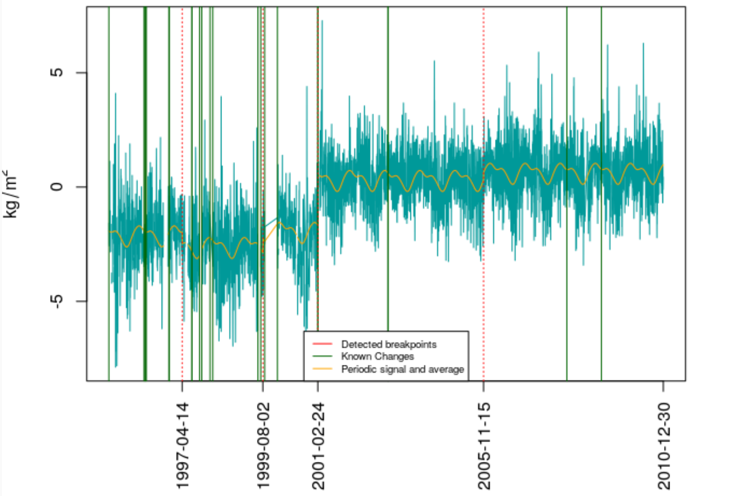
Simulation Results

Breakpoints positioning. Frequencies of all possible breakpoints estimator for $\sigma_1 = 0.5$ and $\sigma_2 = 1.5$ for K_{mBIC} and K_{True}



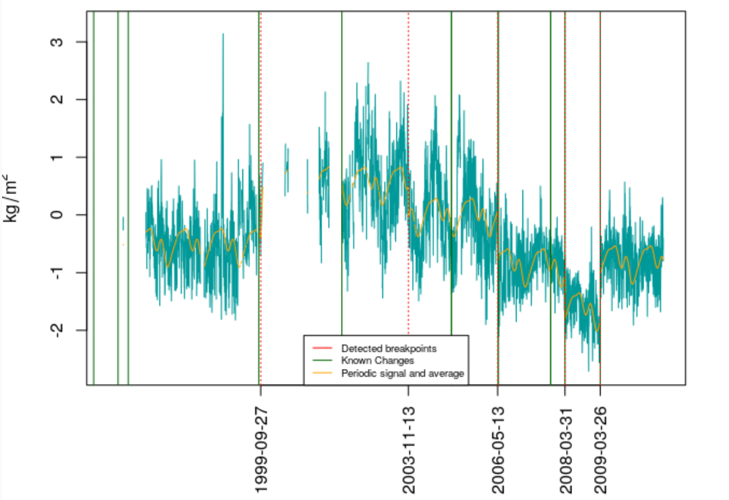
Application of Real Data

Station CCJM with selection criterion BM



Application of Real Data

Station MCM4 with selection criterion BM



Conclusion and Further Developments

Conclusion

1. We proposed a segmentation method for the analysis of GNSS time series: a breakpoint detection in the mean model in which
   - the variance is supposed to be month-dependent
   - a periodic signal is taken into account
2. We developed an R Package, **GNSSseg** available on GitHub: https://github.com/arq16/GNSSseg.git
3. Two articles on the way (Bock in press and Quarello in preparation)

Further developments

- Improvement of the estimation of f: non-parametric approach
- Take into account for a time-dependence that may exist on time series

References

[1] Auger I.E. and Lawrence C.E. (1989) Algorithms for optimal identification of segments neighborhoods, Bull. Math. Biol. 51, 325-337.
[2] Bertin K., Collilieux X., Lebarbier E. and Meza C.(2017) Semi-parametric segmentation of multiple series using a DP-Lasso strategy, Journal of Statistical Computation and Simulation, vol. 87, No. 6, p 1255-1268.
[3] Bock O., Willis P., Lacarra M., Bosser P. (2010) An inter-comparison of zenith tropospheric delays derived from DORIS and GPS data, Adv. Space Res. 46(12), 1648-1660.
[4] Chakar S., Lebarbier E., Lévy-Leduc C. and Robin S. (2017) A robust approach to multiple change-point estimation in an AR(1) process, Bernoulli, vol. 23, No. 2, p 1408-1447.
[5] Lavielle M.(2005) Using penalized contrasts for the change-point problem, Signal Processing 85 (8), 1501-1510.
[6] Lebarbier E.(2005) Detecting multiple change-points in the mean of Gaussian process by model selection, Signal Processing 85,717-736.
[7] Rousseeuw P. J. and Croux C. (1993) Alternative to the Median Absolute Deviation, Journal of the American Statistical Association 88 1273-1283.
[8] Vey S., Dietrich R. ,Fritsche M., Rülke A., Steigenberger P., Rothacher M.(2009) On the homogeneity and interpretation of precipitable water time series derived from global GPS observations, Journal of geophysical research.
[9] Zhang N.R. and Siegmund D.O. (2007) A Modified Bayes Information Criterion with Applications to the Analysis of Comparative Genomic Hybridization Data, Biometrics 63 (1), 22-32.

Merci!

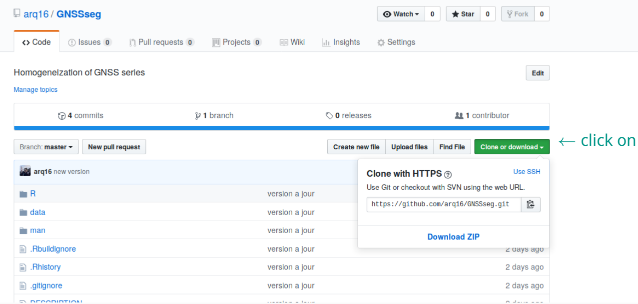
outline

GNSSseg R Package

How to install an R Package from GitHub?

```
library(devtools)
devtools::install_github("https://github.com/arq16/GNSSseg.git")
```

- https://github.com/arq16/GNSSseg.git



← click on "Clone or download"