Homogenization of GNSS-derived IWV time series

Annarosa Quarello[1,2], Olivier Bock[1], Emilie Lebarbier[2]

[1]IGN, LAREG,Univ Paris Diderot, Sorbonne Paris Cité, Paris, France
[2]UMR 518 AgroParisTech / INRA Math. Info. Appli., Paris, France
annarosa.quarello@agroparistech.fr

This presentation participates in OSPP
Outstanding Student Poster & PICO Contest

Introduction

The **integrated water vapor** (IWV), one of the essential climate variables, plays a significant role in climate change and global warming. Ground-based networks of Global Navigation Satellite System (GNSS) receivers provide observations of tropospheric delay and IWV for more than 20 years. However, the time series of GNSS-derived IWV have been affected by abrupt changes (**breakpoints**) due to equipment or environmental changes. The homogenization of such series is a crucial step before any interpretation or analysis [1][2].

Objective

Detect breakpoints in the mean of time series of the difference between IWV from GNSS and IWV from ERA-Interim reanalysis (ΔIWV) with a **new segmentation method** for taking in account the features of ΔIWV: periodic signal and non-stationary variance.

Model and inference procedure

Model

Let be $y = \{y_t\}_{t=1,\dots,n}$ modeled by a gaussian indipendent random process $Y = \{Y_t\}_{t=1\dots,n}$
A segmentation model, in which the **variance** is supposed to be **monthly-dependent** is developed:

$$Y_t = \mu_k + f_t + E_t, \tag{1}$$

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

- f_t is a Fourier decomposition of order 4,

$$f_t = \sum_{i=1}^{4} a_i cos(i\frac{2\pi}{L}t) + \sum_{i=1}^{4} b_i sin(i\frac{2\pi}{L}t) \tag{2}$$

  - L is the length of a year
- $E_t \;\; i.i.d. \sim \mathcal{N}(0, \sigma^2_{month})$

Inference strategy.

For a fixed K, at iteration $[h+1]$:

- estimate f_t on $\tilde{Y}_t = Y_t - \mu_k^{[h]}$
- estimate T, μ and σ^2_{month} on $\tilde{Y}_t = Y_t - f_t^{[h]}$

The number of breakpoints or segments K is chosen using three model selection criteria denoted mBIC [3] , ML [4] and BM [5]

Application on synthetic data

Simulated data.

The simulated series has length of 400 days, 4 years with 2 months by year with stardard deviation σ_1 and σ_2 respectively.

- σ_1 is fixed to 0.6 and σ_2 varies from 0.1 tp 1.5 by step of 0.1.
- the series is affected by $K-1=6$ breakpoints
- the considered function is $f_t = 0.7cos(2\pi t/100)$.

Each configuration (each value of σ_2) is simulated 100 times.

Benchmark dataset.

Daily synthetic data of 120 GNSS stations, created by the COST Action ES1206 (GNSS4SWEC).The length of the series is 6000 days (16.4 years), they include: abrupt changes, seasonal signals and white noise.

Quality criteria

$\star$ $K - \hat{K}$, the difference between true number of segments and the estimated one

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

$\star$ the two components of the Hausdorff distance $d_1(T^\star, \hat{T})$ and $d_2(T^\star, \hat{T})$ where $d_1(a,b) = \max_b \min_a |a-b|$ and $d_2(a,b) = d_1(b,a)$,

$T^\star$ and $\hat{T}$ are the true and estimated breakpoints respectively. A perfect segmentation results in both null d_1 and d_2.

Results

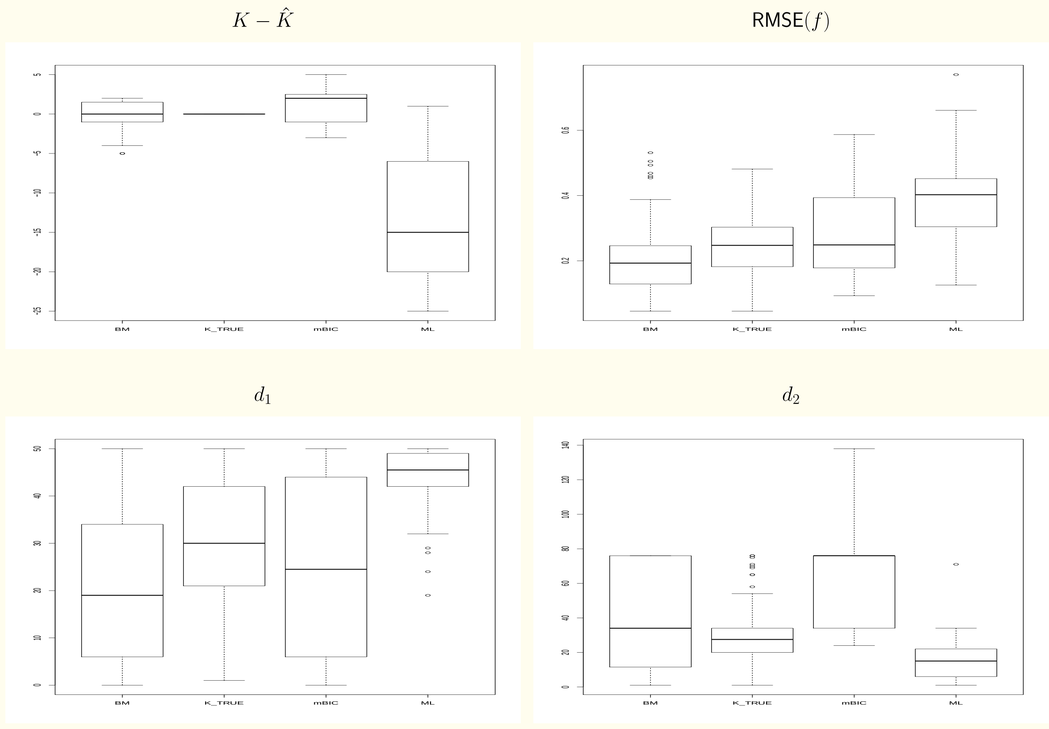

Fig 1. Simulated data with $\sigma_2 = 0.7$. Boxplot of the different quality criteria for all the stations and the true K, the selected one with BM, mBIC and ML respectively (x-axis).

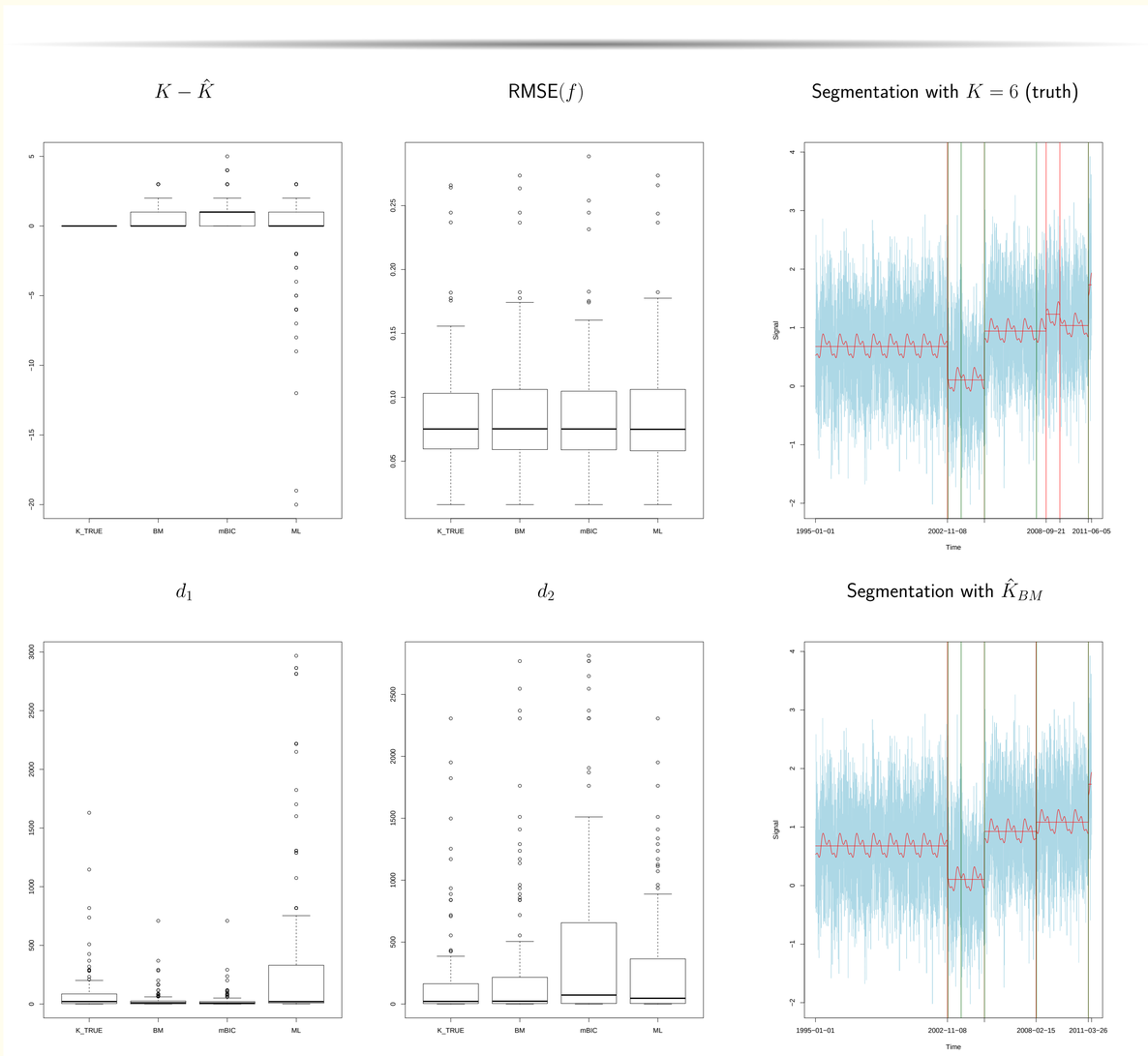

Fig 2. Benchmark data Left: boxplot of the different quality criteria for all the stations and the true K, the selected one with BM, mBIC and ML respectively (x-axis).
Right: obtained segmentation for the particular station POTS with $\hat{K}_{true}$ (top) and $\hat{K}_{BM}$ (bottom). Vertical green lines: true breakpoints; red: estimated mean and breakpoints. The true amplitudes at each true break are $-0.544, -0.004, 0.802, 0.149, 0.690$ respectively.

Conclusions

- The proposed procedure tends to **underestimate** the number of breakpoints. This result was expected since in this case one may prefer to avoid false detections, as generally observed in segmentation problems.
- The criterion of mBIC tends to underestimate more the number of breakpoints, ML to overestimate and not correctly. BM seems to be more appropriate.
- The form of f used in the model is in agreement with the one of the benchmark.

Work in progress:

- Improving the estimation of the function f,
- Taking into account for a time-dependency that may exists in real data,
- Developing **R-package** GNSSseg.

[1] 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.
[2] Bock O., P. Willis, M. Lacarra, P. Bosser (2010) *An inter-comparison of zenith tropospheric delays derived from DORIS and GPS data*,Adv. Space Res. 46(12), 1648-1660.
[3] 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.
[4] Lavielle M.(2005) *Using penalized contrasts for the change-point problem*, Signal Processing 85 (8), 1501-1510.
[5] Lebarbier E.(2005) *Detecting multiple change-points in the mean of Gaussian process by model selection*,Signal Processing 85,717-736.