Homogenization of GNSS-derived IWV time series.

Annarosa Quarello(1), Olivier Bock(1), Emilie Lebarbier(2)

(1) IGN, LAREG,Univ Paris Diderot, Sorbonne Paris Cité, Paris, France (2) AgroParisTech, UMR MIA 518, Paris, France

annarosa.quarello@agroparistech.fr

1. Introduction

The atmospheric water vapor, usually referred to as the integrated water vapor (IWV), is one of the essential climate variables and 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 integrated water vapor (IWV) for more than 20 years. However, the time series of GNSS-derived IWV have been affected by abrupt changes (called breakpoints) due to equipment or environmental changes. The homogenization of such series is a crucial step before any interpretation or analysis [1][2]. For this purpose, segmentation methods are a natural framework. In order to detect correctly these changes, it is necessary to take into account for the specificities of the data. First, the series show high atmospheric variability. To remove it, we propose to work on the corrected series with the ERA-interim analysis denoted by ΔIWV. Despite of this correction, a periodic signal remains in the series. Moreover, it has been observed a non-stationary variance that seems to be month-dependent. Here a new segmentation method is proposed: a change-point detection in the mean model in which a functional is integrated and the variance is month-dependent. The proposed method is tested on a synthetic dataset of 120 stations created by members of a COST action GNSS4SEC.

2. Model and inference procedure

Model.

$$Y_t = \mu_k + f_t + E_t = \mu_k + \sum_{k=1}^{4} a_i cos(i\frac{2\pi}{T}t) + \sum_{k=1}^{4} b_i sin(i\frac{2\pi}{T}t) + E_t, \quad \forall t \in I_k = [t_{k-1}+1, t_k] \cap I_{month} = \{t, date(t) \in month\}, \tag{1}$$

where E_t $i.i.d.$ $\sim \mathcal{N}(0, \sigma^2_{month})$, $T = 365.25$ and $k = 1, \ldots, K$. The parameters to be estimated are the number of breakpoints $K-1$, the $K-1$ breakpoints $T = (t_k)_k$, the K means $\mu = (\mu_k)_k$, the variances $(\sigma^2_{month})_{month}$ and the coefficients $(a_i)_i$ and $(b_i)_i$.

Inference strategy. For a fixed K, an iterative strategy is proposed. At iteration $[h+1]$:

1. estimate f_t on $\tilde{Y}_t = Y_t - \mu_k^{[h]}$, using weighted least-squares with weights $1/\sigma^{2,[h]}_{month}$
2. estimate T, μ and σ^2_{month} on $\tilde{Y}_t = Y_t - f_t^{[h]}$ as follows: (i) estimate robustly σ^2_{month} based on [4] and (ii) segment the signal using a Dynamic Programming algorithm [3]

The number of breakpoints or segments K is chosen using three model selection criteria denoted mBIC [5] , ML [6] and BM [7]

3. Application on synthetic data

The developed method is applied on a benchmark dataset created by the COST Action ES1206 (GNSS4SWEC).

Dataset: daily synthetic data of 120 GNSS stations that include abrupt changes, seasonal signals (annual, semi-annual, 3 & 4 months) and a white noise.

Quality criteria:

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

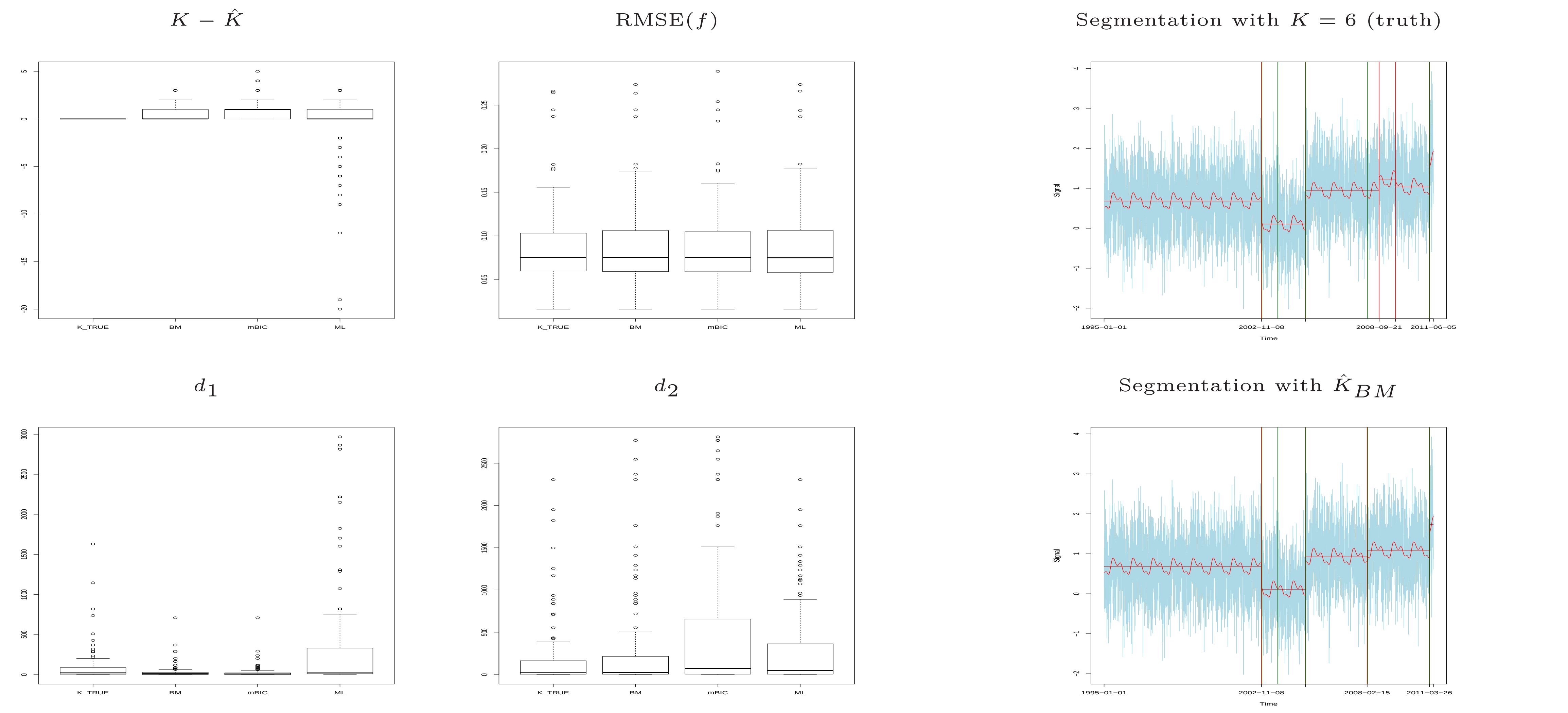
⋆ $\text{RMSE}(f) = \left[\frac{1}{n}\sum_{t=1}^{n}\left\{f_t - \hat{f}_t\right\}^2\right]^{1/2}$

⋆ 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.

4. Results



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

5. Conclusions and perspectives

In general, the proposed procedure tends to underestimate the number of breakpoints leading to a better precision on the breakpoint positioning compared to the true number of breakpoints (smaller d_1, provided that the estimated breakpoints are correctly located). This result was expected since in this case one may prefer to avoid false detections, as generally observed in segmentation problems. This is clearly illustrated on the station POTS where the true breakpoint associated to the amplitude of jump 0.004 is not detected and considering the true number of breakpoints leads to the detection of two false breaks. This underestimation does not alter the estimation of f.
Concerning the three model selection criteria, mBIC tends to underestimate more the number of breakpoints (high d_2) and ML to overestimate a lot (high d_1) and not correctly (high d_2) for some signals. BM seems to be more appropriate.

The form of f used in the model is in agreement with the one of the benchmark. However, on real data, this is not the case and an improvement of the estimation of f could be needed. Moreover, it is known that there exists a time-dependency in the real data that we have to take into account in this model.

6. references

[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 intercomparison of zenith tropospheric delays derived from DORIS and GPS data*,Adv. Space Res. 46(12), 1648-1660.
[3] Auger I.E. and Lawrence C.E. (1989) *Algorithms for optimal identification of segments neighborhoods*, Bull. Math. Biol. 51, 325-337.
[4] Rousseeuw P. J. and Croux C. (1993) *Alternative to the Median Absolute Deviation*, Journal of the American Statistical Association.
[5] 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.
[6] Lavielle M.(2005) *Using penalized contrasts for the change-point problem*, Signal Processing 85 (8), 1501-1510.
[7] Lebarbier E.(2005) *Detecting multiple change-points in the mean of Gaussian process by model selection*,Signal Processing 85,717-736.