A segmentation method for the homogenization of atmospheric GNSS time series.

O. Bock[(a,b)], N. K. Nguyen[(a,b)], E. Lebarbier[(c)] and A. Quarello[(a,b)]

(a) Université Paris Cité, Institut de physique du globe de Paris, CNRS, IGN, F-75005 Paris, France
(b) ENSG-Géomatique, IGN, F-77455 Marne-la-Vallée, France
(c) Modal'X, Université Paris Nanterre, Nanterre, France

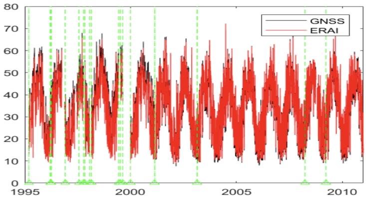
11th Seminar for Homogenization and Quality Control in Climatological Databases and 6th Interpolation, 8–11. May, 2023

Data

▶ Water vapor is a key component of the global hydrologic cycle and plays a major role in many atmospheric processes contributing to the weather and climate.

▶ Recent data: GNSS-derived Integrated Water Vapor daily series (GNSS IWV) (***Bevis et al (1992); Bock (2014)***)

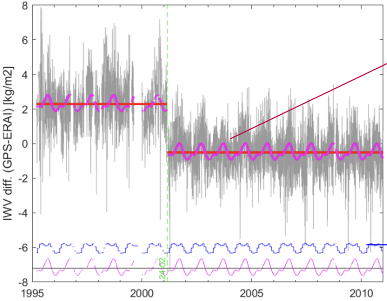
→ show inhomogeneities (abrupt changes)



→ work on the difference between GNSS and ERAI (meteorological reanalysis):

$$\Delta \mathrm{IWV}=\mathrm{IWV}_{GPS}\text{-}\mathrm{IWV}_{ERAI}$$

Features of the data ΔIWV



a periodic signal remains

a non-stationarity of the variability is observed

Objectives

Detection of the abrupt changes (change-points) in the series of difference ΔIWV: a new change-point detection in the mean model taking into account for these features (***Quarello et al (2022)***)

→ this talk

Validation of the detected change-points using the available metadata and study of the sensitivity of the proposed segmentation method to the data properties (***Nguyen et al (2021)***)

→ second talk of Olivier Bock

Attribution of the detected change-points to GNSS or ERAI: a new method using machine learning (paper in revision)

→ third talk of Ninh Nguyen

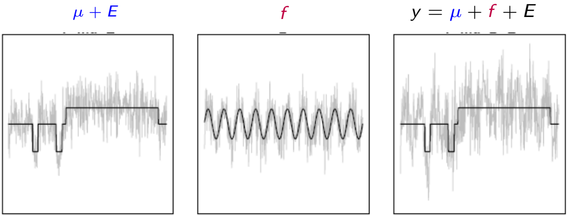
Model

We add a functional part in the model proposed by ***Bock et al (2020)***:

$$y_t \ \text{ind.} \ \sim \ \mathcal{N}(\mu_k + f_t, \sigma^2_{\textbf{month}}) \ \text{ if } t \in r_k^{\textbf{mean}} \cap r_{\textbf{month}}^{\textbf{var}}, \text{ for } k = 1, \dots, K,$$

where

- ⋆ the segments of constant mean $r_k^{\textbf{mean}} = [\![t_{k-1} + 1, t_k]\!]$ are **unknown**,
- ⋆ the segments of constant variance $r_{\textbf{month}}^{\textbf{var}} = \{t; \text{date}(t) \in \text{month}\}$ are **known**



Model

We add a functional part in the model proposed by ***Bock et al (2020)***:

$$y_t \text{ ind.} \sim \mathcal{N}(\mu_k + f_t, \sigma^2_{\textbf{month}}) \text{ if } t \in r_k^{\textbf{mean}} \cap r_{\textbf{month}}^{\textbf{var}}, \text{ for } k = 1, \dots, K,$$

where

- ⋆ the segments of constant mean $r_k^{\textbf{mean}} = [\![t_{k-1} + 1, t_k]\!]$ are **unknown**,
- ⋆ the segments of constant variance $r_{\textbf{month}}^{\textbf{var}} = \{t; \text{date}(t) \in \text{month}\}$ are **known**

▶ Form for f_t? f_t will be approximated using a Fourier series of order 4

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

where L is the mean length of the year ($L = 365.25$ days when time t is expressed in days).

▶ Change-points? we note $T = (t_1, t_2, \dots, t_{K-1})$ the $K-1$ change-points

Maximum likelihood segmentation in K segments

- log-likelihood

$$\log p(y; K, T, f, \mu, \sigma^2) = -\frac{n}{2}\log(2\pi) - \sum_{\text{month}} \frac{n_{\text{month}}}{2}\log(\sigma^2_{\text{month}}) - \frac{1}{2}\sum_{k=1}^{K}\sum_{\text{month}}\sum_{t\in r_k^{\text{mean}}\cap r_{\text{month}}^{\text{var}}} \frac{(y_t-\mu_k-f_t)^2}{\sigma^2_{\text{month}}}$$

- Main challenge: computational issue for the change-points?

$\rightarrow$ σ^2_{month} and f are global parameters (shared by the mean segments)

$\rightarrow$ the classical efficient algorithm (the Dynamic Programming or DP) can not applied

Proposed strategy → to allow the use of DP

- **Step 1: estimation of the variance $\sigma^2_{\textbf{month}}$** using a robust approach (robust to the change-points) → $\hat{\sigma}^2_{\textbf{month}}$ (***Bock et al (2020); Rousseeuw and Croux (1993)***)

- **Step 2: an iterative procedure:** at iteration $[h+1]$:

(1) Estimation of f on $\{y_t - \mu_k^{[h]}\}_t$ using a weigthed least-square regression with weights $1/\widehat{\sigma}^2_{\textbf{month}}$,

$$f^{[h+\mathbf{1}]} = \underset{f}{\operatorname{argmin}} \sum_{k=\mathbf{1}}^{K} \sum_{\textbf{month}} \sum_{t \in r_k^{\textbf{mean}} \cap r_{\textbf{month}}^{\textbf{var}}} \frac{(y_t - f_t - \mu_k^{[h]})^\mathbf{2}}{\widehat{\sigma}^\mathbf{2}_{month}},$$

(2) Estimation of T and μ_k on $\{y_t - f_t^{[h+\mathbf{1}]}\}_t$:

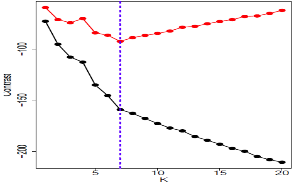
$$(T, \mu)^{[h+\mathbf{1}]} = \underset{T \in \mathcal{M}_n^K, \mu}{\operatorname{argmin}} \sum_{k=\mathbf{1}}^{K} \sum_{\textbf{month}} \sum_{t \in r_k^{\textbf{mean}} \cap r_{\textbf{month}}^{\textbf{var}}} \frac{(y_t - f_t^{[h+\mathbf{1}]} - \mu_k)^\mathbf{2}}{\widehat{\sigma}^\mathbf{2}_{\textbf{month}}}$$

→ DP applies

Choice of K?

- K is chosen as follows

$$\hat{K} = \underset{K}{\operatorname{argmin}} \underbrace{-\log p(y; K, \hat{T}, \hat{f}, \hat{\mu}, \hat{\sigma}^2)}_{\textbf{Fit}} + \underbrace{\beta \ \text{pen}(K)}_{\textbf{Penalty}}$$



- Many criteria have been proposed

Criterion	pen(K)	β
AIC	K	1
BIC	K	$\log(n)/2$
Birge/Massart (BM) (***Birgé and Massart (2001)***)	$c_2 K + c_1 \log(\mathrm{C}_n^K)$	adaptative
Lavielle (Lav) (***Lavielle (2005)***)	K	adaptative
mBIC (***Zhang and Siegmund (2007)***)	$f(K, \sum_k \log n_k)$	$\log(n)/2$

$\rightarrow$ The classical penalties (AIC, BIC) are not theoretically adapted in the segmentation context

$\rightarrow$ Heuristics for the constant penalty calibration: ML, BM1 and BM2

Simulation study

▶ Simulation design.

⋆ $n = 400 = 4$ years of 100 each and 2 months per year

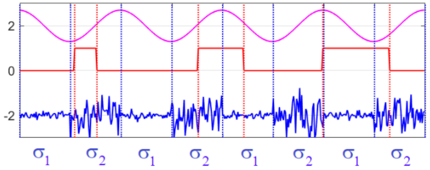
⋆ $\sigma_{\mathbf{1}} = 0.5$

⋆ $\sigma_{\mathbf{2}}$ from 0.1 to 1.5 (by step 0.2)

⋆ $f_t = 0.7cos(2\pi t/100)$

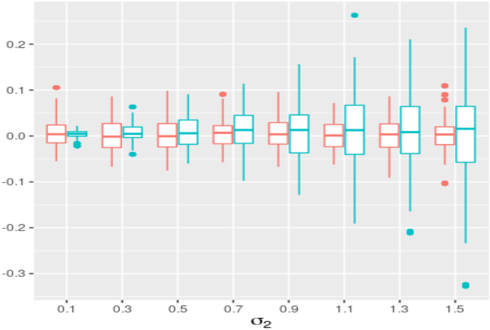
⋆ $T = [55, 77, 177, 222, 300, 366]$ $(K = 7)$

⋆ $\mu = [0, 1, 0, 1, 0, 1, 0]$



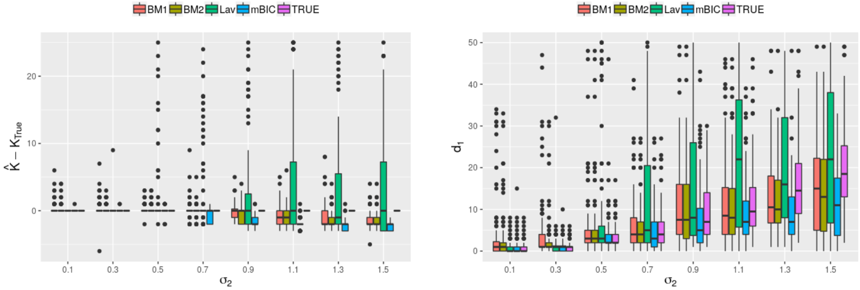
Accuracy of the variance estimates

- ⋆ $\hat{\sigma}_1 - \sigma_1^*$
- ⋆ $\hat{\sigma}_2 - \sigma_2^*$



- ⋆ The variance estimator works well despite the presence of the periodic bias
- ⋆ The dispersion increases when σ_2^* increases

Accuracy of the segmentation estimates



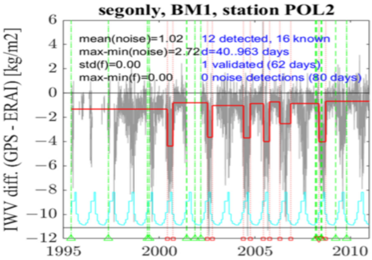
⋆ When the detection is easy (small σ_2^*), all the criteria retrieve the true K and the change-points are well positioned (d_1 small)

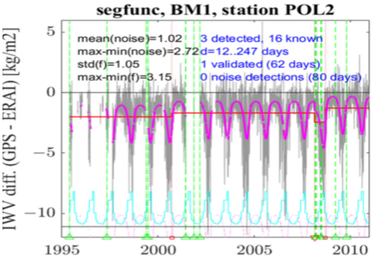
⋆ When the detection is difficult (large σ_2^*):

- Lav tends to give the true K in median, but with large dispersion ,
- BM1, BM2 and mBIC underestimate K

→ but this under-estimation leads to a better precision of the change-point locations (smaller d_1)

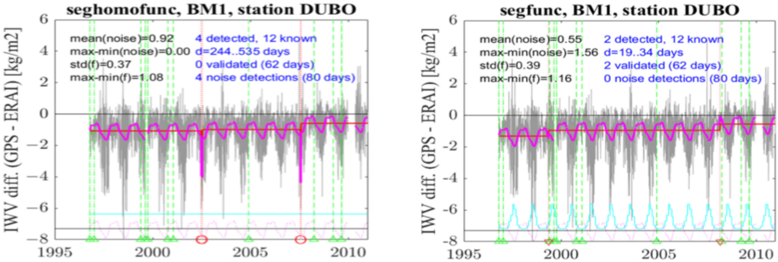
Accounting for the periodic signal?





- ⋆ The series shows a strong periodic variation:
  - without accounting for the periodic signal (segonly), this effect is captured by the segmentation
  - this effect is well fitted with our method (segfunc)

Accounting for the heterogenous variance?



⋆ With a homogenous variance (seghomofunc), we detect 4 change-points corresponding to two spikes

→ they are detected since in this period the real variance is high and here the estimated homogeneous variance is lower

→ they are not validated (using the metadata)

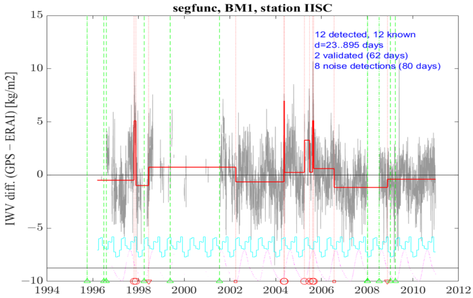
⋆ With a heterogenous variance (segfunc),

→ these two spikes are not detected

→ two other change-points are detected (located in a small variance period) and are validated

Post-processing

- The segmentation method can detect a couple or more change-points located close together.
- They are usually due to spikes in the noise and are unwanted.

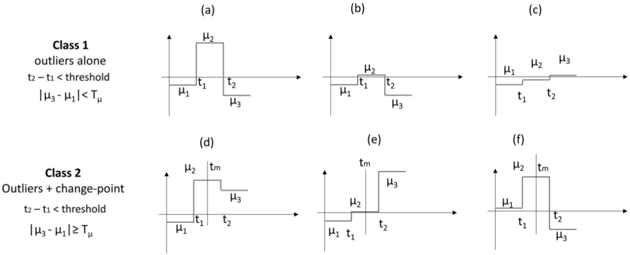


Screening: proposed procedure

⋆ finding clusters of outliers: sets of 'too close' change-points, i.e. in a windows of 80 days (the windows' size has been determinated using a clustering of the length of the segments)

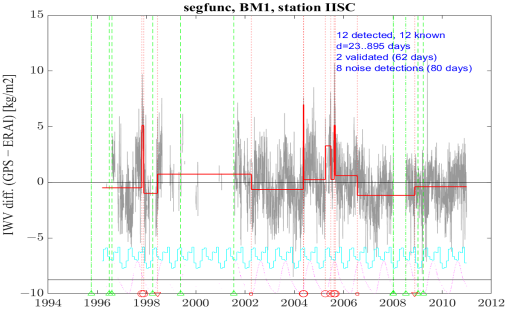
⋆ testing the variations in mean of the segment before and after (using a weighted test of mean comparison)

- if the difference is unsignificant, all the change-points are removed (class 1)
- if the difference is significant, the cluster is replace by one change-point (in the middle) (class 2)



Note that in both cases, the data points in between the two breakpoints are flagged as 'outliers' and are not used in the correction step of the homogenization procedure.

Come back to the real example



⋆ 12 change-points are detected

⋆ three clusters: in October 1997 (2 changes), in May 2004 (2 changes) and in May–August 2005 (4 changes)

⋆ for all these clusters, the test is significant: one change-point is kept per cluster reducing the set of change-points from 12 to 7

Model selection criteria on a real dataset and automatic validation

- Dataset: daily IWV differences for 120 global GNSS stations, for the period from 1 January 1995 to 31 December 2010

- Automatic validation of the change-points: a window of 62 days before or after a documented change (as proposed by ***Van Malderen et al (2020)***)

Model selection criteria on a real dataset and automatic validation

- Dataset: daily IWV differences for 120 global GNSS stations, for the period from 1 January 1995 to 31 December 2010

- Automatic validation of the change-points: a window of 62 days before or after a documented change (as proposed by ***Van Malderen et al (2020)***)

	Before screening				After screening		
	Detections	Outliers	Validations		Detections	Validations	
mBIC	3251	2714	415	13%	1270	263	21%
Lav	474	194	108	23%	341	102	30%
BM1	335	70	93	28%	292	93	32%
BM2	435	113	107	25%	370	105	28%

Model selection criteria on a real dataset and automatic validation

	Before screening				After screening	
	Detections	Outliers	Validations		Detections	Validations	
mBIC	3251	2714	415	13%	1270	263	21%
Lav	474	194	108	23%	341	102	30%
BM1	335	70	93	28%	292	93	32%
BM2	435	113	107	25%	370	105	28%

- Model selection criteria:
  - ⋆ mBIC detects too many change-points and outliers compared to the others
  - ⋆ BM1 has the smallest number of detections and outliers, and the largest percentage of validations (both before and after screening)
- Screening effect:
  - ⋆ as expected, the number of detections is reduced by the screening (strongly for mBIC)
  - ⋆ the number of validations remains the same after the screening for BM1, BM2 and Lav

→ BM1 is the preferred criterion according to this results but BM2 and Lav show close results

A semi-automatic validation method

- The change-points are first checked manually and then validated using the available information
- Results:

	Before	After	Accepted (Manual decision)	Validated (Metadata)	Validated (+TEQC)
BM1	335	292	168 (57%)	99 (58.9%)	105 (62.5%)
BM2	435	370	166 (45%)	99 (59.6%)	105 (63.3%)
Lav	474	194	175 (51%)	103 (58.9%)	109 (62.3%)
Total			187	110 (58.8%)	116 (62.0%)

- With BM1, among the 168 accepted change-points, 99 are validated by the metadata:
  - ⋆ the metadata are not complete (the changes in environment are not included as example),
  - ⋆ some change-points can be du to ERAI → need for the attribution step,
  - ⋆ the segmentation method detects too many change-points.
- Which strategy?
  - ⋆ One specific criterion: BM1 shows the best percentage of accepted change-points (57%) but with the smallest number of change-points (292)
  - ⋆ The special case where all three criteria are consistent and accepted amounts to 58% → not a sufficient condition
  - ⋆ A combining strategy: the accepted change-points from the results of the three criteria: 187 accepted change-points and 116 validated (62%)

Conclusion and improvements

- R packages

→ `GNSSseg` available on the CRAN

→ a faster version `GNSSfast` available on
https://github.com/arq16/GNSSfast.githttps://github.com/arq16/GNSSfast.git

- Some improvements of the segmentation method:
  - ⋆ **the estimation of the functional part:** a non-parametric approach, as a dictionary approach proposed by ***Bertin et al (2016)*** for example,
  - ⋆ **Integrate the presence of the 'outliers' or 'spikes'** in the segmentation by using a specific contrast in the segmentation method as the Huber contrast,
  - ⋆ **Take into account a time-dependence that exists in time series:** use for example the same approach as proposed by ***Chakar et al (2017)***.

Bertin K, Collilieux X, Lebarbier E, Meza C (2016) Semi-parametric segmentation of multiple series using a dp-lasso strategy. Journal of Statistical Computation and Simulation (arXiv:14066627)

Bevis M, Businger S, Herring TA, Rocken C, Anthes RA, Ware RH (1992) Gps meteorology: Remote sensing of atmospheric water vapor using the global positioning system. Journal of Geophysical Research: Atmospheres 97(D14):15,787–15,801

Birgé L, Massart P (2001) Gaussian model selection. J Eur Math Soc 3:203–268

Bock O (2014) Les systèmes de positionnement et de navigation par satellite: Application à la météorologie et à la climatologie. Météorologie (86):38–48

Bock O, Collilieux X, Guillamon F, Lebarbier E, Pascal C (2020) A breakpoint detection in the mean model with heterogeneous variance on fixed time intervals. Statistics and Computing 30:195–207

Chakar S, Lebarbier E, Lévy-Leduc C, Robin S (2017) A robust approach for estimating change-points in the mean of an ar1 process. Bernoulli 23(2):1408–1447

Lavielle M (2005) Using penalized contrasts for the change-point problem. Signal Processing 85(8):1501–1510

Nguyen KN, Quarello A, Bock O, Lebarbier E (2021) Sensitivity of change-point detection and trend estimates to gnss iwv time series properties. Atmosphere 12(9):1102

Quarello A, Bock O, Lebarbier E (2022) Gnssseg, a statistical method for the segmentation of daily gnss iwv time series. Remote Sensing 14(14):3379

Rousseeuw PJ, Croux C (1993) Alternatives to the median absolute deviation. Journal of the American Statistical Association 88(424):1273–1283

Van Malderen R, Pottiaux E, Klos A, Domonkos P, Elias M, Ning T, Bock O, Guijarro J, Alshawaf F, Hoseini M, et al (2020) Homogenizing gps integrated water vapor time series: Benchmarking break detection methods on synthetic data sets. Earth and Space Science 7(5):e2020EA001,121

Zhang NR, Siegmund DO (2007) A modified Bayes information criterion with applications to the analysis of comparative genomic hybridization data. Biometrics 63(1):22–32