An Exploratory Segmentation Method for Time Series

Christian Derquenne

EDF R&D

CHANGER L'ÉNERGIE ENSEMBLE

Outline

- **Issues and motivations**
- The proposed method
- Application : a simulated case
- Contributions, applications and further researches

Issues and motivations

Decomposition of times series

→ Trend, seasonality, volatility and noise

More less regular with respect application case

Evolution of electric consumption for 50 years

⇒ Regular phenomena ⇒ forecasting model at short term (MAPE < 1,5%)

Evolution of financial series (CAC40, S&P 500, …)

⇒ Trend and seasonality occur less regularly and less frequently

⇒ Volatility and irregularly

⇒ Behaviors breaks could characterize series (peaks, level breaks, trend changes, volatility)

⇒ The data modeling is very delicate, to forecast these series can be close to an utopian view

Issues and motivations

Interest to detect behavior breakpoints

→ Building contiguous segments (segmentation)

→ Interesting to detect behavior breakpoints

→ Achieving stationarity of time series with a segmentation model

→ Building symbolic curves to cluster series

→ Modeling multivariate time series

Potential applications

→ Economics, finance, human sequence, meteorology, energy management, etc.

Issues and motivations

Some examples of methods

→ Exploring the segmentation space for the assessment of multiple change-point models [Guédon, Y. (2008)]

→ Inference on the models with multiple breakpoints in multivariate time series, notably to select optimal number of breakpoints [Lavielle, M. et al. (2006)]

→ Sequential change-point detection when the pre- and post-change parameters are unknown [Lai, TL. et al. (2009)]

Common point of these methods

→ Using of dynamic programming to decrease computation complexity of segmentations (total number = 2^{T-1})

→ Complexity is generally in $O(ST^2)$ for the time and in $O(ST)$ for the linear clustered space, but also: $O(T^2)$ and $O(MT^2)$

where T = length of series ; S = number of segments ; M = number of de series

Issues and motivations

Three problems studied by these methods

(i) Change mean with a constant variance

(ii) Change of variance with a constant mean

(iii) Change for overall distribution of time series without change of level, in dispersion and on the distribution of errors

The proposed method

→ Detection of increasing or decreasing trend [Perron & al., 2008]

→ To reduce the computation complexity in $O(KT)$, where K is the smoothing degree, which is generally less than to $\sqrt{T}$

→ Proposition of some solutions of segmentation containing segments with increasing or decreasing trend, constant level and different standard-deviations

Outline

- Issues and motivations
- **The proposed method**
- Application : a simulated case
- Contributions, applications and further researches

Proposed method

Let's $(Y_t)_{t=1,T}$ be a time series, we suppose that it is decomposed in accordance with an heteroskedastic linear model (or variance components) [Rao & al., 1988, Searle & al., 1992]:

$$Y_t = \sum_{s=1}^{S} \left(\beta_0^{(s)} + \beta_1^{(s)} t + \sigma_s \varepsilon_t\right) 1_{[t \in \tau_s]} \qquad (1)$$

where $\beta_0^{(s)}, \beta_1^{(s)}$ and $\sigma_s > 0$, are respectively the level, trend and standard-deviation parameters for the segment τ_s, and ε_t is a $\mathcal{N}(0,1)$

$T_s = \text{card}(\tau_s)$ and $\sum_{s=1}^{S} T_s = T$ then there are $3S$ parameters to estimate and the number S of segments

Inference: OLS ; ML ; REML

→ same solutions for $\beta_0^{(s)}$ and $\beta_1^{(s)}$ with **the three estimators**

→ **ML** and **REML** estimate directly σ_s^2

→ Only **REML** provides an unbiased estimator of σ_s^2

Proposed method

Detailed process: *preparing data*

Step of smoothing: *To keep only the « strong » trends*

→ Using moving median:

$$m_j(t) = \underset{t \in [a_j(t), b_j(t)]}{med}(y_t) \qquad (2)$$

where for j (smoothing degree) fixed: $a_j(t) = t$ et $b_j(t) = t + j - 1$ où $t = 1$ à $T - j + 1$

Remark: The more j increases, the less irregularity of data is taken into account

A little example:

Y_t ~> $\mathcal{N}(5 ; 0{,}01)$ pour $t = 1{,}40$

Y_t ~> $\mathcal{N}(6 ; 0{,}01)$ pour $t = 41{,}100$

Proposed method

Detailed process: *preparing data*

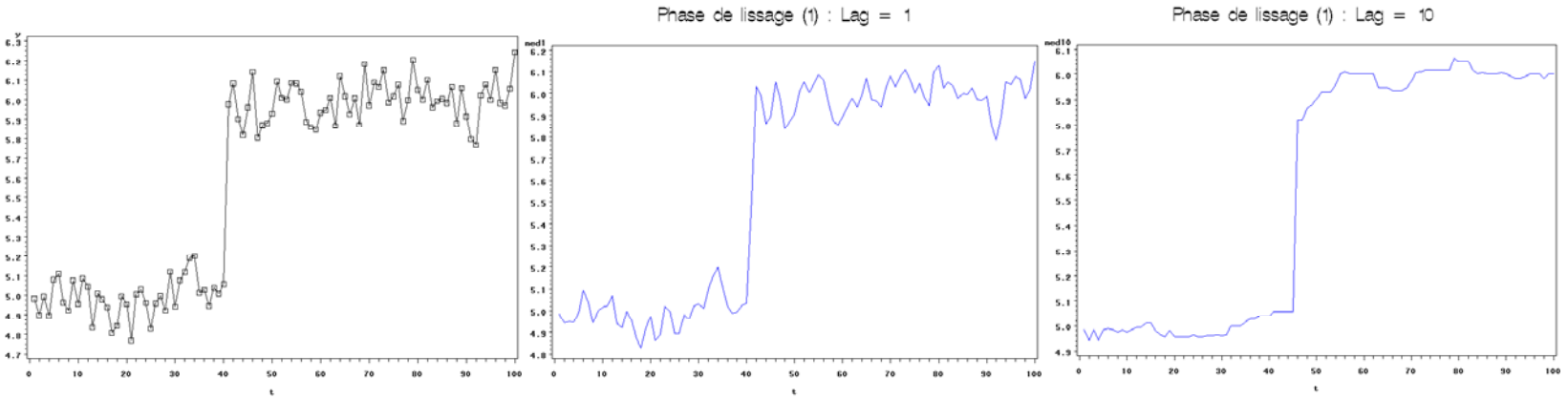
Step of smoothing: *To keep only the « strong » trends*

→ Using moving median:

$$m_j(t) = \underset{t \in [a_j(t), b_j(t)]}{med}(y_t) \qquad (2)$$

where for j (smoothing degree) fixed: $a_j(t) = t$ et $b_j(t) = t + j - 1$ où $t = 1$ à $T - j + 1$

Remark: The more j increases, the less irregularity of data is taken into account



Proposed method

Detailed process: *preparing data*

Differencing step: *to detect the trends of smoothed data*

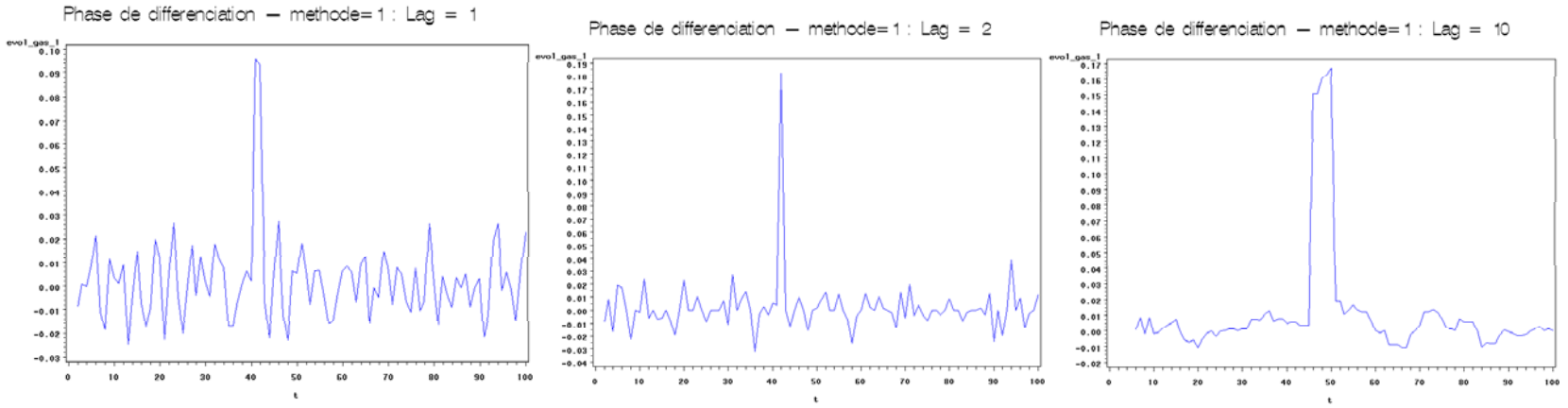
→ Using a relative deviation:

$$d_j(t) = \left(m_j(t) - m_j(t-k)\right) / m_j(t-k) \qquad (3)$$

where $k = t - j/2$ if j is even and $k = t - (j+1)/2$ if j is odd

This differencing must be sufficiently high to reveal trend deviations, but not too much otherwise it could be skipped

Remark: it is only a visual choice and not a theoretical choice



Proposed method

Detailed process: *preparing data*

Step of counting: *number and size of initial segments*

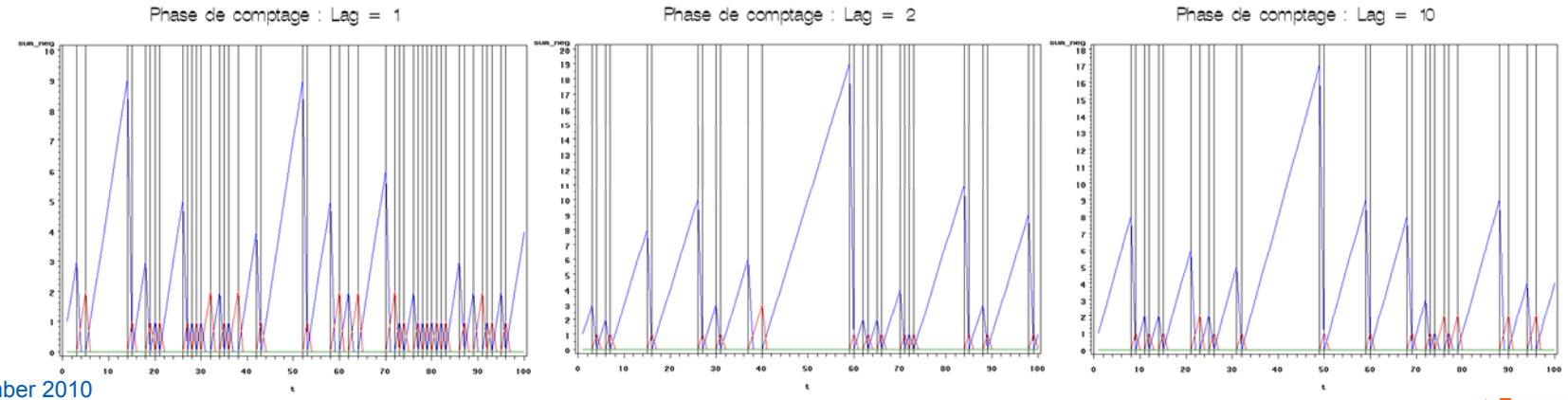
$$T_{j,1}^{(0)} = \mathrm{card}\left(\tau_{j,1}^{(0)}\right) = \sum_{t \geq 2} 1_{\left[\mathrm{sign}\left(d_j(t)\right) = \mathrm{sign}\left(d_j(t-1)\right)\right]} \qquad (4)$$

S segments: $\left(\tau_{j,1}^{(0)}, \ldots \tau_{j,s}^{(0)}, \ldots \tau_{j,S}^{(0)}\right)$ with size $\left(T_{j,1}^{(0)}, \ldots T_{j,s}^{(0)}, \ldots T_{j,S}^{(0)}\right)$ and $\sum_{s=1}^{S} T_{j,s}^{(0)} = T$

Justification:

(i) the nb of values with the same sign is reasonably linked to the smoothing deg.

(ii) The smaller smoothing degrees is, the smaller size of series of differences with same sign is



Proposed method

Detailed process: *preparing data*

Step of counting: *number and size of initial segments*

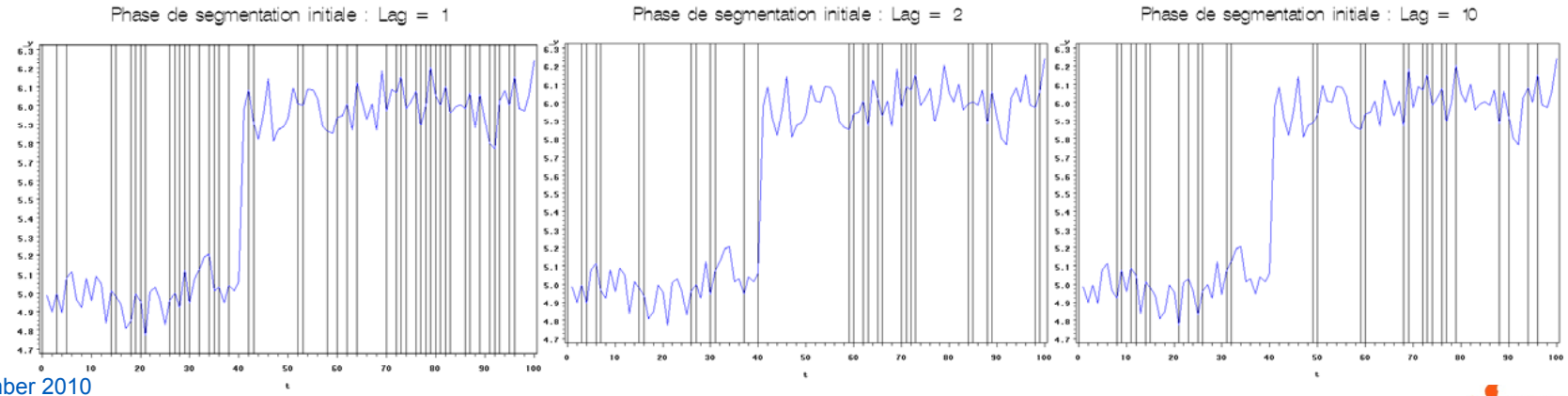
$$T_{j,1}^{(0)} = \mathrm{card}\left(\tau_{j,1}^{(0)}\right) = \sum_{t \geq 2} 1_{\left[\mathrm{sign}\left(d_j(t)\right) = \mathrm{sign}\left(d_j(t-1)\right)\right]} \qquad (4)$$

S segments: $\left(\tau_{j,1}^{(0)}, \ldots \tau_{j,s}^{(0)}, \ldots \tau_{j,S}^{(0)}\right)$ with size $\left(T_{j,1}^{(0)}, \ldots T_{j,s}^{(0)}, \ldots T_{j,S}^{(0)}\right)$ and $\sum_{s=1}^{S} T_{j,s}^{(0)} = T$

Justification:

(i) the nb of values with the same sign is reasonably linked to the smoothing deg.

(ii) The smaller smoothing degrees is, the smaller size of series of differences with same sign is



Proposed method

Detailed process: *modeling data*

Initial step: *To reduce the number of initial segments*

$$Y_t = \sum_{s=1}^{S} \left(\beta_0^{(j,s)} + \beta_1^{(j,s)} t + \sigma_{j,s} \varepsilon_t \right) 1_{\left[t \in \tau_{j,s}^{(0)} \right]} \qquad (5)$$

Inference :

(i) *Estimation of parameters with REML*

(ii) *Homogeneity test of variance (homoskedasticty)*

$\Rightarrow$If H_0 is kept: $$Y_t = \sum_{s=1}^{S} \left(\beta_0^{(j,s)} + \beta_1^{(j,s)} t \right) 1_{\left[t \in \tau_{j,s}^{(0)} \right]} + \sigma_j \varepsilon_t \qquad (6)$$

(iii) *Test of the coefficients:* $\beta_1^{(j,s)} = 0$ for each segment with respect to (ii)

$\Rightarrow$ New model: $$Y_t = \sum_{s=1}^{S} \left(\beta_0^{(j,s)} + \beta_1^{(j,s)} t 1_{\left[\beta_1^{(j,s)} \neq 0 \right]} + \sigma_{j,s} \varepsilon_t \right) 1_{\left[t \in \tau_{j,s}^{(0)} \right]} \qquad (7)$$

Proposed method

Detailed process: *modeling data*

(iv) *Aggregation of consecutive segments (2 by 2)*

(a) **Test of equal variances on**:

$$Y_t = \left(\beta_0^{(j,s)} + \beta_1^{(j,s)} t + \sigma_{j,s}\varepsilon_t\right) 1_{\left[t \in \tau_{j,s}^{(0)}\right]} + \left(\beta_0^{(j,s+1)} + \beta_1^{(j,s+1)} t + \sigma_{j,s+1}\varepsilon_t\right) 1_{\left[t \in \tau_{j,s+1}^{(0)}\right]} \qquad (8)$$

$\Rightarrow$If H_0 is rejected then the both segments are not regrouped

(b) Otherwise, **a test of equal coefficients is applied**: $\beta_1^{(j,s)} = \beta_1^{(j,s+1)}$

$\Rightarrow$If H_0 is rejected then the both segments are not regrouped

(c) Otherwise, **a test of equal intercepts is applied**: $\beta_0^{(j,s)} = \beta_0^{(j,s+1)}$

$\Rightarrow$If H_0 is rejected then the both segments are not regrouped

Proposed method

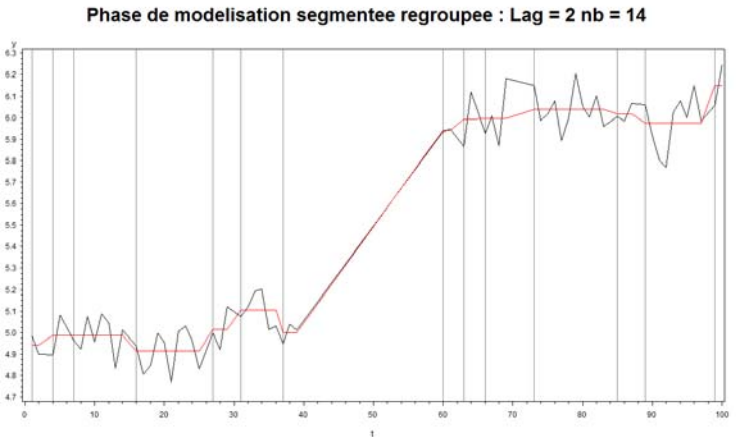
Detailed process: *modeling data*

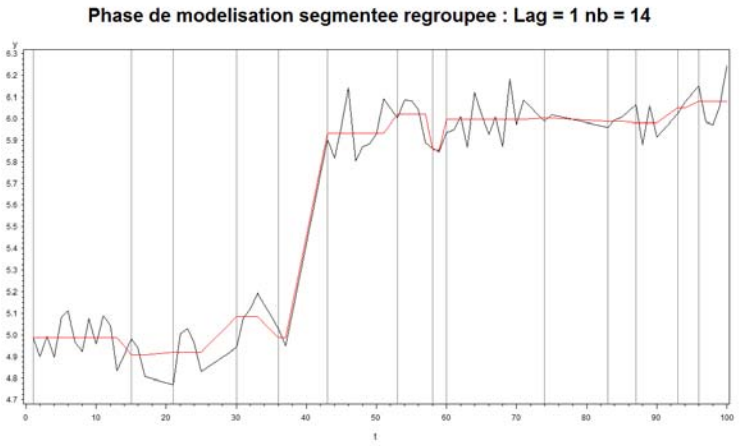
First model:

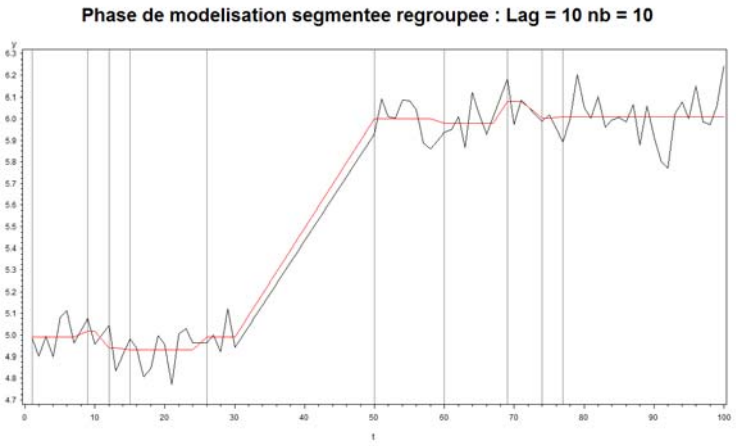
$$Y_t = \sum_{s=1}^{S_1} \left(\beta_0^{(j,s)} + \beta_1^{(j,s)} t + \sigma_{j,s} \varepsilon_t \right) 1_{\left[t \in \tau_{j,s}^{(1)}\right]} \qquad (9)$$

with $S_1 \leq S$

where S_1 is the new number of segments







Proposed method

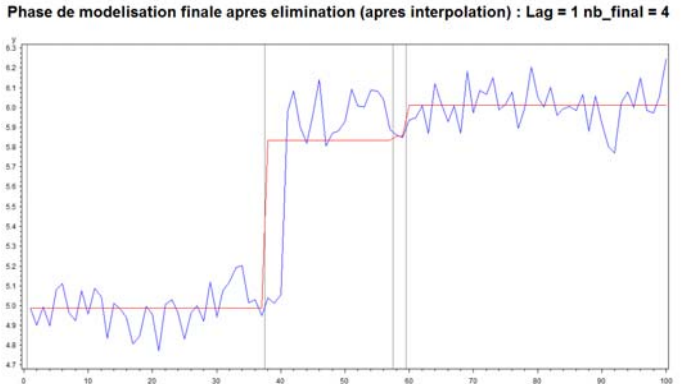
Detailed process: *modeling data*

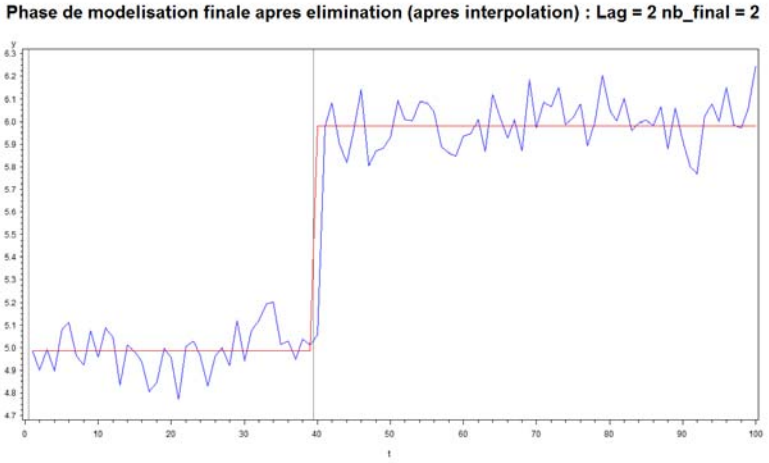
Further steps of modeling: until the number of segments is satisfactory

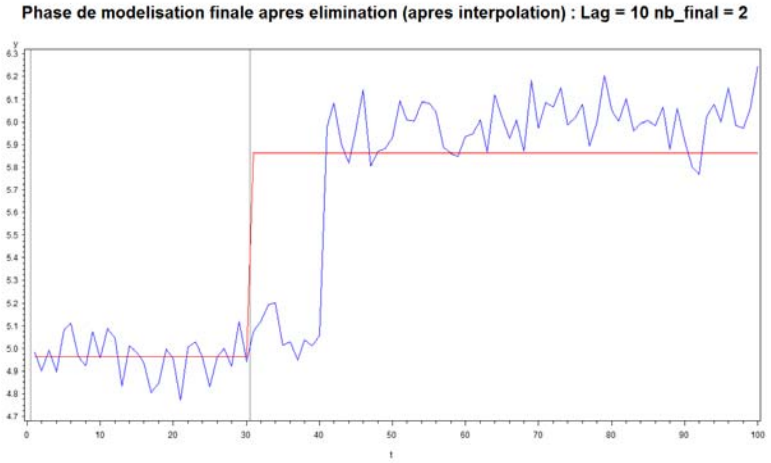
Inference :

(i) The model (9) is submitted to the same process of successive tests as presented previously until the number of segments is satisfactory

(ii) In state of the work, the precise convergence criteria (cf. further researches)







Proposed method

Models assessment

(i) A number K of smoothing degrees is fixed and K segmentations are obtained

(ii) The final model for some segmentations will allow to reconstitute well data and will have a higher probability to provide a good segmentation

(iii) *Remark*: Even if the T smoothing degrees are tried, the optimal segmentation is not guaranteed with a probability equal to one, but the goal of this method is not this one

(iv) *Goal*: to propose some interesting segmentations, in terms of decision aid

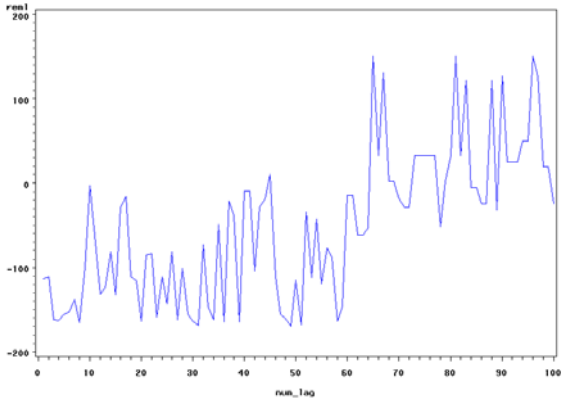
(v) To evaluate each final model and to offer some possible segmentations, REML and MAPE are used. Then the smaller values of these last ones are preferred to decide the quality level of the segmentation

(vi) *Remark*: These measures are heuristic choices because they can have an impact in the process of segmentation, notably to select one or several uninteresting segmentations

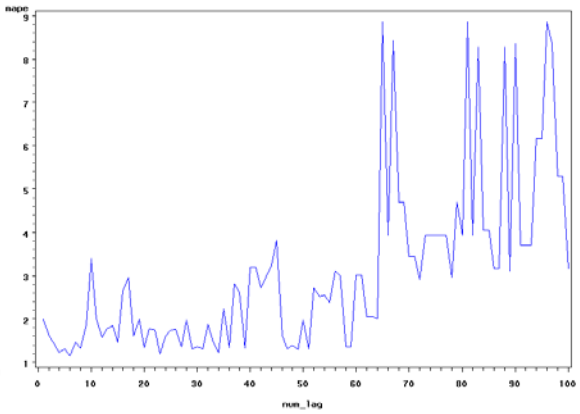
Proposed method

Models assessment

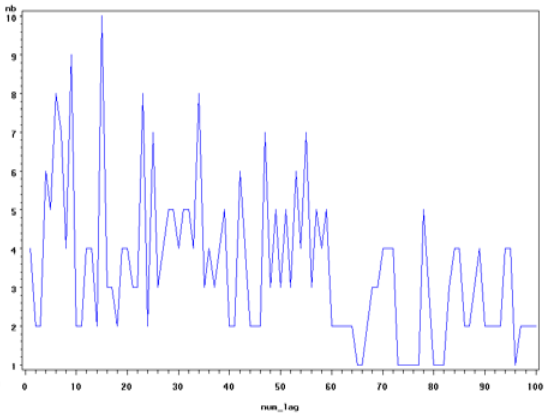
Evolution des REML vs grain de lissage : phase finale apres elimination (apres interpolation)



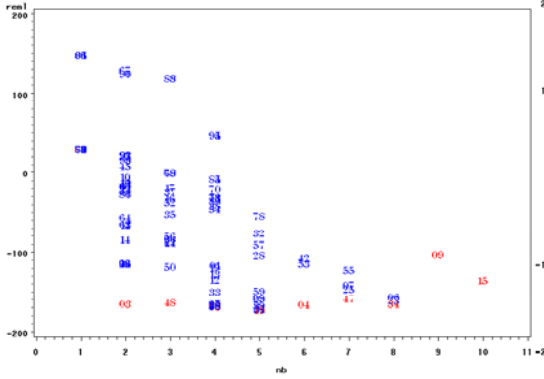
Evolution des MAPE vs grain de lissage : phase finale apres elimination (apres interpolation)



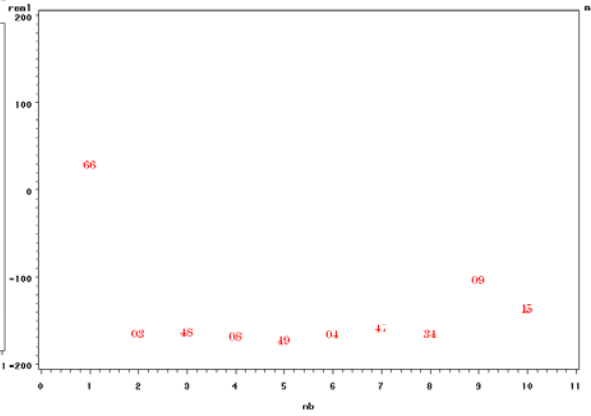
Nombre de segments vs grain de lissage : phase finale apres elimination (avant interpolation)



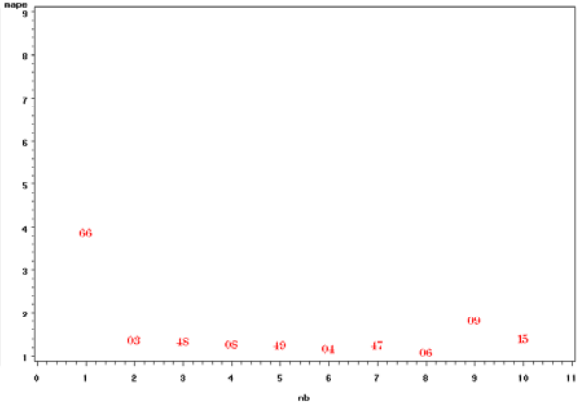
Evolution des REML vs nombre de segments : phase finale apres elimination (apres interpolation



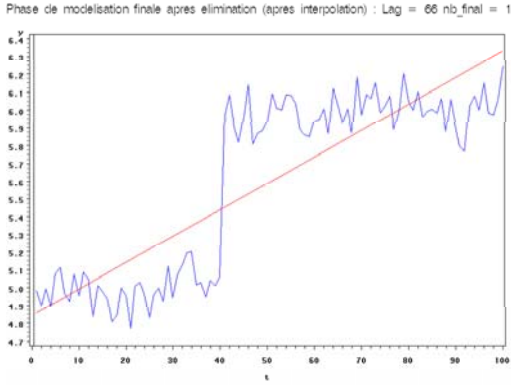
Evolution des REML vs nombre de segments : phase finale apres elimination (apres interpolation)

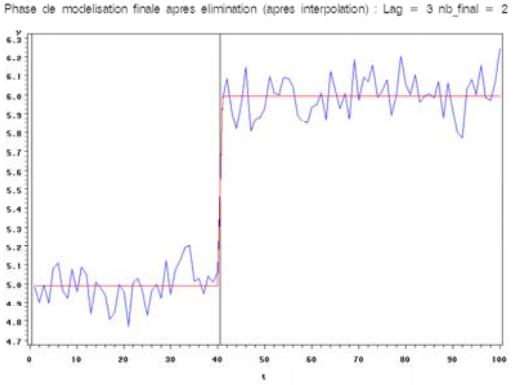


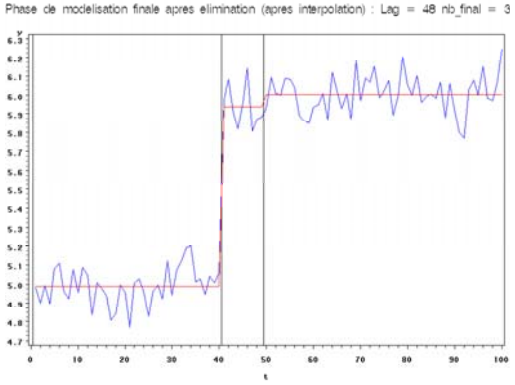
Evolution des MAPE vs nombre de segments : phase finale apres elimination (apres interpolation)

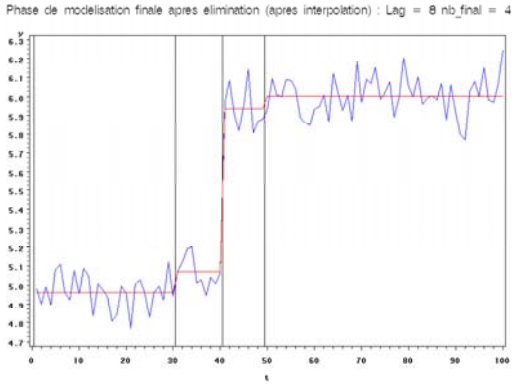


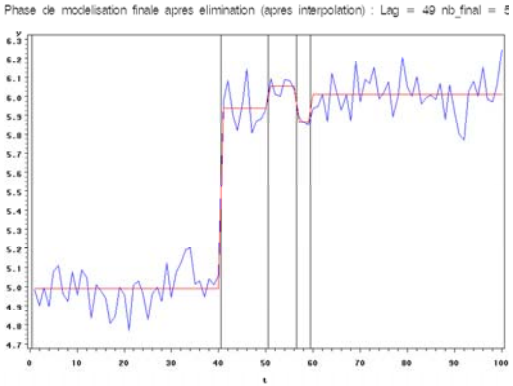
Proposed method

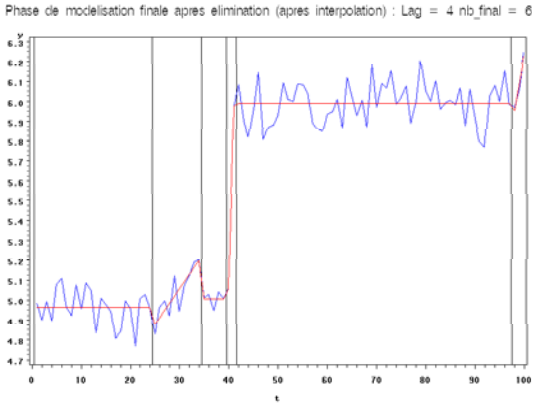


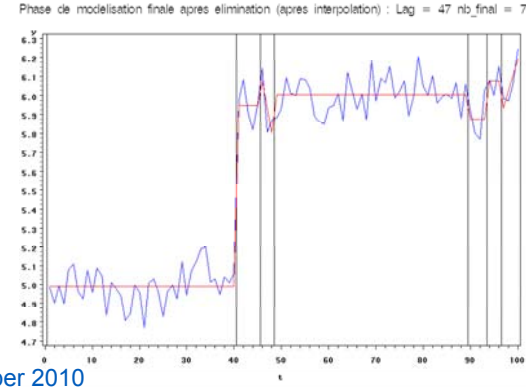


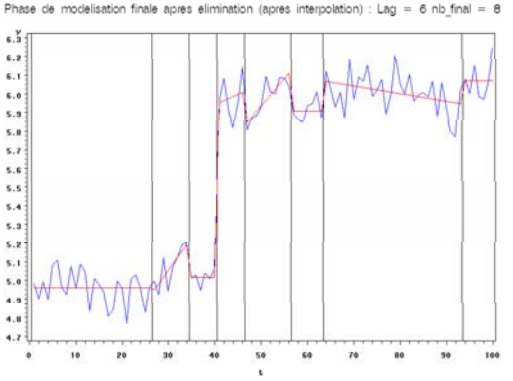


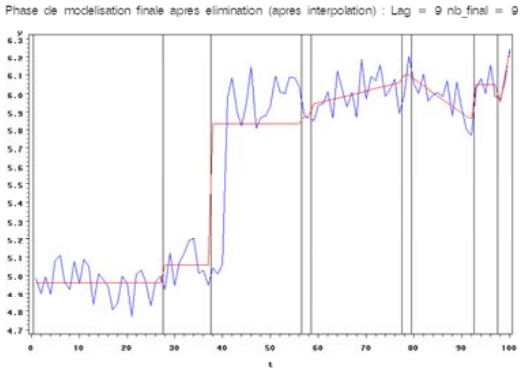












Proposed method

Coveriance Parameter Estimates

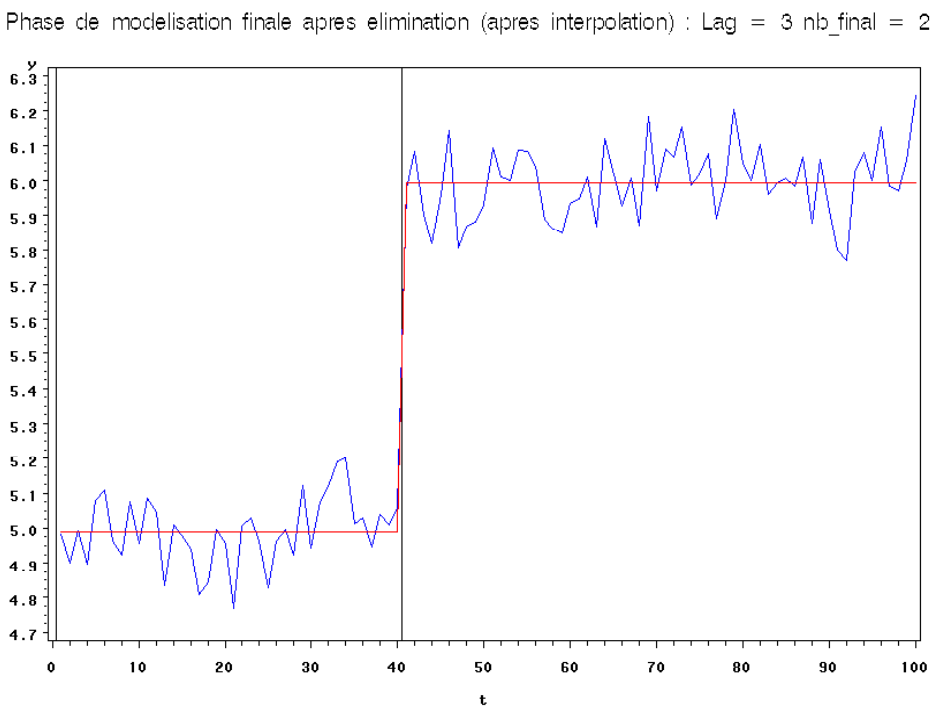
Cov Parm	Group	Estimate
Residual	cls_final 1	0.009506
Residual	cls_final 2	0.01092

Fit Statistics

-2 Res Log Likelihood	-162.2
AIC (smaller is better)	-158.2
AICC (smaller is better)	-158.1
BIC (smaller is better)	-153.0

Null Model Likelihood Ratio Test

DF	Chi-Square	Pr > ChiSq
1	0.22	0.6377



Solution for Fixed Effects

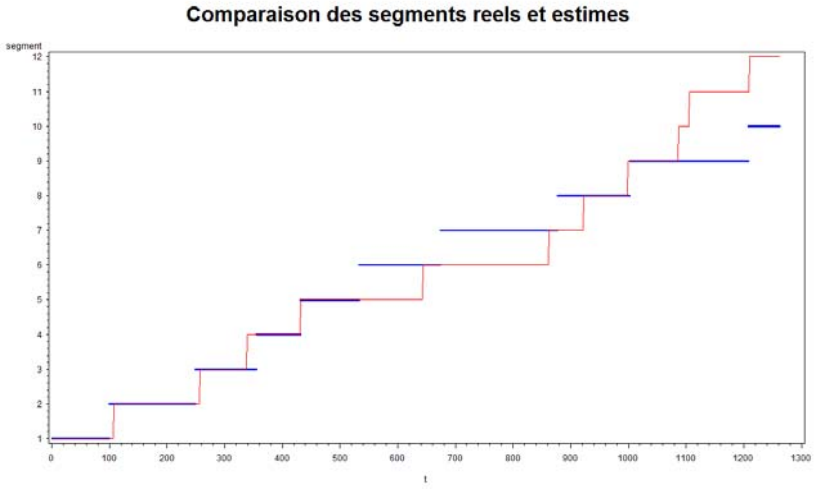
Effect	cls_final	Estimate	Standard Error	DF	t Value	Pr > |t|
t*drap_fin*cls_final	1	0	.	.	.	.
t*drap_fin*cls_final	2	0	.	.	.	.
cls_final	1	4.9904	0.01542	98	323.68	<.0001
cls_final	2	5.9950	0.01349	98	444.45	<.0001

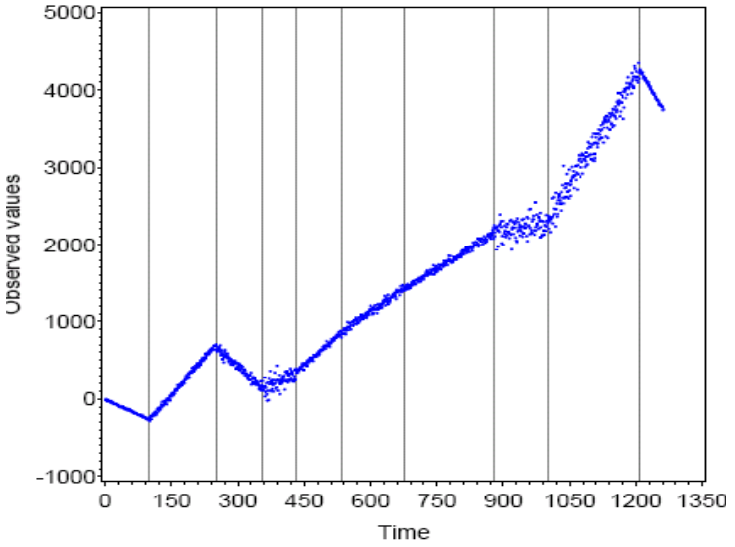
Outline

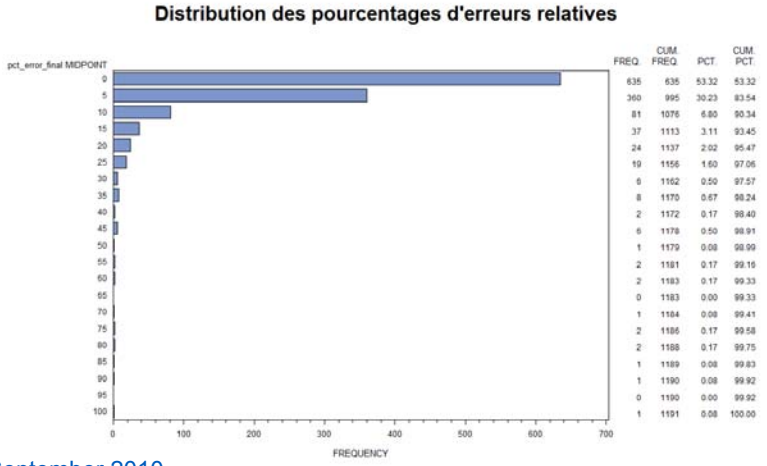
- Issues and motivations
- The proposed method
- **Application : a simulated case**
- Contributions, applications and further researches

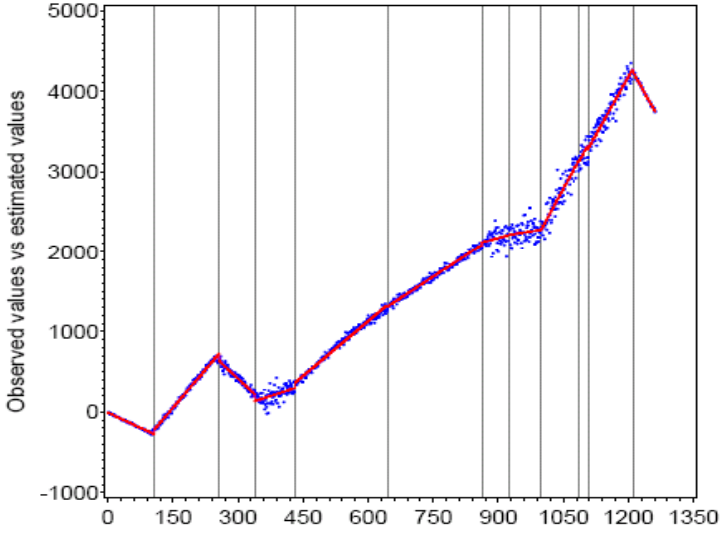
Application: a simulated case

MAPE = 7,1% ; %BCL = 85,3%

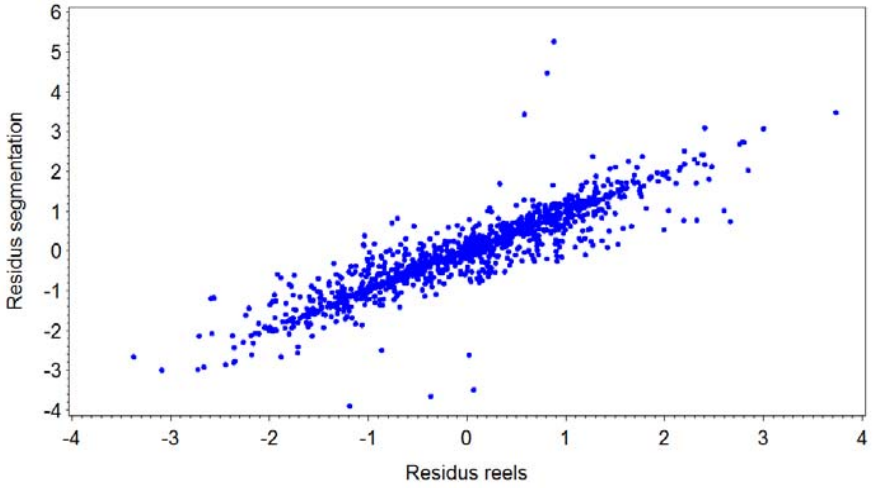


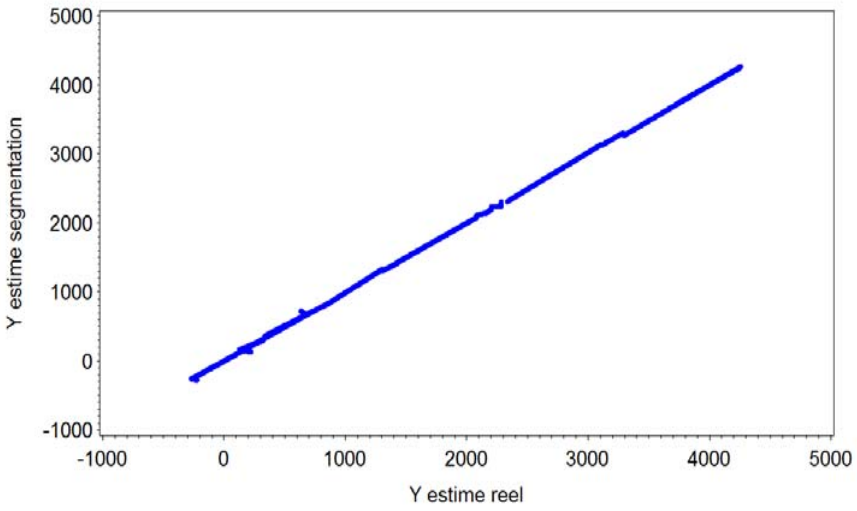


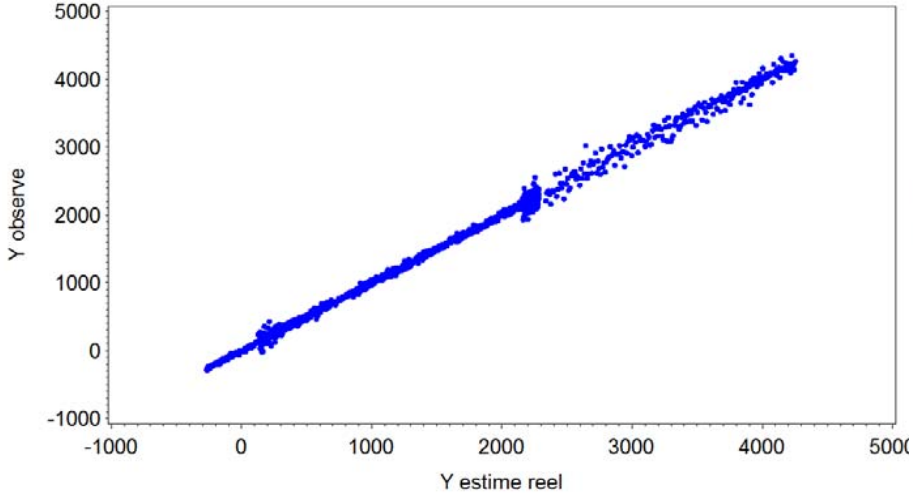


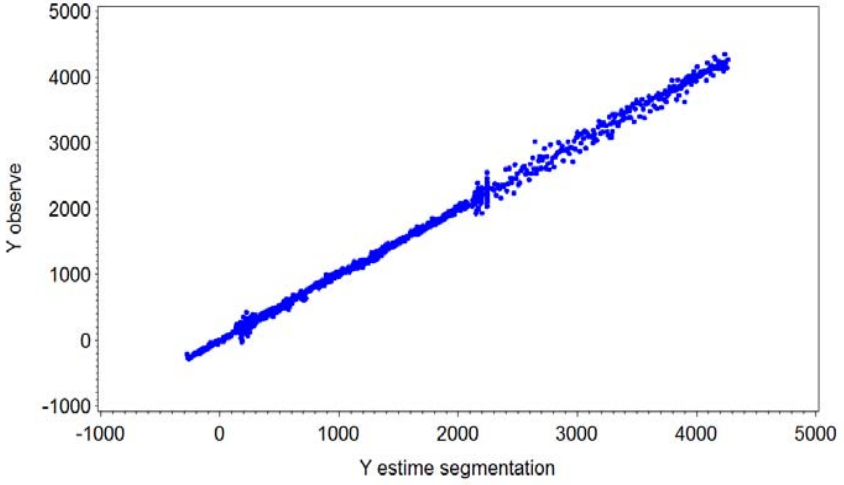


Application: a simulated case



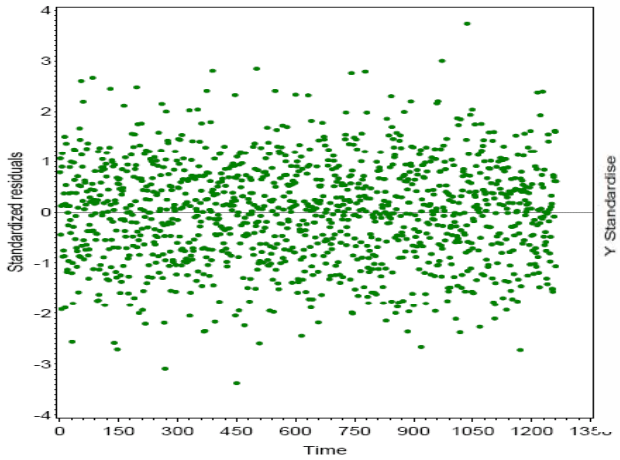


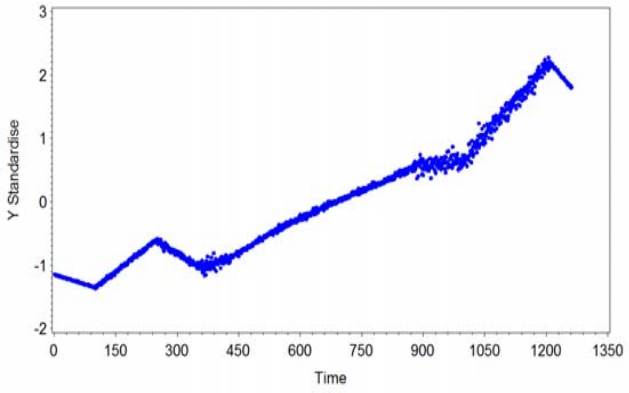


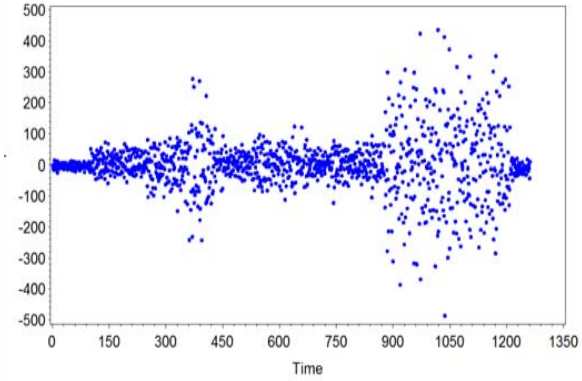


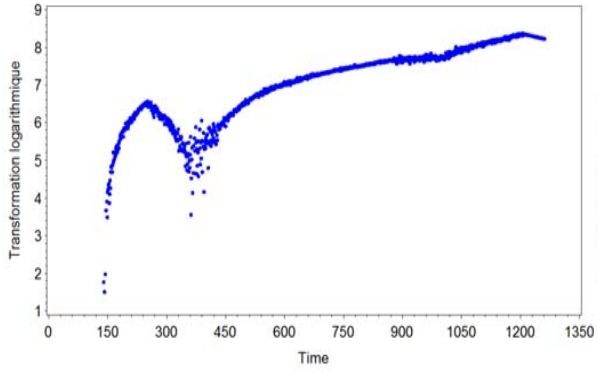
Application: a simulated case

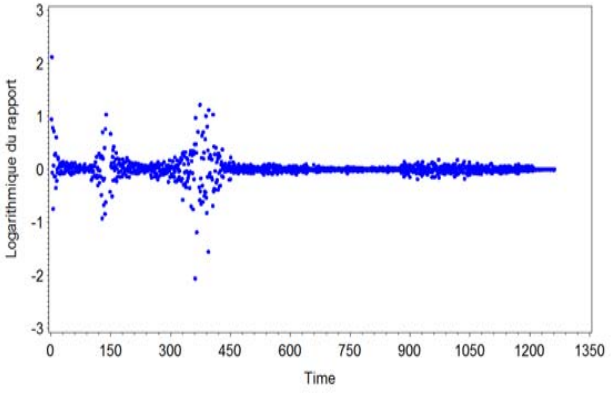
An interest to achieve stationarity a time series

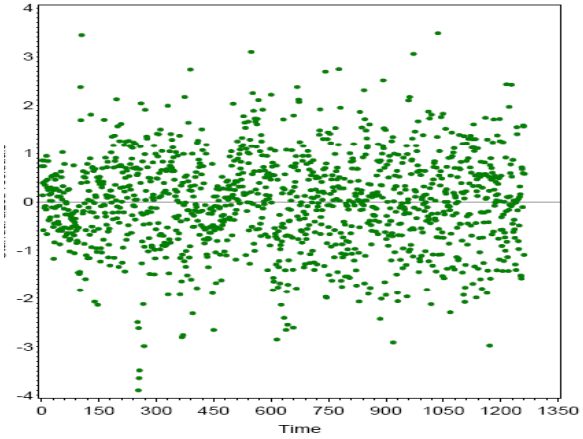












Outline

- Issues and motivations
- The proposed method
- Application : a simulated case
- **Contributions, applications and further researches**

Contributions, applications and further researches

The proposed method allows to segment a time series

→ It offers an original process containing a stage of preparing data which is essential to build the most adequate structure to initialize stage of modelling

→ The modeling step is in accordance with an heteroskedastic linear model including the different trends, levels and variances

→ The goal of this method is not to provide the optimal segmentation as the majority of the methods discussed in introduction, but to provide a decision aid. Indeed, even if the minimum complexity of the other methods is in $\mathcal{O}(T^2)$, it stays high, however

→ The method introduced in this paper uses only assessment criteria, such as values of REML, MAPE and percentage of relative errors less than 10%

Contributions, applications and further researches

→Complexity is in $\mathcal{O}(T)$ for each smoothing degree and the number of this last one is rarely greater than $\sqrt{T}$. Indeed, for high smoothing degree, the quality of segmentations decreases rapidly, because they move away optimality, even if this one is empirical

→This method can be used in a lot of domains of application and for a lot of objectives: searching of segments, achieving stationarity, building of different models on a same time series having different behaviors, simplifying (symbolic approach) of several time series to make clustering of curves, etc

→This method is rather preliminary and we work to improve some steps of this method, particularly on the detection of volatility in data and on the evaluation and the validation tools of segmentations to obtain a better means to have a hierarchy of these last ones

Bibliography

Bartlett, M.S. (1937): Properties of sufficiency and statistical tests. *Proceedings of the Royal Society of London*, Series A **160**, 268-282.

Guédon, Y. (2008): Exploring the segmentation space for the assessment of multiple change-point models. Institut National de Recherche en Informatique et en Automatique, *Cahier de recherche 6619*.

Harville, DA. (1977): Maximum likelihood approaches to variance Component estimation and to related problems. *J Amer Stat Assoc* **72**, 320-340.

Lai, TL. and Xing, H. (2009): Sequential Change-point Detection when the pre- and post-change parameters are unknown. *Technical report 2009-5*, Stanford University, Department of Statistics.

Lavielle, M. and Teyssière, G. (2006): Détection de ruptures multiples dans des séries temporelles multivariées. *Lietuvos Matematikos Rinikinys*, Vol **46**.

Perron, P. and Kejriwal, M. (2006): Testing for Multiple Structural Changes in Cointegrated Regression Models. Boston University, *C22*.