Segmentation de séries temporelles avec prise en compte a priori de composantes de variance

Christian Derquenne

EDF R&D

CHANGER L'ÉNERGIE ENSEMBLE

Plan

- **Contexte, motivations et objectifs**
- Méthode proposée
- Application : un cas simulé
- Application : un cas réel
- Apports, applications et futures recherches

Contexte, motivations et objectifs

Décomposition d'une série temporelle

→ Tendance, saisonnalité, volatilité et bruit

Composantes dépendant du domaine d'application

→ Tendance et saisonnalité moins fréquentes et moins régulières

→ Présence de volatilité et irrégulière

→ Changement de comportement du processus de génération des données (pics, sauts en niveau ou en tendance, en variabilité)

Intérêt de détecter des ruptures de comportement

→ Construction de segments contigus (segmentation)

→ Estimation de sous-modèles pour chaque segment obtenu

→ Modèle de régression tenant compte des points de ruptures

→ Stationnarisation de la série à l'aide de la segmentation

→ Courbes symboliques en vue d'une classification de courbes

Contexte, motivations et objectifs

Méthodes de segmentation :

→ Changement de la moyenne, de la variance, de la distribution

→ **Programmation dynamique** (Lavielle, Guedon, Lai, Perron, …)

→ **Approche exploratoire** (Derquenne, 2010)

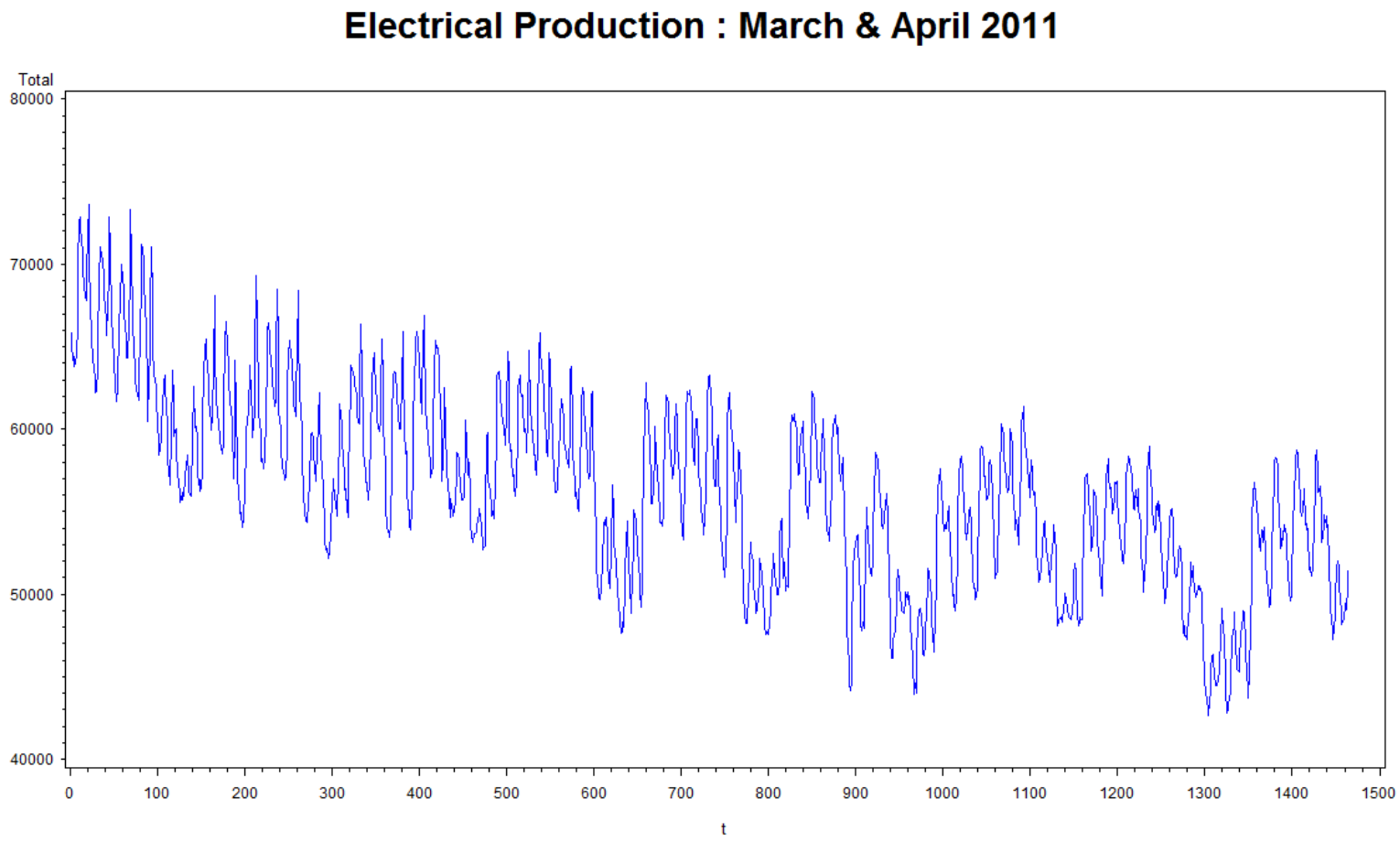
→ **Problème commun à toutes ces méthodes** : Détection délicate de segments contigus avec des niveaux similaires, mais avec des variances différentes

Applications potentielles

→ Economie, finance, séquençage humain, météorologie, management de l'énergie, etc.

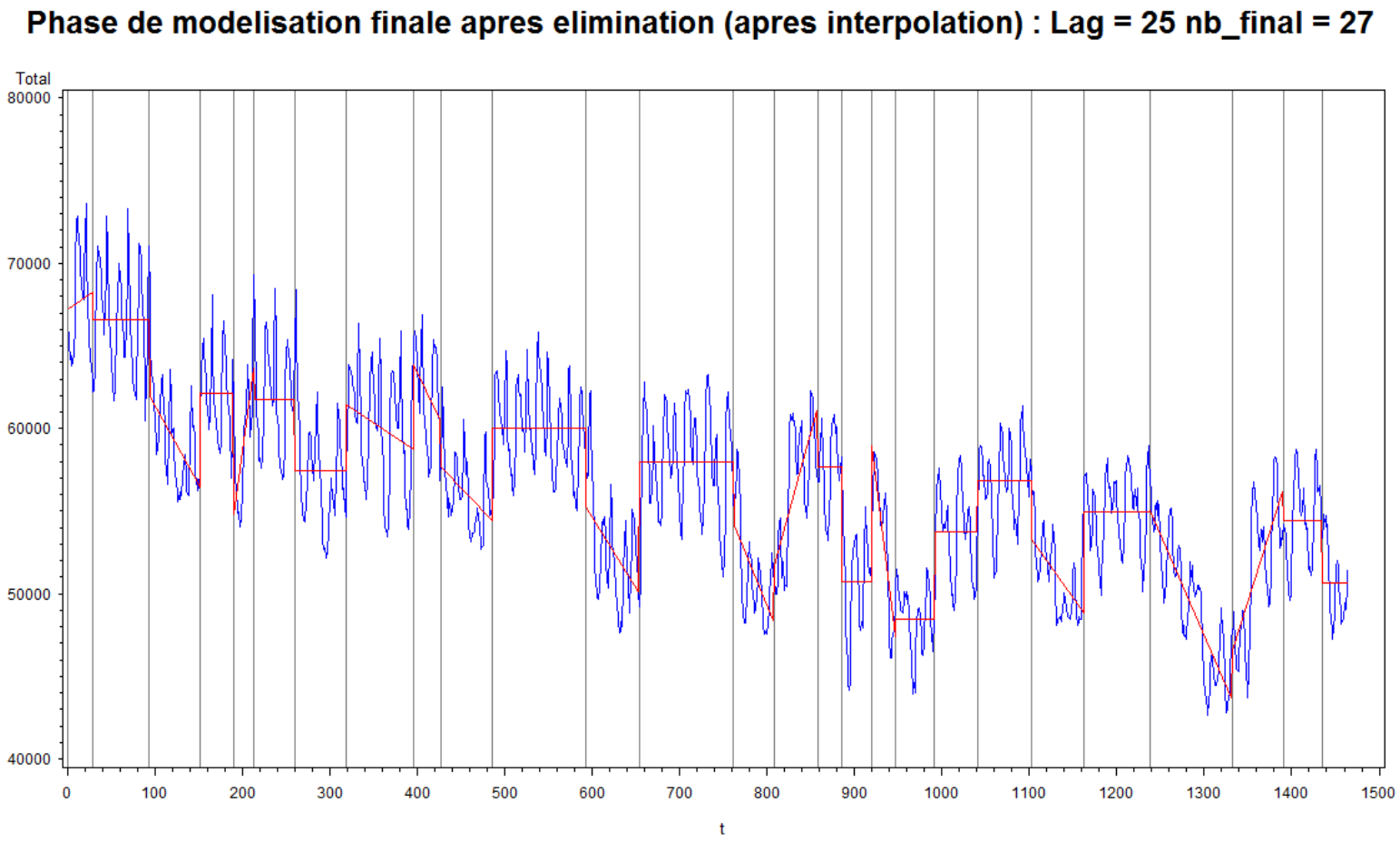
Contexte, motivations et objectifs

Objectif : *Identifier le comportement de la production d'électricité*



Contexte, motivations et objectifs

Segmentation de la production d'électricité



Plan

- Contexte, motivations et objectifs
- **Méthode proposée**
- Application : un cas simulé
- Application : un cas réel
- Apports, applications et futures recherches

Méthode proposée

Processus général de la méthode de segmentation
(Derquenne, 2010)

(i) **Etape de préparation des données $(Y_t)_{t=1,T}$ pour offrir une première segmentation :**
→ **Lissage** pour garder seulement les tendances fortes = médiane mobile (degré de lissage)
→ **Différentiation** : pour faire apparaître les différences issues de la série lissée
→ **Comptage** : pour calculer le nombre et les tailles des segments initiaux

(ii) **Etape de modélisations successives et adaptatives** (tests statistiques : variances, tendances, égalité de coefficients, …) **pour optimiser la segmentation fondée sur un modèle linéaire gaussien hétéroscédastique** estimé par **REML** :

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

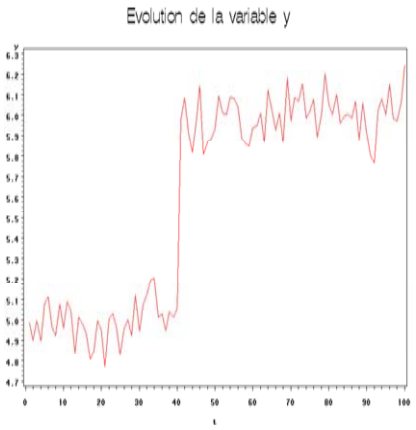
où $\beta_0^{(s)}$, $\beta_1^{(s)}$, $\sigma_s > 0$, sont respectivement le niveau, la tendance et l'écart-type du segment τ_s et $\varepsilon_t \sim> \mathcal{N}(0,1)$ iid ; $T_s = \text{card}(\tau_s)$ et $\sum_{s=1}^{S} T_s = T$

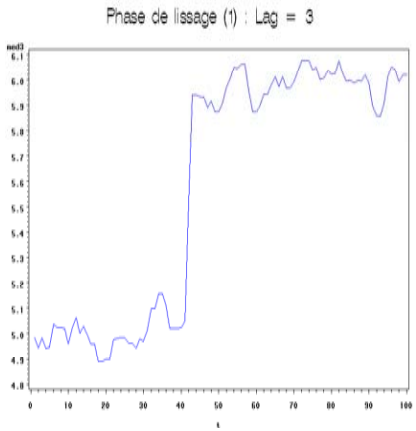
⇒ **Approche non supervisée** : le nombre S de segments, leurs tailles, les $3S$ paramètres à estimer sont inconnus

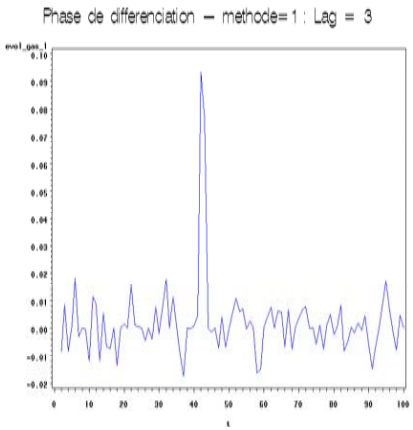
(iii) **Les étapes (i) et (ii) sont répétées pour chaque degré de lissage ($j = 1$ à J)**

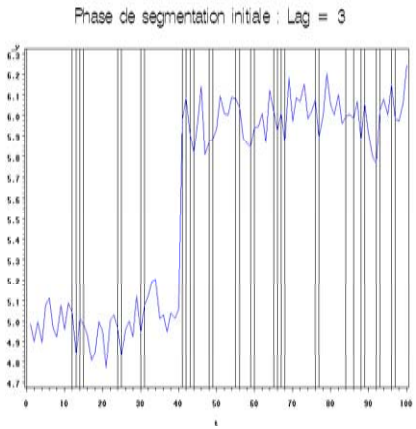
Méthode proposée

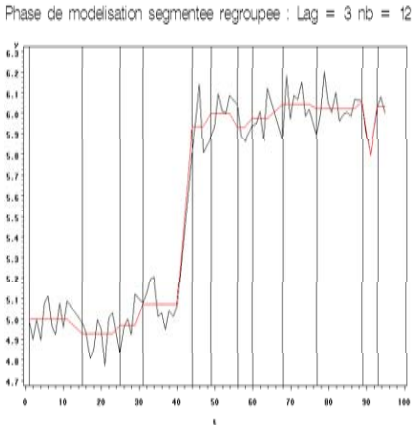
Un exemple simple

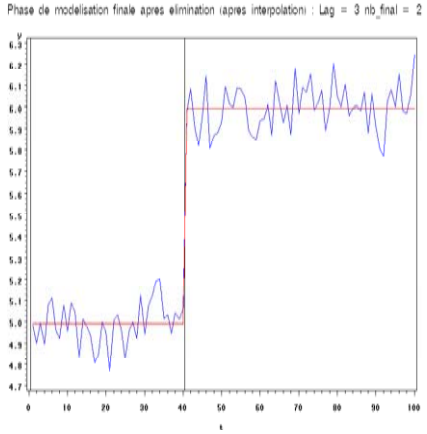












$Y_t \sim> \mathcal{N}(5\ ;\ 0.01)$ for $t = 1{,}40$

$Y_t \sim> \mathcal{N}(6\ ;\ 0.01)$ for $t = 41{,}100$

Covariance Parameter Estimates

Cov Parm	Group	Estimate
Residual	cls_final 1	0.009508
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

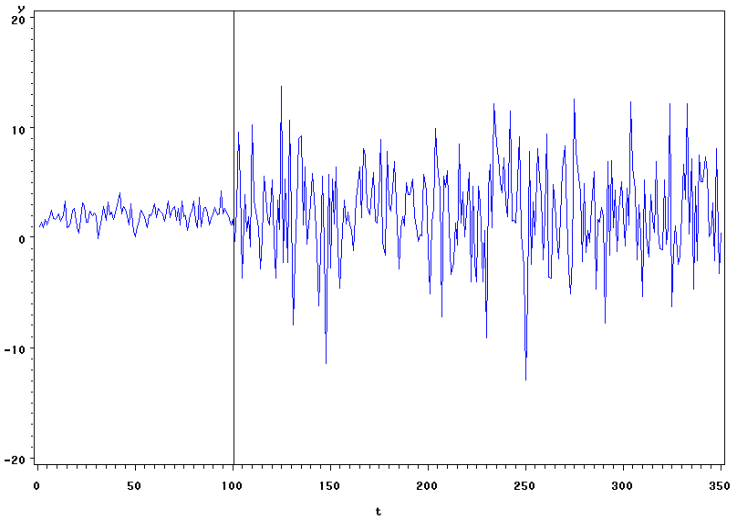
Effect	cls_final	Estimate	Standard Error	DF	t Value	Pr > |t|
t*drop_fin*cls_final	1	0	.	.	.	.
t*drop_fin*cls_final	2	0	.	.	.	.
cls_final	1	4.9904	0.01542	98	323.68	<.0001
cls_final	2	5.9950	0.01349	98	444.45	<.0001

Méthode proposée

Une nouvelle approche pour estimer la variance a priori

Rappel de la problématique :
→ Détection de segments contigus avec niveaux similaires, mais de variances différentes

Solution proposée :
→ Détection la volatilité a priori



Nouvelle méthode : ***Trois phases principales***

(i) Etablir une transformation appropriée de Y_t afin d'obtenir une nouvelle série U_t (or V_t) exhibant le comportement temporel de la dispersion de Y_t

(ii) Segmenter la nouvelle série U_t à l'aide de la méthode de segmentation (Derquenne, 2010) afin d'obtenir des segments de dispersion

(iii) Appliquer la méthode (Derquenne, 2010) sur Y_t mais en tenant compte de la distribution des segments de dispersion estimés U_t intégrée dans le modèle linéaire Gaussien hétéroscédastique

Méthode proposée

Etape 1: Transformation caractérisant la volatilité de la série temp.

Objectif : construire une nouvelle série temporelle pour exhiber la dispersion σ^2 de la série initiale supposée issue du modèle :

$$Y_t = \beta_0 + \beta_1 t + \varepsilon_t \qquad (2)$$

où $\varepsilon_t \sim> \mathcal{N}(0,\sigma^2)$, iid

(i) Utilisation de la différenciation naturelle :

$$Z_t = (I - B)^2 Y_t = Y_t - 2Y_{t-1} + Y_{t-2} = \varepsilon_t - 2\varepsilon_{t-1} + \varepsilon_{t-2} \qquad (3)$$

(ii) Application possible de deux opérateurs sur Z_t:

$$U_t = |Z_t| \qquad V_t = Z_t^2 \qquad (4)$$

(iii) Utilisation de deux théorèmes pour obtenir σ^2 issu de (4)

Méthode proposée

Etape 1 : Transformation caractérisant la volatilité de la série temp.

Théorème 1: Soit Y_t un processus Gaussien indicé dans le temps, de moyenne : $\beta_0+\beta_1 t$ et de variance σ^2, tel que : $Y_t = \beta_0+\beta_1 t+\sigma\varepsilon_t$ où $\varepsilon_t \sim> \mathcal{N}(0,1)$, i.i.d. alors

$$\sigma = \frac{\sqrt{\pi}}{2\sqrt{3}}\mathbf{E}(U_t) = \frac{\sqrt{\pi}}{2\sqrt{3}}\mathbf{E}\left(\left|Y_t - 2Y_{t-1} + Y_{t-2}\right|\right) \qquad (5)$$

Ce théorème fournit une estimation de σ : $\hat{\sigma}_U = \frac{\sqrt{\pi}}{2\sqrt{3}(T-2)}\sum_{t=3}^{T}\left|y_t - 2y_{t-1} + y_{t-2}\right|$

Théorème 2 : Soit Y_t un processus Gaussien indicé dans le temps, de moyenne : $\beta_0+\beta_1 t$ et de variance σ^2, tel que : $Y_t = \beta_0+\beta_1 t+\sigma\varepsilon_t$ où $\varepsilon_t \sim> \mathcal{N}(0,1)$, i.i.d. alors

$$\sigma^2 = \frac{1}{6}\mathbf{E}[V_t] = \frac{1}{6}\mathbf{E}\left[\left(Y_t - 2Y_{t-1} + Y_{t-2}\right)^2\right] \qquad (6)$$

Ce théorème fournit une estimation de σ^2 : $\hat{\sigma}_V^2 = \frac{1}{6(T-2)}\sum_{t=3}^{T}\left(y_t - 2y_{t-1} + y_{t-2}\right)^2$

Méthode proposée

Etape 2: Première segmentation sur la série transformée

→ Les résultats de ces deux théorèmes sont essentiels car ils font apparaître pour chaque observation de la série u_t (ou v_t), les niveaux de dispersion des segments de la série initiale

→ La méthode de segmentation (Derquenne, 2010) est alors appliquée sur u_t (ou v_t)

→ A la fin du processus, la segmentation fournit un ensemble de segments issus du modèle :

$$U_t = \sum_{s=1}^{S_1} \left(\alpha_0^{(s)} + \alpha_1^{(s)} t + \psi_s \zeta_t\right) 1_{[t \in \tau_s^\sigma]} \qquad (7)$$

→ Les segments obtenus peuvent être constants, croissants ou décroissants. Ils fournissent une information supplémentaire sur le comportement des données qui peuvent même encore être hétéroscédastiques sur certains segments

→ Soient $\tau_1^\sigma, \ldots, \tau_{S_1}^\sigma$ les S_1 segments of dispersion obtenus précédemment à partir de u_t, alors le segment τ_s^σ fournit une estimation de T_s valeurs u_t pour $t \in \tau_s^\sigma$ telles que :

$$\hat{u}_t = \hat{\alpha}_0^{(s)} + \hat{\alpha}_1^{(s)} t \qquad (8)$$

Méthode proposée

Etape 3 : Seconde segmentation sur la série initiale

→ **Cette étape fournit une segmentation finale de la série temporelle initiale** Y_t en tenant compte de la dispersion estimée des données lors de l'étape 2

→ **Deux possibilités :**

(i) Chaque valeur y_t est standardisée par $\hat{u}_t$ de façon à éliminer l'effet d'échelle issu de la première segmentation

(ii) Les $\hat{u}_t$ sont seulement intégrés lors de la phase de modélisation pour structurer la matrice de dispersion du modèle linéaire Gaussien hétéroscédastique (1).

⇒ Utilisation de la matrice diagonal de poids pour estimer le modèle à l'aide de l'estimateur REML

Plan

- Contexte, motivations et objectifs
- Méthode proposée
- **Application : un cas simulé**
- Application : un cas réel
- Apports, applications et futures recherches

Application : un cas simulé

Objectif : Comparer les résultats de la nouvelle méthode avec ceux de la précédente (Derquenne, 2010) et ceux issus d'une approche par programmation dynamique (Lavielle, 2009)

Simulation : 10 segments issus du modèle (1). Pour chaque segment, le nbre d'observations, les valeurs des coefficients $(\beta_0 ; \beta_1)$ et les écarts-types σ sont générés aléatoirement

La méthode fondée sur la programmation dynamique :

→ Détecte des points de rupture multiples dans une série temporelle

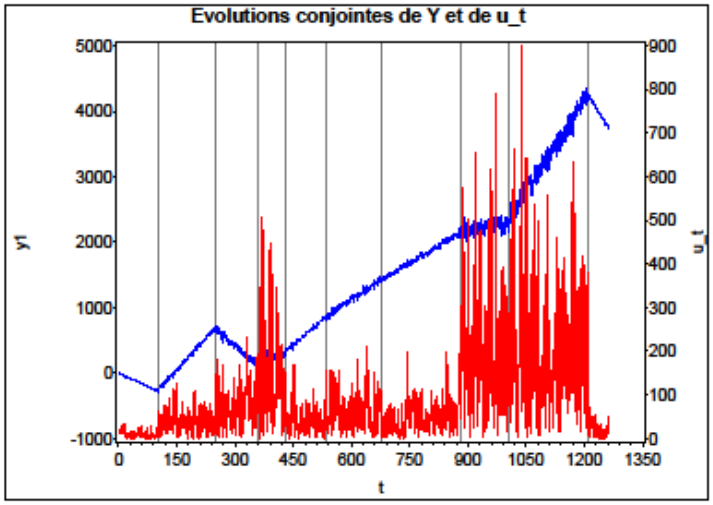
→ Estime le modèle par minimisation sous contraintes de fonctions pénalisées

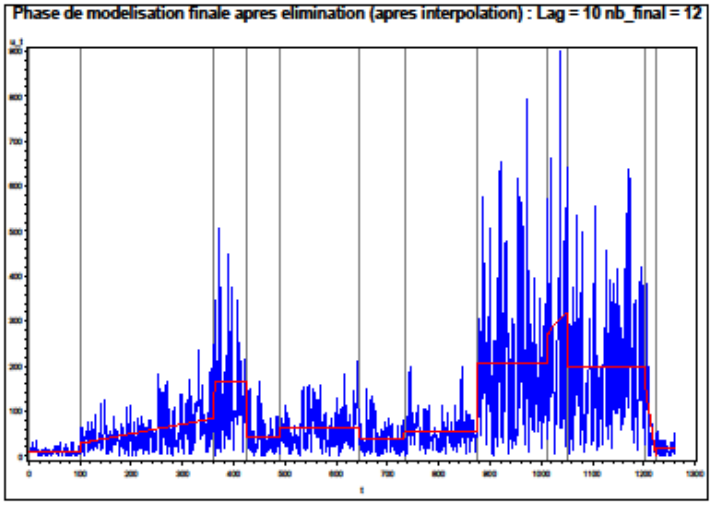
→ Types de ruptures détectés :

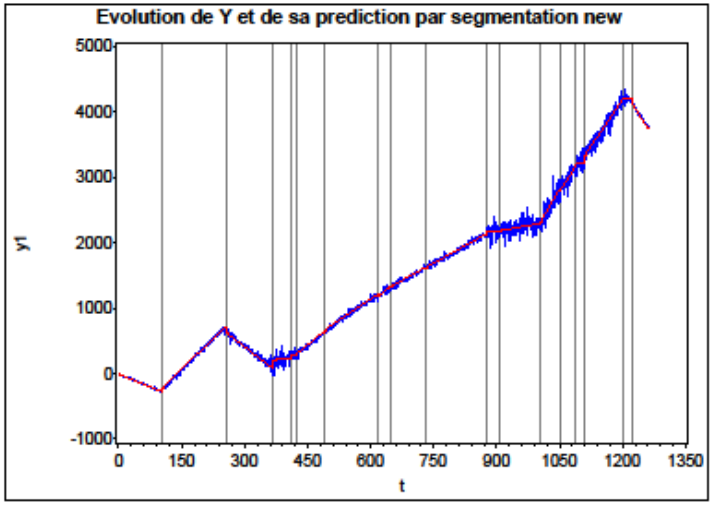
(i) Moyenne, mais variance constante, variance, mais moyenne constante, moyenne et variance non constantes (DCPC3), et avec approche bayesienne (BDCPC3)

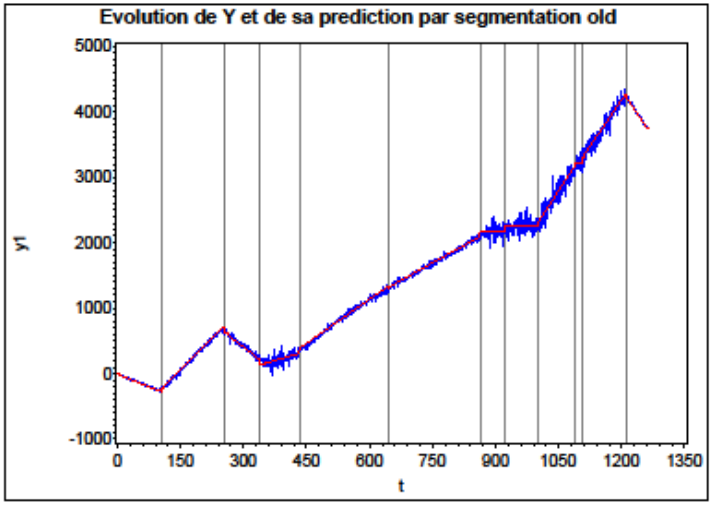
(ii) Distribution (DCPC4) et avec approche bayesienne (BDCPC4)

Application : un cas simulé

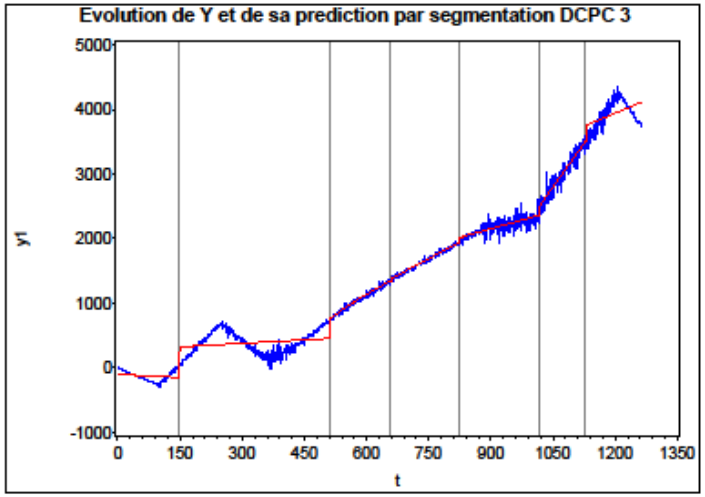


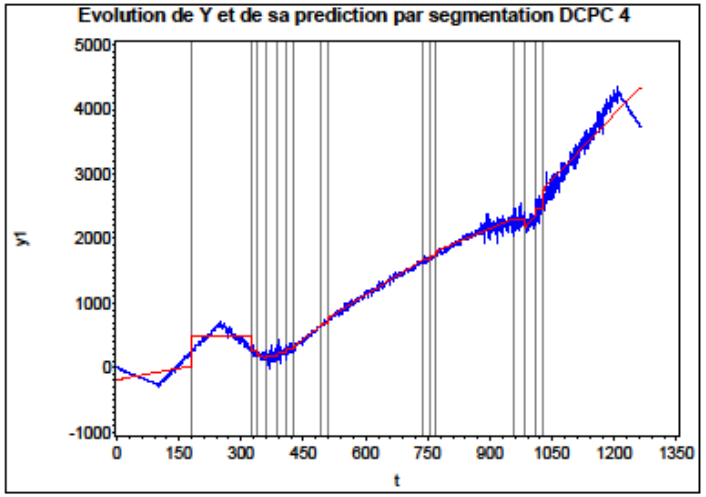


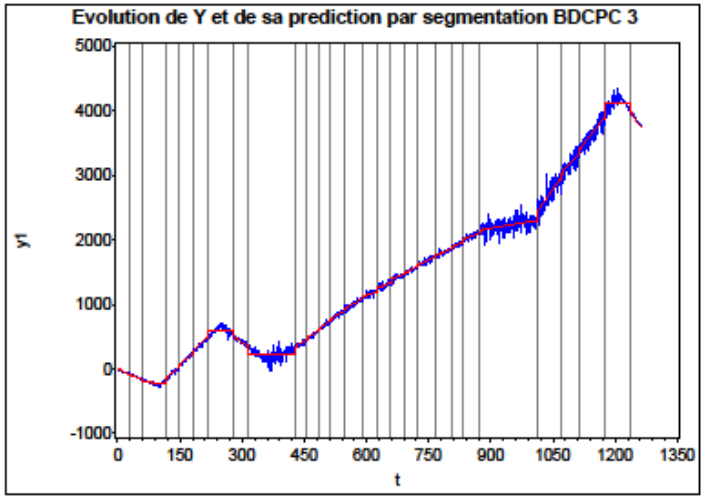


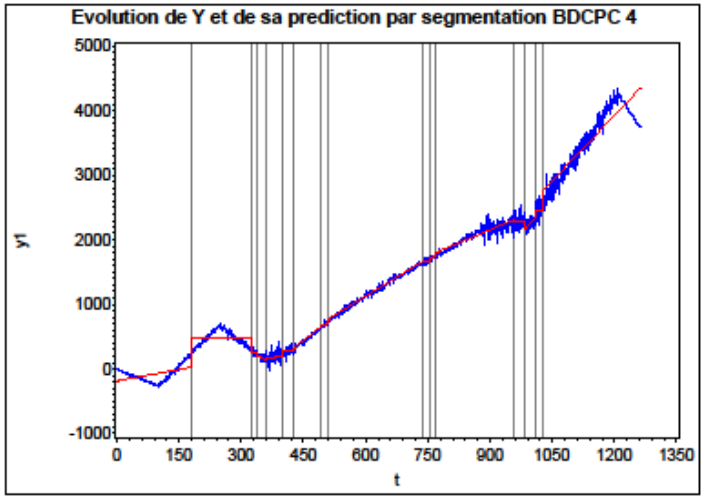


Application : un cas simulé









Application : un cas simulé

Résultats des six méthodes

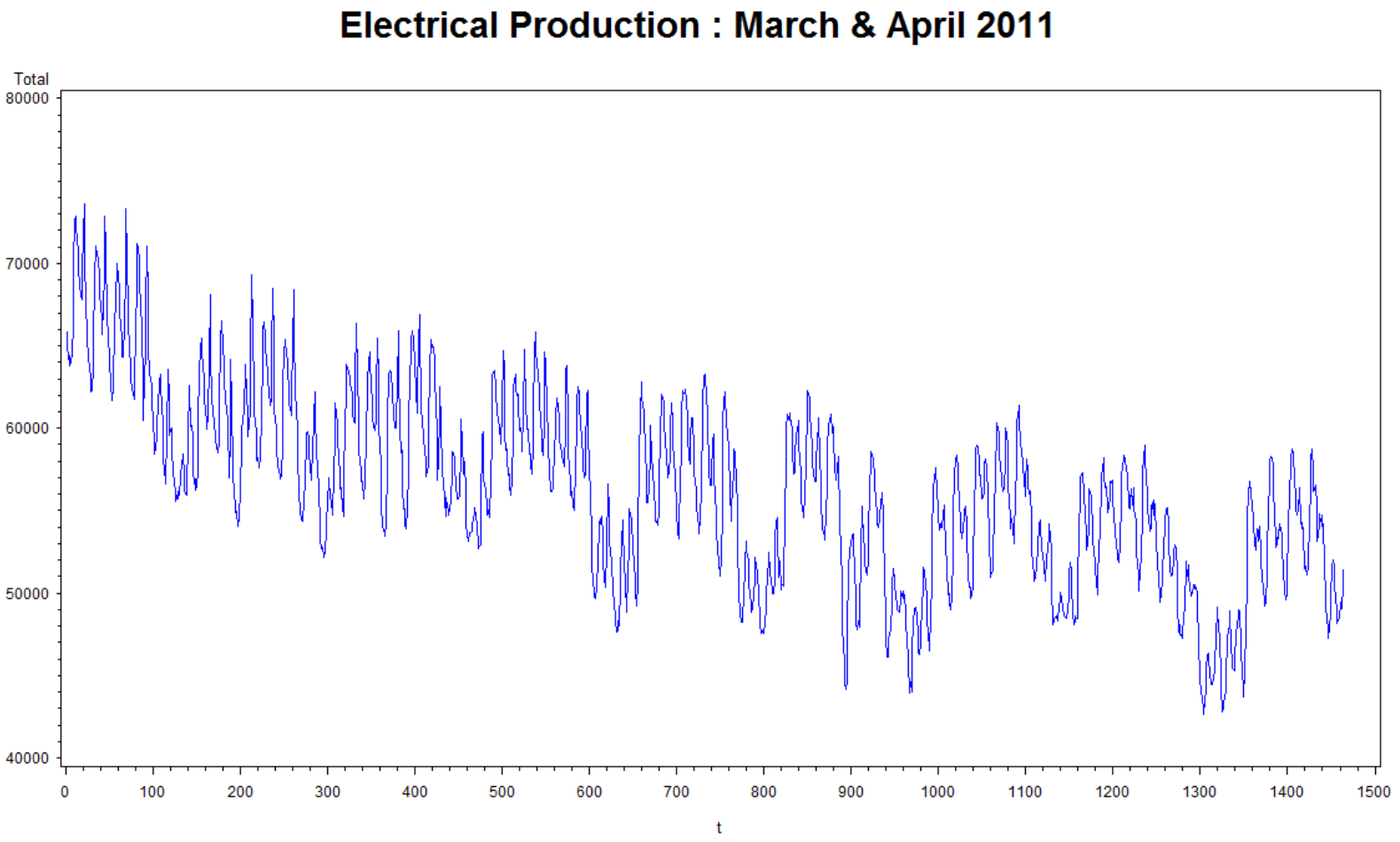
Méthode	*Nb. segments*	*MAPE (%)*	*MED (%)*	*> 10%*	*Seg. détectés*	*Err. seg. (%)*
Simulation	10	9,90	2,32	12,68	n.a.	n.a
1ière méthode	12	11,06	2,23	13,71	6	24,34
Nouvelle méthode	18	10,24	2,30	13,00	6	16,26
DCPC3	7	75,80	4,48	32,69	0	36,77
DCPC4	17	59,72	1,00	29,71	3	52,00
BDCPC3	28	12,42	2,55	16,32	2	35,28
BDCPC4	16	59,85	4,03	10,21	3	48,94

Plan

- Contexte, motivations et objectifs
- Méthode proposée
- Application : un cas simulé
- **Application : un cas réel**
- Apports, applications et futures recherches

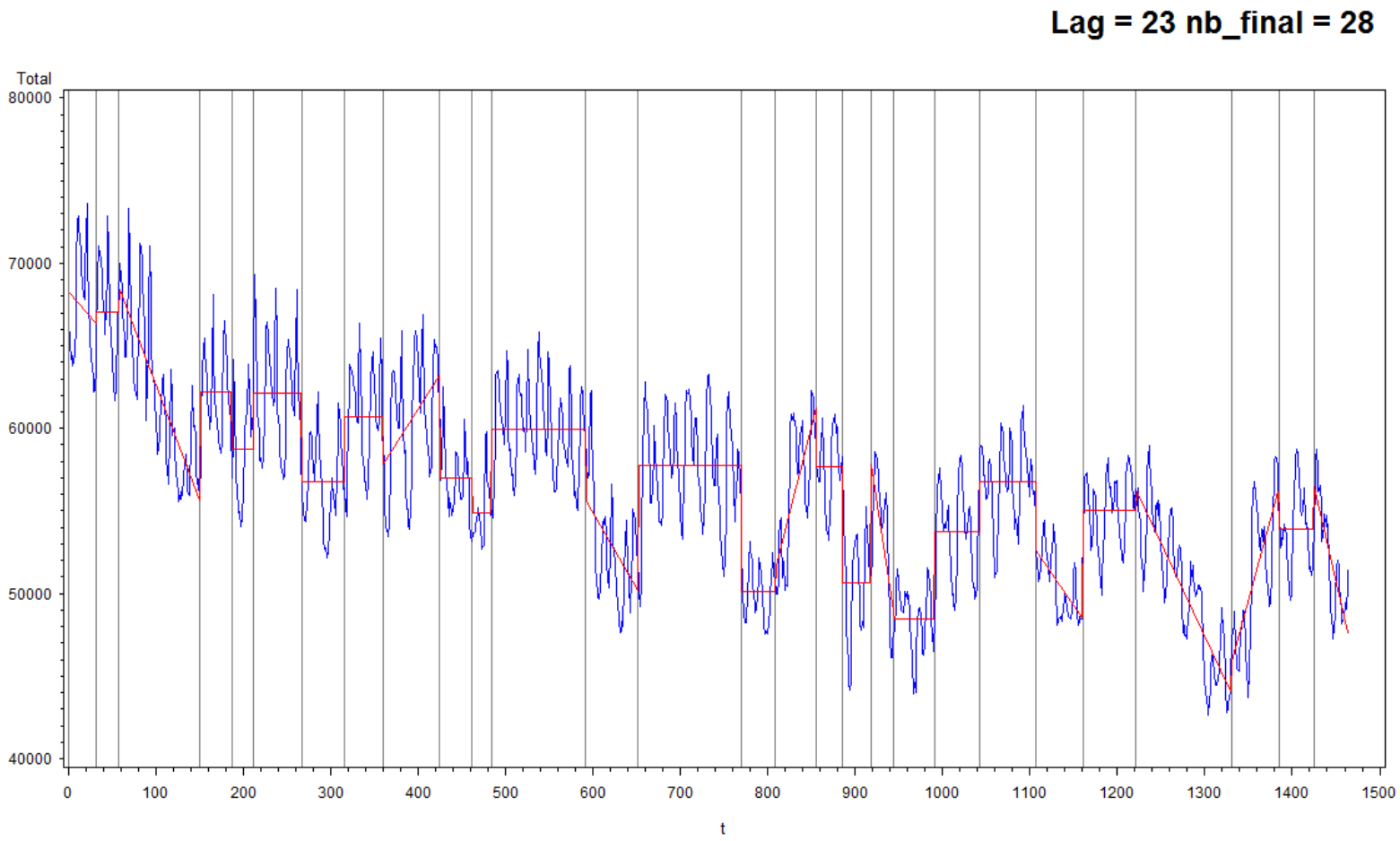
Application : un cas réel

Objectif : *Identifier le comportement de la production d'électricité*



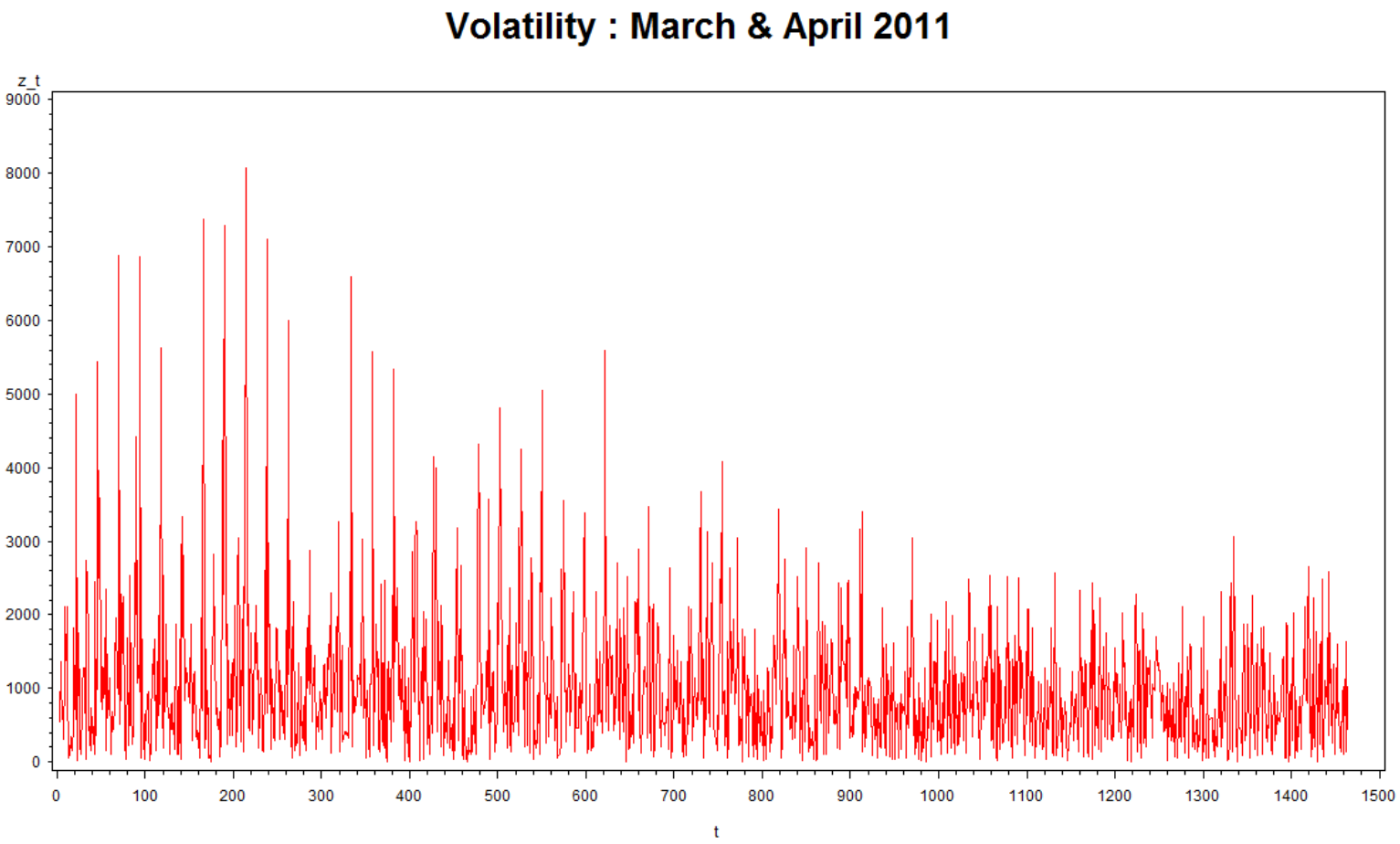
Application : un cas réel

Résultats de la première méthode (Derquenne, 2010)



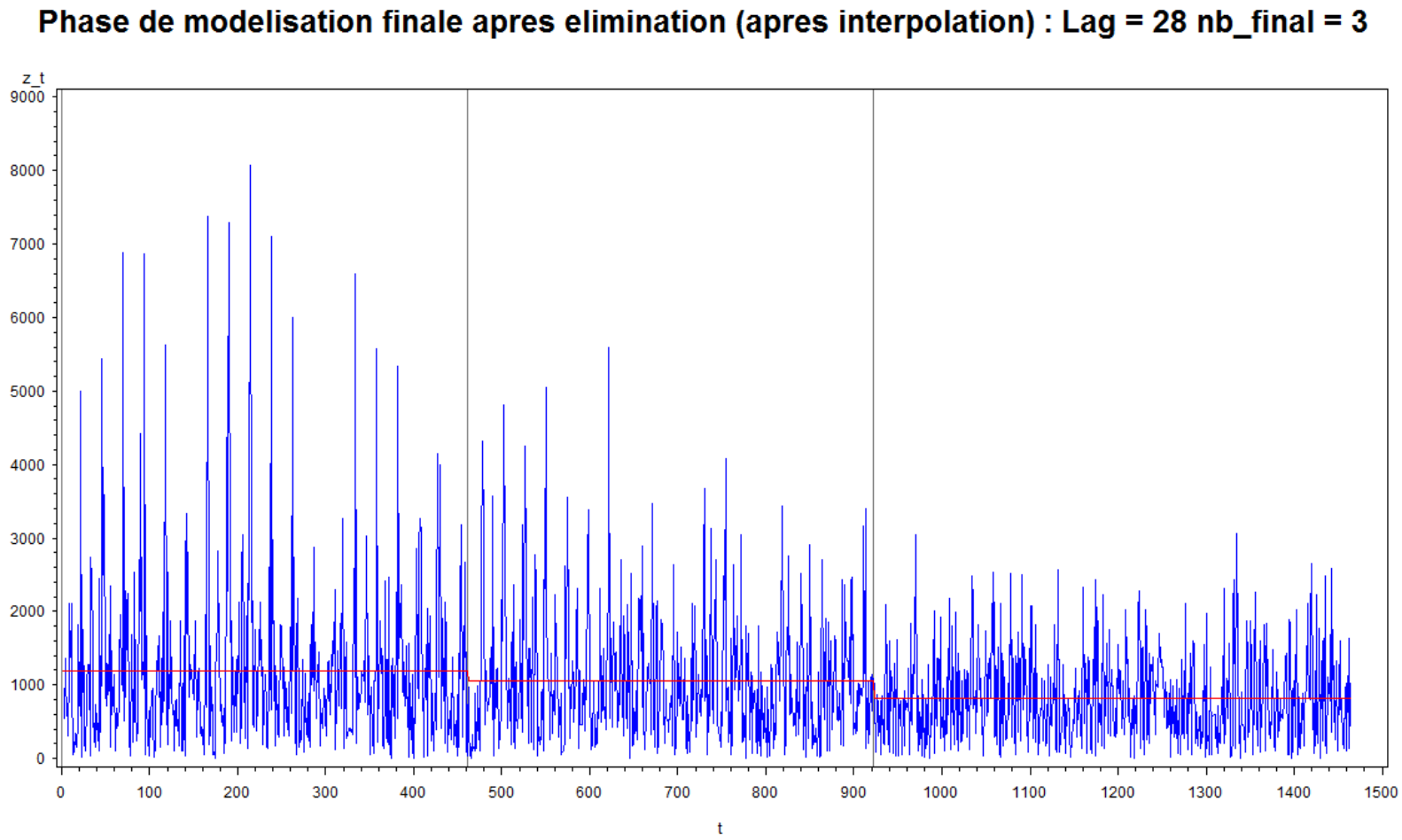
Application : un cas réel

Evolution de la série transformée : u_t



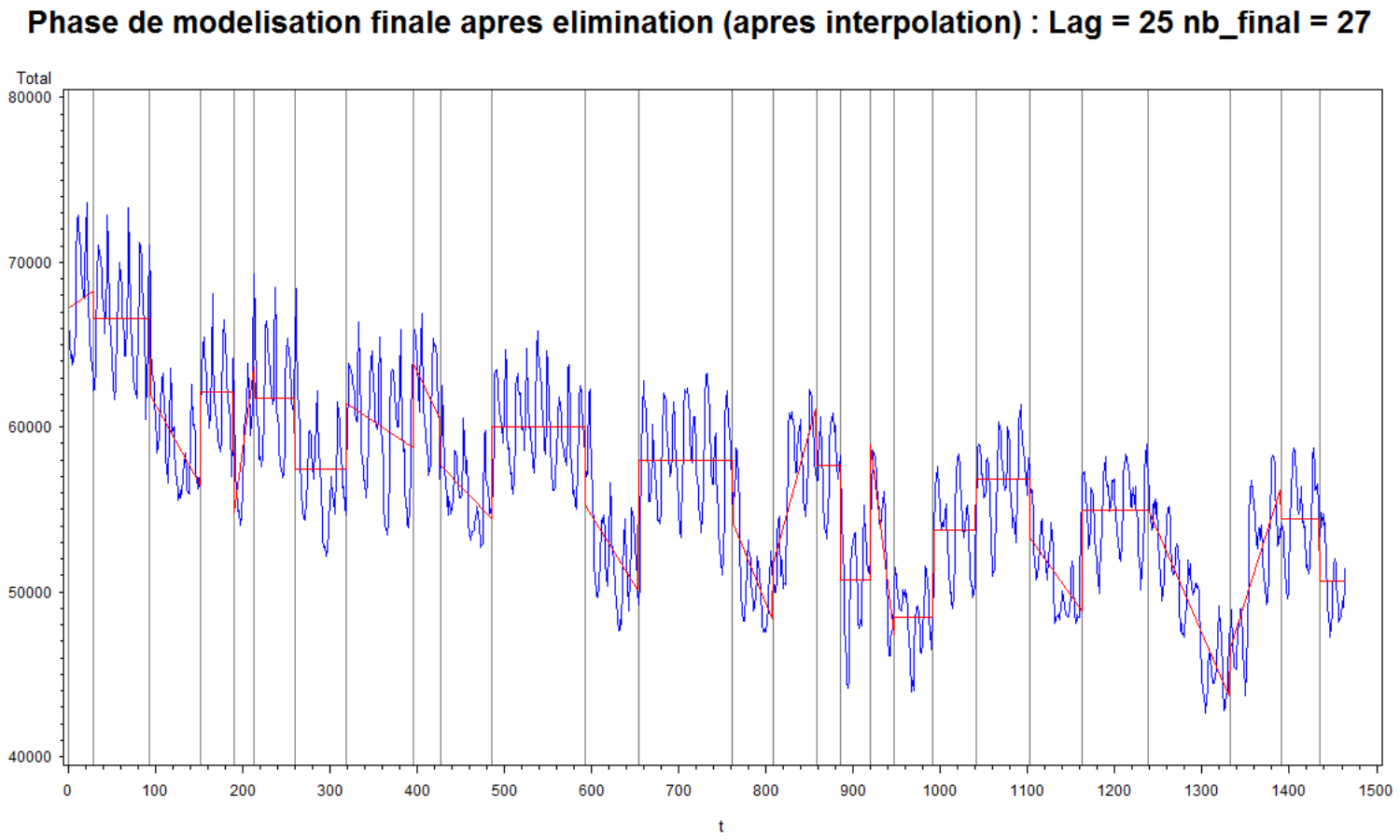
Application : un cas réel

Segmentation de la série transformée : u_t



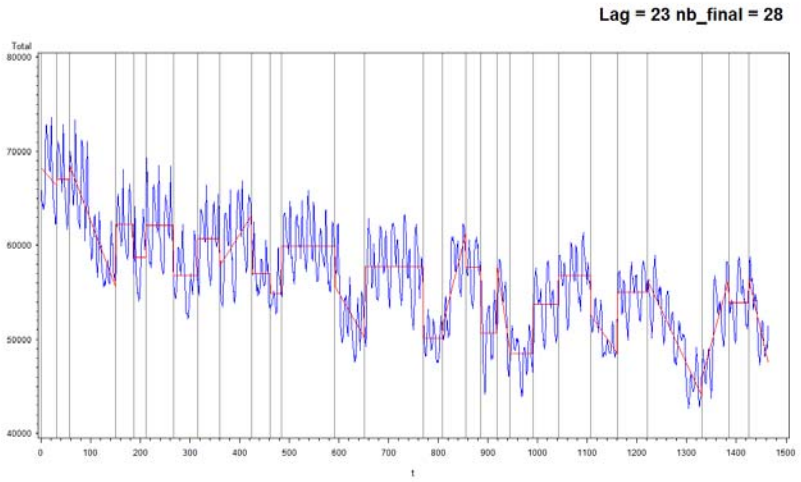
Application : un cas réel

Segmentation de la série initiale y_t en tenant compte de des données transformées : u_t (Derquenne, 2011)



Application : un cas réel

Segmentation de la série initiale y_t en tenant compte de des données transformées : u_t (Derquenne, 2011)



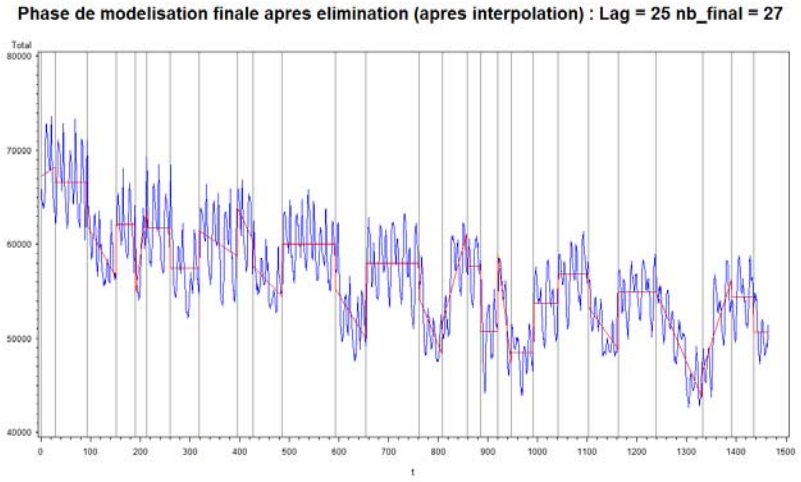
Ancienne méthode (Derquenne, 2010)

MAPE = 3,91% PCT>10% = 2,73%

R2_adj = 0,767 BIC = 26915,4

τ_2 : 2/03 (7:00) to 3/03 (5:00)

τ_3 : 3/03 (6:00) to 6/03 (0:00)



Nouvelle méthode (Derquenne, 2011)

MAPE = 3,93% PCT>10% = 2,60%

R2_adj = 0,769 BIC = 26898,9

τ_2 : 2/03 (4:00) to 4/03 (20:00)

τ_3 : 4/03 (21:00) to 6/03 (0:00)

Plan

- Contexte, motivations et objectifs
- Méthode proposée
- Application : un cas simulé
- Application : un cas réel
- **Apports, applications et futures recherches**

Apports, applications et futures recherches

La démarche de la méthode proposée :

(i) Produit un signal issu des données initiales représentant leur dispersion avec 2 théorèmes

(ii) Etablit une première segmentation de ce signal pour obtenir des segments de dispersion

(iii) Seconde segmentation sur les données initiales avec prise en compte des segments estimés

Apports, applications et futures recherches

Apports de la méthode proposée :

(i) Améliore la démarche introduite dans l'ancienne méthode (Derquenne, 2010)

(ii) Améliore la qualité des résultats sur des données simulées et des données réelles (2010)

(iii) Concurrence fortement la méthode utilisant la programmation dynamique (Lavielle, 2009)

Apports, applications et futures recherches

Domaines d'applications de la méthode proposée :

$\Rightarrow$ Particulièrement utile pour des signaux qui changent de processus à volatilité hétérogène

(i) Finance (CAC40, S&P500, FTSE100, prix de marché de l'énergie, etc.)

(ii) Séquençage humain

(iii) Management de l'énergie (consommation, prix de marché, etc.)

Apports, applications et futures recherches

Futures recherches :

(i) *Comparaison de cette méthode avec une autre méthode* fournissant une segmentation en niveau pour données hétéroscédastiques via une validation croisée (Arlot & al., 2010)

(ii) *Construction d'un méta-modèle* obtenu à l'aide plusieurs segmentations

(iii) *Généralisation des deux théorèmes* à d'autres processus que le processus Gaussien

(iv) *Applications de la segmentation :* modèle (réponse *vs X*'s segmentés) + classification de courbes

Bibliography

Arlot, S. & Celisse, A. (2010): Segmentation of the mean of heteroscedastic data via cross-validation, *Statistics and Computing,* pp. 1-20.

Derquenne, C. (2011): An Explanatory Segmentation Method for Time Series, *in Proceedings of Compstat 2010*, Y. Lechevallier & G. Saporta (eds.), 1st Edition, pp. 935-942.

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*.

Hébrail G., Hugueney B., Lechevallier Y., Rossi F. (2010): Exploratory analysis of functional data via clustering and optimal segmentation. *Neurocomputing* **73** (7-9): pp. 1125-1141.

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

Lavielle, M. (2009): Detection of Changes using a Penalized Contrast (the DCPC algorithm), http://www.math.u-psud.fr/~lavielle/programmes_lavielle.html.

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

Rao, CR. and Kleffe, J. (1988): *Estimation of variance components and applications*. North Holland series in statistics and probability, Elsevier.