SEGMENTATION DE SÉRIES TEMPORELLES AVEC PRISE EN COMPTE A PRIORI DE COMPOSANTES DE VARIANCE

Christian DERQUENNE

EDF R&D

Problématique

Les séries temporelles se décomposent généralement en plusieurs types d'évolution : tendance, saisonnalité, volatilité et bruit. Elles peuvent être plus ou moins régulières selon le domaine d'application. Citons deux exemples à évolution régulières. La courbe de températures climatiques tri-horaires se décompose sous forme d'une sinusoïdale que l'on peut estimer pour obtenir une température normale, l'écart à la normale représentera une approximation du bruit. Même si une tendance peut être soupçonnée, cela reste marginal et de toute façon, il s'agit d'un phénomène lent sur le très long terme. L'évolution de la consommation d'électricité globale[1] française sur 50 ans offre une tendance croissante et une saisonnalité. De nombreux modèles statistiques de prévision de la consommation pour le lendemain ont été mis au point et fournissent une moyenne des erreurs relatives absolues en-dessous de 1,5%. Cependant, il existe de nombreux phénomènes irréguliers, dans le sens où ils sont moins prévisibles, telles que les séries financières : prix de marchés de l'énergie, indices tels que le CAC40, le FTSE 100, le S&P 500, etc. Généralement ces séries possèdent en plus une certaine volatilité qui peut être exhibée à l'aide des rendements. Ici la tendance, et surtout la saisonnalité, apparaissent généralement moins fréquemment et moins régulièrement. Mais ce sont les changements de comportements qui caractérisent principalement ces séries. Ces changements peuvent être soit des pics (prix d'une énergie en situation tendue, mais sur une très courte période), soit des sauts en niveau ou en tendance (rassemblement ou séparation de flux de données), en variabilité (rendement du FTSE 100). La modélisation de ces séries est donc très délicate et demande beaucoup d'expérience dans le domaine d'application. Quant à la prévision ; elle peut friser, dans certains cas, l'utopie. Il peut alors être intéressant de détecter des ruptures des comportements pour de nombreuses applications dans le cadre de pré-traitement ou non des données : construction de sous-modèles sur chaque segment établi, stationnarisation de la série à l'aide de la segmentation, construction de courbes symboliques [11] dans l'optique de réaliser une classification de courbes, modélisation de séries temporelles multivariées, etc.

De nombreuses méthodes ont été et sont développées pour répondre à différentes problématiques en économie, en finance, en séquençage humain, en météorologie, en management de l'énergie, etc. Plusieurs classes de méthodes existent : de l'exploration de l'espace de toutes les segmentations possibles pour un nombre successif de ruptures dans un objectif de validation de modèle [9] à l'inférence sur des modèles à ruptures multiples dans des séries temporelles multivariées [13], en passant par des tests de détection de changements structurels multiples dans des modèles de régression cointégrés [15] ou encore par la détection de ruptures séquentielles lorsque le changement de comportement des paramètres est inconnu [12]. La plupart de ces algorithmes utilisent la programmation dynamique pour diminuer drastiquement le nombre de segmentations possible car il serait bien évidemment complètement illusoire de vouloir les calculer toutes. En effet, le nombre de segmentations pour une série de longueur T et un nombre S fixé de segments vaut $\binom{T-1}{S-1}$ alors que pour l'ensemble de tous les segments de $S=1,T$, le nombre total de segmentations passe à 2^{T-1}. Par exemple, dans le cas d'exploration de l'espace, la complexité est généralement en $\mathrm{O}\left(ST^2\right)$ pour le temps et en $\mathrm{O}(ST)$ pour l'espace (la complexité de cet espace linéaire est d'autant plus élevée que la série est longue). Par contre, cette complexité peut descendre en $\mathrm{O}\left(T^2\right)$ [13], même dans le cadre de ruptures multiples pour M séries temporelles multivariées, alors qu'elle pourrait être en $\mathrm{O}\left(MT^2\right)$. Ces

[1] Cumul des consommations des différents distributeurs du marché de l'électricité en France

méthodes de détection de points de rupture ont pour vocation de résoudre trois problèmes [13] : la détection de changement de moyenne, avec une variance constante, la détection de changement de variance avec une moyenne constante et la détection de changements dans l'ensemble de la distribution du phénomène étudié.

La méthode de segmentation de séries temporelles [3,4] que nous avons introduite permet d'une part de résoudre un quatrième problème qui est de détecter des croissances ou des décroissances dans la série étudiée [15], mais aussi de réduire la complexité du problème en $O(KT)$, où K est le nombre de degrés de lissage dont nous discuterons plus loin, pour proposer des solutions de segmentation (de partitionnement) de la série en une suite de segments pouvant être croissants, décroissants, constants et en ayant des variances différentes ou non.

Notre méthode est originale dans son approche car elle propose, par étapes successives, une aide à la décision pour la segmentation des données. Elle contient deux phases principales : la préparation des données offrant une première segmentation des observations et la modélisation des segments à l'aide d'un modèle linéaire gaussien hétéroscédastique par adaptations successives. Chacune de ses deux phases est répétée un certain nombre de fois en fonction du degré de lissage appliqué aux données. Le degré de lissage peut varier de 1 à T théoriquement. La complexité empirique est en $O(T\sqrt{T})$ et la complexité théorique est en $O(T^2)$. Cette méthode a été testée sur de nombreuses séries et a fourni des résultats encourageants à la fois sur des données simulées afin de juger de la qualité de reconstitution de la série : détection et modélisation des segments, mais surtout sur des données réelles, notamment dans le domaine de la formation des prix de marché de l'énergie.

Cependant pour l'ensemble des méthodes de segmentation qu'elles soient fondées sur la programmation dynamique ou sur une approche exploratoire comme la nôtre, il s'avère que la qualité de la segmentation peut faire défaut lors de la détection de segments contigus quand les niveaux (constants ou pentes) sont proches statistiquement mais ont des variances différentes. Dans ce cas un seul segment sera détecté, alors qu'il y en a deux structurellement. Par conséquent, nous proposons dans cet article, une nouvelle méthode améliorant la précédente. Cette nouvelle approche contient trois phases. La première consiste à établir une transformation adéquate des données afin d'obtenir une nouvelle série caractérisant l'évolution temporelle de la dispersion des observations, la deuxième phase revient à segmenter cette nouvelle série avec le même principe que la méthode [3,4] pour obtenir des segments de dispersion, enfin la troisième phase applique à nouveau [3,4] mais en tenant compte de la distribution des segments de dispersion, notamment lors de la construction du modèle linéaire hétéroscédastique. Afin de tester notre approche, nous avons alors réalisé une étude comparative avec des algorithmes de programmation dynamique proposés en [14]. Comme nous pourrons le constater la qualité des résultats a été considérablement améliorée. Enfin, nous proposons d'étendre les comparaisons avec d'autres méthodes, ainsi que des voies futures de recherche consistant par exemple à généraliser les deux théorèmes introduits dans cet article.

1. La méthode initiale proposée

1.1. Le modèle et son inférence

Soit une série temporelle (Y_t), $t = 1,T$, nous supposons qu'elle se décompose selon le modèle linéaire hétéroscédastique (ou à composantes de variances) [16,17] suivant :

$$Y_t = \sum_{s=1}^{S} \left(\beta_0^{(s)} + \beta_1^{(s)} t + \sigma_s \varepsilon_t\right) 1_{[t \in \tau_s]} \qquad (1.1)$$

où $\beta_0^{(s)}$, $\beta_1^{(s)}$ et σ_s > 0, sont respectivement les paramètres de niveau, de pente et de dispersion pour le segment τ_s, et ε_t est un bruit centré de variance unité. Enfin, le nombre d'observations par segment τ_s est noté T_s, avec $\sum_{s=1}^{S} T_s = T$. Chaque segment τ_s contient l'ensemble des valeurs : Y_t pour $t = U_{s-1} + 1$ à U_s, où $U_s = U_{s-1} + T_s$ et $U_0 = 0$, finalement $U_S = T$. Il y a donc $3S$ paramètres à estimer, sachant que le nombre S de segments et la taille des segments sont inconnus. L'approche proposée est donc complètement non supervisée.

Dans le cas de ce modèle linéaire hétéroscédastique, plusieurs estimateurs sont disponibles : moindres carrés ordinaires (OLS), maximum de vraisemblance (ML) et maximum de vraisemblance restreint ou résiduel (REML) [2,6,7,8,10]. Rappelons, les formes respectives de ces trois estimateurs.

L'estimateur des moindres carrés ordinaires est obtenu en minimisant la quantité suivante pour obtenir le vecteur β des paramètres :

$$L_{OLS}(\beta, \mathbf{y}) = (\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta) \tag{1.2}$$

alors $\hat{\beta}_{OLS} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$, où $\mathbf{y}$ et $\mathbf{X}$ sont respectivement les vecteurs de la variable réponse et des variables candidates à l'explication.

Remarque : l'estimateur OLS n'estimera pas directement les variances σ_1^2, ..., σ_s^2, ..., σ_S^2. Seule la variance globale (non biaisée) sera estimée a posteriori, telle que :

$$\hat{\sigma}^2_{OLS} = \sum_{s=1}^{S}\sum_{t\in\tau_s}\left(y_t - \hat{\beta}_0^{(s)} - \hat{\beta}_1^{(s)}t\right)^2 \Big/ (T - p) \tag{1.3}$$

où p correspond au nombre de paramètres estimés.

Cependant, il est possible d'utiliser les moindres carrés pondérés (WLS) tel que : $\hat{\beta}_{WLS} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{y}$. Si dans la matrice $\mathbf{W}$, les poids des observations sont proportionnels à l'inverse des variances des erreurs, alors WLS sera le meilleur estimateur linéaire sans biais (BLUE). Cela demandera que ces variances soient connues a priori, ce qui est rarement le cas en pratique.

En faisant l'hypothèse que les erreurs sont gaussiennes, l'estimateur du maximum de vraisemblance permet de pallier ce problème. Cet estimateur sera obtenu en prenant l'argument maximum de deux fois la fonction de vraisemblance suivante :

$$L_{ML}(\beta, \mathbf{V}, \mathbf{y}) = -\frac{1}{2}\left[\log|\mathbf{V}| + \mathbf{r}'\mathbf{V}^{-1}\mathbf{r} + n\log(2\pi)\right] \tag{1.4}$$

où $\mathbf{V}$ est la matrice de variances-covariances des erreurs ε. Dans notre cas, $\mathbf{V}$ est une matrice diagonale et aura la forme suivante :

$$\mathrm{V} = \begin{pmatrix} \sigma_1^2 & 0 & \cdots & 0 & & & & \\ 0 & \sigma_1^2 & \cdots & 0 & & & & \\ \vdots & \vdots & \ddots & \vdots & & 0 & & \\ 0 & \cdots & 0 & \sigma_1^2 & & & & \\ & & & & \ddots & & & \\ & & & & & \ddots & & \\ & & & & & & \ddots & \\ & & & & \sigma_S^2 & 0 & \cdots & 0 \\ & & 0 & & 0 & \sigma_S^2 & \cdots & 0 \\ & & & & \vdots & \vdots & \ddots & \vdots \\ & & & & 0 & \cdots & 0 & \sigma_S^2 \end{pmatrix} \begin{matrix} 1 \\ 2 \\ \vdots \\ T_1 \\ . \\ . \\ . \\ 1 \\ 2 \\ \vdots \\ T_S \end{matrix} \tag{1.5}$$

avec T_1 variances σ_1^2, ..., T_s variances σ_s^2, ... et T_S variances σ_S^2, mais il y aura seulement S variances à estimer correspondant à la matrice par blocs diagonaux pour les segments définis dans le modèle (1.1). Dans ce cas, les variances seront estimées directement, par contre celles-ci seront biaisées. Dans le cas de notre modèle (1.1), la variance estimée des erreurs pour le segment τ_s prendra la forme suivante :

$$\hat{\sigma}^2_{ML(s)} = \sum_{t\in\tau_s}\left(y_t - \hat{\beta}_0^{(s)} - \hat{\beta}_1^{(s)}t\right)^2 \Big/ T_s \qquad \text{pour } s = 1 \text{ à } S \tag{1.6}$$

Enfin, le vecteur des résidus du modèle est tel que : $\mathbf{r} = \mathbf{y} - \mathbf{X}\left(\mathbf{X'}\mathbf{V}^{-1}\mathbf{X}\right)^{-}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y} = \mathbf{y} - \mathbf{X}\beta$.

L'avantage de l'estimateur REML sur celui du maximum de vraisemblance est qu'il fournit directement des estimateurs sans biais des composantes de variances et de covariances. La fonction de vraisemblance restreinte prend alors la forme suivante :

$$L_{REML}\ (\beta, \mathbf{V}, \mathbf{y}) = -\frac{1}{2}\left[\log|\mathbf{V}| + \log\left|\mathbf{X'}\mathbf{V}^{-1}\mathbf{X}\right| + \mathbf{r'}\mathbf{V}^{-1}\mathbf{r} + (n-p)\log(2\pi)\right] \tag{1.7}$$

Dans notre cas, les variances estimées sans biais des erreurs vaudront alors :

$$\hat{\sigma}^2_{REML\ (s)} = \sum_{t\in\tau_s}\left(y_t - \hat{\beta}_0^{(s)} - \hat{\beta}_1^{(s)}t\right)^2 \Big/ (T_s - 2) \qquad \text{pour } s = 1 \text{ à } S \tag{1.8}$$

En termes d'estimation des paramètres, les trois estimateurs fourniront les mêmes solutions pour les vecteurs de niveau et de pente : $\left(\beta_0^{(1)},\ldots\beta_0^{(S)}\right)$ et $\left(\beta_1^{(1)},\ldots\beta_1^{(S)}\right)$.

En termes de test sur les paramètres, les valeurs du *t* de Student seront différentes pour les trois estimateurs car les variances associées seront différentes. Premièrement, seuls les estimateurs ML et REML permettront de réaliser de l'inférence en tenant compte des composantes de variance. Deuxièmement, ces dernières seront sans biais, seulement pour l'estimateur REML. Par conséquent, les trois types de *t* de Student associés aux paramètres $\left(\beta_0^{(s)}, \beta_1^{(s)}\right)$ prendront les formes suivantes :

$$stderr_{\ OLS}\left(\hat{\beta}_0^{(s)}\right) = \hat{\sigma}_{OLS}\sqrt{1\Big/T_s + \bar{t}_s^{\,2}\Big/\sum_{t\in\tau_s}\left(t-\bar{t}_s\right)^2} \qquad stderr_{\ OLS}\left(\hat{\beta}_1^{(s)}\right) = \hat{\sigma}_{OLS}\Big/\sqrt{\sum_{t\in\tau_s}\left(t-\bar{t}_s\right)^2} \tag{1.9}$$

$$stderr_{\ ML}\left(\hat{\beta}_0^{(s)}\right) = \hat{\sigma}_{ML\ (s)}\sqrt{1\Big/T_s + \bar{t}_s^{\,2}\Big/\sum_{t\in\tau_s}\left(t-\bar{t}_s\right)^2} \qquad stderr_{\ ML}\left(\hat{\beta}_1^{(s)}\right) = \hat{\sigma}_{ML\ (s)}\Big/\sqrt{\sum_{t\in\tau_s}\left(t-\bar{t}_s\right)^2} \tag{1.10}$$

$$stderr_{\ REML}\left(\hat{\beta}_0^{(s)}\right) = \hat{\sigma}_{REML\ (s)}\sqrt{1\Big/T_s + \bar{t}_s^{\,2}\Big/\sum_{t\in\tau_s}\left(t-\bar{t}_s\right)^2} \qquad stderr_{\ REML}\left(\hat{\beta}_1^{(s)}\right) = \hat{\sigma}_{REML\ (s)}\Big/\sqrt{\sum_{t\in\tau_s}\left(t-\bar{t}_s\right)^2} \tag{1.11}$$

où $\bar{t}_s$ est la moyenne des valeurs *t* appartenant au segment τ_s.

Bien évidemment, le modèle (1.1) est seulement valable statistiquement si l'hypothèse nulle d'homoscédasticité est rejetée. La statistique de test utilisée est celle de deux fois le logarithme du rapport des vraisemblances à variances hétérogènes (modèle hétéroscédastique) et à variance constante (modèle homoscédastique). Sous l'hypothèse nulle, cette statistique suit une loi du χ^2 à (*S*-1) degrés de liberté. Si le modèle est homoscédastique, alors $\forall\ s = 1,S$, $\sigma_s = \sigma$ estimé par $\hat{\sigma}$ pour donner les écarts-types d'erreur des couples de coefficients $\left(\beta_0^{(s)}, \beta_1^{(s)}\right)$: $stderr_{\ OLS}\left(\hat{\beta}_0^{(s)}\right)$ et $stderr_{\ OLS}\left(\hat{\beta}_1^{(s)}\right)$.

1.2. La démarche générale de segmentation

La méthode proposée est essentiellement originale dans sa démarche, c'est-à-dire dans les étapes successives visant à fournir une aide à la décision pour la segmentation des données. En effet, les outils statistiques utilisés pour la modélisation sont tout à fait classiques, comme nous avons pu le constater dans le paragraphe 1.1, mais ils serviront dans une des étapes de la méthode. Il y a deux ensembles d'étapes, le premier correspond à une phase de préparation des données afin d'offrir un moyen raisonnable pour segmenter les données, alors que l'autre correspond à une phase de modélisations successives et adaptatives.

La phase de préparation des données se déroule de la façon suivante. La première étape consiste à « raboter » avec une certaine finesse de grain de ponçage la série temporelle des données, afin d'éliminer les « impuretés ». En d'autres termes, cette étape jouera le rôle que nous pourrions avoir face à une série de données pour la lisser visuellement. Les données étant lissées, il s'agit alors de découvrir les tendances de celles-ci : croissante, décroissante ou constante. Pour cela, une étape de différenciation est nécessaire car elle fournira, en fonction du niveau de lissage, des suites de valeurs positives, négatives ou nulles. Le résultat ainsi obtenu permettra de calculer dans l'étape suivante, la taille de ces suites en comptant le nombre de valeurs de même signe. Ces suites constitueront les segments initiaux et seront d'autant plus nombreuses que la finesse du degré de lissage sera faible.

Ces segments vont être utilisés pour estimer le premier modèle de type (1.1). Celui-ci correspondra à la première étape de la seconde phase de modélisations successives et adaptatives. Ce modèle est estimé par la méthode REML. Un premier test d'homoscédasticité est appliqué, afin de travailler sur des bonnes statistiques de test relatives aux paramètres de pentes et aux constantes. Les résultats issus de ces tests permettront de construire un modèle simplifié dans lequel des coefficients de pentes pourront être mis à zéro. Comme nous l'avons précisé, ce modèle risque d'être constitué de beaucoup de segments obtenus numériquement grâce à la phase de préparation des données. Par conséquent, il peut être intéressant de simplifier le modèle en regroupant des segments. Pour cela, des tests de comparaison des coefficients (pentes et constantes) sont appliqués sur chaque paire de segments successifs. Ce nouveau modèle contiendra de nouveaux segments, moins nombreux, mais peut-être encore trop élevés pour obtenir une segmentation raisonnable. Par conséquent, il repasse par le même processus que le premier : test d'homoscédasticité, tests de significativité des coefficients de pente, regroupement des segments à l'aide des tests de comparaisons successifs des paramètres de pentes et des constantes.

Ce processus est répété un certain nombre de fois. Le modèle final ainsi obtenu doit être le plus proche de la réalité visuelle. Cependant, celui-ci est seulement le résultat pour un degré de lissage choisi au départ et a peu de chance de correspondre à la segmentation optimale. Par conséquent, les deux phases de préparation des données et de modélisation sont effectuées pour un certain nombre de degrés de lissage différents. Ce dernier peut aller théoriquement de 1 à T. En pratique, il est préférable de débuter le processus pour un degré de lissage égal à l'unité, alors que la borne supérieure ne dépasse pas, en général, $\sqrt{T}$. Ce qui fait que la complexité empirique est de $O\left(T\sqrt{T}\right)$ et la complexité théorique est de $O\left(T^2\right)$.

Enfin, une dernière étape permettra d'évaluer l'ensemble des segmentations obtenues à l'aide des différentes degrés de lissage afin d'obtenir une segmentation finale.

1.3. La démarche détaillée de segmentation

Pour suivre cette démarche pas à pas, prenons un exemple volontairement simple. Nous avons généré deux processus gaussiens de même variance (σ^2 = 0,01) et de moyennes respectives égales à 5 (t = 1,40) et 6 (t = 41 à 100), comme montre la figure suivante :

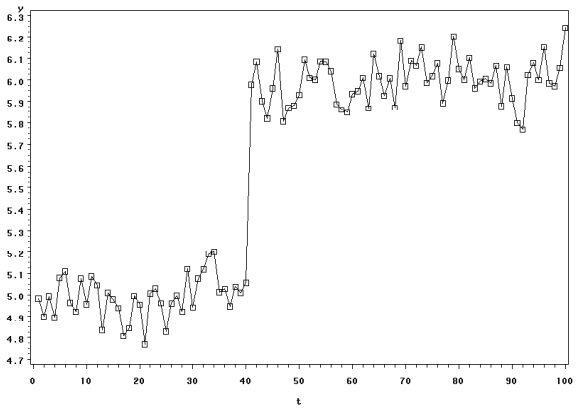

Figure 1.1.1 : données initiales

1.3.1. Phase de préparation des données

1.3.1.1. Etape de lissage

L'objectif est de résumer la série temporelle de façon à ne garder que les tendances fortes de la série afin de préparer les données pour l'étape de différenciation qui suivra. Pour cela, nous avons choisi d'utiliser la médiane mobile car elle est beaucoup plus robuste que la moyenne mobile. Le degré de lissage, noté j, correspond au nombre d'observations incluses dans la médiane mobile $m_j(t)$, pour t=1,T-j. Théoriquement, j =1,T, mais en pratique $j \leq \sqrt{T}$, nous avons :

$$m_j(t) = \underset{t \in [a_j(t), b_j(t)]}{med}(y_t) \qquad \text{pour } t = 1 \text{ à } T\text{-}j \qquad (1.12)$$

où pour un j fixé : $a_j(t) = t$ et $b_j(t) = t + j$ -1, pour t =1 à T-j +1.

Remarque : Plus j croît, moins l'irrégularité des données est prise en compte, comme nous pouvons le constater sur ces quatre figures avec les degrés de lissage suivants : 2, 6, 11 et 20.

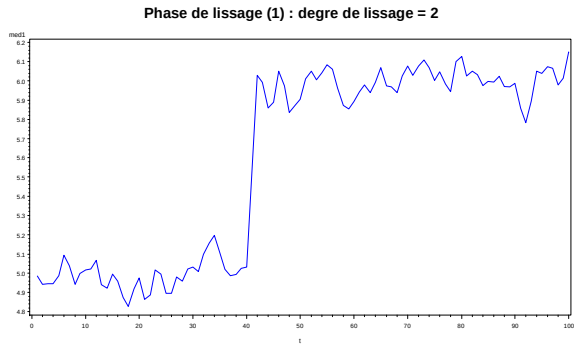

Figure 1.2.1 : Etape de lissage pour j = 2

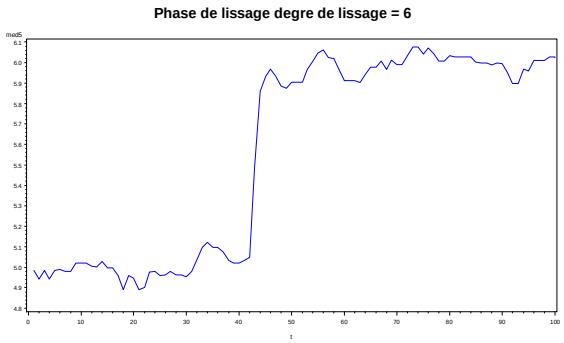

Figure 1.2.2 : Etape de lissage pour j = 6

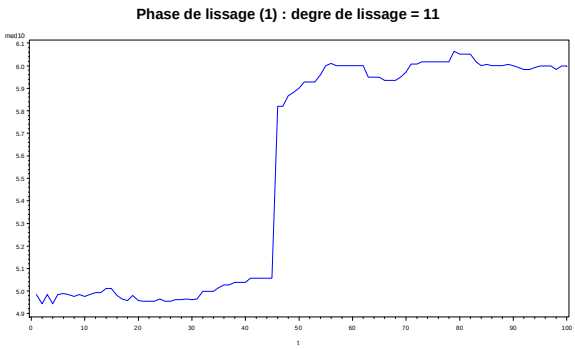

Figure 1.2.3 : Etape de lissage pour j = 11

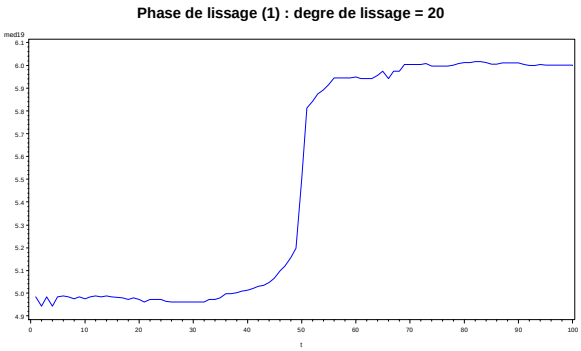

Figure 1.2.4 : Etape de lissage pour j = 20

1.3.1.2. Etape de différenciation

Cette étape permet de détecter les tendances de la série sur laquelle la médiane mobile a été appliquée. La différenciation doit être suffisamment élevée pour faire apparaitre des écarts de tendance, mais pas trop pour ne pas en louper. Pour cela, nous avons choisi de tenir compte de la propriété de la médiane mobile en effectuant une différence au temps t avec le temps t -k, où k =j /2, si j est pair (resp. $k = (j+1)/2$ si j est impair), et $t \geq j$. La différenciation s'effectue de la façon suivante :

$$d_j(t) = \left(m_j(t) - m_j(t-k)\right) \Big/ m_j(t-k) \qquad (1.13)$$

Le dénominateur permet d'obtenir un écart relatif, ce qui est très utile pour raisonner sur des quantités comparables. Mais il s'agit plus d'un choix visuel que d'un choix méthodologique pour la gestion des étapes suivantes. Les quatre figures montrent toutes un pic plus au moins en retard sur la vraie rupture dans les données (t = 41) et plus ou moins large. Sur la figure 1.3.1, le degré de lissage est faible ($j = 2$), le pic se situe vers 40, par contre il est relativement bas comparé à la variation des autres différences. Dans la figure 1.3.2, pour j = 6, le pic est nettement plus élevé et la variation des autres différences a fortement diminué par rapport à celui-ci. Signalons également que le pic a un peu de retard par rapport au précédent. Sur les deux figures suivantes pour des degrés de lissage égaux à 11 et 20, respectivement, le retard des pics a encore augmenté, ainsi que la largeur, alors que la variation a fortement diminué.

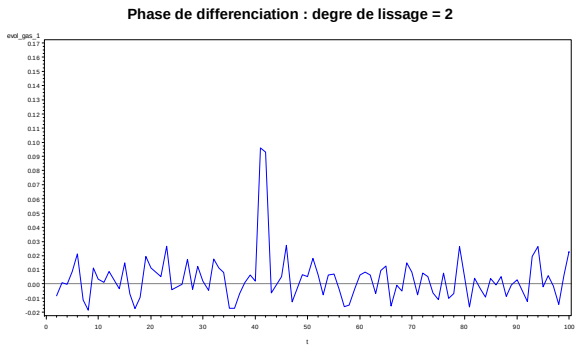

Figure 1.3.1 : Etape de différenciation pour j = 2

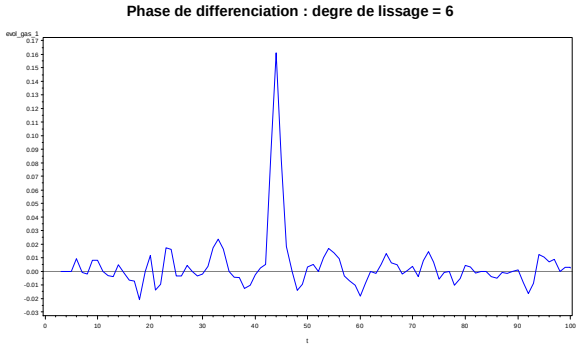

Figure 1.3.2 : Etape de différenciation pour j = 6

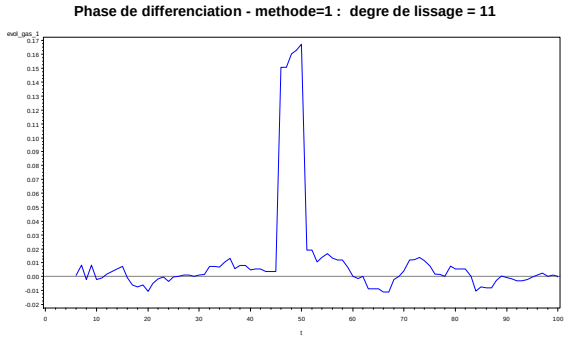

Figure 1.3.3 : Etape de différenciation pour j = 11

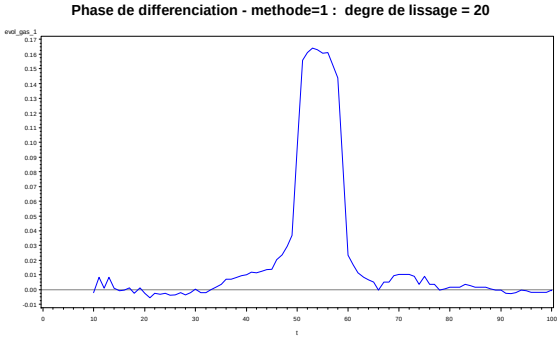

Figure 1.3.4 : Etape de différenciation pour j = 20

1.3.1.3. Etape de comptage

L'étape de différenciation a permis d'établir une suite de différences relatives positives, négatives ou nulles. Le nombre de suites de valeurs de même signe est raisonnablement fonction du degré de lissage. En effet, plus il est faible, plus il y a des chances que la taille des suites de différences de même signe soit petites. Chaque suite correspondra à un segment initial. Le premier segment $\tau_{j,1}^{(0)}$ pour un degré de lissage à j fixé contiendra les $T_{j,1}^{(0)}$ observations ayant le même signe, puis le deuxième segment $\tau_{j,2}^{(0)}$ inclura les $T_{j,2}^{(0)}$ observations ayant le même signe, mais différent de celui de $\tau_{j,1}^{(0)}$, etc. A la fin du processus, nous obtiendrons un vecteur de segments $\left(\tau_{j,1}^{(0)},...\tau_{j,s}^{(0)},...\tau_{j,S}^{(0)}\right)$ de tailles respectives : $\left(T_{j,1}^{(0)},...T_{j,s}^{(0)},...T_{j,S}^{(0)}\right)$, avec $\sum_{s=1}^{S} T_{j,s}^{(0)} = T$. Par exemple, si $d_3(1)$, $d_3(2)$ et $d_3(3)$ sont positifs et $d_3(4)$, $d_3(5)$ sont négatifs, alors $\tau_{31}^{(0)} = \left(y_1, y_2, y_3\right)$ et $\tau_{32}^{(0)} = \left(y_4, y_5\right)$.

Les quatre figures ci-dessous font apparaitre les différents segments créés. Les frontières correspondent aux lignes verticales, alors qu'entre celles-ci, les pentes croissantes en rouge (différences positives), en vert (différences nulles) et en bleu (différences négatives) sont relatives aux

nombres d'observations dans chaque segment. Par exemple, le nombre d'observations dans la figure 1.4.2 (degré de lissage = 6) pour le segment qui va de t = 40 à 46 est de six et correspond à des différences positives, comme nous pouvons le constater sur la figure 1.3.2. On remarquera que le début de ce segment correspond à la rupture sur les données initiales (cf. figure 1.1). Par contre, on peut constater que sur les figures 1.4.3 et 1.4.4, la rupture en t = 41 a été loupée dans la phase de préparation des données.

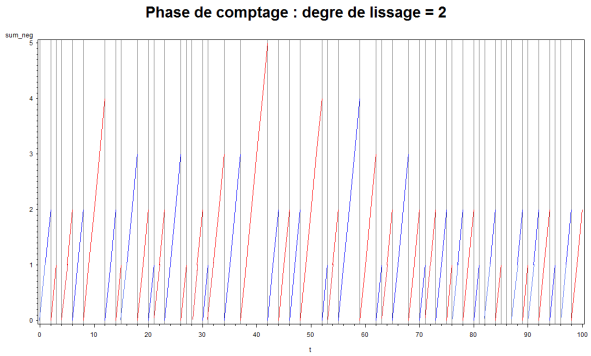

Figure 1.4.1 : Etape de comptage pour j = 2

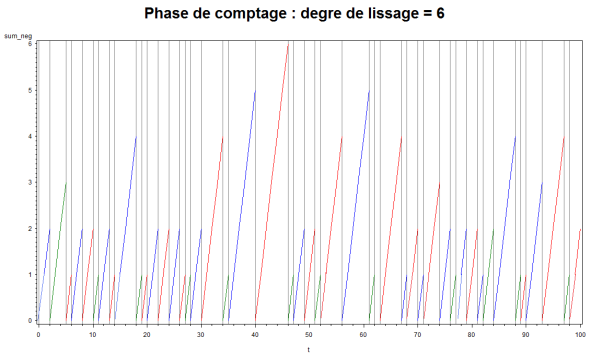

Figure 1.4.2 : Etape de comptage pour j = 6

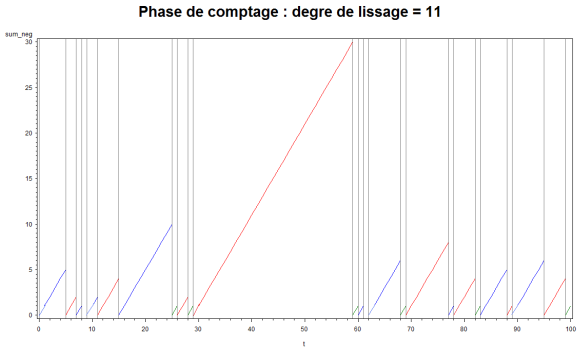

Figure 1.4.3 : Etape de comptage pour j = 11

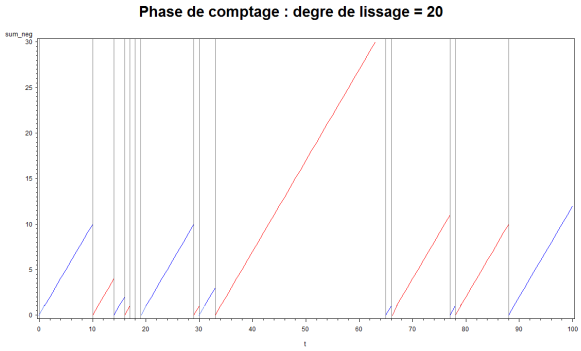

Figure 1.4.4 : Etape de comptage pour j = 20

Les figures suivantes reportent sur les données, les segments initiaux construits à l'aide de l'étape de comptage. Pour un degré de lissage égal à 2 (figure 1.4.5), la rupture en t = 41 qui apparaît est au milieu d'un segment, elle n'a donc pas été détectée. Par contre, pour un degré de lissage valant 6, la rupture est exactement trouvée, comme nous l'avions constaté sur la figure 1.4.2 lors de l'étape de comptage. De façon naturelle, pour les degrés de lissage 11 et 20 (figures 1.4.7 et 1.4.8), les ruptures n'ont bien évidemment pas été détectées.

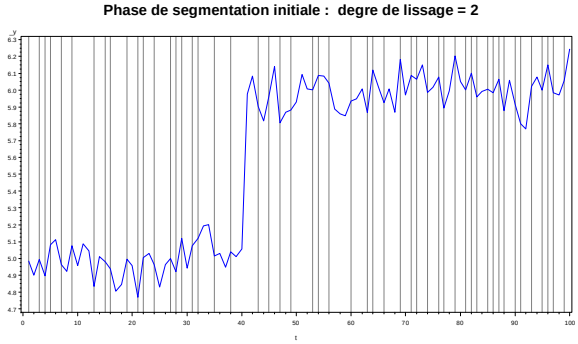

Figure 1.4.5 : Etape de segmentation initiale pour j = 2

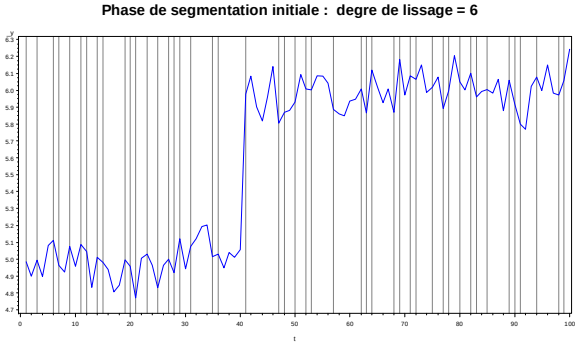

Figure 1.4.6 : Etape de segmentation initiale pour j = 6

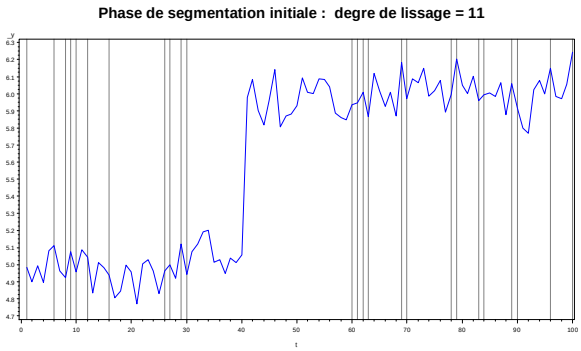

Figure 1.4.7 : Etape de segmentation initiale pour j = 11

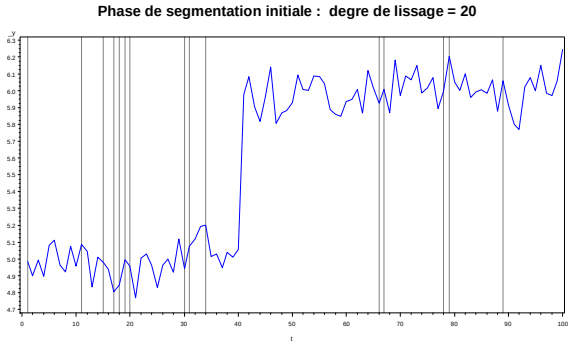

Figure 1.4.8 : Etape de segmentation initiale pour j = 20

1.3.2. Phase de modélisation des données

1.3.2.1. Etape de modélisation initiale

Cette première étape de modélisation contient généralement beaucoup trop de segments, d'autant plus que le degré de lissage est faible. Comme nous l'avons indiqué dans le paragraphe 1.2, chaque étape permet de simplifier le modèle proposé au départ de celle-ci. Par conséquent, elle se décompose en plusieurs sous-étapes : modèle complet, modèle simplifié et modèle à regroupement de segments.

Le ***modèle complet*** pour un degré de lissage *j* est de la forme (1), tel que :

$$Y_t = \sum_{s=1}^{S_0} \left(\beta_0^{(j,s)} + \beta_1^{(j,s)} t + \sigma_{j,s} \varepsilon_t \right) 1_{\left[t \in \tau_{j,s}^{(0)} \right]} \qquad (1.14)$$

Les figures suivantes fournissent tout d'abord, les modèles complets associés aux degrés de lissage : 2, 6, 11 et 20. Ils comportent respectivement 52, 47, 25 et 15 segments initiaux.

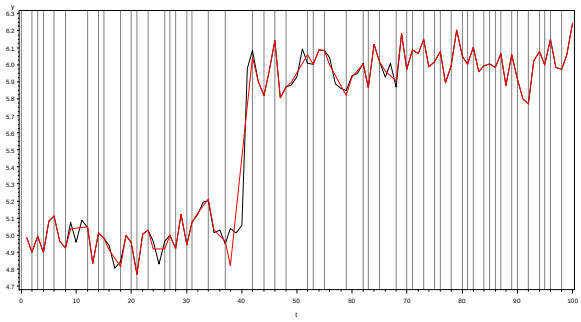

Figure 1.5.1 : Etape de modélisation initiale pour j = 2

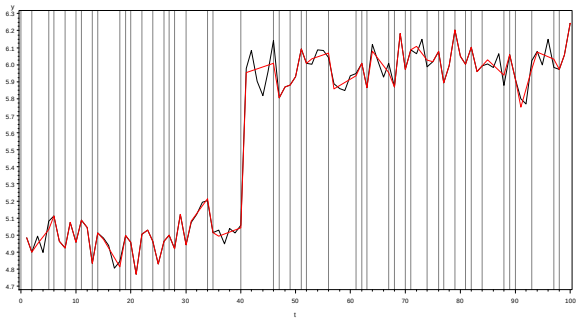

Figure 1.5.2 : Etape de modélisation initiale pour j = 6

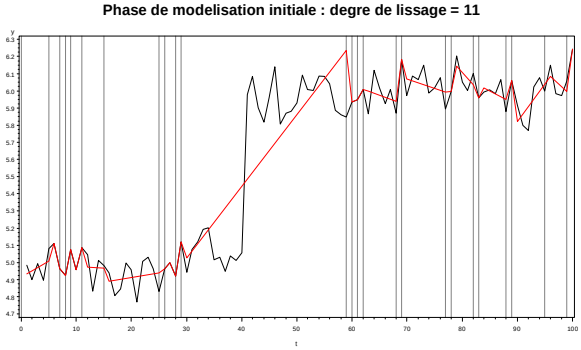

Figure 1.5.3 : Etape de modélisation initiale pour j = 11

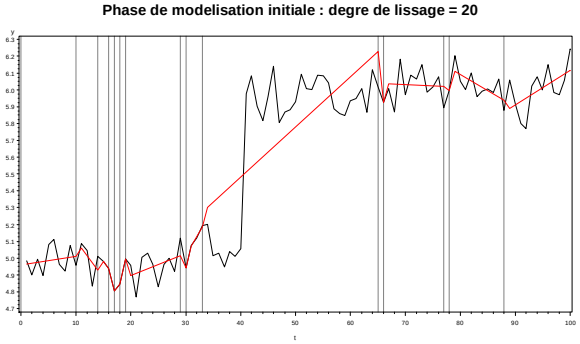

Figure 1.5.4 : Etape de modélisation initiale pour j = 20

Alors ce modèle est estimé par la méthode REML, le test d'homoscédasticité est appliqué. Si l'hypothèse nulle de variance constante n'est pas rejetée, alors le nouveau modèle suivant est estimé :

$$Y_t = \sum_{s=1}^{S_0} \left(\beta_0^{(j,s)} + \beta_1^{(j,s)} t\right) 1_{\left[t \in \tau_{j,s}^{(0)}\right]} + \sigma_j \varepsilon_t \tag{1.15}$$

La structure du ***modèle simplifié*** est construite en réalisant S_0 tests d'égalité à 0 des coefficients $\beta_1^{(j,s)}$ du modèle complet. La statistique de Student, utilisée pour ce test, est de la forme suivante : $\hat{\beta}_1^{(j,s)} / stderr_{REML}\left(\hat{\beta}_1^{(j,s)}\right)$ si le modèle est hétéroscédastique, sinon l'écart-type d'erreur du coefficient est remplacé par $stderr_{OLS}\left(\hat{\beta}_1^{(j,s)}\right)$ comme nous l'avons déjà indiqué dans le paragraphe 1.1. Ce modèle a donc la forme :

$$Y_t = \sum_{s=1}^{S_0} \left(\beta_0^{(j,s)} + \beta_1^{(j,s)} t\, 1_{\left[\beta_1^{(j,s)} \neq 0\right]} + \sigma_{j,s} \varepsilon_t\right) 1_{\left[t \in \tau_{j,s}^{(0)}\right]} \tag{1.16}$$

Enfin, le modèle obtenu précédemment est encore simplifié pour obtenir le ***modèle à regroupement de segments***. Pour cela, seuls les segments successifs dans le temps : $\tau_{j,s}^{(0)}$ et $\tau_{j,s+1}^{(0)}$ sont comparés dans le but de les regrouper s'ils sont identiques statistiquement. Chaque segment se caractérise par trois paramètres si le modèle est hétéroscédastique : $\left(\beta_0^{(j,s)}, \beta_1^{(j,s)}, \sigma_{j,s}\right)$ ou par le couple $\left(\beta_0^{(j,s)}, \beta_1^{(j,s)}\right)$ si le modèle est homoscédastique. Dans le premier cas, le premier test concerne l'égalité des variances $\sigma_{j,s}^2$ et $\sigma_{j,s+1}^2$, cela revient à tester l'homoscédascité sur le modèle suivant :

$$Y_t = \left(\beta_0^{(j,s)} + \beta_1^{(j,s)} t + \sigma_{j,s} \varepsilon_t\right) 1_{\left[t \in \tau_{j,s}^{(0)}\right]} + \left(\beta_0^{(j,s+1)} + \beta_1^{(j,s+1)} t + \sigma_{j,s+1} \varepsilon_t\right) 1_{\left[t \in \tau_{j,s+1}^{(0)}\right]} \tag{1.17}$$

Si les variances sont égales et si les deux coefficients de $\beta_1^{(j,s)}$ et $\beta_1^{(j,s+1)}$ sont différents de zéro, alors le test (1.18) suivant les compare. Si l'hypothèse n'est pas rejetée, alors les coefficients $\beta_0^{(j,s)}$ et $\beta_0^{(j,s+1)}$ sont comparés, à l'aide de la statistique de test de Student :

$$\left|\hat{\beta}_1^{(j,s)} - \hat{\beta}_1^{(j,s+1)}\right| / stderr_{OLS}\left(\hat{\beta}_1^{(j,s)}, \hat{\beta}_1^{(j,s+1)}\right) \tag{1.18}$$

où $stderr_{OLS}\left(\hat{\beta}_1^{(s)}, \hat{\beta}_1^{(s+1)}\right) = \hat{\sigma}_{j,s,s+1} \sqrt{\left(1 / \sum_{t \in \tau_{j,s}^{(0)}} \left(t - \bar{t}_{j,s}\right)^2 + 1 / \sum_{t \in \tau_{j,s+1}^{(0)}} \left(t - \bar{t}_{j,s+1}\right)^2\right)}$, avec $\bar{t}_{j,s}$ la moyenne des t dans le segment $\tau_{j,s}^{(0)}$ et $\hat{\sigma}_{s,s+1}^2 = \left(\left(T_{j,s}^{(0)} - 2\right)\hat{\sigma}_{j,s}^2 + \left(T_{j,s+1}^{(0)} - 2\right)\hat{\sigma}_{j,s+1}^2\right) / \left(T_{j,s}^{(0)} + T_{j,s+1}^{(0)} - 4\right)$.

Si ces deux coefficients $\beta_1^{(j,s)}$ et $\beta_1^{(j,s+1)}$ sont égaux, alors un test d'égalité des constantes $\beta_0^{(j,s)}$ et $\beta_0^{(j,s+1)}$ est pratiqué. Si l'hypothèse nulle de ce test n'est pas rejetée, alors les segments $\tau_{j,s}^{(0)}$ et $\tau_{j,s+1}^{(0)}$ sont regroupés. Dans le cas, où les coefficients $\beta_1^{(j,s)}$ et $\beta_1^{(j,s+1)}$ sont égaux à zéro, seuls les tests d'homoscédasticité et de comparaison des constantes sont mis en œuvre. A la fin de ce processus, le nombre de groupes obtenus $S_1 \leq S_0$ correspondra aux nouveaux segments à incorporer dans le modèle à regroupement de segments, tel que :

$$Y_t = \sum_{s=1}^{S_1} \left(\beta_0^{(j,s)} + \beta_1^{(j,s)} t\, 1_{\left[\beta_1^{(j,s)} \neq 0\right]} + \sigma_{j,s} \varepsilon_t\right) 1_{\left[t \in \tau_{j,s}^{(1)}\right]} \tag{1.19}$$

La figure 1.5.5 correspond au modèle simplifié du modèle complet (figure 1.5.1) pour un degré de lissage égal à 2. Le nombre de segments est passé de 52 à 19. Par exemple, les deux segments

allant de t = 35 à 43 ont été regroupés. Pour le degré de lissage valant 6 (figure 1.5.6), la rupture en t = 41 qui avait été détectée lors de la phase de préparation des données, reste bien évidemment et le nombre de segments a diminué de 47 à 18. Enfin, nous pouvons constater que pour les degrés de lissage égaux à 11 et 20, pour lesquels la rupture avait été complètement ratée, les nombres de segments sont passés de 25 à 8, et de 15 à 5, respectivement.

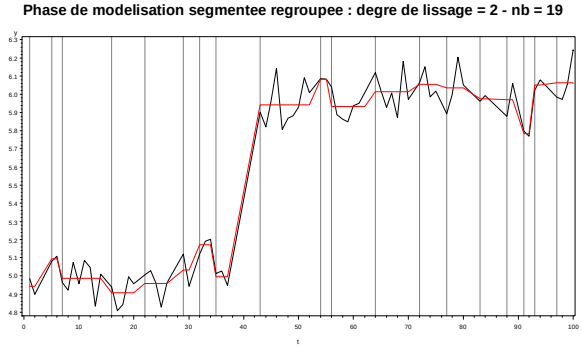

Figure 1.5.5 : Etape de modélisation regroupée pour j = 2

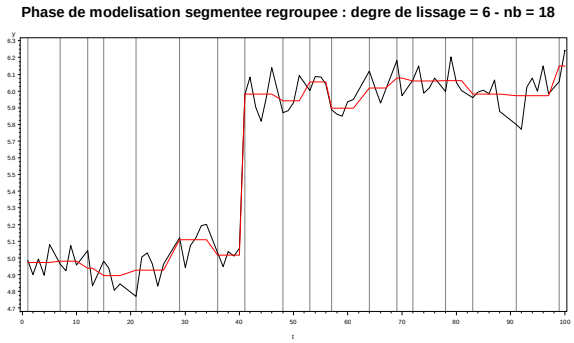

Figure 1.5.6 : Etape de modélisation regroupée pour j = 6

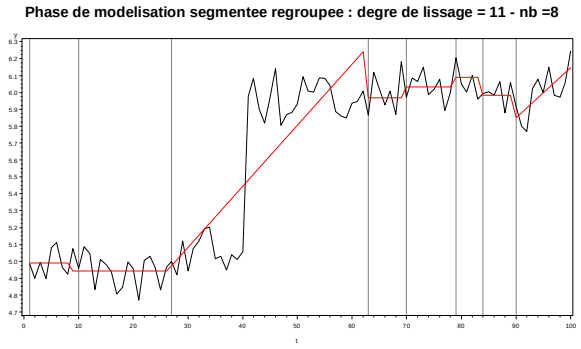

Figure 1.5.7 : Etape de modélisation regroupée pour j = 11

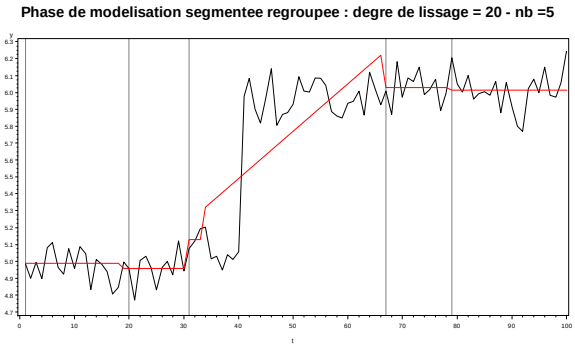

Figure 1.5.8 : Etape de modélisation regroupée pour j = 20

1.3.2.2. Etape de modélisation suivante

Dans l'étape suivante, le modèle (1.19) passe par le même processus de tests successifs que ceux établis dans le paragraphe précédent et S_2 segments sont obtenus. A la suite de cette étape, un nouveau modèle est proposé et contient généralement moins de segments (=S_3). Enfin, ce modèle passe à nouveau dans le même processus pour obtenir le modèle final avec S_4 segments. Cependant, si l'on juge que ce modèle n'est pas suffisamment simplifié, il est possible d'effectuer une étape supplémentaire, mais elle risque d'éliminer beaucoup d'informations pertinentes, d'autant plus que le degré de lissage sera élevé.

Le modèle final fournit donc S_4 segments : $\left(\tau_{j,1},\ldots\tau_{j,s},\ldots\tau_{j,S_4}\right)$ contenant respectivement $\left(T_{j,1},\ldots T_{j,s},\ldots T_{j,S_4}\right)$ observations temporelles et le modèle final est :

$$Y_t = \sum_{s=1}^{S_4}\left(\beta_0^{(j,s)} + \beta_1^{(j,s)} t\, 1_{\left[\beta_1^{(j,s)}\neq 0\right]} + \sigma_{j,s}\varepsilon_t\right) 1_{\left[t\in\tau_{j,s}\right]} \qquad (1.20)$$

Pour j = 2, quatre segments ont été construits, le troisième segment correspond à un petit saut qui aurait dû normalement être regroupé avec le deuxième et le quatrième, par contre la rupture a bien été détectée. Pour j = 6, le modèle final correspond exactement au modèle généré. Par contre, comme nous l'avons souligné précédemment, le modèle final correspondant au degré de lissage égal à j = 11, donne 2 segments, mais le second débute à t = 29, à la place de t = 41. Enfin, pour le degré

de lissage j = 20, il y a 3 segments, dont le deuxième segment qui est petit, puis le début du troisième rate à nouveau la rupture avec t = 33.

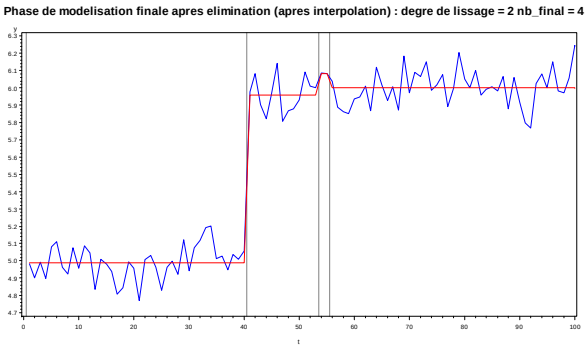

Figure 1.6.1 : Etape de modélisation finale pour j = 2

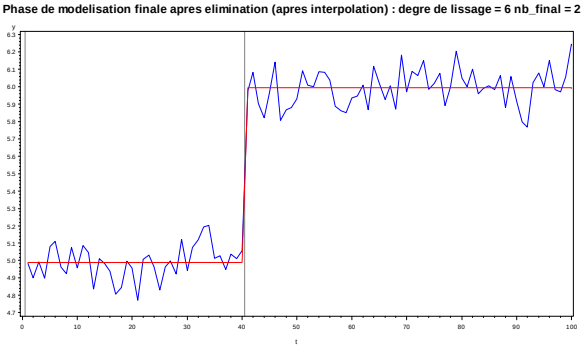

Figure 1.6.2 : Etape de modélisation finale pour j = 6

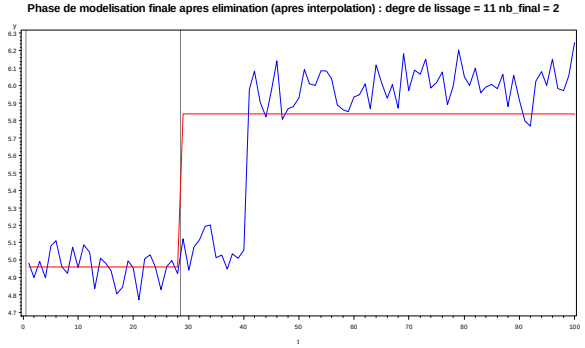

Figure 1.6.3 : Etape de modélisation finale pour j = 11

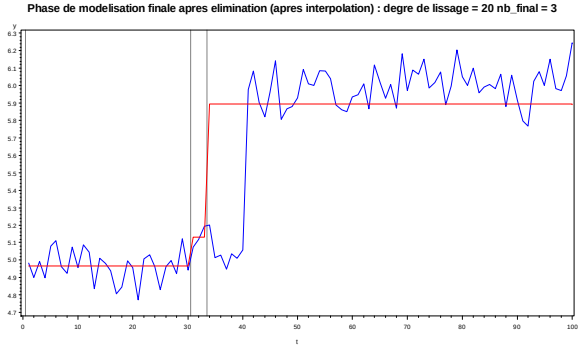

Figure 1.6.4 : Etape de modélisation finale pour j = 20

1.4. Evaluation des modèles

Comme nous l'avons indiqué les deux phases (préparatoire et modélisation) sont effectuées pour chaque degré de lissage, pouvant aller de 1 à T. Pour certains d'entre eux, le modèle final, permettra de mieux reconstituer les données ; il aura d'autant plus de chances pour fournir la meilleure segmentation possible. Signalons que même si les T degrés de lissage sont essayés, cela ne garantit pas l'obtention d'une segmentation optimale avec une probabilité de 1, comme cela pourrait éventuellement se présenter pour un algorithme de programmation dynamique dans des cas relativement simples. Cependant, la méthode de segmentation proposée n'a pas pour objectif incontournable d'offrir une telle solution optimale. En effet, comme le modèle est relativement complexe car il permet de révéler les tendances, les niveaux et les dispersions pour chaque segment de la série temporelle, le but est plutôt de proposer un certain nombre de segmentations candidates. Pour cela, nous travaillons avec quelques mesures statistiques permettant d'offrir un choix de segmentations raisonnables.

Tout d'abord nous avons choisi la mesure naturelle qui a permis d'estimer le modèle : la valeur de la vraisemblance résiduelle estimée (REML), ainsi que les critères d'information, telles que l'AIC et le BIC. Par ailleurs, sous l'angle : estimateur des moindres carrés, nous utilisons également le RMSE et le R^2, avec comme critère de qualité, le R^2 ajusté. Nous calculons également le MAPE (Mean of Absolute Percentage Errors) qui correspond à la moyenne des valeurs absolues des erreurs relatives ($\left|\left(y_t - \hat{y}_t\right)/y_t\right|$), ainsi que la distribution de celles-ci. Pour les premières mesures, le maximum des valeurs obtenues pour l'ensemble des degrés de lissage a de fortes chances de fournir la « meilleure » segmentation, alors qu'il s'agira du minimum pour le MAPE. Un certain nombre de segmentations pourront alors être proposées, en ordonnant ces mesures.

Nous avons calculé l'ensemble des 100 segmentations possibles, afin d'évaluer de façon détaillée les résultats obtenus. La figure 1.7.1 montre l'évolution du critère BIC en fonction du degré de lissage, la

valeur minimum est de -153 (REML=-162,2) et est obtenue pour différents degrés de lissage : 6, 23, 29, …, pour lesquels les deux segments générés ont été parfaitement retrouvés (cf. figure 1.7.5), alors que sur la figure 1.7.2 qui affiche les MAPE, le minimum est donné pour 1,36% (1,42% pour la précédente segmentation) pour un degré de lissage égal à 24 avec 4 segments (cf. figure 1.7.6). De plus, cela correspond au R^2 ajusté le plus élevé qui vaut 0,9657 (figure 1.7.3) contre 0,9598 pour la première segmentation. Enfin, il y a seulement 11% d'erreurs relatives absolues supérieures à 3% pour la segmentation à 4 segments (figure 1.7.6) et 14% pour celle à 2 segments (figure 1.7.5).

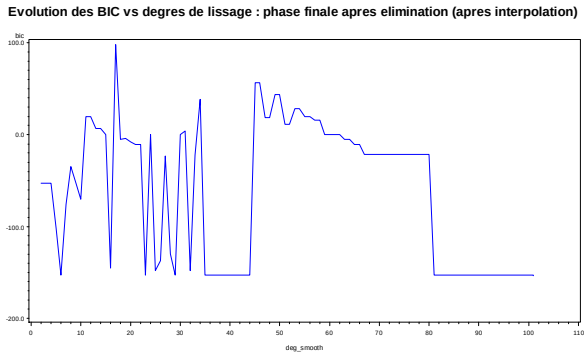

Figure 1.7.1 : Evolution de l'indicateur BIC

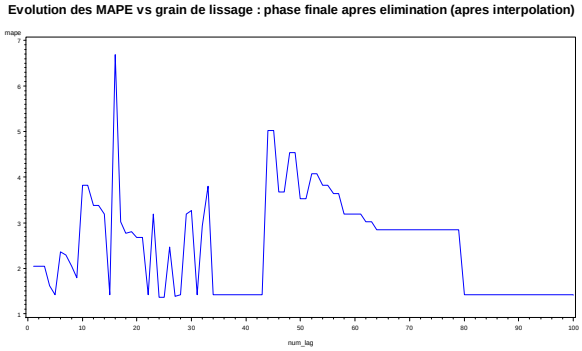

Figure 1.7.2 : Evolution du MAPE

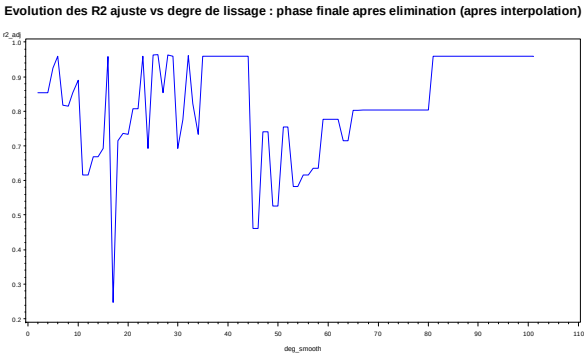

Figure 1.7.3 : Evolution du R2 ajusté

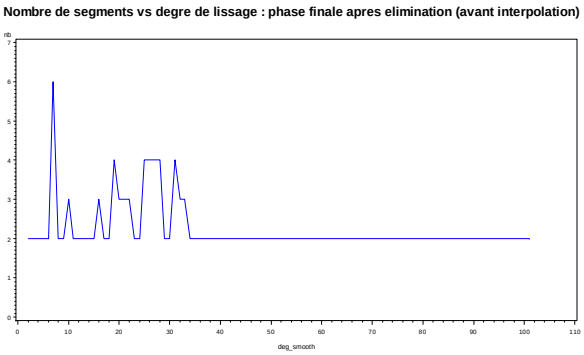

Figure 1.7.4 : Evolution du nombre de segments

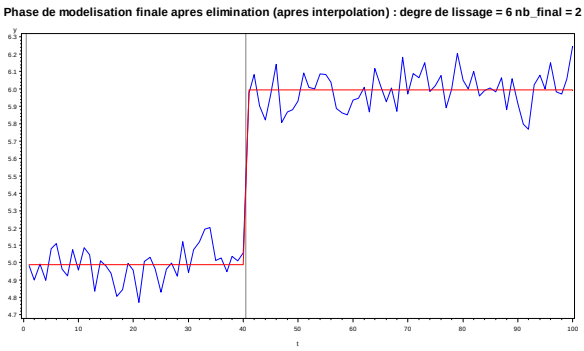

Figure 1.7.5 : Segmentation avec 2 segments

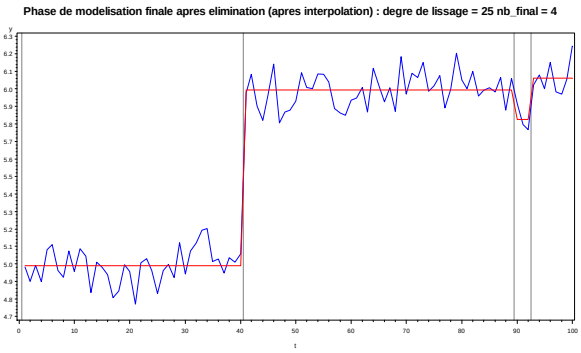

Figure 1.7.6 : Segmentation avec les 4 segments

1.5. Un exemple sur un cas simulé

Nous avons appliqué la méthode proposée sur un jeu de données simulées. Les données temporelles ont été générées sur 10 segments, selon le modèle (1.1). Pour chacun des 10 segments, le nombre d'observations, les valeurs des coefficients β_0 et β_1, et la dispersion σ associée sont générés. Pour juger de la qualité des segmentations proposées par rapport à la segmentation générée, nous comparons les distributions des segments simulés et des segments estimés à l'aide du V de Cramer, du τ_b de Kendall et du τ_c de Stuart, ainsi que le pourcentage d'observations mal attribuées. La

segmentation estimée retenue contient 12 segments (figure 1.8.2) *vs* 10 segments pour la segmentation simulée (figure 1.8.1). La figure 1.8.2 montre une très bonne adéquation entre les deux segmentations. En effet, les instants de ruptures réels et estimés coïncident assez bien et le modèle estimé (trait plein rouge) reconstitue bien les données (points bleus). D'autre part, les résidus standardisés issus des modèles sur les 2 segmentations ont des comportements très similaires (figures 1.8.3 et 1.8.4). Ce résultat montre l'intérêt de cette méthode pour stationnariser une série.

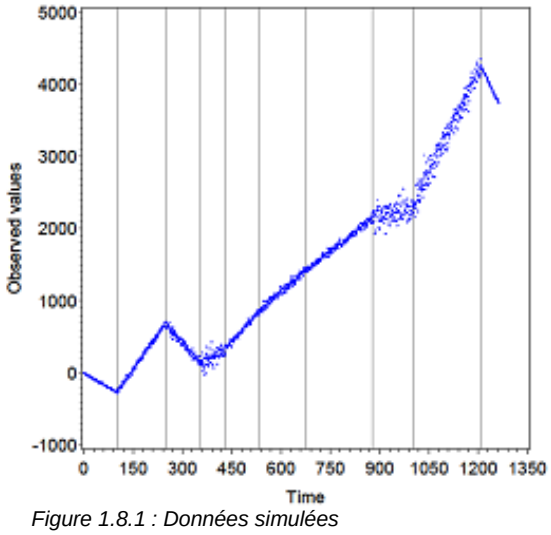

Figure 1.8.1 : Données simulées

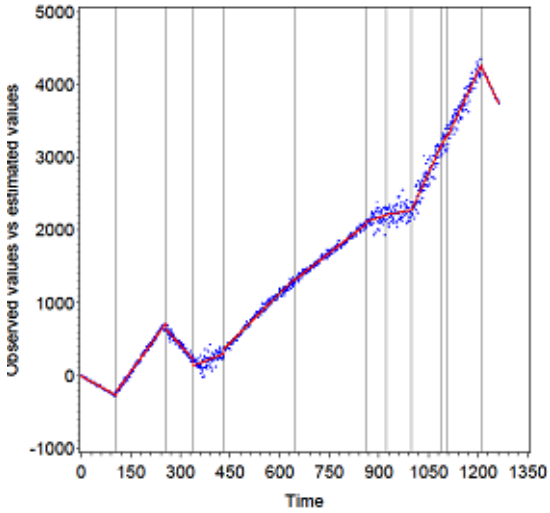

Figure 1.8.2 : Segmentation des données simulées

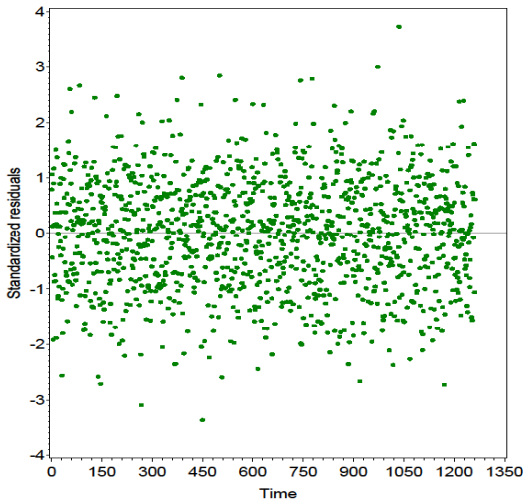

Figure 1.8.3 : Erreurs observées standardisées

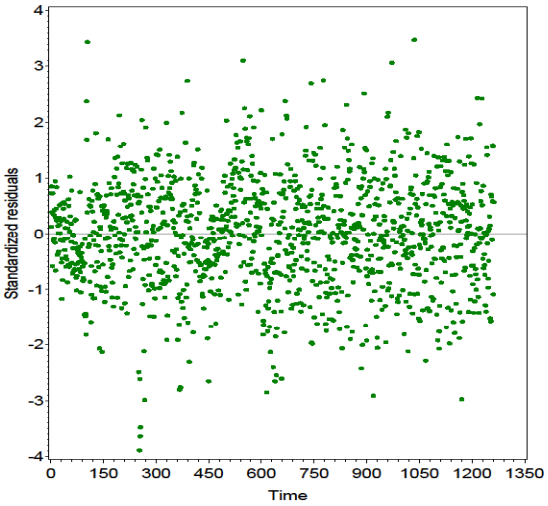

Figure 1.8.4 : Erreurs estimées standardisées

Par ailleurs, le MAPE sur les données simulées lorsque que l'on applique directement le modèle généré vaut 9,90%, alors que celui issu de la segmentation vaut 11,06%. De plus, il y a 12,68% d'erreurs relatives absolues supérieures à 10% pour la segmentation simulée et 13,71% pour la segmentation estimée. On peut donc constater qu'il y a une forte ressemblance entre les valeurs des indicateurs de la segmentation estimée par rapport à la segmentation simulée. Ces très bons résultats sont confirmés par le fait que sur les neuf ruptures de la segmentation simulée, six ont été découvertes par la segmentation estimée proposée.

2. Une nouvelle approche par estimation préalable de la dispersion

Comme indiqué, la méthode [3,4] a fourni des résultats encourageants sur différents types de séries temporelles qu'elles soient simulées ou réelles, par rapport à des méthodes fondées sur la programmation dynamique. Cependant pour l'ensemble des méthodes de segmentation développées qu'elles soient fondées sur une approche par programmation dynamique ou sur une approche exploratoire comme la nôtre, il s'avère que la qualité de la segmentation peut faire défaut dans le cas de figure suivant. Lors de la détection de segments contigus : les niveaux (constants ou linéaires) sont proches statistiquement mais ont des variances différentes, dans ce cas un seul segment sera détecté à la place de deux structurellement. La nouvelle méthode proposée ici a notamment pour objectif de pallier ce problème.

Cette nouvelle approche contient trois phases principales. La première consiste à établir une transformation adéquate des données afin d'obtenir une nouvelle série caractérisant l'évolution temporelle de la dispersion des observations, la deuxième phase revient à segmenter cette nouvelle série avec le même principe que la méthode [3,4] pour obtenir des segments de dispersion, enfin la troisième phase applique à nouveau [3,4] mais en tenant compte de la distribution des segments de dispersion, notamment lors de la construction du modèle linéaire hétéroscédastique.

Illustrons cette nouvelle approche sur l'exemple suivant. Nous avons généré deux processus gaussiens de même moyenne égale à 5, mais de variances différentes : $\sigma_1^2 = 0{,}05$ (t = 1,40) et $\sigma_2^2 = 0{,}01$ (t = 41 à 100), comme le montre la figure suivante :

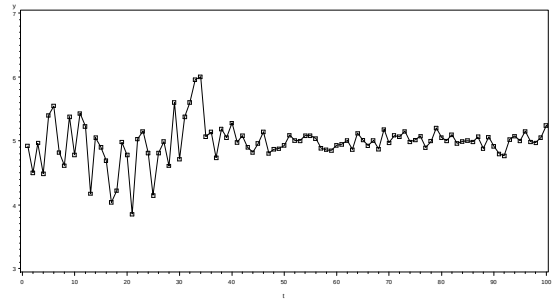

Figure 2.1.1 : données initiales

2.1. Phases 1 et 2 : Transformation caractérisant la dispersion des données temporelles et première segmentation

L'objectif est de construire une nouvelle série temporelle permettant d'exhiber la volatilité des observations de la série temporelle dont on suppose qu'elles sont régies par le modèle (1.1). La transformation la plus naturelle est une différenciation à l'ordre 2, telle que : $Z_t = (I - B)^2 Y_t$, où Y_t est la série temporelle originale. Il est alors possible d'appliquer deux opérateurs sur Z_t, soit $U_t = |Z_t|$, soit $V_t = Z_t^2$. Les deux théorèmes suivants permettent d'obtenir la variance σ^2 à partir de ces transformations.

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 ε_t ~> $\mathcal{N}(0,1)$, i.i.d. alors $\sigma = \left(\sqrt{\pi}/2\sqrt{3}\right)\mathbf{E}\left(|Y_t - 2Y_{t-1} + Y_{t-2}|\right)$.

Démonstration : 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 ε_t ~> $\mathcal{N}(0,1)$, i.i.d. Posons $Z_t = Y_t - 2Y_{t-1} + Y_{t-2}$ alors il vient directement que $Z_t = \sigma(\varepsilon_t - 2\varepsilon_{t-1} + \varepsilon_{t-2})$ est de moyenne nulle et de variance égale à $6\sigma^2$.
Calculons maintenant la loi : $U_t = |Z_t| = |Y_t - 2Y_{t-1} + Y_{t-2}|$.

La fonction de répartition associée à U_t : F_{U_t} telle que :

$$F_{U_t}(u_t) = \Pr[U_t < u_t] = \Pr[|Z_t| < u_t] = \Pr[-u_t < Z_t < +u_t] = 2\Pr[Z_t < u_t] - 1 = 2F_{Z_t}(u_t) - 1$$

Alors la fonction de densité associée est : $f_{U_t} = 2f_{Z_t}$

Calculons enfin l'espérance de U_t : $\mathbf{E}(U_t) = \frac{2}{\sqrt{6}\sigma\sqrt{2\pi}} \int_0^{+\infty} u_t e^{-\frac{u_t^2}{12\sigma^2}} du_t$

En posant $w_t = \frac{u_t^2}{12\sigma^2}$ d'où $dw_t = \frac{u_t}{6\sigma^2} du_t$. Par conséquent, nous obtenons :

$\mathbf{E}(U_t) = \frac{2\sqrt{3}\sigma}{\sqrt{\pi}} \int_0^{+\infty} e^{-w_t} dw_t = \frac{2\sqrt{3}\sigma}{\sqrt{\pi}} \left[-e^{-w_t}\right]_0^{+\infty} = \frac{2\sqrt{3}\sigma}{\sqrt{\pi}}$, cqfd.

Le résultat de ce premier théorème permet de calculer une estimation de σ, telle que :

$$\hat{\sigma}_U = \frac{\sqrt{\pi}}{2\sqrt{3}} \frac{1}{(T-2)} \sum_{t=3}^{T} |y_t - 2y_{t-2} + y_{t-2}| \qquad (2.1)$$

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 ε_t ~> $\mathcal{N}(0,1)$, i.i.d. alors $\sigma^2 = \frac{\mathbf{E}\left[(Y_t - 2Y_{t-1} + Y_{t-2})^2\right]}{6}$.

Démonstration : 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 ε_t ~> $\mathcal{N}(0,1)$, i.i.d. Posons $Z_t = Y_t - 2Y_{t-1} + Y_{t-2}$ alors il vient directement que $Z_t = \sigma(\varepsilon_t - 2\varepsilon_{t-1} + \varepsilon_{t-2})$ est de moyenne nulle et de variance égale à $6\sigma^2$.

Calculons maintenant la loi : $V_t = Z_t^2 = (Y_t - 2Y_{t-1} + Y_{t-2})^2$.

On pose la fonction de répartition associée à V_t : F_{V_t} telle que :

$$F_{V_t}(v_t) = \Pr[V_t < v_t] = \Pr[Z_t^2 < v_t] = \Pr[-\sqrt{v_t} < Z_t < +\sqrt{v_t}] = 2\Pr[Z_t < \sqrt{v_t}] - 1 = 2F_{Z_t}(\sqrt{v_t}) - 1$$

Alors la fonction de densité associée est : $f_{v_t} = \frac{1}{\sqrt{v_t}} f_{Z_t}$

Calculons enfin l'espérance de V_t : $\mathbf{E}(V_t) = \frac{1}{\sqrt{6}\sigma\sqrt{2\pi}} \int_0^{+\infty} v_t \times v_t^{-1/2} e^{-\frac{v_t^2}{12\sigma^2}} dv_t$

On pose $w_t = \frac{v_t^2}{12\sigma^2}$ d'où $dw_t = \frac{v_t}{6\sigma^2} dv_t$. Par conséquent, nous obtenons :

$\mathbf{E}(V_t) = \frac{12\sigma^2}{\sqrt{\pi}} \int_0^{+\infty} w_t^{1/2} e^{-w_t} dw_t$. Alors il suffit de remarquer que $\frac{2}{\sqrt{\pi}} w_t^{1/2} e^{-w_t} = \frac{1}{\frac{1}{2}\Gamma\left(\frac{1}{2}\right)} w_t^{1/2} e^{-w_t}$ n'est autre que la fonction de densité d'une loi Gamma de paramètre θ= 3/2. Par conséquent, il vient : $\mathbf{E}(V_t) = \frac{6\sigma^2}{\frac{1}{2}\Gamma\left(\frac{1}{2}\right)} \int_0^{+\infty} w_t^{1/2} e^{-w_t} dw_t = 6\sigma^2$, cqfd.

Le résultat de ce second théorème permet de calculer une estimation de σ^2, telle que :

$$\hat{\sigma}_V^2 = \frac{1}{6(T-2)} \sum_{t=3}^{T} (y_t - 2y_{t-2} + y_{t-2})^2 \qquad (2.2)$$

Les résultats obtenus à l'aide des deux théorèmes précédents sont essentiels pour la deuxième phase car ils permettent de faire apparaitre dans chaque série observée u_t ou v_t, les niveaux de dispersion de segments candidats de la série temporelle. En effet, la démarche de segmentation expliquée dans le paragraphe 1.3. est alors appliquée, soit sur la série u_t, soit sur la série v_t. A la fin du processus, la segmentation sélectionnée offrira un ensemble de segments caractérisés par le modèle (1.1). Les segments linéaires obtenus peuvent être constants, croissants ou décroissants, ce qui offre une information supplémentaire intéressante sur le comportement des données qui peut être hétéroscédastique même sur un segment.

Soient maintenant τ_1^σ,..., $\tau_{S_1}^\sigma$, les S_1 segments de dispersion obtenus précédemment sur la série u_t. Alors le segment τ_s^σ fournira une estimation des T_s valeurs des u_t, telle que : $\hat{u_t} = \hat{\alpha}_0^{(s)} + \hat{\alpha}_1^{(s)} t$, pour $t \in \tau_s^\sigma$. Dans notre exemple développé avec deux segments dans la première partie, la dispersion était la même sur l'ensemble de la série, avec : $\sigma = 0{,}1$, dans ce cas, $\forall t, u_t = \alpha_0^{(1)} = 0{,}1\sqrt{\pi}/\left(2\sqrt{3}\right) = 0{,}1954$, alors que l'estimation sur notre jeu de données est : $\forall t, \hat{u_t} = \hat{\alpha}_0^{(1)} = 0{,}1934$.

Dans l'exemple développé dans cette partie avec deux variances différentes, mais de même moyenne, la meilleure segmentation retenue sur l'ensemble des critères : BIC, REML, MAPE = 3,05, RMSE = 0,2244, et R^2 ajusté = 0,4995, contient 9 segments pour un degré de lissage égal à 5. Elle a la forme suivante :

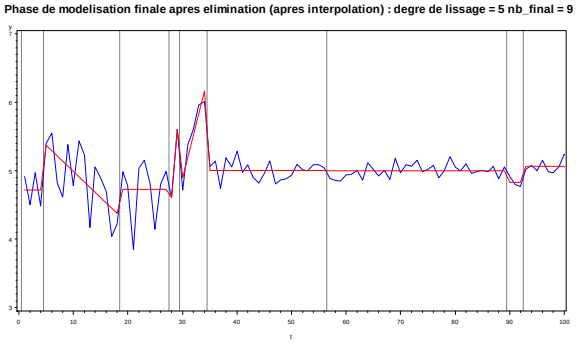

Figure 2.2.1 : Segmentation directe retenue sur les y_t (1ière approche)

Nous pouvons constater que la segmentation réelle n'a pas été découverte, notamment il y a trop de segments qui sont pollués par la forte variation dans le premier segment généré (cf. figure 2.1.1). Par conséquent, nous avons appliqué la segmentation sur les quantités transformées fournies par : $U_t = |Z_t| = |Y_t - 2Y_{t-1} + Y_{t-2}|$. La figure 2.2.2. montre la meilleure segmentation selon le critère BIC. Le premier segment estimé va de t = 1 à 39 à la place de t = 1 à 40 pour le segment généré. Le modèle de segmentation estimé est donc de la forme :
$\hat{u_t} = \hat{\alpha}_0^{(1)} \times 1_{[t=1,39]} + \hat{\alpha}_0^{(2)} \times 1_{[t=40,100]} = 0{,}8971 \times 1_{[t=1,39]} + 0{,}1943 \times 1_{[t=40,100]}$. Dans le cas où l'on considère que le modèle est gaussien, alors les deux dispersions estimées sont les suivantes : $\hat{\sigma}_1 = \hat{\alpha}_0^{(1)}\sqrt{\pi}/\left(2\sqrt{3}\right) = 0{,}4590$ et $\hat{\sigma}_2 = \hat{\alpha}_0^{(2)}\sqrt{\pi}/\left(2\sqrt{3}\right) = 0{,}0994$, à la place de σ_1 = 0,5 et σ_2 = 0,1.

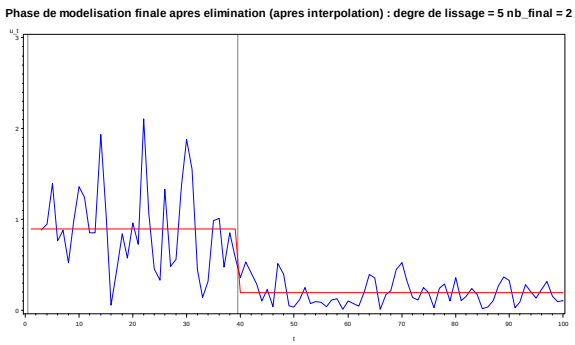

Figure 2.2.2 : Segmentation retenue sur les u_t

2.2. Phase 3 : 2sd segmentation en tenant compte de la dispersion

Cette troisième et dernière phase a pour objectif de fournir une segmentation finale de la série temporelle initiale Y_t en tenant compte de la dispersion a priori des données qui est estimée à l'aide de la phase 2. Pour cela, nous proposons deux solutions.

Soit chaque valeur y_t est réduite par $\hat{u}_t$, afin d'éliminer l'effet de dispersion dés le départ de la segmentation, c'est-à-dire dés l'étape de lissage dans la phase de préparation des données. Le modèle sous-jacent fourni à la seconde application de la méthode de segmentation est le suivant :

$$y_t / \hat{u}_t^{(1)} = \beta_0^{(1)} / \hat{u}_t^{(1)} + \left(\sigma_1 / \hat{u}_t^{(1)}\right)\varepsilon_t \quad \Leftrightarrow \quad \tilde{y}_t = \tilde{\beta}_0^{(1)} + \tilde{\sigma}_1 \varepsilon_t \quad (t = 1 \text{ à } 39)$$

$$y_t / \hat{u}_t^{(2)} = \beta_0^{(1)} / \hat{u}_t^{(2)} + \left(\sigma_1 / \hat{u}_t^{(2)}\right)\varepsilon_t \quad \Leftrightarrow \quad \tilde{y}_t = \tilde{\beta}_0^{(1.2)} + \tilde{\sigma}_{1.2} \varepsilon_t \quad (t = 40)$$

$$y_t / \hat{u}_t^{(2)} = \beta_0^{(2)} / \hat{u}_t^{(2)} + \left(\sigma_2 / \hat{u}_t^{(2)}\right)\varepsilon_t \quad \Leftrightarrow \quad \tilde{y}_t = \tilde{\beta}_0^{(2)} + \tilde{\sigma}_2 \varepsilon_t \quad (t = 41 \text{ à } 100)$$

Par conséquent, afin d'obtenir les estimations finales des y_t , il faudra donc procéder à une nouvelle transformation des valeurs estimées issues de la seconde segmentation :

$$\hat{\tilde{y}}_t = \hat{\tilde{\beta}}_0^{(1)} \quad \Leftrightarrow \quad \hat{y}_t = \hat{\tilde{\beta}}_0^{(1)} \hat{u}_t^{(1)} = \hat{\beta}_0^{(1)} \quad (t = 1{,}39)$$

$$\hat{\tilde{y}}_t = \hat{\tilde{\beta}}_0^{(1.2)} \quad \Leftrightarrow \quad \hat{y}_t = \hat{\tilde{\beta}}_0^{(1.2)} \hat{u}_t^{(2)} = \hat{\beta}_0^{(1.2)} \quad (t = 40)$$

$$\hat{\tilde{y}}_t = \hat{\tilde{\beta}}_0^{(2)} \quad \Leftrightarrow \quad \hat{y}_t = \tilde{\beta}_0^{(2)} \hat{u}_t^{(2)} = \hat{\beta}_0^{(2)} \quad (t = 41{,}100)$$

Nous avons appliqué cette première transformation sur les données initiales. Comme nous pouvons le voir sur la série de données transformées en bleu (figure 2.3.1), deux segments apparaissent très nettement en moyenne. Nous avons donc appliqué une seconde segmentation sur ces données. La meilleure qui a été obtenue est fournie dans la figure 2.3.2.

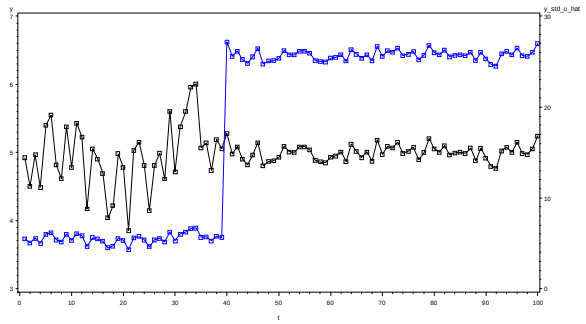

Figure 2.3.1 : Transformation des données avec les u_t estimés

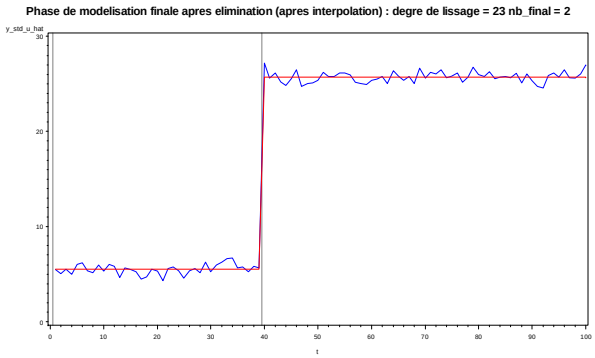

Figure 2.3.2 : segmentation associée

Les équations estimées sont les suivantes (cf. figure 2.3.3) :

$$\hat{\tilde{y}}_t = 5{,}5101 \quad \Leftrightarrow \quad \hat{y}_t = 5{,}5101 \times 0{,}8971 = 4{,}9432 \quad (t = 1{,}39)$$

$$\hat{\tilde{y}}_t = 25{,}7330 \quad \Leftrightarrow \quad \hat{y}_t = 25{,}7330 \times 0{,}1943 = 4{,}9999 \quad (t = 40{,}100)$$

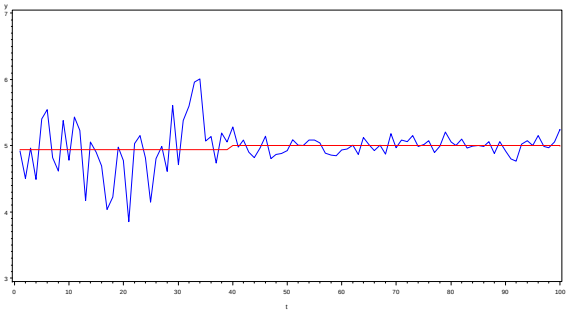

Figure 2.3.3 : Segmentation finale en intégrant les u_t dés le départ (1ière solution)

La valeur du MAPE est de 4,06%.

Soit les $\hat{u_t}$ sont seulement intégrés dans la phase de modélisation pour structurer la matrice de dispersion du modèle linéaire gaussien hétéroscédastique (1.1). Cela revient à utiliser une matrice diagonale de matrice de variance-covariance des erreurs $\mathbf{V}$ lors de l'estimation des paramètres du modèle à l'aide de l'estimateur REML, ou la matrice de poids $\mathbf{W}$ dans l'estimateur WLS.

Le modèle de segmentation estimé possède un seul segment, pour un degré de lissage égal à 11, nous obtenons : $\hat{y}_t = \hat{\beta}_0^{(11)} = 4{,}9778$, le MAPE vaut 4,11%, à comparer avec 4,06% obtenu pour la segmentation précédente.

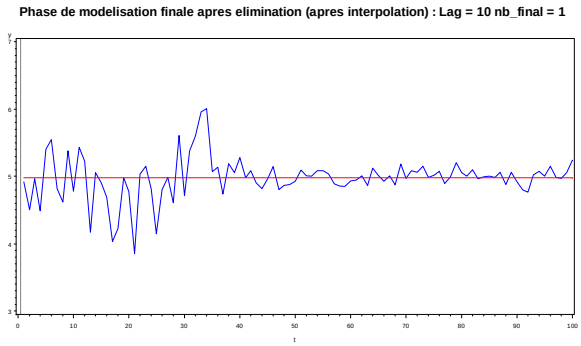

Figure 2.3.4 : Segmentation finale en intégrant les u_t seulement sous forme de poids (2^{sd} solution)

2.3. Application sur le cas simulé

Nous reprenons notre exemple sur les données simulées (cf. § 1.5) sur lesquelles nous avons appliquées directement notre méthode de segmentation sans tenir compte au préalable de la dispersion des données. Nous avons maintenant traité ces données avec la nouvelle approche proposée.

Par ailleurs, nous avons comparé nos résultats avec ceux obtenus en utilisant des algorithmes de programmation dynamique développée dans [6]. Ils peuvent détecter plusieurs points de rupture dans une série chronologique. Ils sont estimés en minimisant une fonction de contraste pénalisé. Cette méthode permet de détecter les types de changements suivants : moyennes différentes avec variance constante, variances différentes avec moyenne constante, moyennes et variances différentes (nommé DCPC3 en figure 2.4.5), distributions différentes (DCPC4, 2.4.6). En outre, une version a été développée avec une approche bayésienne sur les mêmes types de changement (BDCPC3 et BDCPC4, les figures 2.4.7 et 2.4.8). Dans ce cas, les points de rupture sont estimés par minimisation d'une distribution a posteriori. Le mode de cette distribution a posteriori correspond à l'estimation minimale du contraste pénalisé.

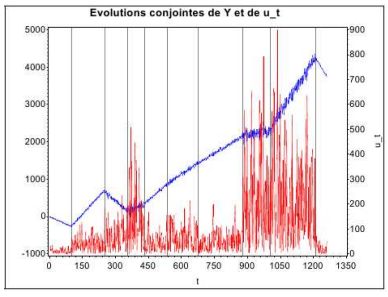

Figure 2.4.1 : Données simulées & transformations en u_t

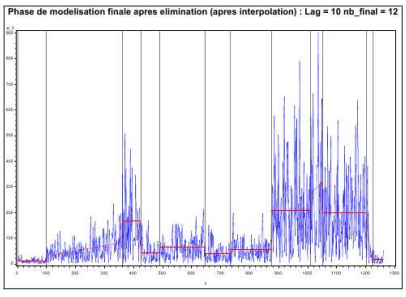

Figure 2.4.2 : Segmentation des u_t

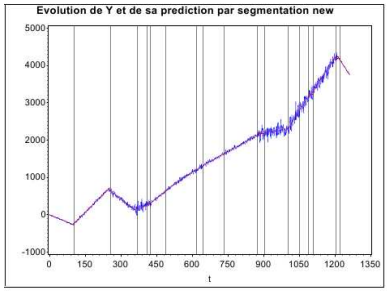

Figure 2.4.3 : Segmentation avec la nouvelle approche

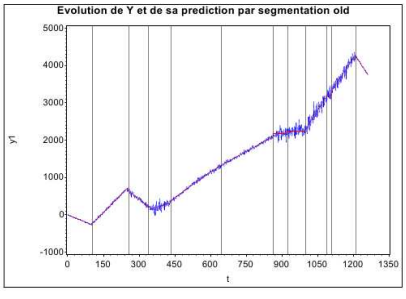

Figure 2.4.4 : Segmentation avec l'ancienne approche

Nous utilisons les quantités u_t dans les phases 1 et 2, et les valeurs estimées de ces poids pour le modèle linéaire gaussien hétéroscédastique en phase 3. La figure 2.4.1 montre la série simulée (bleu), les segments générés et l'évolution temporelle des associés u_t (rouge) de la phase 1. On peut noter que les u_t ne sont pas constants, en moyenne, ce qui signifie que des segments importants de niveau différent de la dispersion existent. La figure 2.4.2 présente les résultats issus de la première segmentation sur les u_t (phase 2) sur lequel 12 segments de dispersion ont été détectés. Nous pouvons voir que tous les segments ne sont pas constants, car les deuxième et neuvième présentent des pentes croissantes. La figure 2.4.3 exhibe la segmentation finale (phase 3), qui se compose de 18 segments dans lesquels tous les points de rupture semblent être à peu près bien détectés. Il en est d'ailleurs de même pour les résultats issus de la méthode de segmentation appliquée directement sur la série temporelle, dans laquelle 12 segments ont été identifiés (figure 2.4.4). Visuellement, les qualités respectives des deux segmentations sont comparables. Cependant, bien que cela ne soit pas très visible, les segments supplémentaires qui apparaissent grâce à la nouvelle approche semblent mieux découper la variation. Par exemple, le quatrième segment de la méthode [3,4] est divisé en deux sous-segments avec la nouvelle approche. En effet, nous pouvons constater qu'il y a plus de variabilité dans le premier sous-segment que dans le second. Par contre, la méthode de programmation dynamique (figures 2.4.5 à 2.4.8) fournit des résultats beaucoup moins satisfaisants. En effet, dans la figure 2.4.5 (DCPC3), les ruptures des trois premiers et du dernier segment ne sont pas détectés par l'algorithme même sur les sept segments estimés. Ce problème apparait également sur les types de changement DCPC4 (figure 2.4.6) et BDCPC4 (figure 2.4.8) qui devraient être plus précis, car le type de détection est plus riche. Il y a par exemple 16 segments dans les résultats issus de DCPC4. Enfin l'approche BDCPC3 (figure 2.4.7) qui donne un nombre prohibitif de segments (28 segments), échoue également pour identifier correctement les trois premiers points de rupture, ainsi que le dernier, comme les trois autres méthodes de programmation dynamique.

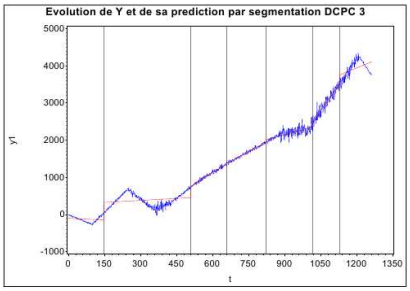

Figure 2.4.5 : Segmentation DCPC3

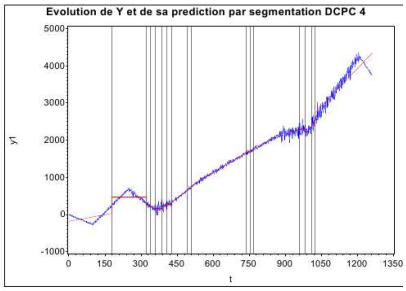

Figure 2.4.6 : Segmentation DCPC4

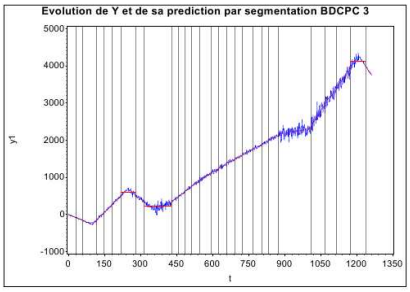

Figure 2.4.7 : Segmentation BDCPC3

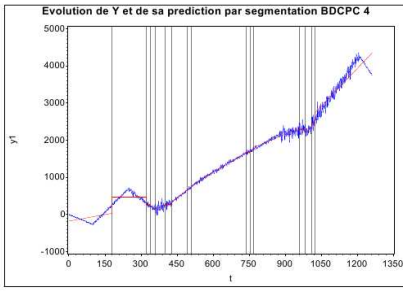

Figure 2.4.8 : Segmentation BDCPC4

Pour terminer cette comparaison entre les six méthodes, nous avons évalué la qualité de détection des points de rupture et l'adéquation des segmentations estimées aux données simulées (cf. tableau 2.1). Pour cela, nous constatons que les valeurs du MAPE de nos méthodes anciennes et nouvelles sont très similaires à celle de la segmentation générée, alors que ce n'est pas le cas pour DCPC3, DCPC4 et BDCPC4. Il en est de même pour la médiane des distributions de pourcentage d'erreurs relatives absolues (MED). La colonne suivante indique le pourcentage d'erreurs de plus de 10%, sur laquelle apparaît BDCPC4 comme le plus efficace (10,21%), bien que le MAPE soit très élevé (59,85%). Enfin, les deux dernières colonnes fournissent le nombre de segments de chaque méthode identifiés au même endroit que ceux de la segmentation générée. Encore une fois, nos deux approches sont les plus efficaces sur ces données que les quatre autres, avec 6 segments trouvés, contre 0, 2 ou 3 pour les méthodes de programmation dynamique. Ces résultats sont complétés par l'erreur d'estimation de la distance des segments aux segments générés. La méthode proposée dans cet article obtient un pourcentage d'erreur relativement faible (16,26%) par rapport à l'ancienne (24,34%), mais surtout à l'égard des quatre autres approches, la meilleure qualité est pour BDCPC3 avec 35,28%.

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

Tableau 2.1

3. Application sur un cas réel : la production d'électricité

L'objectif de l'étude présente est d'étudier le comportement de l'évolution de la production française d'électricité sur les mois de mars et avril 2011. Les questions que l'on peut se poser sont les suivantes : Y a-t-il des ruptures de comportement ? Peut-on identifier un phénomène latent, répétitif ? Est-ce que la variabilité de la production change au cours du temps ? … La figure 3.1, ci-dessous montre l'évolution de la production d'électricité horaire française.

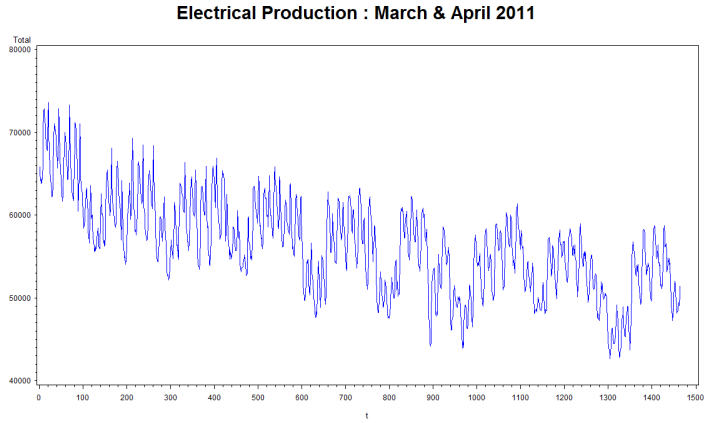

Figure 3.1 : Evolution de la production française d'électricité

Comme nous pouvons le constater sur la figure 3.2, la variabilité des données (transformation en u_t) semble décroitre au fur et à mesure du temps. Cet effet est principalement dû à la diminution des pics, ce qui peut d'ailleurs se voir sur la chronique observée. La première segmentation sur la volatilité (figure 3.3) propose trois segments, avec un degré de lissage égal à 28. Les deux ruptures se situent, le samedi 19 mars à 8 heures et le mercredi 7 avril à 8 heures, mais cela ne correspond pas spécialement à des événements marquants de la production. Il faut aussi garder à l'esprit que les constantes associées à chaque segment sont égales à la moyenne des u_t et que cette statistique est sensible aux valeurs extrêmes qui correspondent ici aux pics de variabilité comme nous l'avons indiqué précédemment. Cependant ce « défaut » prend justement toute sa valeur dans le cadre de l'analyse de la variabilité.

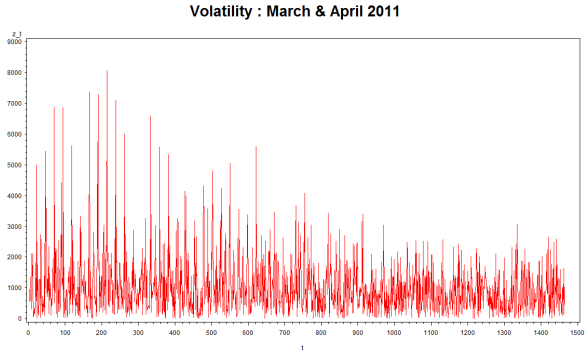

Figure 3.2 : Transformation des données avec les u_t estimés

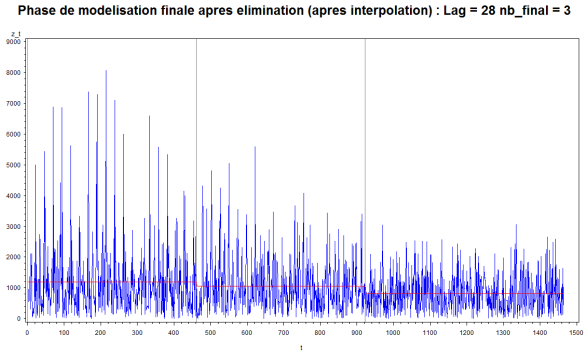

Figure 3.3 : Segmentation associée

Pour la seconde segmentation (phase 3) sur les données initiales en tenant de la variabilité de la série, nous avons choisi d'introduire les estimations des u_t dans la matrice de poids de l'estimateur REML. La figure 3.4 montre bien des ruptures de comportements dans les jours ouvrables *vs* les week-end. En effet, il y a généralement une première rupture vers le vendredi à minuit, puis une seconde, le lundi vers 6h00 du matin. Par exemple, le premier week-end de cette chronique est le 5 et 6 mars 2011. La rupture du début de ce week-end est en t = 96 et la fin se situe en t = 150 que nous pouvons voir sur la figure 3.4. Par ailleurs, la figure 3.5 qui affiche la segmentation construite à l'aide de l'ancienne approche montre peu de différence visuelle et en termes de nombre de segments : 28 *vs* 27. De plus, le MAPE pour l'ancienne méthode est égal à 3,91% contre 3,93% pour la nouvelle. Il en est de même pour le pourcentage d'erreurs supérieur à 10% : 2,73% *vs* 2,60%. Il est logique qu'il y ait peu de différence de résultats car la variation temporelle des données n'est pas très marquée.

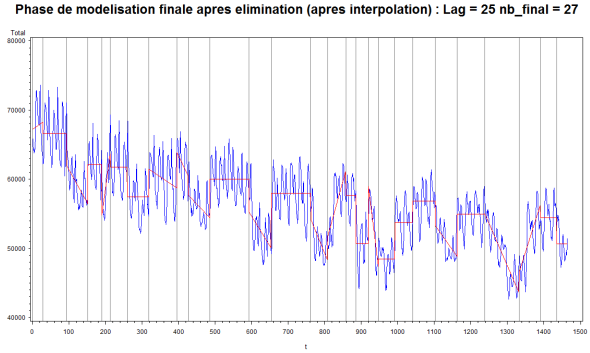

Figure 3.4 : Nouvelle approche

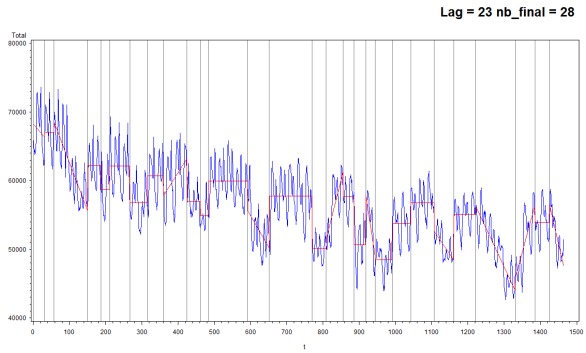

Figure 3.5 : Ancienne approche

4. Apports, critiques, applications et voies futures

La méthode proposée ici [5], qui permet de segmenter une série temporelle, a pour objectif d'améliorer une démarche que nous avions introduite en 2011 [3,4]. Celle-ci avait obtenu des résultats encourageants sur des données simulées et sur des données réelles. Elle concurrence fortement les approches fondées sur la programmation dynamique. L'amélioration proposée consiste à exhiber un signal à partir des données brutes représentant leur dispersion grâce à deux théorèmes, puis de réaliser une première segmentation du signal afin de tirer des segments de dispersion, et enfin d'inclure ces derniers comme des poids dans une seconde segmentation lors de la phase de modélisations successives. Sur des exemples simulés, cette nouvelle approche permet à la fois d'améliorer l'ancienne méthode, mais aussi de montrer qu'elle est plus performante que des approches par programmation dynamique. Cette méthode est notamment très intéressante pour des applications dans lesquelles les signaux changent de processus. La poursuite de nos travaux ira dans quatre directions. Premièrement, nous comparerons notre méthode à celle développée dans Arlot et al. [1] qui utilise une approche par validation croisée, deuxièmement, les deux théorèmes introduits pour un signal gaussien seront généralisés à d'autres lois, comme par exemple des processus de données catégorielles, ou de données de comptage, troisièmement, nous développerons une méta-segmentation permettant de prendre en compte le « meilleur » de chaque segmentation construite pour les différents degrés de lissage, enfin nous appliquerons ces différents développements à la modélisation d'une variable réponse à l'aide de variables explicatives segmentées, d'une part et à la classification de séries temporelles chacune étant résumées par leurs formes segmentées, d'autre part.

Bibliographie

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

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

[3] Derquenne, Ch. (2010), Une méthode de segmentation pour le traitement de séries temporelles, 42ièmes *Journées de Statistique,* Marseille, France.

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

[5] Derquenne, Ch. (2011), Segmentation of Time Series with Heteroskedastic Components, *58th World Statistical Congress of ISI, Dublin, Ireland*.

[6] Foulley J.L., Delmas C., Robert-Granié C. (2002), Méthodes du maximum de vraisemblance en modèle linéaire mixte, *Journal de la Société Française de Statistique*, 143, pp. 5-52.

[7] Foulley J-L. (2003), Algorithme EM : Théorie et application au modèle mixte, *Journal de la Société Française de Statistique*, **143,** (3-4), pp. 57-109.

[8] Hartley H.O. & Rao J.N.K, (1967), Maximum Likelihood Estimation for the Mixed Analysis of Variance Models, *Biometrika*,**54,** pp. 93-108.

[9] 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*.

[10] Harville, DA. (1977), Maximum likelihood approaches to variance component estimation and to related problems. JASA **72**, pp. 320-340.

[11] 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.

[12] 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.

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

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

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

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

[17] Searle, SR., Casella, G. and Mc Culloch, CE. (1992), *Variance components*, Wiley & sons, New-York.