Une méthode de segmentation pour le traitement de séries temporelles

Christian Derquenne

Une méthode de segmentation pour le traitement de séries temporelles

Christian Derquenne

Electricité de France - R&D - 1, avenue du Général de Gaulle - 92141 Clamart Cedex

Résumé : La méthode proposée permet de segmenter une série temporelle. Elle offre une démarche originale contenant une phase essentielle de préparation des données afin de produire la structure la plus adéquate possible pour initialiser la phase de modélisation selon un modèle linéaire hétéroscédastique incluant des tendances, des constantes et des dispersions différentes. Cette méthode peut être utilisée dans de nombreux domaines d'applications, mais aussi pour de nombreux objectifs : stationnarisation, recherche de segments, construction de différents modèles sur une même série ayant des comportements différents, simplification de séries dans le but de réaliser de la classification de courbes, etc.

Abstract: The method proposes to segment a time series. It offers an original process with a first step of preparing data which is crucial to build the most adequate structure to initialize the second step of modelling an heteroskedastic linear model including the different trends, levels and variances. This method can be used in a lot of domains and to set up several objectives. Building of sub-models on each detected segment, achieving stationarity of time series with a segmentation model, building symbolic curves to cluster series, modelling multivariate time series are so many examples in this context.

Mots-clés : segmentation, points de rupture, séries temporelles, composantes de variance.

1 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. L'évolution de la consommation d'électricité globale 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 des MAPE 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 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, arrivent 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 : construction de sous-modèles sur chaque segment établi, stationnarisation de la série à l'aide de la segmentation, construction de courbes symboliques 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 [2] à l'inférence sur des modèles à ruptures multiples dans des séries temporelles multivariées [5], en passant par des tests de détection de changements structurels multiples dans des modèles de régression cointégrés [6] ou encore par la détection de ruptures séquentielles lorsque

le changement de comportement des paramètres est inconnu [4]. 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, elle est généralement en $O(ST^2)$ pour le temps et en $O(ST)$ 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 $O(T^2)$ [5], même dans le cadre de ruptures multiples pour M séries temporelles multivariées, alors qu'elle pourrait être en $O(MT^2)$. Ces méthodes de détection de points de rupture ont pour vocation de résoudre trois problèmes [5] : la détection de changement de la 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, sans distinguer des changements en niveau, en variabilité et sur la distribution des erreurs. La méthode développée ici permet d'une part de résoudre un quatrième problème qui est de détecter des croissances ou des décroissances dans la série étudiée [6], 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 mêlant des croissances, des décroissances, des constantes et pouvant avoir des variances différentes.

2 La méthode proposée

2.1 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) [7,8] suivant :

$$Y_t = \sum_{s=1}^{S} (\beta_0^{(s)} + \beta_1^{(s)} t + \sigma_s \epsilon_t) 1_{[t \in \tau_s]} \tag{1}$$

où $\beta_0^{(s)}$, $\beta_1^{(s)}$ et $\sigma_s > 0$ sont respectivement les paramètres de niveau, de pente et de dispersion pour le segment τ_s, et ϵ_t suit une Normale standard. 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 $U_s = U_{s-1} + T_s$, finalement $U_S = T$. Il y a donc $3S$ paramètres à estimer, sachant que le nombre S de segments est inconnu. 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) [2,3]. Ces trois estimateurs fournissent les mêmes solutions pour les vecteurs de niveau et de pente : $\beta_0^{(s)}$ et $\beta_1^{(s)}$. Par contre, seuls les estimateurs ML et REML permettent d'inclure directement dans l'estimation le vecteur des dispersions : $(\sigma_1, ..., \sigma_S)$. Mais ce vecteur est calculable a posteriori pour OLS avec :

$$\hat{\sigma_s} = \sqrt{\sum_{t \in \tau_s} (y_t - \hat{\beta}_0^{(s)} - \hat{\beta}_1^{(s)} t)^2 / (T - 2)} \tag{2}$$

qui est un estimateur sans biais de σ_s pour $s = 1, S$. L'avantage de la méthode REML sur celle du maximum de vraisemblance est qu'elle fournit des estimateurs sans biais des composantes de variances et de covariances qui sont exactement ceux obtenus en (2). En termes de test sur les paramètres, là encore les estimateurs ML et REML permettent de réaliser de l'inférence en tenant compte des composantes de variance. Par contre, les écarts-types d'erreurs associés aux paramètres $(\beta_0^{(s)}, \beta_1^{(s)})$ de la statistique de test de Student issus de ces deux méthodes sont différents. Pour REML, ils prennent la forme suivante :

$$\hat{\sigma}^{(REML)}(\hat{\beta}_0^{(s)}) = \hat{\sigma}_s^{(REML)} \sqrt{1/T_s + \bar{t}_s^2 / \sum_{t \in \tau_s} (t - \bar{t}_s)^2} \tag{3}$$

$$\hat{\sigma}^{(REML)}(\hat{\beta}_1^{(s)}) = \hat{\sigma}_s^{(REML)} / \sqrt{\sum_{t \in \tau_s} (t - \bar{t}_s)^2} \tag{4}$$

où $\bar{t}_s$ est la moyenne des valeurs $t \in \tau_s$ et $\hat{\sigma}_s^{(REML)}$ correspond à (2), estimateur sans biais de σ_s, alors que pour l'estimateur ML, $\hat{\sigma}_s^{(REML)}$ est remplacé par $\hat{\sigma}_s^{(ML)}$, estimateur biaisé de σ_s. Le modèle (1) est seulement valide 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 hétérogénes (modèle hétéroscédastique) et à variance constante (modèle homoscédastique). Sous l'hypothèse nulle, cette statistique suit une χ^2_{S-1}. Si le modèle est homoscédastique, alors pour $s = 1, S$, $\sigma_s = \sigma$ estimé par $\hat{\sigma}$ pour obtenir les écarts-types d'erreur des couples de coefficients $\beta_0^{(s)}$ and $\beta_1^{(s)}$ pour OLS.

2.2 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, mais ils serviront dans une des phases de la méthode. Il y a deux types d'étapes, le premier correspond à une phase de préparation des données afin d'offrir un moyen raisonnable pour segmenter les données, alors que la seconde correspond à une phase de modélisations successives et adaptatives. Chacune de ses deux phases sont répétées un nombre de fois en fonction du degré de lissage appliqué aux données (cf. étape de lissage, ci-après). Le degré de lissage peut varier de 1 à T théoriquement. 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 en $O(T\sqrt{T})$ et la complexité théorique est en $O(T^2)$.

2.3 La démarche détaillée de segmentation

2.3.1 Phase de préparation des données

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 $t = 1, T - j$. Plus j croît, moins l'irrégularité des données est prise en compte.

$$m_j(t) = \underset{t \in [a_j(t), b_j(t)]}{med}(y_t) \tag{5}$$

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

Etape de différenciation Elle 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. Nous avons choisi de tenir compte de la propriété de la médiane mobile avec une différence au temps t et au temps $k = t - j/2$ si j est pair, et $k = t - (j+1)/2$ si j est impair. La différenciation s'effectue de la façon suivante :

$$d_j(t) = (m_j(t) - m_j(t-k))/m_j(t-k) \tag{6}$$

Le dénominateur permet d'obtenir un écart relatif, 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.

Etape de comptage L'étape de différenciation a permis d'établir une suite de différences relatives positives, négatives ou nulles. Le nombre 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 $\tau_{j,1}^{(0)}$ 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 $(\tau_{j,1}^{(0)}, ..., \tau_{j,s}^{(0)}, ..., \tau_{j,S}^{(0)})$, de tailles $(T_{j,1}^{(0)}, ..., T_{j,s}^{(0)}, ..., T_{j,S}^{(0)})$ et $\sum_{s=1}^{S} T_{j,s}^{(0)} = T$.

2.3.2 Phase de modélisation

Etape initial de modélisation 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 2.2, chaque étape permet de simplifier le modèle proposé au départ de celle-ci. Par conséquent, elle se décompose en plusieurs sous-étapes. Le modèle complet pour un degré de lissage j est de la forme (1), tel que :

$$Y_t = \sum_{s=1}^{S} (\beta_0^{(j,s)} + \beta_1^{(j,s)} t + \sigma_{j,s}\epsilon_t) 1_{[t \in \tau_{j,s}^{(0)}]} \tag{7}$$

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 :

$$Y_t = \sum_{s=1}^{S} (\beta_0^{(j,s)} + \beta_1^{(j,s)} t) 1_{[t \in \tau_{j,s}^{(0)}]} + \sigma_j \epsilon_t \tag{8}$$

Le modèle simplifié est construit en réalisant S tests d'égalité à 0 des coefficients $\beta_1^{(j,s)}$ du modèle complet. La statistique de test de Student est : $\hat{\beta}_1^{(j,s)} / \hat{\sigma}^{REML}(\hat{\beta}_1^{(j,s)})$, si le modèle est hétéroscédastique, sinon l'écart-type d'erreur du coefficient est remplacé par $\hat{\sigma}^{OLS}(\hat{\beta}_1^{(j,s)})$. Le modèle général est :

$$Y_t = \sum_{s=1}^{S} (\beta_0^{(j,s)} + \beta_1^{(j,s)} t 1_{[\beta_1^{(j,s)} \neq 0]} + \sigma_{j,s}\epsilon_t) 1_{[t \in \tau_{j,s}^{(0)}]} \tag{9}$$

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 : $(\beta_0^{(j,s)}, \beta_1^{(j,s)}, \sigma_{j,s})$, ou par $(\beta_0^{(j,s)}, \beta_1^{(j,s)}, \sigma_j)$, si le modèle est homoscédastique. Le premier test concerne l'égalité des variances $\sigma_{j,s}^2$ et $\sigma_{j,s+1}^2$ sur le modèle suivant :

$$Y_t = (\beta_0^{(j,s)} + \beta_1^{(j,s)} t + \sigma_{j,s}\epsilon_t) 1_{[t \in \tau_{j,s}^{(0)}]} + (\beta_0^{(j,s+1)} + \beta_1^{(j,s+1)} t + \sigma_{j,s+1}\epsilon_t) 1_{[t \in \tau_{j,s+1}^{(0)}]} \tag{10}$$

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 un test sur l'égalité des coefficients $\beta_0^{(j,s)}$ et $\beta_0^{(j,s+1)}$ est appliqué, à l'aide de la statistique de test :

$$\mid \hat{\beta_1}^{(j,s)} - \hat{\beta_1}^{(j,s+1)} \mid / \hat{\sigma}^{OLS}(\hat{\beta_1}^{(j,s)}, \hat{\beta_1}^{(j,s+1)}) \tag{11}$$

où

$$\hat{\sigma}^{OLS}(\hat{\beta_1}^{(j,s)}, \hat{\beta_1}^{(j,s+1)}) = \hat{\sigma}_{j,s,s+1} \sqrt{1/\sum_{t \in \tau_{j,s}^{(0)}} (t - \bar{t}_{j,s})^2 + 1/\sum_{t \in \tau_{j,s+1}^{(0)}} (t - \bar{t}_{j,s+1})^2}$$

avec $\hat{\sigma}^2_{j,s,s+1} = (T^{(0)}_{j,s}\hat{\sigma}^2_{j,s} + T^{(0)}_{j,s+1}\hat{\sigma}^2_{j,s+1})/(T^{(0)}_{j,s} + T^{(0)}_{j,s+1})$. Si ces deux coefficients sont égaux, alors les segments $\tau^{(0)}_{j,s}$ et $\tau^{(0)}_{j,s+1}$ 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 oeuvre. A la fin de ce processus, le nombre de groupes obtenus $S_1 \leq S$ correspondra aux nouveaux segments à incorporer dans le modèle à regroupement de segments, tel que :

$$Y_t = \sum_{s=1}^{S_1} (\beta_0^{(j,s)} + \beta_1^{(j,s)} t + \sigma_{j,s}\epsilon_t) 1_{[t \in \tau_{j,s}^{(1)}]} \tag{12}$$

Etapes de modélisation suivantes Dans l'étape suivante, le modèle (12) passe par le même processus de tests successifs que ceux établis dans le paragraphe précédent. Ce processus de simplification du modèle est répété jusqu'à ce le nombre de segments soit satisfaisant. Dans l'état actuel de notre travail, le critère d'arrêt est encore empirique (en général, 4 itérations).

2.3.3 Evaluation des modèles

Les deux phases (préparation 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 et aura d'autant plus de chances de fournir la meilleure segmentation possible. Signalons que même si les T degrés de lissage sont essayés, cela ne garantit pas d'obtenir la segmentation optimale avec une probabilité de 1, mais l'objectif n'est pas celui-ci dans le cadre de la méthode proposée. En effet, comme le modèle est relativement complexe car il permet de révéler la tendance, le niveau et la dispersion de la série temporelle, l'objectif est plutôt de proposer un certain nombre de segmentations candidates. Pour cela, nous travaillons avec quelques mesures statistiques pour offrir un choix de segmentations raisonnables. Tout d'abord nous avons choisi la valeur du REML qui a permis d'estimer le modèle, puis le MAPE (moyenne des erreurs relatives en valeurs absolues) et la distribution de celles-ci. Pour REML et MAPE le minimum des valeurs obtenues pour l'ensemble des degrés de lissage a de fortes chances de fournir la "meilleure" segmentation. Un certain nombre de segmentations pourront alors être proposées, en ordonnant les deux mesures.

3 Application

Nous avons appliqué la méthode proposée sur un jeu de données simulées. Les données temporelles ont été simulées sur 10 segments, selon le modèle (1). Pour chacun des 10 segments, le nombre d'observations, les valeurs des coefficients β_0 et β_1, et la dispersion σ associés 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 (fig. 2) *vs* 10 segments pour la segmentation simulée (fig. 1). La figure 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 (fig. 3 et 4). Ce résultat montre l'intérêt de cette méthode pour stationnariser une série.

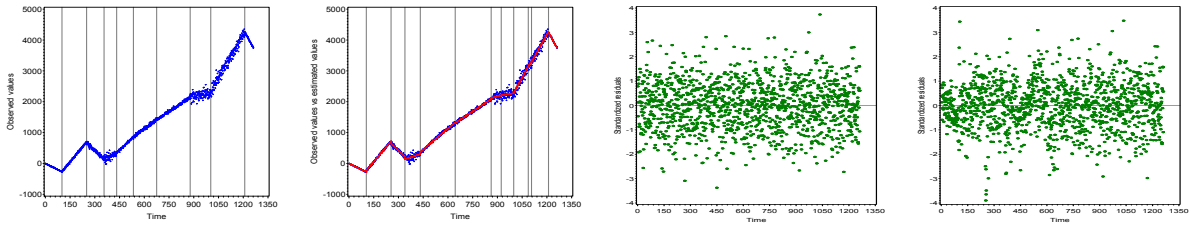

Figure 1: Segment. obs. – Figure 2: Segment. est. – Figure 3: Erreurs obs. – Figure 4: Résidus est.

4 Apports, critiques, applications et voies futures

La méthode proposée permet de segmenter une série temporelle. Elle offre une démarche originale contenant une phase essentielle de préparation des données pour produire la structure la plus adéquate possible pour initialiser la phase de modélisation selon un modèle linéaire hétéroscédastique incluant des tendances, des constantes et des dispersions différentes. Cette méthode n'a pas pour objectif de fournir la segmentation optimale comme dans la plupart des méthodes discutées dans l'introduction, car même si la complexité est généralement en $O(T^2)$, au minimum ; elle reste néanmoins élevée. La méthode introduite a pour objectif de proposer un certain nombre de segmentations candidates. Pour cela elle utilise des critères d'évaluation, tels que la vraisemblance de l'estimateur REML, le MAPE et le pourcentage d'erreurs relatives inférieures à 10%, par contre sa complexité est en $O(T)$ pour chaque degré de lissage et leur nombre dépasse très rarement $\sqrt{T}$. En effet, la qualité des segmentations estimées se déprécie rapidement lorsqu'elles s'éloignent de l'optimalité, même si celle-ci est empirique. Comme nous l'avons indiqué, cette méthode peut être utilisée dans de nombreux domaines d'applications, mais aussi pour de nombreux objectifs : stationnarisation, recherche de segments, construction de différents modèles sur une même série ayant des comportements différents, simplification de séries dans le but de réaliser de la classification de courbes, etc. Enfin, un certain nombre de travaux d'amélioration restent à effectuer, notamment en termes de validation des segmentations obtenues, par exemple, en ayant une approche d'évaluation par segment, ou encore en estimant la volatilité a priori.

Bibliographie

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

[2] 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*.

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

[4] 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.

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

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

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

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