# Formation d'un phénomène temporel à l'aide d'un méta-modèle via la segmentation

Christian Derquenne

*Electricité de France - Recherche et Dévelopement - 1, avenue du Général de Gaulle - 92141 Clamart Cedex - christian.derquenne@edf.fr*

**Résumé.** Est-ce qu'un phénomène temporel est systématiquement expliqué par les mêmes prédicteurs sur toute la longueur de la série ? Cette question se pose notamment dans le cadre de séries irrégulières non stationnaires, comme par exemple en finance (prix de marché, CAC40, ...), afin de comprendre comment se forme la réponse. L'approche proposée permet non seulement de répondre à ce problème, mais fournit également la contribution statistique de chaque prédicteur à la réponse. Une méthode de segmentation de série temporelle est appliquée au préalable sur la réponse pour détecter la non-stationnarité, puis pour chaque segment, différentes stratégies de régression multiple sont utilisées afin d'obtenir un méta-modèle.

**Mots-clés.** Série temporelle, segmentation, méta-modèle, modéle linéaire, régression PLS.

**Abstract.** Does a response series is systematically explained by the same inputs along the entire length of the series? This question arises particularly in the context of non-stationary irregular series, such as finance (market prices, CAC40, ...), to understand how to shape the response. The proposed approach can not only answer to this problem, but also provides the statistical contribution of each predictor to the response. For this, a time series segmentation method is used to detect non-stationary, then for each segment of the response, we apply a multiple regression model using different strategies to obtain a meta-model.

**Keywords.** Time series, segmentation, meta-model, linear model, PLS regresssion.

## 1 Contexte - objectif

Les séries temporelles se décomposent généralement en plusieurs types d'évolution : tendance, saisonnalité, dispersion et bruit. Elles peuvent être plus ou moins régulières selon le domaine d'application. Les changements de comportements qui caractérisent principalement ces séries sont de plusieurs types : pics (prix d'une énergie en situation tendue, mais sur une très courte période), sauts en niveau ou en tendance (rassemblement ou séparation de flux de données), sauts 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. Il peut alors être intéressant de détecter des ruptures de comportement pour de nombreuses applications dans le cadre du prétraitement des données : 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 de segmentation ont été et sont développées, notamment par Arlot (2010), Guédon (2008), Lavielle et Teyssière (2006) pour répondre à différentes problématiques en économie, en finance, en séquençage humain, en météorologie, en management de l'énergie,

etc. La plupart de ces méthodes reposent sur l'utilisation de la programmation dynamique pour diminuer drastiquement le nombre de segmentations possible : $2^{T-1}$, où $T$ est la longueur de la série. Ces méthodes de détection de points de rupture ont pour vocation de résoudre trois problèmes : la détection de changement de la moyenne, la détection de changement de variance et la détection de changements en distribution du phénomène étudié.

Nous avons introduit une méthode de segmentation dans Derquenne (2011a). Cette méthode permet non seulement de réduire la complexité par rapport à d'autres méthodes, mais surtout propose des solutions de segmentation de la série contenant des segments croissants, décroissants, constants et des dispersions différents. 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 de la série temporelle et la modélisation des segments à l'aide d'un modèle linéaire gaussien hétéroscédastique par adaptations successives. Chacune de ces deux phases est répétée un certain nombre de fois en fonction du degré de lissage $j$ appliqué aux données jusqu'à convergence du nombre de segments. Le degré de lissage correspond au nombre d'observations incluses dans la médiane mobile utilisée dans la phase de préparation des données. Ce degré de lissage peut varier de 2 à $T$ théoriquement. Cette méthode a été améliorée grâce à une meilleure prise en compte de la variabilité des données (Derquenne, 2011b), puis à l'aide d'une approche de méta-segmentation (Derquenne, 2012) sélectionnant les meilleurs segments issus des différents degrés de lissage $j$ disponibles. 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.

Si les méthodes de segmentation permettent principalement de détecter des ruptures de comportements, il est également important de répondre à d'autres questions pratiques : (i) Existent-ils des comportements similaires entre ces séries, notamment y a t-il des ruptures communes ? (ii) Existent-ils des groupes de traits communs et/ou de traits différents ? (iii) Existent-ils des leviers inobservables d'un ou plusieurs phénomènes temporels à expliquer ? (iv) Est-ce qu'une réponse temporelle est systématiquement expliquée par les mêmes facteurs ? (v) Corollairement, quels sont les poids respectifs de ces facteurs explicatifs dans la formation de cette réponse ? Nous avons proposé des approches pour répondre à (i), (ii) et (iii) dans Derquenne (2013,2014). L'objectif de cette contribution est d'introduire une méthode pour répondre à (iv) et (v). La section 2 formalise cette approche, puis la section 3 est consacrée à une application sur les prix de marché de l'énergie, enfin la dernière section conclut sur les apports, les améliorations potentielles de l'approche proposée, ainsi que sur les voies futures de recherche.

## 2 Formation d'un phénomène temporel par méta-modèle

L'objectif est de modéliser une réponse à temps discret $(Y_t)_{t=1,T}$, à l'aide de préditeurs temporels $(\mathbf{X}_t)_{t=1,T} = (X_{1,t}, ..., X_{p,t})_{t=1,T}$, candidats à l'explication. Pour cela, de nombreuses approches ont été proposées dans la littérature, comme par exemple des ARMAX ou des modèles de cointégration. Dans ce type de modèles, les séries sont préalablement différenciées selon un

certain ordre, dépendant de la nature des données, afin de les rendre les plus stationnaires possible. Cependant, si une ou plusieurs séries sont non stationnaires par morceaux, l'application d'ordre de différenciation identique par chronique peut pénaliser la qualité de l'ajustement, et par conséquent la compréhension de la formation d'une réponse ou de sa prévision future. Une solution raisonnable, pour pallier ce problème, est de détecter les ruptures de comportements dans la série temporelle à l'aide d'une méthode de segmentation. En effet, la forme de chaque segment de la réponse $(Y_t)_{t=1,T}$ peut être seulement liée à tout ou partie, voire même à aucun préditeur parmi les $p$ proposés. Cela permet de répondre à la question (iv) posée dans l'introduction ; "Est-ce qu'une réponse temporelle est systématiquement expliquée par les mêmes facteurs ?". De plus, dans le cas où il y a plusieurs prédicteurs significatifs pour un même segment de la réponse, il peut alors être fructueux de hiérarchiser leurs contributions individuelles à la formation de celle-ci. Cela permet alors de répondre à la question complémentaire : "Quels sont les poids respectifs des facteurs explicatifs dans la formation de cette réponse ?

Soient $(\tau_1^{(Y)}, ..., \tau_Q^{(Y)})$, les $Q$ segments de la réponse $(Y)$ et soit $T_q^{(Y)}$, le nombre d'observations $(y_t)$ de $\tau_q^{(Y)}$, avec $\sum_{q=1}^{Q} T_q^{(Y)} = T$. Au segment $\tau_q^{(Y)}$ correspond un certain ensemble de segments notés $(\tau_{q,1}^{(X_j)}, ..., \tau_{q,S_q}^{(X_j)})$ associés à chaque prédicteur $X_j$. Signalons que le premier et le dernier segments de cet ensemble peuvent ne pas être complets. En effet, le segment $\tau_{q,1}^{(X_j)}$ peut correspondre à la fin du segment $\tau_{q-1,S_{q-1}}^{(X_j)}$ du précédent segment dans la segmentation de $X_j$. Le principe de construction du méta-modèle reposant sur la segmentation est le suivant. Soit $\tau_1^{(Y)}$, le premier segment de $(Y)$ contenant $T_1^{(Y)}$ observations $(y_1, ..., y_{T_1^{(Y)}})$, à celles-ci correspondent $p$ séquences contenant le même nombre d'observations associées aux $p$ prédicteurs temporels $X_j$ découpés eux-mêmes en un certain nombre de segments, comme indiqué précédemment. Il y a alors deux possibilités pour modéliser ce premier segment de $(Y)$, soit tenir compte de la structure en segments de chaque prédicteur, soit postuler un comportement commun des observations, comme s'il y avait un seul segment associé à chaque prédicteur. Dans les deux cas, un pré-traitement de chaque couple $(Y, X_j)$ sera nécessaire avant la modélisation de la réponse à l'aide des $p$ prédicteurs temporels. Dans le premier cas (plusieurs segments), pour chaque $X_j$, le segment $\tau_{1,1}^{(X_j)}$ sera regroupé avec $\tau_{1,2}^{(X_j)}$ et/ou $\tau_{1,S_1}^{(X_j)}$ sera rassemblé avec $\tau_{1,S_1-1}^{(X_j)}$ si leur taille respective est strictement inférieure à 3. Dans le second cas (un seul segment), comme il est probable que les séquences d'observations associées à chaque prédicteur soient relativement hétérogènes, alors une régression robuste linéaire avec le $M$-estimateur introduit par Huber (1973) sera appliquée afin de se prémunir d'éventuels points atypiques. Cette méthode fournit des poids associés à chaque observation en fonction de son degré d'atypicité. Si une observation $(y_t, x_{j,t})$ obtient un poids en-dessous d'un seuil fixé, alors la valeur $x_{j,t}$ prendra le statut de donnée manquante. En d'autres termes, celle-ci ne participera pas à l'ajustement du modèle de régression multiple au sein de ce premier segment. Mais comme ce choix, d'élimination d'observations, risque de pénaliser à son tour l'ajustement par régression multiple, la régression PLS (Partial Least Squares) introduite par S. Wold (1984) pourra être utilisée. En effet, celle-ci permet de construire un modèle de régression en préservant le plus d'observations possible, à condition que la proportion de données manquantes ne soit pas trop élevée. Cette méthode est également intéressante car la taille de certains segments de $(Y)$ peut être faible. A la suite de ces

deux types de pré-traitement, nous proposons quatre stratégies de modélisation par régression multiple. Chacune d'elle peut être appliquée soit sur les données brutes, soit sur les données standardisées à l'aide de la segmentation. Cette transformation revient à stationnariser par morceaux (les segments) la série temporelle. La première stratégie consiste à appliquer une méthode de sélection pas à pas des prédicteurs temporels à l'aide de l'estimateur MCO (Moindres Carrés Ordinaires), puis d'utiliser ce modèle pour ajuster le premier segment de la réponse $(Y)$. La deuxième stratégie modélise les données à l'aide de la régression PLS avec les prédicteurs sélectionnés précédement. La régression PLS permet de pallier le problème de multicolinéarité qui peut être présent parmi les variables explicatives. La troisième stratégie sélectionne tout d'abord les préditeurs en fonction de leurs corrélations marginales avec la variable réponse. Pour cela, le test du rapport de Fisher sera utilisé sur le premier segment de $(Y)$ associé à chaque prédicteur segmenté (premier cas) ou à l'aide d'un test de corrélation linéaire simple de Pearson (second cas). Puis une régression PLS est appliquée sur les prédicteurs retenus. Enfin, la quatrième stratégie consiste à modéliser la réponse $(Y)$ par l'ensemble des $p$ prédicteurs à l'aide de la régression PLS. Lorsque le premier segment est modélisé, alors la même démarche est appliquée sur le deuxième segment de $(Y)$, etc. La construction du méta-modèle se termine lorsque le dernier segment de $(Y)$ est ajusté. Les formes respectives du méta-modèle pour le premier et le second cas sont fournies par les équations (1) et (2), quelle que soit la stratégie.

$$y_{t\in\tau_q^{(Y)}} = \sum_{s=1}^{S_q}(\sum_{j=1}^{p}\beta_j^{(q,s)}1_{[p-val(\beta_j^{(q,s)})<\alpha]}x_{j,t\in\tau_{q,s}} + \beta_0^{(q,s)}) + \epsilon_t \quad (1)$$

où $\beta_j^{(q,s)}$ désigne le coefficient de régression associé au $s$-ième segment $\tau_{q,s}^{(X_j)}$ du prédicteur $X_j$ au sein du $q$-ième segment de la réponse $(Y)$ et $\beta_0^{(q,s)}$ est la constante associée.

$$y_{t\in\tau_q^{(Y)}} = \sum_{j=1}^{p}\beta_j^{(q)}1_{[p-val(\beta_j^{(q)})<\alpha]}x_{j,t} + \beta_0^{(q)} + \epsilon_t \quad (2)$$

où $\beta_j^{(q)}$ désigne le coefficient de régression associé au prédicteur $X_j$ au sein du $q$-ième segment de la réponse $(Y)$ et $\beta_0^{(q)}$ est la constante associée.

Le coefficient de régression $\beta_j^{(q,s)}$, respectivement $\beta_j^{(q)}$ sera présent dans l'équation de régression, seulement si sa $p$-valeur associée au test de Student est inférieure à un seuil $\alpha$ fixé, par exemple 0,05, sinon il sera égal à 0 et dans ce cas, seule la constante $\beta_0^{(q,s)}$, respectivement $\beta_0^{(q)}$ sera estimée. Signalons que cette règle est seulement valide si le modèle a été calculé à l'aide de l'estimateur des MCO. Dans le cas de la régression PLS, tous les coefficients associés aux prédicteurs sélectionnés garderont leurs propres valeurs estimées. En effet, il n'y pas de test statistique disponible, et raisonnable, associé aux coefficients estimés au moyen de la régression PLS car ceux-ci sont construits à l'aide d'un certain nombre de composantes PLS permettant de résumer l'espace des prédicteurs. Signalons également que des tests de Student sur les coefficients des composantes PLS ne sont pas valables non plus car trop dépendants de la réponse $(Y)$ dans leurs constructions. Il sera alors nécessaire d'utiliser des tests de validation croisée. Terminons, en répondant à la question "Quels sont les poids respectifs des facteurs explicatifs dans la formation de la réponse ?" afin de hiérarchiser les prédicteurs de la variable réponse $(Y)$. Pour cela, des poids associés à chaque variable explicative sont calculés pour chaque segment $\tau_q^{(Y)}$ de $(Y)$. Ces poids sont obtenus à partir de la décomposition du coefficient de détermination $R^2$ pour chaque segment $\tau_q^{(Y)}$ qui prend la forme suivante :

$$R_q^2 = (\beta^{(q)'}\mathbf{X}'\mathbf{X}\beta^{(q)})/S_Y^2 = (\beta^{(q)'}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'Y)/S_Y^2 = \tilde{\beta}^{(q)'}r_{\mathbf{X}Y} = \sum_{j=1}^{p_q}\tilde{\beta}_j^{(q)}cor(Y,X_j) \quad (3)$$

où $r_{\mathbf{X}Y}$ est le vecteur des corrélations linéaires simples entre la réponse et les prédicteurs, $\tilde{\beta}^{(q)}$ est le vecteur des coefficients de régression standardisés de taille $p_q$ (les prédicteurs sélectionnés pour le segment $\tau_q^{(Y)}$) et $S_Y^2$ est la variance de $(Y)$. La contribution d'un prédicteur est alors $\%CTR(X_j) = (\tilde{\beta}_j^{(q)}cor(Y,X_j)/R_q^2) \times 100$ et doit varier entre 0 et 100. En effet, les signes de l'ensemble des coefficients de régression multiple et des corrélations simples qui constituent $R_q^2$ doivent être identiques pour que cet indice de contribution puisse être raisonnablement utilisé.

## 3 Application de la méthode

Nous disposons de 6 prédicteurs $(X_1, ..., X_6)$ et d'une réponse $Y$ sous forme de séries temporelles, auxquelles sont associées leurs segmentations. Ici, la réponse $Y$ est découpée en 20 segments. Ces données réelles sur des prix de marché de l'énergie ont été anonymisées pour confidentialité.

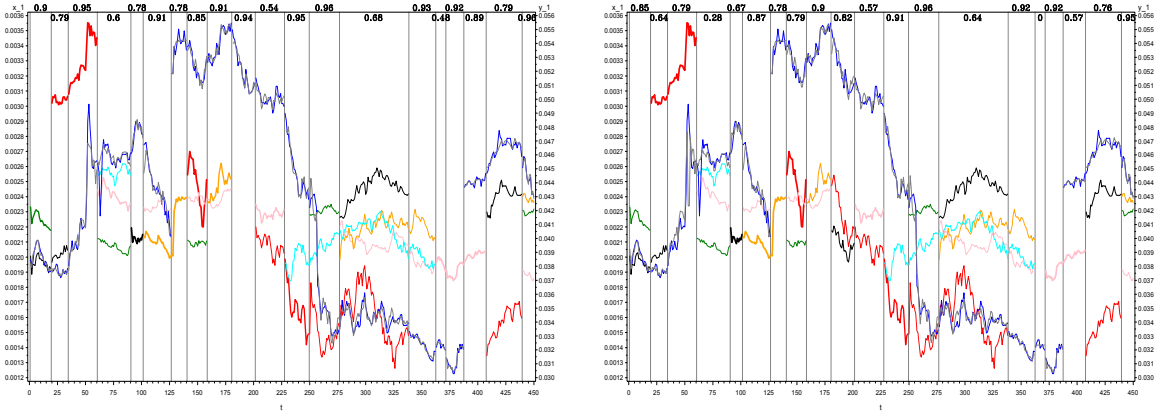

Figure 1: (a) Contribution des prédicteurs par MCO (b) Contributions de prédicteurs par PLS

Nous présentons ici seulement les résultats des stratégies (i) et (ii) sur les données brutes, sans tenir compte des segmentations des prédicteurs. Sur la figure 1(a), les courbes bleue et grise correspondent respectivement à la série observée de $(Y)$ et à sa prédiction, les autres courbes correspondent aux six séries de prédicteurs. Par exemple, les séries verte et noire sont les seules à être significatives pour le premier segment. La courbe verte est plus épaisse que la courbe noire, cela signifie que le premier prédicteur a une contribution plus élevée que celle du second (68,6% *vs* 31,4%). Le $R^2$ de ce premier segment est égal à 0,90 (cf. valeur indiquée en haut du graphique) ce qui représente une très bonne qualité d'adéquation du modèle aux données. Dans le troisième segment, seul le prédicteur rouge est significatif, mais il possède un pouvoir explicatif élevé puisque $R^2 = 0,95$. Il est intéressant de constater que dans le segment 10 ($t$=181 à 201), il n'y a pas de courbe car il y avait un problème de multicollinéarité : quelques signes des coefficients de régression multiple ne correspondaient pas à ceux de la corrélation simple

(cf. formule (3)). Cependant ce problème a été pris en compte grâce à la régression PLS (cf. figure 1(b)). En effet, les préditeurs rouge et noir se partagent à peu près la contribution à la formation de $(Y)$, 56,2% *vs* 43,8%. Les séries orange et rouge n'apparaissent quasiment jamais ensembles pour les 20 segments, sauf dans le segment 14 ($t$=277 à 338) qui prend en charge 5 prédicteurs sur 6. Visuellement, plus le $R^2$ d'un segment est élevé, plus les courbes des préditeurs significatifs sont similaires à la courbe de la réponse observée. Enfin, les stratégies (iii) et (iv) permettent aussi de bien ajuster la réponse, mais sont plus sensibles à la multicolinéarité entre prédicteurs.

## 4 Apports, applications et voies futures

L'approche proposée permet de comprendre quels prédicteurs forment une réponse temporelle non stationnaire par morceaux et quels sont leurs poids respectifs. Les résultats obtenus dans l'application apportent des informations très fructueuses pour les experts du domaine d'application, à la fois pour repérer à quel moment certains prédicteurs influencent la formation de la réponse, mais aussi quels niveaux de contribution ils apportent. Les futurs travaux de recherche seront consacrés à une meilleure prise en compte de la multicolinéarité des prédicteurs et à la prévision à court terme d'une réponse temporelle à l'aide de l'approche méta-modèle.

## Bibliographie


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

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

[3] Derquenne, C. (2011b): Segmentation of Time Series with Heteroskedastic Components, $58^{th}$ *World Statistical Congress of ISI*, Dublin, Ireland.

[4] Derquenne, C. (2012): Meta-segmentation of time series for searching a better segmentation, *in Proc. of Compstat 2012*, Limassol, Cyprus, pp. 191-204.

[5] Derquenne, C. (2013): Clustering of time series via a segmentation approach, IFCS 2013, *Program and Book of Abstracts*, Tilburg, The Netherlands p. 134.

[6] Derquenne, C. (2014): Modelling multivariate time series by structural equations modelling and segmentation approach, *in Proc. of Compstat 2014*, Geneva, Switzerland, pp. 459-466.

[7] 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*.

[8] Huber P.J., (1973): Robust Regression: Asymptotics, Conjectures and Monte-Carlo, Statistics, *Ann. Stat*, Vol 1, **5**, 799-821.

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

[10] Wold S., Ruhe A., Wold H. and Dunn III W.J., (1984): The collinearity problem in linear regression. The partial least squares (PLS) approach to generalized inverses, *SIAM J. Sci. Stat. Comput.*, **5**, n°3, 735-743.