PREVISIONS DE PROCESSUS AUTOREGRESSIFS HILBERTIENS

J. DAMON

MEDIAMETRIE

Résumé

Cet article présente des méthodes d'estimation de processus fonctionnels. Un intérêt tout particulier est porté sur l'estimation de processus autorégressifs Hilbertiens par lissage spline. Ceux-ci sont comparés aux régressions par noyau scalaire et fonctionnelle, et à des méthodes paramétriques de type SARIMA. Un exemple d'application est donné à partir de l'étude du taux moyen d'audience de la télévision.

Mots Clés : ARH(1), processus autorégressif fonctionnel, processus lisse, lissage spline, SARIMA, prédiction d'audience.

1 Introduction

L'emploi de modèles en temps discret sur les observations d'un processus dont le vrai modèle est à temps continu peut négliger des caractéristiques essentielles de ce processus et nuire aux prévisions. Il faut garder à l'esprit que si les observations sont toujours en temps discret, il n'en n'est pas de même pour le processus qui les a générées. L'étude d'observations en temps discret n'est pas alors une justification valable pour l'emploi d'un modèle en temps discret.

En effet, on peut perdre ainsi la notion de continuité du temps (et souvent aussi celle du processus). L'emploi d'un modèle autorégressif discret permet de rendre compte de la dépendance temporelle mais ne doit pas s'appliquer à la discrétisation d'un processus

[1]Etudiant en thèse rattaché au Laboratoire de Statistiques Théoriques et Appliquées, Université Pierre et Marie Curie, 75292 Paris cedex 05, FRANCE (E-mail : jdamon@mediametrie.fr)

autorégressif fonctionnel tel qu'un ARH(1) défini au paragraphe 2.3. Dans l'hypothèse où le vrai modèle est en temps continu et vérifie des propriétés de continuité des fonctions ou des dérivées, on se prive, en employant des modèles en temps discret, de caractéristiques pouvant améliorer les prédictions.

De tels processus sont fréquemment rencontrés lors de l'étude de phénomènes à variations lentes tels les cycles climatologiques, ou lors de l'observation à haute fréquence de processus, comme cela est le cas à Médiamétrie. L'audience de la télévision est mesurée seconde par seconde, ce qui fournit une longue série d'observations.

Historiquement, l'étude de données fonctionnelles a été explorée dans le cadre factoriel avec, entre autres, Deville [DEV74] qui réalise une Analyse en Composantes Principales (ACP) sur données fonctionnelles. Besse et Ramsay [BR86] montrent ensuite qu'une ACP fonctionnelle correspond à une norme modifiée. Par la suite Silverman [SIL96] et Besse, Cardot, et Ferraty [BCF97] réalisent une ACP fonctionnelle avec contrainte de lissage sur la norme.

L'objet de notre étude est de comparer différents modèles de prédiction sur le processus lié à l'audience globale de la télévision. Dans un premier temps nous préciserons les notations et présenterons les différents modèles employés : régression ponctuelle par noyau d'une part, régression fonctionnelle par noyau et modèles autorégressifs fonctionnels lisses d'autre part. Nous discuterons ensuite de la mise en œuvre et des résultats obtenus, en particulier des différences obtenues entre les modèles fonctionnels et les modèles ponctuels, non paramétriques ou, plus traditionnellement, paramétriques de type SARIMA.

2 Modélisations

On considère une série chronologique à valeurs réelles $(X_k)_{k\in\mathbb{Z}}$ observée p fois sur n périodes $\{x_1, \ldots, x_{np}\}$. Cette série pourra être considérée par la suite comme les observations $\{y_i(t_j) = x_{(i-1)p+j}\ ;\ i = 1, \ldots, n\ ; j = 1, \ldots, p\}$ de n fonctions $(y_i)_{i=1,\ldots,n}$; les $(t_j)_{j=1,\ldots,p}$ décrivent les temps d'observation sur chaque période (cf. figure 1).

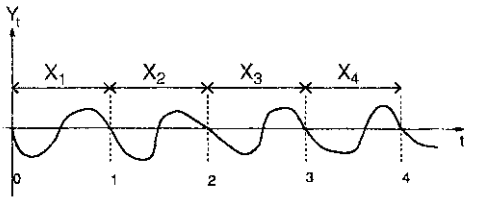

Figure 1 - Modélisation d'un processus à temps continu par un processus fonctionnel.

Cette approche du processus par suites de fonctions présente l'avantage de réduire la complexité du modèle employé en faisant passer dans les observations $(y_i)_{i=1,\ldots,n}$ l'aspect temps continu. En outre, il est possible de choisir les $(y_i)_{i=1,\ldots,n}$ avec une certaine souplesse. Dans le cas présent nous employons les valeurs de $(X_k)_{k\in\mathbb{Z}}$ à la suite, mais la modélisation peut être constituée de fonctions contigües (comme sur la figure 1) ou non, qui se chevauchent ou non, sur un intervalle ou plusieurs. Là encore la complexité n'est pas entre les $(y_i)_{i=1,\ldots,n}$ mais au sein de chaque fonction.

2.1 Régression scalaire par noyau

La prévision non-paramétrique d'un processus réel à partir de son r-historique (ses r valeurs précédentes) est une approche possible pour l'étude de la série $(X_k)_{k\in\mathbb{Z}}$. Notons[1]

$$\mathbf{X}_t^r = (X_t, ..., X_{t-r+1}) \in \mathbb{R}^r$$

le vecteur r-historique et s l'horizon de prévision $(0 < s \leq p)$.

L'autorégression à horizon s à partir de l'historique $(X_k)_{k=1,\ldots,T}$ est définie par

$$f_s(\mathbf{x}) = \mathbb{E}\left[X_{T+s} \,|\mathbf{X}_T^r = \mathbf{x}\right]$$

L'estimateur à noyau de $f_s(\mathbf{x})$ à partir des observations $\{x_1, \ldots, x_T\}$ est

$$f_{T,s}(\mathbf{x}) = \frac{\sum_{t=r}^{T-s} \mathbf{x}_{t+s} \cdot K\left(\frac{\mathbf{x}-\mathbf{x}_t^r}{h_T}\right)}{\sum_{t=r}^{T-s} K\left(\frac{\mathbf{x}-\mathbf{x}_t^r}{h_T}\right)} \tag{1}$$

où K est un noyau de dimension r et h_T la taille de la fenêtre. La pratique montre que le choix du noyau n'est pas crucial dans ce type de modèle (cf. Bosq [BOS98]), prenons donc pour K le noyau gaussien

$$K(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{r}{2}}} e^{-\frac{\|\mathbf{x}\|^2}{2}}, \ \mathbf{x} \in \mathbb{R}^r$$

La prévision à l'horizon s est alors donnée par

$$X_{T+s|T} = f_{T,s}(\mathbf{x}_T^r)$$

Le choix de h_T est déterminé par validation croisée selon

$$h_T = \arg\min \sum_{s=1}^{p} CV_s(h)$$

[1] Les matrices et vecteurs sont notés en gras.

avec

$$CV_s(h) = \sum_{k=r}^{T-pm-s} (x_{k+s} - f_{s,h,-k}(\mathbf{x}_k^r))^2$$

où $f_{s,h,-k}$ est la fonction d'autorégression à l'horizon s obtenue à partir des observations $\{x_1, \ldots, x_{k-1}, x_{k+1}, \ldots, x_{T-pm}\}$ et m est le nombre de périodes gardées afin de tester la qualité de la prédiction.

Cet estimateur dispose de bonnes propriétés asymptotiques avec une vitesse de convergence optimale sous des hypothèses de stationnarité, de mélangeance, de propriétés de Markov ou d'ergodicité (cf. Bosq [BOS98]).

2.2 Régression fonctionnelle par noyau

Si maintenant on considère l'approche fonctionnelle, les observations sont issues d'un processus Hilbertien stationnaire du second ordre $(Y_i)_{i\in\mathbb{Z}}$ Markovien. Afin de considérer des fonctions 'lisses', on se place sous l'hypothèse que l'espace Hilbertien est l'espace de Sobolev W^d des fonctions définies sur $[t_1, t_p]$ et telles que

$$\left\{f, f', ..., f^{(d-1)} \text{ sont absolument continues}; f^{(d)} \in L^2([t_1, t_p])\right\}$$

Par la suite on considèrera le cas particulier $d = 2$, très employé en pratique.

Une méthode classique permettant d'approximer ce type de fonction par minimisation d'une norme de W^2 est l'interpolation par des fonctions splines. Naturellement, si les hypothèses de régularité précédentes ne s'appliquent pas à l'objet d'étude, il peut être plus judicieux d'employer d'autres type d'interpolations (telles que linéaire ou par ondelettes).

Bosq [BOS83] propose de réaliser la prédiction de tels processus via l'espérance conditionnelle $\rho(y) = \mathbb{E}[Y_i \,|Y_{i-1} = y]$. L'opérateur ρ n'étant pas nécessairement linéaire en y, on peut l'estimer au moyen d'une régression non paramétrique par noyau.

En pratique, partant des observations discrètes $\{y_i(t_j) = x_{(i-1)p+j}\ ;\ i = 1, \ldots, n\ ;\ j = 1, \ldots, p\}$, on approxime les courbes y_i par les interpolations $\tilde{y}_i$ par splines obtenues suivant

$$\tilde{y}_i = \arg\min \left\|D^2 \tilde{y}_i\right\|_{L^2}^2 \text{ sous la contrainte } \tilde{y}_i(t_j) = y_i(t_j) \text{ pour } j = 1, \ldots, p$$

où D est l'opérateur de différenciation et $\|.\|_{L^2}$ est la norme de $L^2([t_1, t_p])$. Dans ce cas, les fonctions splines employées sont des polynômes de degré 3 par morceaux.

On réalise alors l'approximation de ρ par l'estimateur à noyau

$$\hat{\rho}_{h_n}(y) = \frac{\sum_{i=1}^{n-1} \tilde{y}_{i+1} \cdot K\left(\frac{\|\tilde{y}_i - y\|_{L^2}}{h_n}\right)}{\sum_{i=1}^{n-1} K\left(\frac{\|\tilde{y}_i - y\|_{L^2}}{h_n}\right)} \tag{2}$$

qui fournit la prédiction

$$\hat{y}_{n+1} = \hat{\rho}_{h_n}(y_n)$$

La valeur de h_n est obtenue par validation croisée

$$h_n = \arg\min CV(h)$$

avec

$$CV(h) = \sum_{k=n-m-r}^{n-m-1} \|\hat{\rho}_{h,n-r-m}(y_k) - \bar{y}_{k+1}\|_{L^2}^2$$

où $\hat{\rho}_{h,n-m-r}$ est l'estimation obtenue à partir des observations $\{y_1, \ldots, y_{n-m-r}\}$. Les m dernières courbes sont gardées afin de tester la qualité de la prédiction.

2.3 Prédiction de modèles autorégressifs Hilbertiens

Les processus autorégressifs Hilbertiens d'ordre 1 (ARH(1)) sont des processus Markoviens particuliers introduits par Bosq [BOS91].

On suppose ici que $(Y_i)_{i\in\mathbb{Z}}$ est un ARH(1) à valeurs dans l'espace de Hilbert H, il est défini par

$$\forall i \in \mathbb{Z}, \quad Y_i = \rho Y_{i-1} + \varepsilon_i$$

avec $\mathbb{E}Y_i = 0$ et $\mathbb{E}\|Y_i\|_H^2 < +\infty$. L'opérateur d'autocorrélation ρ est supposé compact et de norme strictement inférieure à 1. Les erreurs $(\varepsilon_i)_{i\in\mathbb{Z}}$ à valeurs dans H sont supposées de moyenne nulle, indépendantes, identiquement distribuées telles que $\mathbb{E}\|\varepsilon_i\|_H^2 = \sigma^2 < +\infty$.

Afin de réaliser la prédiction d'un ARH(1), on cherche à estimer ρ et les méthodes proposées rendent nécessaire l'estimation et l'inversion de l'opérateur Hilbertien de covariance défini ci-dessous. La difficulté de la démarche se situe dans cette inversion car l'opérateur de covariance est a priori non borné. Les différentes méthodes d'estimations de prédicteurs résolvent ce problème en projettant le processus dans un sous-espace permettant le calcul de l'inverse. Bosq [BOS91] propose ainsi de réduire l'espace d'estimation à celui associé aux valeurs propres les plus fortes de l'opérateur. La méthode que nous présentons ici, proposée par Besse et Cardot [BC96], anticipe cette réduction de dimension de l'espace d'estimation.

Notons par $\langle .,. \rangle_H$ le produit scalaire de H et notons $x \otimes y$ l'opérateur de rang 1 tel que

$$\forall (x, y, z) \in H^3, \quad [x \otimes y](z) = \langle x, z \rangle_H \, y$$

Soient $\Gamma = \mathbb{E}(Y_i \otimes Y_i)$ et $\Delta = \mathbb{E}(Y_i \otimes Y_{i+1})$ les opérateurs de covariance et de covariance retardée du processus ; on a la propriété suivante

$$\rho\Gamma = \Delta$$

qui permet d'estimer ρ, en inversant Γ, à partir des versions empiriques de Γ et Δ

$$\hat{\Gamma} = \frac{1}{n}\sum_{i=1}^{n} y_i \otimes y_i \text{ et } \hat{\Delta} = \frac{1}{n-1}\sum_{i=1}^{n-1} y_i \otimes y_{i+1}$$

De même que pour le modèle précédent, les observations des $(Y_i)_{i\in\mathbb{Z}}$ étant discrètes, il faut approximer les courbes. Comme par ailleurs, l'inversion de Γ demande une projection dans un sous-espace, Besse et Cardot [BC96] proposent de réaliser l'approximation spline en contraignant les fonctions splines en terme de lissage (paramètre l) et de rang (paramètre q), ce qui revient à projeter dans un sous-espace de dimension q en assurant une approximation optimale par fonctions splines. Pour cela, on effectue la minimisation suivante

$$\min_{\hat{y}_i \in H_q} \left[\frac{1}{n}\sum_{i=1}^{n} \left(\frac{1}{p}\sum_{j=1}^{p} (y_i(t_j) - \hat{y}_i(t_j)) \right)^2 + l \left\| D^2 \hat{y}_i \right\|_{L^2}^2 \right] \tag{3}$$

où H_q est un sous-espace de H de dimension q, à estimer.

3 Ecritures matricielles

3.1 Splines de lissage

La première étape d'estimation fonctionnelle consiste à associer à nos observations discrètes des fonctions dans un espace continu. Nous avons choisi ici de travailler dans l'espace de Sobolev W^2 avec une représentation par des fonctions splines dans un sous-espace S_p de dimension p, plutôt qu'une approximation linéaire ou par ondelettes, pour la qualité d'approximation de celles-ci. La base de fonctions splines choisie est une base de noyaux reproduisants (cf. Wahba [WAH90]).

Soit $\mathbf{Y}$ la matrice des observations $y_i(t_j)$ de taille $(n \times p)$ de vecteurs lignes $\mathbf{y}_i$. On note $\mathbf{M}$ et $\mathbf{N}$ les matrices associées respectivement à la projection de la norme de $L^2([t_1, t_p])$ et à la semi-norme de W^2 dans S_p, c'est-à-dire telles que

$$\|\tilde{\mathbf{y}}_i\|_{L^2}^2 = \mathbf{y}_i' \mathbf{M} \mathbf{y}_i = \|\mathbf{y}_i\|_{\mathbf{M}}^2$$
$$\left\| D^2 \tilde{\mathbf{y}}_i \right\|_{L^2}^2 = \mathbf{y}_i' \mathbf{N} \mathbf{y}_i = \|\mathbf{y}_i\|_{\mathbf{N}}^2$$

où $\tilde{\mathbf{y}}_i$ est l'approximation spline de $\mathbf{y}_i$ dans S_p obtenues par une minimisation similaire à celle de la formule (3).

Notons $\mathbf{A}(l)$ la matrice de lissage définie par

$$\mathbf{A}(l) = (\mathbf{M} + l\mathbf{N})^{-1}$$

3.2 Régression fonctionnelle par noyau

Dans le cadre de l'interpolation spline (qui est un lissage spline avec $l = 0$), la formule (2) du prédicteur par régression fonctionnelle s'écrit

$$\hat{\rho}_{h_n}(\mathbf{y}) = \frac{\sum_{i=1}^{n-1} \mathbf{y}_{i+1} \cdot K\left(\frac{(\mathbf{y}_i - \mathbf{y})'\mathbf{M}(\mathbf{y}_i - \mathbf{y})}{h_n}\right)}{\sum_{i=1}^{n-1} K\left(\frac{(\mathbf{y}_i - \mathbf{y})'\mathbf{M}(\mathbf{y}_i - \mathbf{y})}{h_n}\right)} \tag{4}$$

et sa prédiction est

$$\hat{\mathbf{y}}_{n+1} = \hat{\rho}_{h_n}(\mathbf{y}_n)$$

3.3 Estimation d'un ARH(1)

Soit $\mathbf{U}$ la matrice des trajectoires centrées de vecteurs lignes $\mathbf{u}_i$ $(\mathbf{u}_i = \mathbf{y}_i - \frac{1}{n}\sum_{j=1}^{n} \mathbf{y}_j)$, et $\mathbf{S}$ la matrice de covariance lissée donnée par

$$\mathbf{S} = \frac{1}{n}\mathbf{A}\,(l)^{1/2}\,\mathbf{U}'\mathbf{U}\mathbf{A}\,(l)^{1/2}$$

dont les éléments propres classés par ordre décroissant des valeurs propres sont notés $(\lambda_1, \mathbf{v}_1)$, ..., $(\lambda_p, \mathbf{v}_p)$.

Soit $\mathbf{V}_q$ la matrice orthogonale contenant les q vecteurs propres associés aux q premières valeurs propres de $\mathbf{S}$. Alors

$$\hat{\mathbf{y}}_i = \mathbf{A}\,(l)^{1/2}\,\mathbf{V}_q\mathbf{V}_q'\mathbf{A}\,(l)^{1/2}\,\mathbf{y}_i \quad , \quad i = 1, ..., n$$

Enfin, on estime les opérateurs de covariance par

$$\hat{\Gamma}_{q,l} = \frac{1}{n}\sum_{i=1}^{n} \hat{\mathbf{y}}_i\hat{\mathbf{y}}_i'\mathbf{M}$$

$$\hat{\Delta}_{q,l} = \frac{1}{n-1}\sum_{i=1}^{n-1} \hat{\mathbf{y}}_{i+1}\hat{\mathbf{y}}_i'\mathbf{M}$$

d'où

$$\hat{\rho}_{q,l} = \hat{\Delta}_{q,l}\hat{\Gamma}_{q,l}^{-1}$$

et la prévision s'écrit

$$\hat{\mathbf{y}}_{n+1} = \hat{\rho}_{q,l}\mathbf{y}_n + \mathbf{A}\,(l)\,\bar{\mathbf{y}}$$

4 Etude du taux moyen total TV

L'audience de la télévision est typiquement un processus à temps continu. Elle est mesurée à Médiamétrie à partir d'un panel d'individus, âgés de 4 ans et plus, répartis dans 2800 foyers à la fréquence d'une observation par seconde, ce qui nous incite à explorer des modélisations en temps continu sur celle-ci.

Pour chaque individu du panel, on sait, à la seconde près, s'il regarde la télévision et, dans ce cas, quelle chaîne il a choisi. On peut alors construire tout un ensemble de résultats à partir de cette information élémentaire, et en particulier le taux d'audience (par exemple des 4 ans et plus) qui correspond à la proportion des individus du panel regardant, pendant une période de temps considéré, la télévision ou une chaîne particulière suivant le cas.

Nous allons nous intéresser à la modélisation du taux moyen total TV (noté TTV) sur la période du début de soirée[2]. Le choix du TTV est lié au caractère univarié de la série obtenue, contrairement aux taux moyens des chaînes. L'intérêt du début de soirée est tant intrinsèque que dicté par le choix d'une période sur laquelle l'audience n'est pas constante au fil des jours (alors qu'en milieu de journée, de 15h à 17h par exemple, l'audience se répète quasi à l'identique d'une semaine sur l'autre). Notons toutefois que l'audience de la télévision est déterminée par un grand nombre de facteurs (exogènes) qui ne sont pas pris en compte dans les modèles présentés ici. On peut citer, par exemple, le type de programme sur chacune des chaînes, les conditions météorologiques, ou encore les évènements liés à l'actualité. Ce choix se justifie principalement par l'ignorance de certaines des valeurs de ces variables exogènes au moment de la réalisation des prédictions.

4.1 Présentation des données

Les caractéristiques des séries employées sont les suivantes :

$\Longrightarrow$ Objet d'étude : le taux moyen TTV.

$\Longrightarrow$ Nombre de séries : 7, une par jour nommé (les lundis, les mardis, ...).

$\Longrightarrow$ Tranche horaire : 20h45 - 22h45.

$\Longrightarrow$ Fréquence d'observation : toutes les 2 minutes (soit 61 observations par jour).

$\Longrightarrow$ Période d'étude : du 01/10/1998 au 31/05/2000 (88 semaines).

$\Longrightarrow$ Période d'apprentissage : du 01/10/1998 au 12/04/2000 (80 semaines).

$\Longrightarrow$ Période de validation croisée : du 24/02/2000 au 12/04/2000 (7 semaines).

$\Longrightarrow$ Période de test : du 13/04/2000 au 31/05/2000 (8 semaines).

[2]Partie de la grille de programme recouvrant "l'émission principale" de la soirée, et correspondant à l'heure de la plus forte écoute de la télévision.

Le choix fait ici est de séparer les audiences par jour nommé en raison du comportement fortement différencié des audiences suivant le jour de la semaine. Cette caractéristique de l'audience découle principalement de l'offre des programmes télévisés qui est très différente d'un jour nommé à l'autre. Par contre, étant relativement stable quand on étudie une série pour un jour nommé donné, elle permet de vérifier des hypothèses de stationnarité.

La fréquence des observations retenue a été choisie de manière à garder un maximum d'information sur l'audience tout en réduisant les problèmes d'implémentation informatique. En effet, si le nombre de points formant chaque courbe est trop important, l'estimation SARIMA et ARH(1) pose parfois problèmes (telle que la non convergence de l'algorithme de maximisation de la vraisemblance ou la manipulation de matrices numériquement non inversibles). Nous avons donc préféré travailler avec une observation toutes les 2 minutes obtenue comme la moyenne des secondes qui les composent. Ce choix effectue un léger lissage qui fait disparaître une partie du bruit mais correspond à des audiences très proches en pratique.

A partir de ces séries, nous nous intéresserons, naturellement, à la prédiction des courbes d'audience du taux moyen TTV, mais nous regarderons aussi les prédictions obtenues sur la moyenne journalière de cette audience que nous noterons Txm. Cette dernière série présente en effet un intérêt plus important concernant les applications possibles de prévisions d'audience.

4.2 Méthodologie

Pour chacune des 7 séries, on réalise les 6 modélisations ci-dessous, les 4 premières modélisant les courbes d'audience et les 2 dernières le Txm.

1. Un modèle SARIMA de saisonnalité égale à la longueur d'une courbe (61 points) permettant de prédire à une semaine. Les modèles obtenus sont généralement du type $(1,1,0)\,(0,1,0)_{61}$ (voir [GM95] pour plus de détails).
2. Une régression scalaire par noyau obtenu par la formule (1) avec un nombre de retards r valant 61.
3. Une régression par noyau fonctionnel obtenue par la formule (4) en considérant une courbe par jour.
4. Une modélisation par un modèle ARH(1) obtenue par splines de lissage. La dimension du sous-espace de projection ainsi que le paramètre de lissage sont obtenus par validation croisée.
5. Un modèle SARIMA du Txm généralement de la forme $(1,1,0)$.
6. Une régression par noyau du Txm obtenue par la formule (1) avec un nombre de retards r valant 1.

En outre, on comparera les résultats obtenus avec le modèle très simple de persistance qui consiste à prédire l'audience de la semaine suivante par celle de la semaine précédente (c'est-à-dire $\hat{Y}_{i+1} = Y_i$).

	lundi	mardi	mercredi	jeudi	vendredi	samedi	dimanche
Noyau	11,4	10,5	10,3	12,3	7,15	6,48	6,7
Noyau fonctionnel	5,4	37,1	13,4	5,96	30,7	11,4	34,2
ARH(1) l=	1058	0,267	89,3	181	$6{,}38\ 10^{-06}$	18,5	2280
ARH(1) q=	2	2	2	2	2	1	2
Noyau Txm	1,53	2,09	1,78	3,39	2,33	1,26	2,43

Tableau 1 – Valeur des paramètres des modèles obtenus par validation croisée.

On trouvera dans le tableau 1 les paramètres obtenus par validation croisée pour les modèles de régression par noyau (paramètre h) et de lissage spline (paramètres q et l) pour la modélisation ARH(1). On remarque que si les paramètres de lissage sont relativement stables d'une série à l'autre pour les régressions scalaires par noyau, la régression fonctionnelle par noyau, dans une moindre mesure, et surtout la modélisation ARH(1) voient de très fortes variations dans ces paramètres. Notons toutefois que le paramètre le plus sensible d'un lissage spline avec contrainte de rang est la dimension q du sous-espace dans lequel sont projetées les courbes. On constate que celui-ci est constant à 2 à l'exception de la série du samedi pour laquelle il vaut 1.

4.3 Comparaison

La période de test des 8 dernières semaines est employée afin de quantifier les résultats des différents modèles. Les indicateurs sont la moyenne des erreurs au carré (MEC) et la moyenne des erreurs absolues relatives (MEAR) pour les courbes d'audience et pour le Txm (MEC2 et MEAR2) définis ci-après, où $\bar{Y}$ désigne la moyenne de $\mathbf{Y}$:

$$MEC = \frac{1}{\tau}\sum_{t=1}^{\tau}\left(\hat{\mathbf{Y}}_{T+t|T} - \mathbf{Y}_{T+t}\right)^2 \quad MEAR = \frac{1}{\tau}\sum_{t=1}^{\tau}\frac{\left|\hat{\mathbf{Y}}_{T+t|T} - \mathbf{Y}_{T+t}\right|}{\left|\mathbf{Y}_{T+t}\right|}$$

$$MEC2 = \frac{1}{\tau}\sum_{t=1}^{\tau}\left(\widehat{\bar{Y}}_{T+t|T} - \bar{Y}_{T+t}\right)^2 \quad MEAR2 = \frac{1}{\tau}\sum_{t=1}^{\tau}\frac{\left|\widehat{\bar{Y}}_{T+t|T} - \bar{Y}_{T+t}\right|}{\left|\bar{Y}_{T+t}\right|}$$

4.3.1 Courbes d'audience

Le tableau 2 récapitule pour l'ensemble des modèles les erreurs obtenues (les erreurs les plus faibles y figure en gras). La première constatation est qu'uniformément les erreurs sont faibles (MEAR de l'ordre de 4 %) comme le confirme l'exemple de prévisions de la figure 2, mais on remarque aussi l'absence d'un modèle uniformément meilleur que les autres. Suivant le jour nommé les erreurs MEC et MEAR sont plus faibles pour le modèle

ARH(1) (les lundi, mardi, et dimanche), la régression scalaire par noyau (les mercredi et samedi) ou pour la régression fonctionnelle par noyau (les jeudi et vendredi). Les erreurs de prévision des modèles SARIMA sont par contre toujours les plus importantes. Ce n'est pas réellement une surprise puisque ces modèles ne sont pas adaptés à la prévision de courbes. En effet, l'hypothèse de stationnarité des courbes $\mathbf{Y}_i$ est insuffisante pour avoir la stationnarité du processus ponctuel $y_i(t_j)$.

Les autres modèles font mieux que le modèle de persistance avec un gain pouvant atteindre 25 %. Le gain obtenu n'est cependant pas constant par jour nommé ni par modèle, ni en considérant le meilleur modèle à chaque fois. On retrouve ici la forte différenciation par jour nommé des séries d'audience.

Si on regarde les exemples d'erreurs présentées sur la figure 2, on s'aperçoit qu'elles ont des formes similaires, liées au fait que les modèles réalisent des prédictions proches de celle du modèle de persistance. La principale cause d'erreurs de prévision semble être liée à des ruptures de programmes (fin d'un ou de plusieurs films, etc) qui, variant d'une semaine à l'autre, perturbent la série de courbes d'audience. Ainsi, le vendredi, qui est un jour où sont programmées des émissions avec des cases horaires[3] fixes, le modèle ne prévoit pas le niveau d'audience d'un programme alors qu'est diffusée une publicité, et ainsi les erreurs de prévision sont plus faibles et plus constantes sur la tranche horaire étudiée.

4.3.2 Taux moyen d'audience 20h45 - 22h45

Même si les modèles fonctionnels ne sont pas conçus dans ce but, on remarque, à l'étude des erreurs MEC2 et MEAR2 figurant elles aussi dans le tableau 2, que les prévisions du Txm obtenues par ces modèles sont généralement meilleures que celles obtenues par les modèles ponctuels (SARIMA Txm et noyau Txm). Il n'est clairement jamais équivalent de prédire une série puis de l'agréger, et de commencer par agréger la série avant de réaliser des prédictions. Dans le cas présent, il semblerait que le fait de travailler directement sur le Txm masque des propriétés de la série utiles à la réalisation de prédictions. Contrairement à la prédiction des courbes d'audience, il semble que les modèles de régression par noyau soient les plus performants dans ce domaine, à l'exception de la série des mardis où le modèle ARH(1) est nettement le plus performant.

La figure 3 confirme une bonne performance de modèles de lissage (noyau). On peut remarquer par ailleurs que les modèles ARH(1) ont une évolution plus variable que les modèles de lissage, ce qui provoque soit de bonnes prédictions (comme le mardi) soit de mauvaises prédictions (comme le jeudi).

[3]Une case horaire est une période déterminée par une heure de début et une durée.

	lundi	mardi	mercredi	jeudi	vendredi	samedi	dimanche
Persistance							
MEC	2,13	2,89	2,59	2,4	2,08	2,28	3,44
MEAR	3,97 %	4,82 %	5,18 %	5,06 %	4,78 %	4,9 %	6,55 %
MEC2	1,57	1,41	1,72	1,81	1,70	1,79	3,13
MEAR2	2,69 %	2,72 %	3,41 %	3,97 %	4,18 %	3,85 %	5,98 %
SARIMA							
MEC	3,66	4,6	4,53	5,24	4,44	2,83	4,69
MEAR	6,44 %	8,03 %	9,07 %	9,41 %	7,97 %	5,83 %	9,7 %
MEC2	1,67	1,67	2,17	1,51	3,96	2,49	3,9
MEAR2	3,71 %	3,13 %	4,54 %	2,95 %	5,72 %	4,54 %	7,38 %
Noyau							
MEC	1,99	2,63	**2,04**	2,31	2,12	**1,80**	3,37
MEAR	3,79 %	4,6 %	**4,25 %**	4,78 %	4,70 %	**4,01 %**	6,89 %
MEC2	**1,32**	1,48	1,24	1,44	1,75	**1,22**	2,92
MEAR2	**2,38 %**	2,73 %	2,75 %	2,77 %	4,09 %	**2,87 %**	6,1 %
Noyau fonctionnel							
MEC	2,07	2,47	2,12	**1,99**	**1,46**	1,95	3,35
MEAR	4,08 %	4,61 %	4,42 %	**4,05 %**	**3,14 %**	4,96 %	6,45 %
MEC2	1,55	1,61	1,24	1,39	**0,943**	1,56	2,71
MEAR2	2,79 %	3,23 %	2,72 %	2,61 %	**2,08 %**	3,99 %	**5,04 %**
ARH(1)							
MEC	**1,91**	**2,18**	2,22	2,15	1,87	1,89	**3,02**
MEAR	**3,73 %**	**3,76 %**	4,53 %	4,33 %	4,02 %	4,36 %	**6,03 %**
MEC2	1,36	**1,03**	1,45	1,67	1,55	1,53	**2,60**
MEAR2	2,44 %	**1,91 %**	2,93 %	3,09 %	3,37 %	3,25 %	5,26 %
SARIMA Txm							
MEC2	1,66	1,56	1,62	1,74	1,46	1,45	3,22
MEAR2	2,82 %	3,04 %	3,28 %	3,78 %	3,48 %	2,95 %	6,62 %
Noyau Txm							
MEC2	1,48	1,59	**1,22**	**1,29**	1,05	1,78	2,73
MEAR2	2,76 %	3,23 %	**2,71 %**	**2,43 %**	2,2 %	4,56 %	5,19 %

Tableau 2 - Comparaison des erreurs des différents modèles (les meilleurs résultats sont en gras).

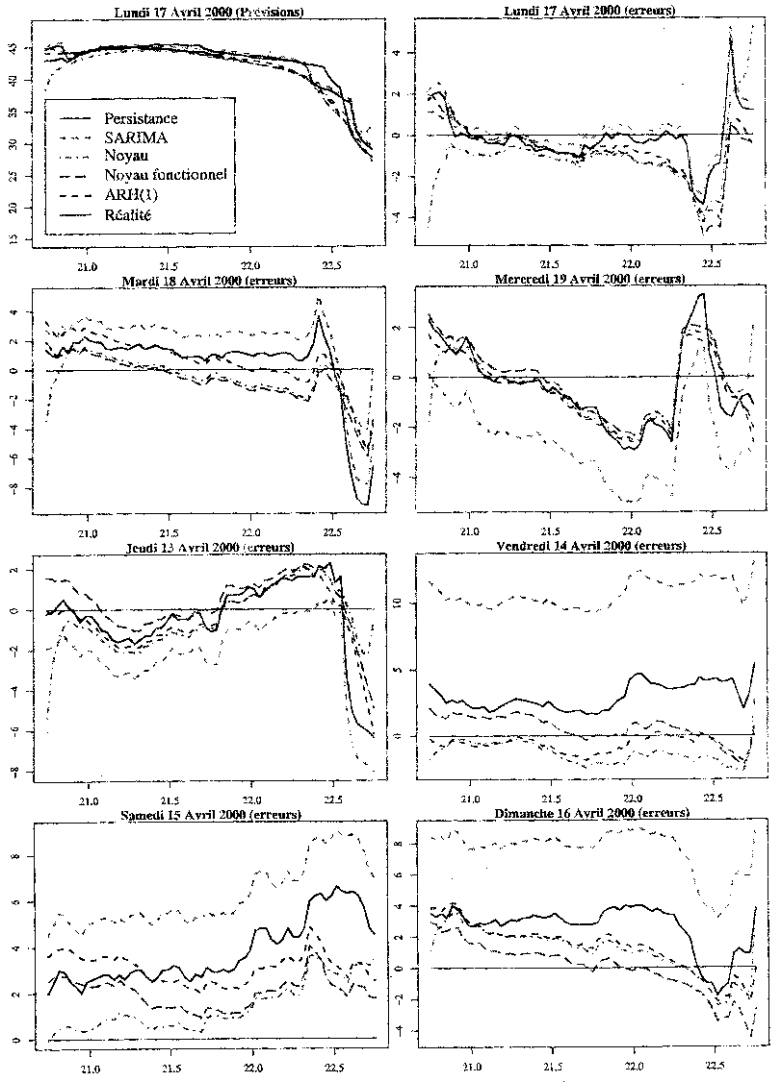

Figure 2 – Exemples de prévisions ($1^{ère}$ vignette) et d'erreurs obtenues au moyen des différents modèles, par jour nommé.

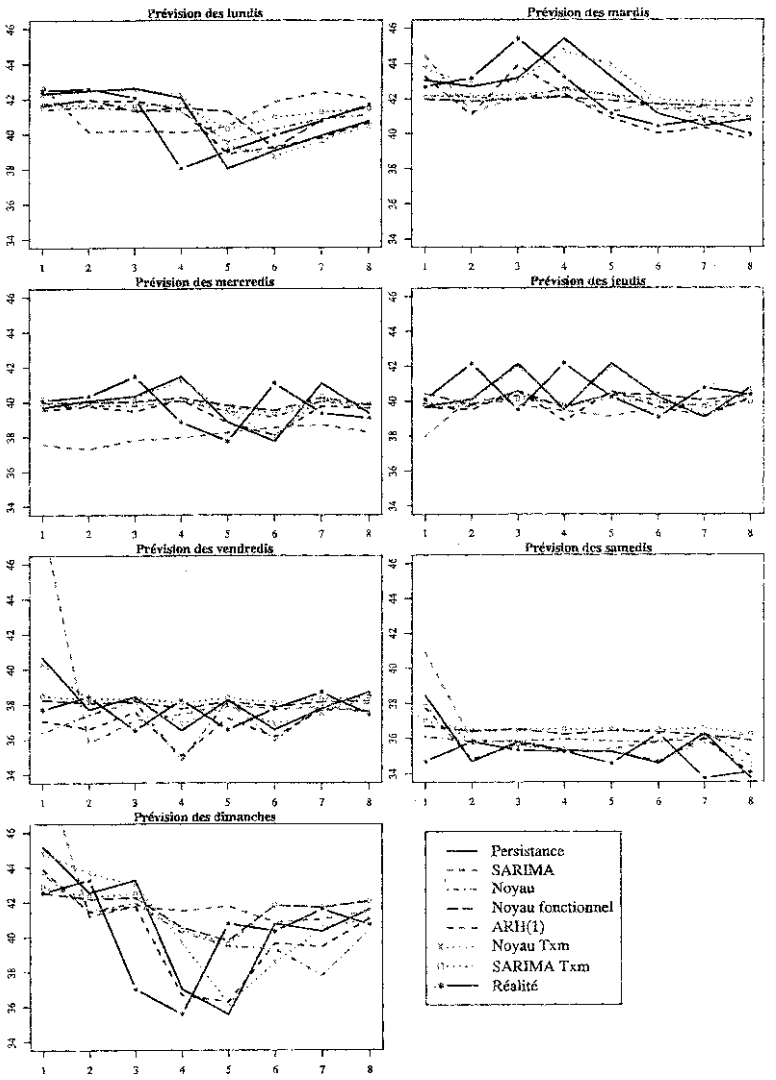

Figure 3 Prévisions du taux moyen 20h45 - 22h45 obtenues au moyen des différents modèles, par jour nommé.

5 Conclusion

Les modèles à temps continu montrent ici leur capacité à fournir des prévisions de courbes d'audience autant que de taux moyens avec des erreurs inférieures ou équivalentes à celles de modèles plus classiques.

La régression fonctionnelle par noyau fournit ainsi un bon prédicteur non paramétrique tout en étant simple à mettre en œuvre et rapide à calculer avec les ordinateurs actuels. Les modèles ARH(1) permettent de prendre en compte des contraintes de régularité des processus et fournissent de bons résultats. Leur mise en œuvre n'est, par contre, pas toujours aisée, en particulier le calcul des matrices de lissage qui s'effectue avant la projection sur un sous-espace.

Dans notre exemple, la réduction de dimension, menée par validation croisée, nous apporte une information supplémentaire sur nos données. En effet, le TTV étant obtenu par la somme des audiences des différentes chaînes, il aurait été compréhensible que la dimension du sous-espace spline fût 6 ou 5, ce qui aurait correspondu à l'influence de chacune des chaînes sur l'audience totale de la télévision. Or, cette dimension est généralement de 2, ce qui montre un comportement du téléspectateur divisé en 2 composantes, dont on peut se demander si elles correspondent aux comportements de 2 chaînes ? Tout du moins, cela suggère l'idée que pour réaliser des prévisions d'audience par chaîne, il est possible d'employer des modèles en 2 étapes : d'abord on effectue la prévision de l'audience télévisuelle, puis on estime la part d'audience de chaque chaîne.

Références

[BC96] P. BESSE et H. CARDOT. Approximation spline de la prévision d'un processus fonctionnel autorégressif d'ordre 1. *Revue Canadienne de Statistique*, 24 : 467–487, 1996.

[BCF97] P. BESSE, H. CARDOT, et F. FERRATY. Simultaneaous nonparametric regressions of unbalanced longitudinal data. *Computational Statistics and Data Analysis*, 24 : 255–270, 1997.

[BD91] P.J. BROCKWELL et R.A. DAVIS. *Times Series : theory and methods.* Springer-Verlag, 1991.

[BER90] A.R. BERGSTRÖM. *Continuous time econometric modelling.* Oxford, 1990.

[BES94] P. BESSE. Simultaneous non-parametric regressions. Dans *Publication Du Laboratoire de Statistique et Probabilités de Toulouse*, 1994.

[BJ76] G. BOX et G. JENKINS. *Time series analysis.* Holden Day, 1976.

[BOS83] D. BOSQ. *Non Parametric Prediction in Stationary Processes*, volume 16 de *Lectures Notes in Statistics.* Berlin-Heidelberg-New York, 1983. pages 69-84.

[BOS91] D. BOSQ. *Nonparametric Functional Estimation and Related Topics*, chapitre Modelization, Non-Parametric Estimation and Prediction for Continuous Time Processes, pages 509–529. Roussas, G., 1991.

[BOS98] D. BOSQ. *Nonparametric Statistics for Stochastic Processes, Estimation and Prediction*, volume 110 de *Lecture Notes in Statistics.* Springer-Verlag, New York, 2 édition, 1998.

[BOS00] D. BOSQ. *Linear Processes in Function Spaces : Theory and Applications*, volume 149 de *Lecture Notes in Statistics.* Springer-Verlag, New York, 2000.

[BR86] P. BESSE et J. RAMSAY. Principal component of sampled curves. *Psychometrika*, 51 : 285–311, 1986.

[CAR97] H. CARDOT. *Contribution à l'Estimation et à la Prévision Statistique de Données Fonctionnelles.* Thèse, Université de Toulouse 3, 1997.

[CD93] M. CARBON et M. DELECROIX. Nonparametric forecasting in time series : A computational point of view. *Applied Stochastic Models and Data Analysis*, 9(3) : 215–229, 1993.

[DEV74] J. DEVILLE. *Méthodes Statistiques et Numériques de l'Analyse Harmonique*, volume 15 de *Annales de l'INSEE.* INSEE, 1974.

[EPA69] V.A. EPANECHNIKOV. Nonparametric estimation of a multidimensional probability density. *Theory of probability and applications*, 14 : 153–158, 1969.

[GM95] C. GOURIEROUX et A. MONFORT. *Séries Temporelles et Modèles Dynamiques.* Economica, Paris, 2 édition, 1995.

[GS94] P. GREEN et B. SILVERMAN. *Nonparametric Regression and Generalized Linear Models.* Chapman and Hall, 1994.

[GUE94] D. GUEGAN. *Séries chronologiques non linéaires à temps discret.* Economica, 1994.

[HAR90] W. HARDLE. *Applied nonparametric regression.* Cambridge University Press, 1990.

[IG96] IHAKA et GENTLEMAN. R : A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, 5 : 299–314, 1996.

[MER96] F. MERLEVEDE. *Processus Linéaires Hilbertiens : Inversibilité, Théorèmes Limites, Estimation et Prévision.* Thèse, Université Paris 6, 1996.

[MOU95] T. MOURID. *Contribution à la statistique des processus autorégressifs à temps continu.* Thèse de doctorat d'état, Université Paris 6, 1995.

[PUM92] B. PUMO. *Estimation et Prévision de Processus Autoregressifs Fonctionnels. Applications aux Processus à Temps Continu.* Thèse, Université Paris 6, 1992.

[RD91] J. RAMSAY et C. DALZELL. Some tools for functional data analysis (with discussion). *Journal of the Royal Statistical Society (B)*, 53(3) : 539–572, 1991.

[RS97] J. RAMSAY et B. SILVERMAN. *Functional Data Analysis.* Springer-Verlag, 1997.

[SIL96] B. SILVERMAN. Smoothed functional principal components analysis by choice of norm. *The Annals of Statistics*, 24 : 1-24, 1996.

[VR00] W.N. VENABLES et B.D. RIPLEY. *S Programming.* Springer-Verlag, 2000.

[WAH90] G. WAHBA. *Spline Models for Observational Data.* SIAM, 1990.