

REVUE DE STATISTIQUE APPLIQUÉE

JEAN-BAPTISTE DURAND

LAURENT BOZZI

GILLES CELEUX

CHRISTIAN DERQUENNE

Analyse de courbes de consommation électrique par chaînes de Markov cachées

Revue de statistique appliquée, tome 52, n° 4 (2004), p. 71-91

http://www.numdam.org/item?id=RSA_2004__52_4_71_0

© Société française de statistique, 2004, tous droits réservés.

L'accès aux archives de la revue « *Revue de statistique appliquée* » (<http://www.sfds.asso.fr/publicat/rsa.htm>) implique l'accord avec les conditions générales d'utilisation (<http://www.numdam.org/conditions>). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM

*Article numérisé dans le cadre du programme
Numérisation de documents anciens mathématiques*

<http://www.numdam.org/>

ANALYSE DE COURBES DE CONSOMMATION ÉLECTRIQUE PAR CHAÎNES DE MARKOV CACHÉES

Jean-Baptiste DURAND*, Laurent BOZZI**, Gilles CELEUX***,
Christian DERQUENNE**

* INRIA Rhône-Alpes, Montbonnot, France

** EDF-R&D, Clamart, France

*** INRIA Futurs, Orsay, France

RÉSUMÉ

Nous présentons une méthodologie basée sur les chaînes de Markov cachées pour la modélisation et l'analyse statistique de courbes de consommation électrique. Suite à une analyse de variance qui permet d'estimer l'effet sur la log-consommation de facteurs contrôlés (mois, jour, heure, type de contrat et puissance maximale souscrite), nous modélisons les résidus par une chaîne de Markov cachée. Ensuite, les états cachés sont restaurés puis interprétés grâce à un tableau d'intensité les mettant en relation avec la consommation de différents appareils électriques (ou usages) lorsque celle-ci est disponible. Nous montrons comment cette approche, par l'usage conjoint du tableau d'intensité et des états cachés restaurés, permet une estimation de la consommation de chaque usage dans le cas où elle est inconnue. La prise en compte d'informations *a priori* sur la consommation rend cette estimation plus réaliste. Enfin, nous abordons le problème de la sélection de modèles. Les critères classiques conduisent à des modèles complexes qui rendent très difficile l'interprétation des états cachés par leur mise en correspondance avec les usages. Nous présentons et commentons des méthodes de sélection alternatives qui tiennent compte de l'objectif de mise en relation des états cachés et des classes de consommation.

Mots-clés : Chaînes de Markov cachées, Consommation électrique, Déconvolution, Sélection de modèles.

ABSTRACT

We develop an approach based on hidden Markov chains for the modelling and statistical analysis of electric consumption curves. A preliminary analysis of variance leads to the estimation of effects of fixed factors (month, day, hour, type of contract and maximal power) on the log-consumption. Then the residuals are modelled with a hidden Markov chain. The hidden states are restored and interpreted using an intensity table relating the restored hidden states to the consumption of various electrical devices when available. We show how the consumption induced by each device can be estimated, from the restored states and the intensity table. The estimates are made more realistic by taking advantage of prior information on the consumption. Finally, we deal with model selection issues. The usual criteria select too complex models, and thus do not help relating hidden state to the device consumption in a clear way. We present

alternative selection methods taking into account the purpose of linking hidden states and consumption groups.

Keywords : *Hidden Markov Chains, Electric Consumption, Deconvolution, Model Selection.*

1. Introduction

Cette étude, réalisée dans le cadre d'une collaboration entre le département «clientèle» de EDF-R&D Clamart et l'INRIA Rhône-Alpes, a pour objet l'analyse statistique de courbes de consommation électrique journalières. Il s'agit de déterminer les facteurs expliquant les variations de la consommation, d'estimer la part de la consommation due aux différents appareils électriques (ou *usages*) et de comparer les usages entre eux. Le but est également de disposer de modèles stochastiques adaptés à la consommation à des fins d'analyse, de simulation et de prévision.

L'approche retenue pour la modélisation repose sur l'hypothèse que la consommation électrique d'un foyer s'explique essentiellement par l'existence de périodes homogènes du point de vue de la consommation électrique et qu'à l'intérieur de chacune de ces périodes, les fluctuations de la consommation sont dues au hasard. Ce principe conduit à considérer un modèle basé sur les chaînes de Markov cachées.

Au préalable, les variations de la consommation électrique s'expliquent en partie grâce à une analyse de variance, qui permet de mettre en évidence l'effet de facteurs contrôlés : mois, heure, type de contrat et puissance maximale souscrite. Ce sont alors les résidus de cette analyse de variance, interprétés comme l'information non prise en compte par ce modèle, qui sont modélisés par une chaîne de Markov cachée.

Les états cachés, dont la succession entraîne l'existence de périodes homogènes de consommation, sont restaurés puis interprétés grâce à un tableau d'intensité les mettant en relation avec la consommation de différents appareils électriques (ou *usages*) lorsque celle-ci est disponible. Un objectif de cette étude est de permettre l'estimation de la consommation de chaque usage dans le cas où elle est inconnue. Nous donnons une méthode pour estimer ces consommations qui utilise le tableau d'intensité croisant les états cachés restaurés et les usages. Mais les courbes de consommation estimées manquent de réalisme, en particulier à cause de l'absence, pour certains foyers, d'appareils électriques connus, et de la non prise en compte de l'ordre de grandeur des usages. Nous montrons comment la prise en compte d'informations *a priori* sur la consommation permet de rendre cette estimation plus réaliste.

Enfin, nous abordons le problème de la sélection de modèles. Les critères classiques conduisent à des modèles complexes qui rendent très difficile l'interprétation des états cachés par leur mise en correspondance avec les usages. C'est pourquoi nous présentons une méthode de sélection privilégiant la facilité d'interprétation du modèle (*i.e.* des états cachés), quantifiée par la mesure de l'écart à l'indépendance entre usages et états cachés.

La problématique de l'étude et la nature des données sont précisées dans la partie 2. L'analyse de variance préalable à la mise en œuvre de ces modèles est présentée en partie 3. La partie 4 est dédiée à la modélisation par chaînes de Markov

cachées. Les états cachés du modèle sont interprétés par une analyse factorielle des correspondances décrite en partie 5. La mise en correspondance des usages et des états cachés, à la base de la méthode d'estimation des usages, est présentée en partie 6. La sélection de modèles est traitée en partie 7; elle est suivie d'une discussion et de remarques de conclusion en partie 8.

2. Données et problématique

Il s'agit d'exploiter les résultats d'une campagne de mesure des principaux usages domestiques de la clientèle résidentielle d'EDF : le chauffage électrique (désigné par l'abréviation CHA), l'eau chaude sanitaire (ECS), l'éclairage halogène (H), l'éclairage non halogène (NH), les autres types d'éclairage (ECL), le lave-vaisselle (LV), le lave-linge (LL), le sèche-linge (SL), la cuisson classique (CUI), les équipements de télévision et chaîne hi-fi (TVH), les appareils de réfrigération (FRD), les autres appareils électriques (AUT). La consommation totale du logement, tous usages et appareils confondus, est également mesurée. Cette campagne concerne 100 clients dont les usages sont mesurés sur une année entière, toutes les dix minutes.

Pour des raisons de confidentialité des données et pour tester la robustesse de notre méthode, les courbes de charge fournies par EDF ne sont pas des mesures de la consommation réelle. Au logarithme des valeurs mesurées est ajouté un bruit gaussien centré, de variance égale à 10 % de celle estimée à partir des données ayant mêmes mois et heure. De plus, seules 856 courbes, choisies pour leur représentativité sur la base d'une classification des clients puis d'une stratification par la classe et le mois, sont sélectionnées parmi les 18 250 initialement disponibles. Nous disposons ainsi d'autant de courbes de charge, donnant la consommation totale journalière du logement (une mesure toutes les dix minutes). Ces courbes sont utilisées pour estimer les paramètres des modèles d'analyse de variance et de chaîne de Markov cachée (parties 3 et 4). D'autre part, nous disposons également de 444 courbes, de même pas de temps que ci-dessus et bruitées suivant un procédé similaire, où la décomposition par usage est disponible. Ces courbes sont dédiées à l'estimation de la part de la consommation due à chaque usage (partie 6). De nombreux relevés de consommation sont manquants, aussi bien dans les courbes de consommation totale que dans celles correspondant aux usages.

Les apports attendus d'un modèle statistique pour l'analyse de ces courbes de consommation sont l'étude des variations de la consommation en fonction de différents facteurs contrôlés, parmi le mois, le jour de la semaine ou du mois, l'heure, le type de tarif (normal, nuit ou EJP) et la puissance souscrite. Pour le tarif EJP (Effacement Jour de Pointe), l'année est partagée en 22 jours où le prix du kWh est élevé, et 343 jours normaux – où ce prix est moins cher qu'en tarif de base. Nous désirons aussi réaliser l'estimation des usages quand seule la consommation totale est connue, de même que la prévision de la consommation future de logements, et la simulation de scénarii de consommation électrique, d'où l'intérêt d'un modèle stochastique.

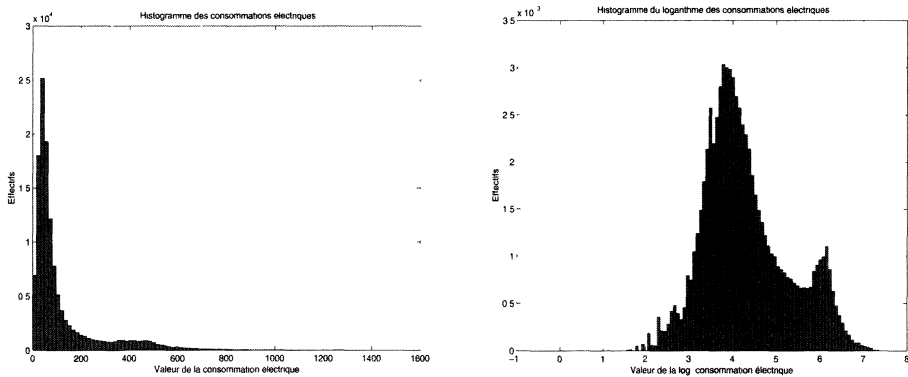
Dans ce qui suit, nous nous intéressons principalement à l'étude de facteurs contrôlés ayant un effet sur la consommation électrique, à la modélisation de cette consommation par chaînes de Markov cachées et à la décomposition de la

consommation totale suivant les usages. L'utilisation des chaînes de Markov cachées présente, entre autres avantages, l'intérêt de rendre simple la simulation et la prévision de la consommation électrique une fois les paramètres du modèle estimés.

3. Prétraitement des données

Préalablement, nous effectuons une analyse de variance en fonction de différents facteurs contrôlés. Effectivement, un modèle à structure cachée vise à mettre en évidence des effets non mesurables directement à partir des covariables disponibles. Il est donc important de travailler orthogonalement aux effets contrôlés. Ainsi, la modélisation par chaînes de Markov cachées doit se faire sur les résidus d'un modèle validé d'analyse de variance. Notons que de la sorte, on consolide l'hypothèse de stationnarité du processus observé (sa non stationnarité implique celle du processus caché).

Dans l'analyse de variance, les variables aléatoires sont supposées gaussiennes. L'histogramme de gauche sur la figure 1 représente la répartition de tous les relevés de consommation électrique globale $(G_t)_t$, tandis que celui de droite représente celle de leur logarithme. Nous constatons visuellement que l'hypothèse d'une loi gaussienne ou de mélange gaussien est nettement plus plausible pour les $(\log(G_t))_t$ que pour les $(G_t)_t$. Cependant, l'assimilation de la courbe des log-consommations à une loi gaussienne est imparfaite. On peut donc s'attendre à ce que les résidus de l'analyse de variance ne soient pas réductibles à du bruit et contiennent en fait de l'information que nous allons modéliser par une chaîne de Markov cachée.



Consommations électriques

Logarithme des consommations électriques

FIGURE 1

Histogramme des consommations électriques et leur logarithme

Les facteurs contrôlés disponibles sont les suivants : *type de contrat* (ou *tarif*) (c), *puissance souscrite* (p), *mois* (m), *heure du jour* (h), *jour du mois* et *jour de la semaine*. La partie concernant la sélection des facteurs et de leurs interactions dans l'analyse de variance est reportée dans la partie 7, dédiée à la sélection de modèles.

Nos conclusions sont que la prise en compte des effets *jour du mois* et *jour de la semaine* contribue peu à réduire la déviance par rapport aux autres facteurs contrôlés; ces effets sont donc ignorés. En outre, la prise en compte d'interactions entre facteurs autres que le mois et le tarif n'est pas pertinente non plus. De plus, les tarifs «nuit» et «EJP» sont regroupés dans une même modalité de l'effet c – l'autre modalité étant le tarif «normal».

En définitive, nous considérons le modèle suivant d'analyse de variance pour le logarithme de la consommation électrique globale $G_{m,h,p,c,t}$:

$$\log(G_{m,h,p,c,t}) = \alpha + \beta_m + \gamma_h + \delta_p + \eta_c + \zeta_{c,m} + \varepsilon_t, \quad (1)$$

le terme $\zeta_{c,m}$ représentant une interaction entre le *tarif* et le *mois*. Les estimateurs des coefficients associés aux facteurs *mois* et *heure* sont représentés figure 2. Par convention, pour chaque effet contrôlé, la valeur du paramètre pour la première modalité est nulle.

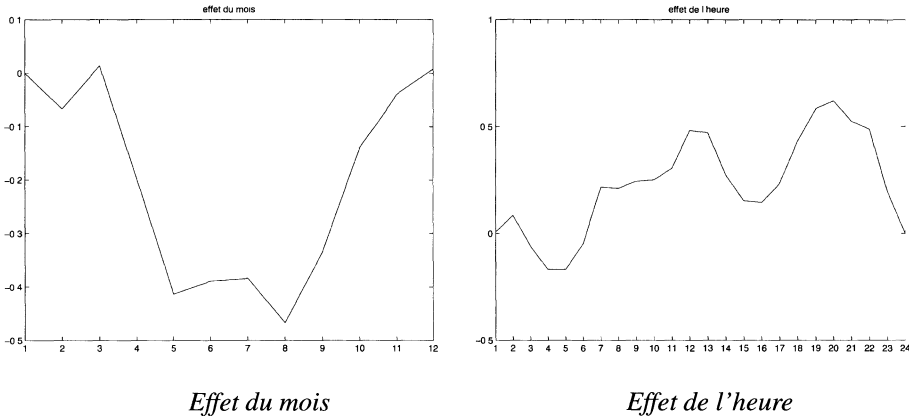


FIGURE 2
Estimateurs de paramètres de l'analyse de variance
pour les effets dus au mois et à l'heure

Nous proposons ensuite de modéliser le processus des $(\varepsilon_t)_t$ par une chaîne de Markov cachée, de sorte que par cohérence avec les notations usuelles utilisées dans la partie 4, le résidu ε_t est noté Y_t . Dans le modèle d'analyse de variance, les résidus $(\varepsilon_t)_t$ sont supposés indépendants et de loi normale de moyenne nulle et de variance σ^2 . L'idée présidant à notre tentative de modélisation est précisément que ces hypothèses d'indépendance et de normalité ne sont pas vérifiées dans le contexte de courbes de consommation électrique.

4. Modélisation

Nous précisons les hypothèses conduisant à considérer des modèles de Markov cachés puis rappelons les méthodes d'inférence spécifiques à ces modèles.

4.1. Définition des chaînes de Markov cachées et application à la modélisation de la consommation

L'idée à l'origine de la modélisation de la consommation électrique par des chaînes de Markov cachées est que la consommation électrique d'un logement s'explique essentiellement par l'existence de périodes homogènes et qu'à l'intérieur de chacune de ces périodes, les fluctuations de la consommation sont dues au hasard. Par exemple, pour un ménage, les périodes de comportement homogènes peuvent être interprétées en termes d'activité domestique induisant différents niveaux de consommation électrique : repos, lessive, préparation de repas, veillée avec utilisation de la télévision ou de la chaîne hi-fi, etc. Les transitions entre ces activités sont modélisées par une chaîne de Markov. Or nous ne disposons pas d'information sur l'activité des ménages, c'est pourquoi les valeurs de la chaîne de Markov sont cachées. Du fait que, dans une période homogène donnée, les fluctuations de la consommation sont dues au hasard uniquement, les consommations instantanées sont conditionnellement indépendantes sachant l'état caché et suivent une loi de type connu (par exemple une loi gaussienne).

Pour tout processus $(z_t)_{t \geq 1}$ et tout couple d'entiers (t_1, t_2) avec $t_1 \leq t_2$, nous désignons la séquence $(z_{t_1}, \dots, z_{t_2})$ par $z_{t_1}^{t_2}$. Formellement, le modèle est le suivant. La consommation électrique est décrite par deux processus : le processus observé (Y_1, \dots, Y_n) de la consommation électrique proprement dite, à valeurs dans \mathbb{R}^+ et noté \mathbf{Y}_1^n , et un processus caché (S_1, \dots, S_n) , à nombre fini d'états, à valeurs dans $\{1, \dots, K\}$ et noté \mathbf{S}_1^n , qui sont tels que :

- \mathbf{S}_1^n est une chaîne de Markov homogène, irréductible et stationnaire de matrice de transition P et de distribution stationnaire $\pi = (\pi_1, \dots, \pi_K)$ avec $\pi_j \geq 0$ et $\sum_{j=1}^K \pi_j = 1$;
- les Y_t sont indépendants conditionnellement aux S_t ;
- sachant $S_t = j$, Y_t suit une loi de densité f_{θ_j} appartenant à une famille paramétrée $(f_{\theta})_{\theta \in \Theta}$.

Il s'ensuit que la loi de \mathbf{Y}_1^n est gouvernée par une chaîne de Markov cachée. Les paramètres de ce modèle sont notés $\lambda = (\pi, P, \theta_1, \dots, \theta_K)$.

Notons que le choix de la famille $(f_{\theta})_{\theta \in \Theta}$ des lois d'émission, qui modélise la nature du bruit, est important. Nous choisirons une famille gaussienne, de sorte que $\theta = (\mu, \sigma^2)$ où μ représente la moyenne et σ^2 la variance. Ce choix est justifié dans la partie 3 par le fait que les résidus Y_t semblent suivre une loi de mélange gaussien plutôt qu'une loi gaussienne (voir figure 1), puis il est validé dans la sous-partie 5.2. Sous les hypothèses ci-dessus, pour chaque t compris entre 1 et n , le mélange définissant la loi marginale de Y_t a pour densité :

$$f_{Y_t}(y) = \sum_{j=1}^K \pi_j f_{\theta_j}(y). \quad (2)$$

4.2. Estimation des paramètres

Nous considérons, dans un premier temps, les 856 courbes de charge journalière de la consommation électrique totale, les usages n'étant pas disponibles. Nous disposons de plusieurs séquences de n relevés de consommation électrique pour l'estimation des paramètres. Ceux-ci sont estimés par maximum de vraisemblance, en utilisant l'algorithme EM pour les chaînes de Markov cachées dû à Baum *et al.* (1970). Cet algorithme itératif part d'une valeur initiale $\lambda^{(0)}$ des paramètres et crée une suite $(\lambda^{(m)})_{m \geq 0}$ qui converge vers la solution consistante des équations de vraisemblance, si $\lambda^{(0)}$ est proche de la solution optimale.

Chaque itération de l'algorithme EM consiste à :

- calculer la loi conditionnelle de chaque paire d'états cachés consécutifs, sachant les observations (*étape E*);
- réestimer les paramètres en utilisant ces probabilités de sorte à maximiser l'espérance conditionnelle des données complètes, constituées du processus observé Y_1^n et du processus caché S_1^n , sachant Y_1^n et la valeur courante des paramètres du modèle (*étape M*).

L'estimateur obtenu peut dépendre fortement de la valeur initiale $\lambda^{(0)}$, c'est pourquoi nous considérons trois valeurs initiales tirées au hasard et effectuons 20 itérations de l'algorithme pour chacune d'entre elles. La valeur ayant la vraisemblance maximale est utilisée comme valeur initiale d'une exécution supplémentaire de l'algorithme EM, stoppée lorsque 10 000 itérations sont effectuées ou lorsque la croissance relative de la log-vraisemblance passe en dessous du seuil 10^{-6} .

Les lois conditionnelles intervenant dans l'*étape E* sont calculées par une récursion *avant-arrière* inspirée de celle de Devijver (1985) mais prenant en compte l'existence d'observations Y_t manquantes. Cette récursion a été développée dans Celeux et Durand (2002) et est détaillée dans Durand (2003) dans le contexte de la sélection de modèles de Markov cachés par demi-échantillonnage.

Dans le cas de mesures réelles de la consommation dans des logements, on peut faire l'hypothèse que les consommations journalières de différents logements sont des réalisations de plusieurs modèles distincts. Cependant, nous disposons en l'occurrence de courbes bruitées obtenues à partir d'un sous-ensemble des courbes réelles, que nous supposons être des réalisations mutuellement indépendantes d'un même modèle. Nous disposons ainsi, pour l'estimation des paramètres, de 856 courbes de longueur 144 (courbes journalières où les mesures sont effectuées avec un pas de temps de dix minutes) supposées indépendantes.

5. Interprétation des états cachés

Une compréhension fine du modèle passe par l'interprétation des états cachés. D'une part, celle-ci permet d'associer un niveau de consommation moyen à chaque état. En outre, la restauration des états cachés permet d'obtenir des histogrammes reflétant la distribution empirique conditionnelle de la consommation dans un état

donné. Enfin, la mise en correspondance des usages et des états cachés permet leur interprétation en termes d'activité.

Dans ce qui suit, nous considérons un modèle de chaîne de Markov cachée à $K = 7$ états cachés. Ce choix est expliqué dans la partie 7, dédiée à la sélection de modèles.

5.1. Restauration des états cachés

Les états cachés $(\hat{s}_t)_t$ les plus probables sont calculés pour chaque courbe séparément par l'algorithme du Maximum A Posteriori, ou MAP (dit également *algorithme de Viterbi*, pour les chaînes de Markov cachées, voir Forney, 1973). Cet algorithme détermine les états cachés maximisant la probabilité conditionnelle de la chaîne cachée entière, sachant les observations. La figure 3 permet la comparaison, pour une courbe donnée (courbe numéro 1), de la consommation électrique globale $(c_t)_{1 \leq t \leq 144}$ (cadre supérieur), son logarithme $(\log(c_t))_{1 \leq t \leq 144}$ (cadre du milieu) et les résidus $(y_t)_{1 \leq t \leq 144}$, superposés aux états cachés restaurés par l'algorithme de Viterbi (cadre inférieur).

Les sept états cachés se distinguent par la valeur de la moyenne et de la variance des lois conditionnelles gaussiennes. Les moyennes estimées $(\hat{\mu}_j)_{1 \leq j \leq 7}$ sont données par l'équation (3). On peut renommer $(C1, \dots, C7)$ les indices des états cachés en ordonnant ces derniers par valeur décroissante de $\hat{\mu}_j$.

$$\begin{bmatrix} \hat{\mu}_1 \\ \hat{\mu}_2 \\ \hat{\mu}_3 \\ \hat{\mu}_4 \\ \hat{\mu}_5 \\ \hat{\mu}_6 \\ \hat{\mu}_7 \end{bmatrix} = \begin{bmatrix} \hat{\mu}_{C3} \\ \hat{\mu}_{C2} \\ \hat{\mu}_{C7} \\ \hat{\mu}_{C4} \\ \hat{\mu}_{C6} \\ \hat{\mu}_{C1} \\ \hat{\mu}_{C5} \end{bmatrix} \approx \begin{bmatrix} 0,4 \\ 1,0 \\ -1,4 \\ 0,1 \\ -0,7 \\ 1,8 \\ -0,3 \end{bmatrix} \quad (3)$$

La variance estimée est du même ordre de grandeur pour les différents états cachés, à savoir d'environ 0,065 sauf pour l'état 3 (aussi appelé C7 et correspondant aux consommations les plus faibles), où elle est trois fois plus élevée. Ainsi, les états cachés sont essentiellement liés aux différents niveaux de la log-consommation, corrigée des variations saisonnières et de celles dues au type de contrat.

5.2. Estimation de la densité marginale et des densités conditionnelles

Le nombre d'états cachés peut être vu comme le nombre de composants de la loi marginale de Y_t , puisqu'il s'agit d'une loi de mélange (voir équation (2)). Pour apprécier visuellement la qualité de l'estimation de cette densité marginale, nous superposons dans la figure 4 l'histogramme des valeurs observées $(y_t)_t$ avec la densité théorique en utilisant les paramètres estimés.

De plus, la restauration des états cachés permet de tracer l'histogramme des résidus y_t dont l'état le plus probable est $S_t = k$ (au sens du MAP). Nous superposons ensuite cet histogramme à la densité conditionnelle estimée de Y_t sachant $S_t = k$, en l'occurrence la loi $\mathcal{N}(\hat{\mu}_k, \hat{\sigma}_k^2)$ – voir figure 5. Nous constatons une bonne coïncidence

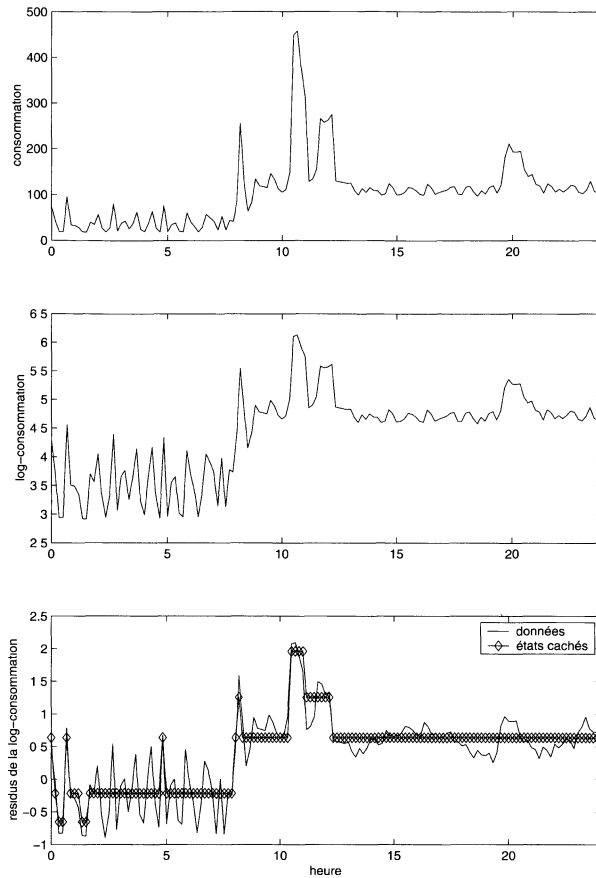


FIGURE 3

Consommation, log consommation, résidus et états cachés pour la courbe numéro 1. En haut, la consommation électrique globale; au milieu, la log-consommation; en bas, les résidus de l'analyse de variance et les états cachés restaurés

entre les deux courbes pour chacun des états cachés. Ceci tend à justifier l'hypothèse que les lois d'émission sont en effet log-normales, si l'on considère la consommation électrique (nous étudions en réalité son logarithme). D'autre part, l'interprétation des états cachés en est facilitée. En revanche, une non coïncidence n'apporte pas d'information sur la qualité du modèle, car l'algorithme de Viterbi utilise une procédure d'affectation par le principe du MAP qui ne respecte pas nécessairement la répartition des composants, notamment lorsqu'ils sont peu séparés.

5.3. Mise en correspondance des usages et des états cachés

Nous exposons maintenant comment tirer parti des états cachés restaurés $(\hat{s}_t)_t$ et de la connaissance des usages pour certaines courbes (en nombre $R = 444$), pour

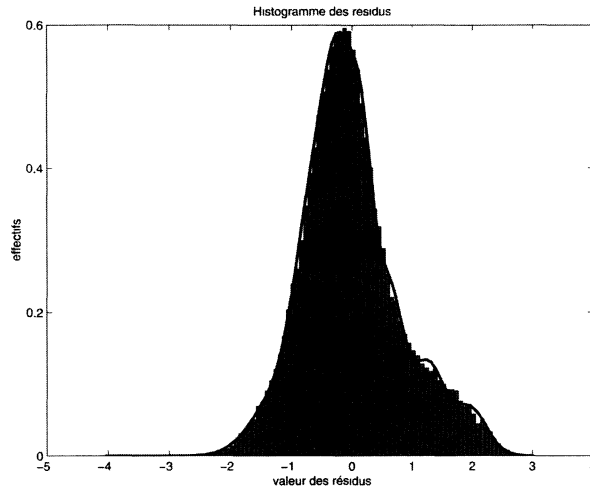


FIGURE 4
Densité marginale théorique et empirique des résidus

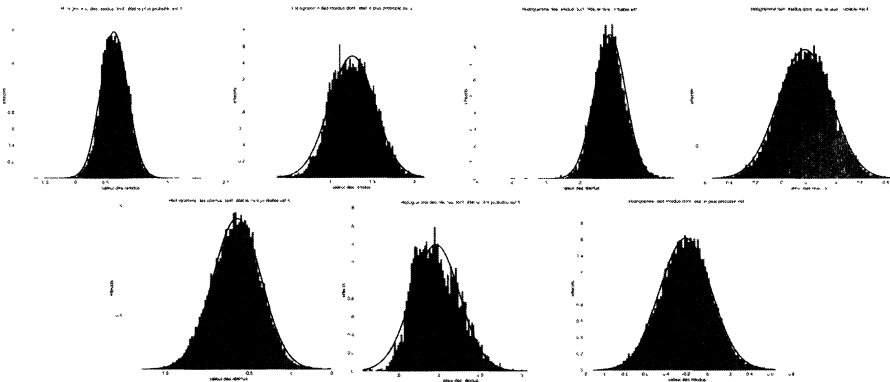


FIGURE 5
Histogrammes des résidus conditionnellement à l'état caché le plus probable

interpréter les états cachés en termes d'usages privilégiés. En effet, à chaque instant t , la mesure de la consommation électrique est affectée à l'un des sept états cachés s_t . Dans un premier temps, nous utilisons les données concernant les usages pour calculer la quantité suivante, pour chaque usage u et chaque valeur de l'état caché j :

$$n_{ju} = \sum_{r=1}^R \sum_{t=1}^{144} \text{conso}_u^r(t) \mathbb{1}_{\{s_t=j\}}$$

où $\text{conso}_u^r(t)$ représente la consommation à l'instant t due à l'usage u , pour la courbe de numéro r . La consommation globale pour l'usage u , toutes courbes confondues,

s'exprime donc par

$$\sum_j \sum_r \sum_t \text{conso}_u^r(t) \mathbb{1}_{\{\hat{s}_t=j\}}.$$

On obtient un tableau d'intensité $(n_{ju})_{j,u}$ (représenté tableau 1) dont les lignes j correspondent aux états et les colonnes u aux usages, et dont les valeurs dépendent de l'unité de puissance adoptée. Pour ce tableau, on peut calculer la statistique du Φ^2 , définie par

$$\frac{1}{N} \sum_j \sum_u \frac{\left(n_{ju} - \frac{n_j \cdot n_u}{N} \right)^2}{\frac{n_j \cdot n_u}{N}}, \text{ où } N = \sum_{j,u} n_{ju}.$$

Cette statistique du Φ^2 mesure l'écart à l'indépendance des douze usages et des K états; multipliée par N , elle serait la statistique du χ^2 si N était un effectif. Nous normalisons cet indice descriptif par $11(K - 1)$ afin de rendre les indices descriptifs comparables entre eux pour différentes valeurs de K . L'interprétation des états cachés par rapport aux usages passe par une analyse factorielle des correspondances.

TABLEAU 1
Tableau d'intensité pour $K = 7$

États cachés	Usages					
	ECL	ECS	FRD	H	LL	CUI
1	1,03e+05	1,07e+05	1,73e+05	2,42e+04	2,65e+04	5,73e+04
2	5,52e+04	3,12e+05	8,59e+04	2,83e+04	5,88e+04	1,09e+05
3	4,32e+03	1,66e+02	3,50e+04	1,31e+02	1,37e+02	1,03e+03
4	8,72e+04	1,82e+04	2,56e+05	6,32e+03	8,84e+03	1,62e+04
5	3,03e+04	5,33e+03	1,57e+05	1,38e+03	2,64e+03	1,27e+04
6	5,20e+04	6,22e+05	6,72e+04	3,57e+03	1,67e+05	1,05e+05
7	5,82e+04	6,77e+03	2,23e+05	1,56e+03	4,36e+03	7,94e+03
	LV	NH	SL	TVH	CHA	AUT
1	8,92e+03	2,02e+03	8,64e+03	4,39e+04	3,73e+05	5,00e+05
2	4,95e+04	2,18e+02	8,60e+04	2,71e+04	5,14e+05	4,47e+05
3	2,10e+01	6,00e+00	4,50e+01	6,32e+03	5,16e+03	6,16e+04
4	1,23e+03	3,33e+07	1,20e+03	6,17e+04	2,09e+05	5,23e+05
5	1,68e+03	2,37e+02	7,13e+02	2,58e+04	3,36e+04	2,68e+05
6	1,52e+05	1,79e+02	1,32e+05	1,81e+04	3,46e+05	3,38e+05
7	6,22e+03	2,78e+03	1,39e+03	5,02e+04	8,07e+04	3,98e+04

L'inertie du nuage des $(n_{ju})_{j,u}$ a un pourcentage de représentation de 99 % dans le premier plan principal de l'analyse des correspondances (figure 6). Tous les états et tous les usages sauf l'éclairage non halogène NH sont bien représentés dans ce plan. Les états sont situés sur une parabole sur laquelle ils sont ordonnés, suivant la valeur de la moyenne de leur loi d'émission. Il s'agit d'une manifestation de l'effet Guttman, dû au fait que les états ont une signification quantitative et qu'il existe une relation non linéaire entre les deux axes principaux (au sujet de l'effet Guttman, voir par exemple Lebart *et al.*, 1995). Les usages sont pratiquement situés sur cette parabole. Ainsi, le premier axe principal est très fortement lié à l'intensité de la consommation électrique et oppose les valeurs maximales (état C1) – associées principalement à l'eau chaude mais aussi au lave-vaisselle, lave-linge et sèche-linge – aux valeurs minimales (états C7, C6 et C5) – associées à la réfrigération. Le deuxième axe principal oppose les consommations intermédiaires – liées à l'halogène, qui a un profil atypique, et au chauffage – aux consommations extrémales. Le quadrant «sud-ouest» correspond essentiellement aux consommations nocturnes où la réfrigération représente l'usage dominant, ainsi que la télévision quand elle est en veille.

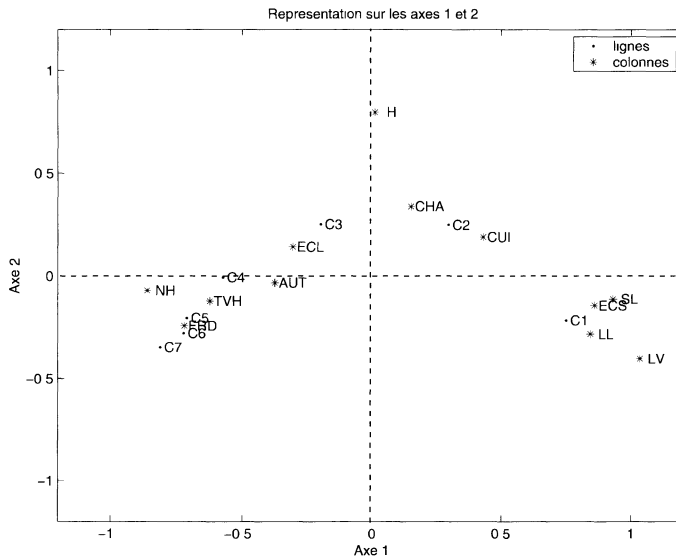


FIGURE 6

Représentation des états et des usages dans le premier plan principal de l'AFC (modèle à 7 états cachés)

6. Estimation de la consommation due aux usages

Nous présentons, dans cette partie, une méthode permettant d'estimer la part de la consommation globale due à chaque usage, en utilisant la restauration des états cachés et le tableau d'intensité mettant en correspondance les usages et les états cachés

restaurés. Étant donné une courbe de la consommation électrique journalière totale $(g_t)_{1 \leq t \leq 144}$ d'un ménage, éventuellement avec des relevés manquants, on souhaite estimer la consommation $(g_t^u)_{1 \leq t \leq 144}$ due à chaque usage u . Pour ce faire, nous calculons les résidus $(y_t)_t$ des $(\log(g_t))_t$ par l'analyse de variance (1). Nous utilisons ensuite le modèle de chaîne de Markov cachée estimé pour restaurer les états cachés s_1^{144} par l'algorithme de Viterbi. Le tableau d'intensité croisant usages et états cachés restaurés permet ensuite le calcul de la loi conditionnelle des usages sachant l'état caché le plus probable. Cette probabilité conditionnelle constitue une estimation de la part de la consommation globale due à l'usage u . On pose pour tout état j et tout usage u

$$f_{ju} = \frac{n_{ju}}{\sum_{j'} \sum_{u'} n_{j'u'}}$$

La quantité f_{ju} est un estimateur de la probabilité jointe d'être dans l'état j pour l'usage u . La probabilité conditionnelle de l'usage u sachant l'état j est estimée par

$$f_{u|j} = \frac{f_{ju}}{\sum_{u'} f_{u'j}}$$

Enfin, si à l'instant t on a $\hat{s}_t = j$, alors l'estimateur de la consommation due à l'usage u est $\hat{g}_t^u = f_{u|j} g_t$. La figure 7 contient la courbe de consommation totale réelle, puis celles des résidus et des états cachés, tandis que la figure 8 présente les courbes restaurées par cette méthode avec, pour comparaison, la consommation réelle pour quatre des usages (AUT, LV, ECS et ECL), qui est connue pour chacune des 444 courbes considérées pour la construction du tableau d'intensité. Rappelons cependant qu'il s'agit de consommations bruitées. Les différences importantes entre ces deux dernières courbes mettent en évidence l'intérêt de prendre en compte des informations extérieures sur les habitudes de consommation des foyers.

L'information extérieure la plus facile à prendre en compte est l'absence d'un appareil électrique dans un logement. Par exemple, dans le cas de la figure 8, le logement ne comporte pas de lave-linge, de sèche-linge, de chauffage de l'eau et de chauffage électrique, ni d'éclairage non halogène. Ceci est pris en compte en imposant que les n_{ju} soient égaux à 0 pour toutes les valeurs de j et pour les valeurs de u correspondant aux usages absents. Les estimateurs des probabilités conditionnelles sont alors réactualisés en conséquence, en utilisant la même formule que dans la partie précédente. Les trois courbes de la figure 8 illustrent l'effet de la suppression des cinq usages cités ci-dessus. Il est également possible de supprimer certains usages à des heures précises de la journée et non plus forcément pour la journée entière. Ceci ne change pas le raisonnement précédent, valable pour des instants t quelconques.

Sur la figure 8, la part de la consommation due aux usages, en particulier l'éclairage, est mieux estimée quand on prend en compte les usages absents. Cependant, cette méthode d'estimation possède encore les inconvénients suivants : les grandeurs du tableau d'intensité correspondent à des grandeurs de type consommation instantanée \times temps, qui s'expriment en watt heure. Ces watt heures sont répartis globalement suivant les différents usages, mais c'est la consommation instantanée que nous cherchons à reconstituer. Par conséquent, l'estimation des usages se

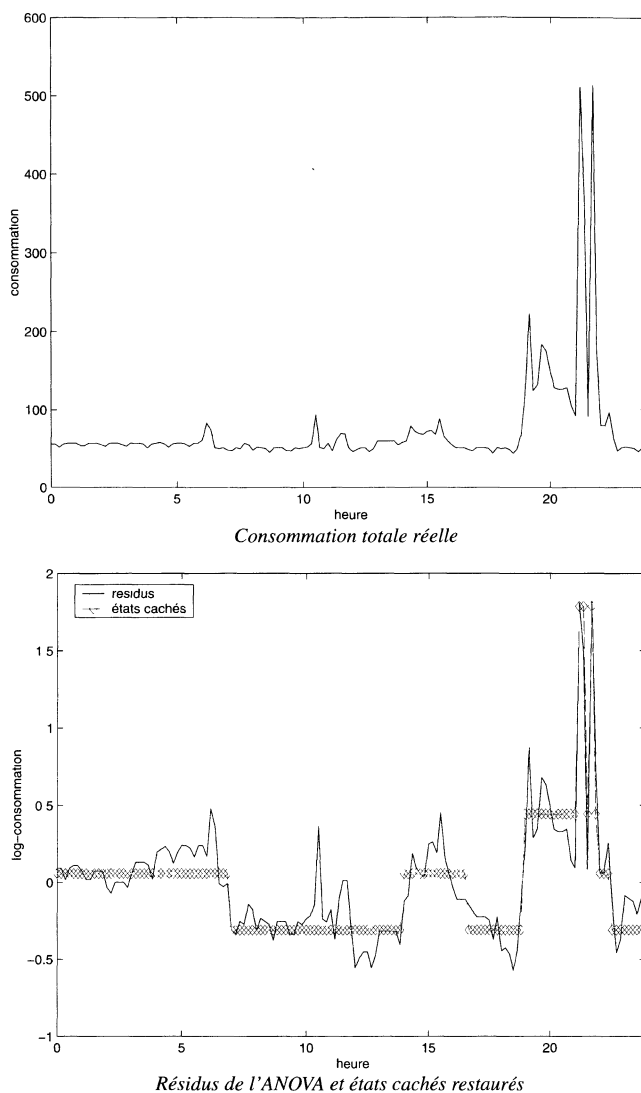
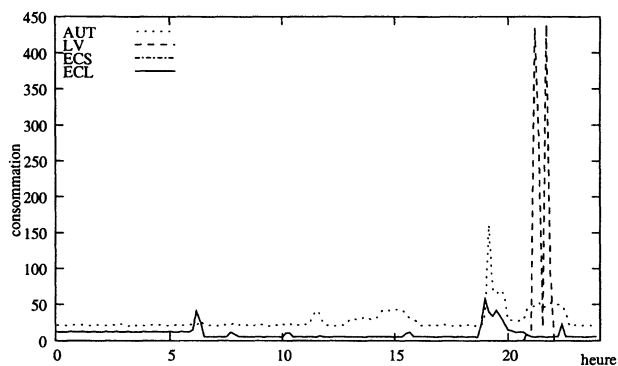
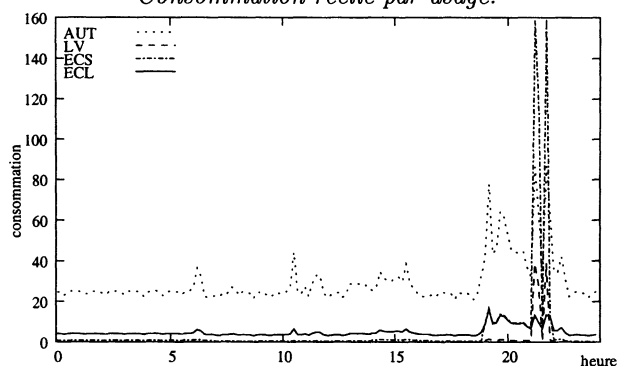


FIGURE 7
Consommation totale réelle et résidus de l'ANOVA

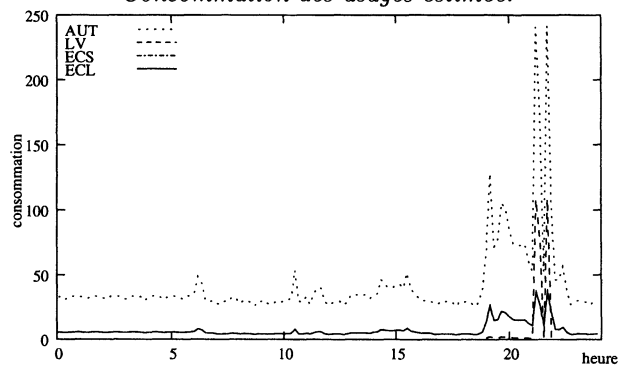
fait sur la base de la consommation moyenne et de manière proportionnelle à la consommation totale. Ceci ne prend pas en compte le fait que la plupart des appareils qui consomment beaucoup fonctionnent brièvement (sèche-linge et lave-linge par exemple). L'estimation par notre modèle de la consommation de ces appareils conduit, au contraire, à des valeurs plus faibles (due à la répartition proportionnelle de la consommation entre les usages) mais pendant une durée plus longue que la durée réelle. D'autre part, l'usage *autres appareils* (AUT) est le plus représenté. De plus, il



Consommation réelle par usage.



Consommation des usages estimée.



Estimation prenant en compte l'absence de certains appareils électriques.

FIGURE 8

Estimation de la consommation due aux usages sèche-linge, eau chaude, éclairage et autres, sur des données réelles bruitées

n'est associé à aucun état spécifique. La part de la consommation affectée à cet usage est donc toujours élevée (voir par exemple le pic de la consommation estimée pour cet usage, vers 22 heures, figure 8).

Nous avons donc tenté de tenir compte des heures de pointe, de l'ordre de grandeur relatif de la consommation de chaque appareil et de la saisonnalité des usages, à l'aide d'une courbe type $(\tilde{g}_t^u)_i$ de la consommation pour chaque usage. Le choix d'un poids $p_u \in [0, 1]$ détermine l'importance donnée à la courbe type par rapport à la courbe estimée. La consommation à l'instant t due à l'usage u est estimée dans un premier temps par

$$\hat{g}_t^u = p_u f_{u|j} g_t + (1 - p_u) \tilde{g}_t^u$$

si $\hat{s}_t = j$. Comme on souhaite que la contrainte $\sum_u \hat{g}_t^u = g_t$ soit vérifiée, on normalise ainsi l'estimation

$$\hat{g}_t^u = g_t \frac{\hat{g}_t^u}{\sum_u \hat{g}_t^u}.$$

En réalité, l'application de ce principe ne nous a pas permis d'améliorer l'estimation. D'une part, ceci est dû à la difficulté d'établir des courbes types pertinentes. D'autre part, il reste le problème, évoqué plus haut, lié à l'affectation des watt heures aux différents usages alors qu'on veut estimer une consommation instantanée. Le problème de l'estimation de pics de consommation et de leur affectation à un appareil unique, plutôt qu'à un groupe d'appareils, demeure et ne peut être résolu par la modélisation proposée. Il paraît donc plus réaliste de tenter d'estimer la consommation due à un groupe d'usages similaires.

7. Sélection de modèles

7.1. Sélection du modèle d'analyse de variance

Le premier problème de sélection de modèles rencontré dans cette étude concerne la détermination du modèle d'analyse de variance, pour déterminer les effets contrôlés qui ont une influence sur la valeur moyenne de la log-consommation. Les modèles considérés *a priori* sont ceux faisant intervenir tout ou partie des facteurs suivants : *tarif, puissance souscrite, mois, jour et heure*, ainsi que toutes les interactions d'ordre un entre ces facteurs. Dans un premier temps, nous avons essayé de baser la sélection sur des tests. Cette méthode conduit à rejeter systématiquement le modèle le moins complexe. Cela est sans doute dû au nombre élevé de données (une centaine de milliers) qui rend le test très puissant, et au fait que l'hypothèse H_0 est peu réaliste vu les données (voir partie 3). Nous avons considéré également le critère BIC, dont l'application aux modèles linéaires est discutée dans Kass et Raftery (1995). Le critère BIC conduit au même résultat, ce qui est lié au fait que la pénalité reste minimale par rapport à la log-vraisemblance et par rapport à sa croissance suivant la complexité du modèle.

En réalité, il est suffisant, dans le cadre de cette étude, de disposer d'un modèle simple mettant en évidence les effets les plus flagrants, vu qu'il s'agit uniquement de calculer les résidus de l'analyse de variance. Or ceux-ci sont apparus peu sensibles à la complexification du modèle, à partir d'un certain seuil. C'est pourquoi nous nous

sommes contentés de sélectionner le modèle offrant un bon rapport entre la déviance expliquée et le nombre de degrés de liberté. Pour ce faire, nous avons considéré des modèles de plus en plus complexes jusqu'à ce que l'accroissement de ce rapport devienne inférieur à 0,1 %. Le modèle finalement retenu met en jeu les facteurs *type de contrat* (ou *tarif*) c , *mois* m , *heure* h , *puissance* p et l'interaction entre *tarif* et *mois*. De plus, cette méthode nous a permis de regrouper les types de tarif « nuit » et « EJP », qui ont des effets similaires sur la moyenne de la log-consommation.

7.2. Sélection du nombre d'états cachés

Dans cette étude, le principal paramètre à déterminer pour la sélection d'un modèle de chaîne de Markov cachée est son nombre d'états K . L'ensemble des valeurs possibles pour K est choisi comme suit : la valeur maximale est fixée à douze, ce qui correspond au nombre d'usages disponibles. On espère pouvoir mettre en correspondance, de cette manière, les usages et les états cachés, par la méthode indiquée en partie 5.3. La valeur minimale est fixée à sept par les experts d'EDF, au vu de l'histogramme des résidus $(y_t)_t$ (voir figure 4).

Nous envisageons tout d'abord un choix parmi les différents modèles, caractérisés par une valeur de K entre sept et douze, basé sur des critères de sélection BIC, ICL (proposé par Biernacki *et al.*, 2001 dans le cadre des mélanges indépendants) et le demi-échantillonnage. Comme les différentes courbes sont supposées mutuellement indépendantes, cette dernière technique utilise la moitié des séquences pour l'estimation du paramètre λ , l'autre moitié servant à évaluer la vraisemblance du paramètre, utilisée comme critère associé au demi-échantillonnage.

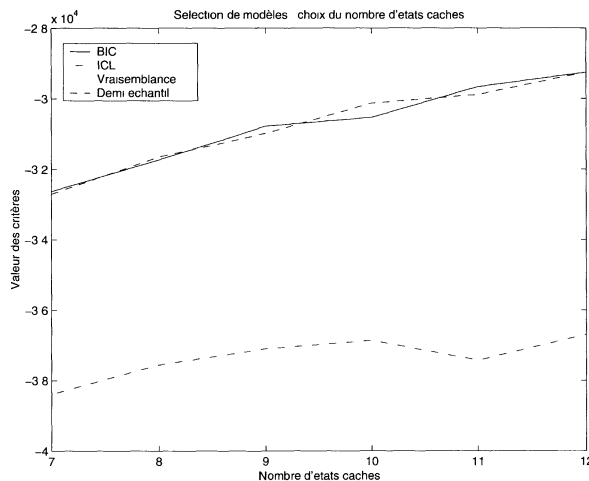


FIGURE 9

Critères de choix d'un modèle pour la consommation électrique :
BIC, ICL et demi échantillonnage.

La vraisemblance est tracée pour permettre la comparaison avec ces critères

La figure 9 représente les courbes pour ces critères (la vraisemblance est également tracée). Ces critères ont un comportement analogue à celui concernant la sélection d'un modèle pour l'analyse de variance, en ce qu'ils choisissent le modèle le plus complexe. Leur croissance est approximativement linéaire par rapport au nombre d'états cachés. Notons néanmoins le comportement original d'ICL par rapport aux autres critères, puisqu'il pénalise fortement un modèle à onze états pour lequel deux états se trouvent être pratiquement confondus (même moyenne et même variance). Ceci s'explique par le fait qu'ICL favorise les modèles qui aboutissent à une partition nette des données. D'autre part, les autres critères ont une croissance très similaire à celle de la log-vraisemblance. En réalité, la pénalité de BIC est négligeable par rapport à la croissance de la log-vraisemblance en fonction de K . Mais cela indique que le modèle de Markov caché est en définitive un modèle peu réaliste pour les résidus de la log-consommation électrique. Cela est évidemment lié au fait que le nombre de données disponible est considérable (plus d'une centaine de milliers). Nos objectifs d'interprétation des états, dont ne tiennent pas compte les critères ci-dessus, imposent des modèles plus parcimonieux. C'est pourquoi nous avons envisagé une méthode de sélection de modèles permettant de prendre en compte cet objectif d'interprétation des états cachés par rapport aux usages.

Rappelons que l'interprétation des états cachés se base avant tout sur le tableau d'intensité mettant en relation les usages et les états cachés restaurés. Les modèles les plus facilement interprétables, du point de vue de la mise en relation des usages et des états cachés, sont ceux qui maximisent l'écart à l'indépendance entre ces deux facteurs. Le tableau d'intensité dépend de la valeur de K : nous calculons l'indice descriptif du Φ^2 normalisé $\Phi_N^2(K)$ pour chacune d'entre elles. Cet indice mesurant l'écart des données à l'indépendance vis-à-vis des douze usages et des K états, l'idée est de sélectionner le modèle (*i.e.* la valeur de K) qui le maximise. La figure 10 représente cet indice en fonction de K , non seulement pour les valeurs de K entre 7 et 12 mais aussi, pour information, pour des valeurs comprises entre 1 et 6.

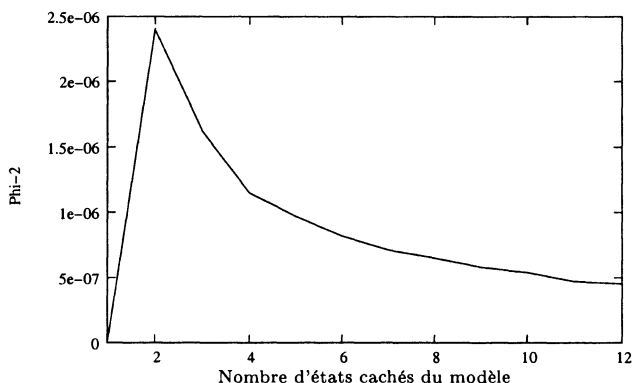


FIGURE 10

L'indice du Φ^2 normalisé en fonction du nombre d'états cachés

La courbe de la figure 10 conduit à rejeter les modèles les plus complexes. Autrement dit, l'augmentation du nombre d'états cachés rend plus difficile leur interprétation vis-à-vis des usages. Ceci est illustré par les résultats de l'analyse factorielle des correspondances pour un modèle à dix états cachés (figure 11), présentés à titre de comparaison avec la figure 6. Alors que les conclusions énoncées dans la partie 5.3 ne sont pas modifiées, l'interprétation des états est rendue plus difficile par leur plus grand nombre.

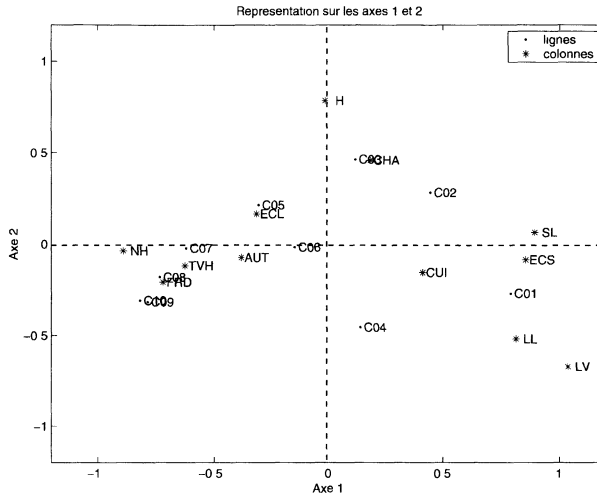


FIGURE 11

Représentation des états et des usages dans le premier plan principal de l'AFC [modèle à 10 états cachés]

Nous sommes conduits, en utilisant le critère du Φ^2 normalisé, à accepter un modèle trivial à deux états cachés, opposant les consommations faibles et les consommations élevées – qui est effectivement le plus facilement interprétable... Mais ces critères ne tiennent pas compte de l'adéquation entre le modèle et les données, cependant essentielle pour la validité de l'inférence et de l'analyse des courbes de consommation électrique à partir du modèle. La sélection de modèles ne peut donc se faire en utilisant ces critères seuls. En définitive, nous utilisons le modèle le plus simple parmi ceux sélectionnés par les experts, pour que l'interprétation des états cachés soit la plus facile possible, soit un modèle avec $K = 7$.

8. Discussion et perspectives

Bien que les chaînes de Markov cachées constituent un modèle très simplifié de la consommation électrique, vu les données disponibles, elles ont un intérêt important, du point de vue de l'interprétation des états cachés en termes de niveaux de consommation électrique; de plus, ces niveaux peuvent être mis en relation avec

les usages disponibles par un tableau d'intensité. On en déduit une méthode pour l'estimation des usages.

8.1. Sélection de modèles

Ce manque de réalisme des chaînes de Markov cachées, ainsi que le nombre considérable de données, rend inopérante la sélection par les critères classiques (comme BIC, ICL et la validation croisée), d'un modèle à la fois parcimonieux et interprétable. Clairement, la méthode retenue en définitive pour choisir un modèle n'est pas très satisfaisante : elle revient à déterminer, de manière *ad hoc*, un modèle réalisant un bon compromis entre l'adéquation aux données et l'interprétation. Il serait souhaitable de pouvoir quantifier ce compromis, ce qui pourrait être fait par un critère sélectionnant un modèle de Markov caché conduisant à une classification stable des données. Une possibilité de définir un critère de cette nature dans le cas où une partition des instants est disponible, chaque classe étant associée à l'utilisation d'un usage (ou d'un type d'usages) dominant, est décrite dans Durand (2003).

L'application de ce type d'approche à notre problème de sélection de modèles passe par le choix d'une partition *a priori* des données observées. Les classes d'une telle partition pourraient être associés à un usage ou à type d'usages dominant pendant une période donnée.

8.2. Extensions du modèle

L'approche présentée a consisté, dans un premier temps, à estimer et supprimer les effets fixes (tarif, mois, puissance, heure) par une analyse de variance dont les hypothèses ne sont pas vérifiées (non normalité et dépendance des résidus). Puis les résidus sont modélisés par une chaîne de Markov cachée pour prendre en compte la dépendance résiduelle (hypothèse de Markov sur la chaîne cachée) et la non normalité (hypothèse d'une loi de mélange). Ainsi les états cachés visent à modéliser les tendances non prises en compte par les effets fixes du fait que nous travaillons orthogonalement à ces effets.

D'autres approches pourraient autoriser l'analyse conjointe des effets fixes et des états cachés en se basant sur des modèles markoviens de changement de régime. Un tel modèle peut par exemple être défini comme suit :

- $(S_t)_t$ est une chaîne de Markov cachée;
- conditionnellement à $S_t = k$, $\log(G_t)$ suit une loi normale $\mathcal{N}(\mu_{m,h,c,p,k}; \sigma_k^2)$, où par exemple $\mu_{m,h,c,p,k} = \alpha_k + \beta_k^m + \gamma_k^h + \delta_k^c + \eta_k^p + \zeta_k^{c,m}$;
- sachant $\{S_t = s_t\}_t$, les $(\log(G_t))_t$ sont conditionnellement indépendants.

Ainsi, les états cachés sont ici déterminés conjointement par la consommation électrique globale et par les facteurs contrôlés. On peut également envisager des modèles où les états cachés auraient une signification différente suivant le foyer (effet contrôlé ayant éventuellement une incidence sur la variance) ou suivant un type de foyer, également caché. Il s'agirait alors d'un mélange des modèles ci-dessus. Tous ces modèles sont plus riches que celui que nous avons utilisé, mais plus difficiles à estimer de manière fiable et aussi à interpréter.

Plus simplement, il est possible, dans un état d'esprit exploratoire, d'effectuer une analyse factorielle (AFC ou ACP) des courbes de consommation, suivie d'une classification, puis d'identifier un modèle pour chaque classe retenue, cette analyse factorielle et cette classification pouvant également se faire sur les résidus de l'analyse de variance.

Remerciements : Les auteurs remercient le Comité de Rédaction de la Revue de Statistique Appliquée pour ses remarques et suggestions concernant cet article.

Références

- [1] BAUM L.E., PETRIE T., SOULES G. et WEISS N. (1970), A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov chains. *The Annals of Mathematical Statistics*, vol. 41, n° 1, pp. 164-171.
- [2] BIERNACKI C., CELEUX G. et GOVAERT G. (2001), Assessing a Mixture Model for Clustering with the Integrated Completed Likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, n° 7, pp. 719-725.
- [3] CELEUX G. et DURAND J.-B. (2002), Choosing the order of a hidden Markov chain through cross-validated likelihood. In : *Compstat2002. Berlin (Allemagne)*, 24-28 août.
- [4] DEVIJVER P.A. (1985), Baum's forward-backward Algorithm Revisited. *Pattern Recognition Letters*, vol. 3, pp. 369-373.
- [5] DURAND J.-B. (2003), *Modèles à structure cachée : inférence, sélection de modèles et applications*. Thèse de doctorat, Université Grenoble 1 – Joseph Fourier, janvier.
- [6] FORNEY Jr. G.D. (1973), The Viterbi Algorithm. In : *Proceedings of the IEEE*, mars, pp. 268-278.
- [7] KASS R.E. et RAFTERY A.E. (1995), Bayes factors. *Journal of the American Statistical Association*, vol. 90, n° 430, pp. 773-795.
- [8] LEBART L., MORINEAU A. et PIRON M. (1995), *Statistique exploratoire multidimensionnelle*. Dunod.