COURBOTREE : UNE METHODE DE CLASSIFICATION DE COURBES APPLIQUEE AU LOAD PROFILING

Véronique Stéphan

EDF R&D
1, av. du Général de Gaulle
92 141 Clamart Cédex
Veronique.Stephan@edf.fr

Résumé
Depuis le 1er Juillet 2004, l'ensemble des clients industriels et professionnels peuvent choisir leur fournisseur d'électricité. Pour la majeure partie de ces clients, EDF ne dispose pas des courbes de consommation électrique mais seulement d'index qui permettent de calculer le volume total consommé entre 2 index consécutifs. Un profilage (load profiling), c'est-à-dire une estimation de la courbe de consommation électrique des clients, point par point sur cette période, est alors nécessaire. Celui-ci peut être réalisé en prenant en compte les connaissances métiers et par l'analyse des données récoltées à partir d'un échantillon de clients télérelevés. Ainsi, face à ces nouveaux enjeux commerciaux, des besoins spécifiques en techniques classificatoires appliquées aux courbes apparaissent. Nous proposons dans cet article une méthode de classification de courbes, appelée Courbotree, s'intégrant dans une démarche plus globale de load profiling. Cette méthode repose sur les techniques d'arbres de régression multivariée. Elle répond à un double objectif de classification et de prédiction de courbes. Dans le contexte multivarié, la construction de l'arbre est similaire à celle des méthodes AID et CART. La seule différence réside dans le choix du critère de coupure qui est celui de l'inertie calculée sur les composantes des courbes. A partir d'un échantillon de clients, sur lesquels on dispose d'une part de leurs caractéristiques et d'autre part d'une courbe de consommation, Courbotree fournit une classification de ces courbes directement interprétable par l'utilisateur en terme de règles d'affectation métiers. Elle permet ainsi de profiler tout nouveau client selon ses valeurs observées sur les variables explicatives.
Mots-Clés : classification de courbes, load profiling, arbre de régression, méthode AID, classification divisive.

Abstract
Since July 2004, industrial and professional sector can choose their electric power supplier. For the most part of these customers, load curves aren't enabled and EDF just has a total load consumption over a large period. A load profiling appproach, which consists in estimating customer load curve, is therefore necessary. This technique may be performed taking into account business knowledge and analyzing data from telemetered customers. Faced to new marketing challenges, dedicated needs in clustering techniques appear. This article details a curves clustering technique, called Courbotree, integrated in a global load profiling approach. We present how applying multivariate regression trees meets these aims. These ones answer both problems of clustering and prediction applied to curves. Considering multivariate regression trees, tree building is performed in a similar way than AID and CART methods do. The main difference lies in the cutting criterion which is based on inertia computed on the curve components. Applied to a sample of customers, which are described by a load curve and a set of characteristics, Courbotree offers both a curve sample partition and a set of identification rules to affect a curve profile to a new individual according to his explanatory values.
Key-Words : curves partitioning, regression trees, AID method, divisive partitionning.

1. Introduction

Depuis le 1er Juillet 2004, l'ensemble des clients industriels, tertiaires et professionnels peuvent choisir leur fournisseur d'électricité. Face à l'ouverture du marché, une connaissance fine des profils de consommation des clients est ainsi devenu un enjeu stratégique pour proposer des offres tarifaires adaptées et anticiper les besoins électriques de chacun des segments clientèles. En effet, dans une situation de monopole, une estimation de la consommation d'électricité au niveau France était suffisante. L'apparition d'offres de prix différentiées par type de clients nécessite une connaissance de ceux-ci à une maille fine. EDF a notamment besoin de construire une prévision de leur courbe de consommation à un horizon d'une année, le plus souvent au pas horaire d'une part pour prévoir la consommation d'un client en particulier, mais aussi et surtout prévoir la consommation d'un ensemble de clients (d'une zone géographique par exemple) afin de pouvoir s'approvisionner au plus près de la consommation réelle.
Pour les clients consommant plus de 7 GWh, les méthodes de prévision s'appuient sur des réalisations passées puisque les courbes de charge sont connues par la télé relève du compteur. Cependant pour les sites plus petits, EDF ne dispose pas des courbes de consommation électrique mais seulement d'index qui permettent de calculer le volume total consommé entre 2 index consécutifs. Pour la majeure partie de ces clients, un profilage (load profiling), c'est-à-dire une estimation de la courbe de consommation électrique des clients, point par point sur cette période, est alors nécessaire. Une première solution consiste à utiliser des profils types réglementaires fournis par la Commission de Régulation de l'Electricité (CRE). Ces profils ont été développés pour permettre le règlement des écarts sur le marché de l'électricité. Une alternative serait d'exploiter, si les performances sont intéressantes, un profilage plus précis, tirant parti des connaissances métier d'EDF et des données récoltées sur les clients.

EDF R&D s'applique donc à définir et appliquer des méthodes qui permettent d'attribuer à un client dont on ne connaît pas la courbe de charge, un profil de consommation qui se rapproche le plus possible de sa courbe de consommation réelle inconnue. Cette attribution se fait en ne connaissant sur le client que les données qui sont généralement disponibles. Il s'agit des données du contrat, de l'activité professionnelle ainsi que les données de facturation des années précédentes pour les anciens clients. Les méthodes développées ici supposent donc au préalable de définir des règles générales à partir d'un Panel de clients dont on connaît la courbe de charge. Traditionnellement, on effectue une classification sur les courbes des clients de ce Panel, puis une discrimination des classes obtenues et une détermination de la procédure d'affectation en fonction des données de contrat et de facturation entre autres. Une courbe de charge-type peut ainsi être affectée à tout nouveau client.
Nous proposons dans cet article une méthode de classification de courbes, appelée Courbotree (nom également de l'outil informatique). Elle répond à un double objectif de classification et de prédiction de courbes. Pour un échantillon donné, chaque individu est décrit par une courbe (composée de q variables quantitatives Y_1, ..., Y_q) et par un ensemble de variables explicatives. Courbotree repose sur les techniques d'arbres de régression multivariés, sans prise en compte de l'aspect curviligne. Dans le contexte multivarié, la construction de l'arbre est similaire à celle des méthodes AID et CART. La seule différence réside dans le choix du critère de coupure qui est celui de l'inertie calculée sur les composantes des courbes. A partir d'un échantillon de clients, sur lesquels on dispose d'une part de leurs caractéristiques et d'autre part d'une courbe de consommation, Courbotree fournit une classification de ces courbes directement interprétable par l'utilisateur en terme de règles d'affectation métiers. Elle permet ainsi d'attribuer à un client dont on ne connaît pas la courbe de charge, un profil de consommation qui se rapproche le plus possible de sa courbe de consommation réelle inconnue.

Dans une première partie, nous rappelons les techniques d'arbres binaires de régression et leur généralisation au cas multivarié pour la classification de courbes. La partie suivante décrit l'algorithme implémenté dans Courbotree. Enfin une application portant sur des données d'hôtels est présentée.

2. Problématique

1.1. Rappel sur les arbres binaires de régression

L'objectif de la technique d'arbre binaire de régression est de prédire la valeur d'une variable quantitative *Y* en fonction d'un ensemble de variables explicatives de nature quelconque. La construction de l'arbre binaire s'effectue par partitionnement récursif de l'ensemble d'apprentissage. Comparée aux techniques d'arbre de segmentation, les spécificités de la méthode sont les suivantes :

- l'objectif est ici de prédire pour un individu sa valeur sur la variable *Y* en fonction du noeud terminal (feuille) dans lequel il tombe. Cette valeur est égale à la moyenne des valeurs observées sur *Y* dans la feuille ;
- le critère de coupure de l'arbre n'est pas évalué à partir des notions d'impureté mais il est défini comme une minimisation de la variance intra-groupe de la variable *Y*.

1.2. Généralisation au problème de classification de courbes

Dans le cas où l'on s'intéresse non pas à une seule variable à expliquer mais à un ensemble de variables à expliquer (cas multidimensionnel), la méthode d'arbre de régression est dite multivariée. Nous proposons dans cet article d'appliquer une telle méthode au problème de classification puis de prédiction d'une courbe définie sur q composantes Y_1, ..., Y_q. Le principe est analogue, seul le critère de coupure est modifié (voir partie 2.2.). Dans ce contexte, l'aspect curviligne des données disparaît pour être traité comme du multivarié.
L'exemple suivant est issu du domaine commercial d'EDF. Supposons que l'on dispose d'un échantillon de clients hôtels pour lesquels nous observons à la fois la courbe de consommation détaillée ainsi qu'un ensemble de variables décrivant leur activité, leurs usages et leur type de contrat d'électricité. CourboTree a été appliquée sur cet échantillon en retenant 20 classes de courbes. Un extrait de l'arbre en sortie est présenté dans la figure suivante (Fig. 1).

Figure 1 : extrait de l'arbre de régression des courbes

Nous interprétons directement la ***classe 1*** comme le groupe des hôtels dont le ratio heures creuses / heures pleines est compris entre 0 et 0.9 (heure_creuse=[0,0.9]). La courbe moyenne de la classe (avec un intervalle de confiance de 10%) est présentée dans la figure 2.

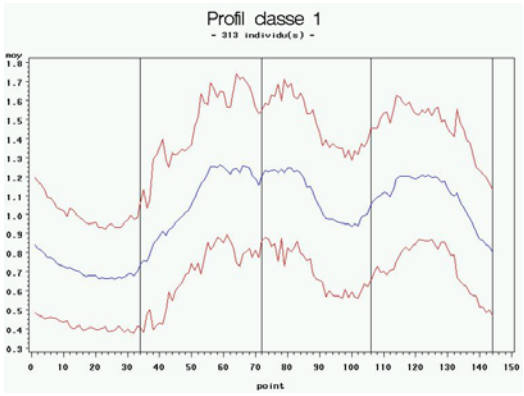

Figure 2 : Profil journalier moyen de la classe 1

Appliqué aux courbes, l'arbre de régression multivarié répond à deux objectifs complémentaires :

- **Classification de courbes** : il s'agit d'obtenir une partition des courbes directement interprétable en terme de variables supplémentaires dites explicatives ;
- **Prédiction d'un profil de courbe** : dans certains cas, on ne dispose de courbes que sur un échantillon de la population (par exemple un panel, les clients télérelevés, ...). L'arbre de régression, construit à partir des données de l'échantillon, fournit alors directement les règles d'affectation d'un profil de courbe à un individu quelconque de la population.

1.3. Travaux relatifs

L'utilisation d'une structure d'arbre de régression a été proposée dans l'algorithme AID de Morgan et Sonquist (1963). Elle a été par la suite développée par Breiman et al. (1984). Les auteurs soulignent le fait que cette structure permet de dégager par ordre d'importance les variables prédictives. L'introduction de l'aspect multivarié a été étudié par Zhang (1998) et par Segal (1992).

H. Zhang propose une généralisation de la méthode de segmentation par arbre dans le cas de plusieurs variables à expliquer binaires. M. Segal généralise les méthodes d'arbre de régression au cas multivarié pour des données longitudinales. Enfin, une application de cette technique dans le contexte de données fonctionnelles est proposée dans Yu et al. (2000). Dans l'étude, les arbres de régression multivariés sont construits à la suite d'une réduction de dimension des données fonctionnelles.

Notre étude se place naturellement dans le cadre des méthodes classificatoires non supervisées. Les techniques principalement utilisées sont les nuées dynamiques et les cartes auto-organisatrices de Kohonen. Une application des cartes auto-organisatrices aux données de consommation d'énergie est présentée par Debrégeas et al. (1998). Dans le cadre classificatoire, nous pouvons également citer le travail de Chavent (1997) qui propose un algorithme de classification divisive sur des données de nature quelconque. Notre étude se place dans le cadre classificatoire des courbes dans la mesure où l'algorithme fournit en sortie un ensemble de feuilles correspondant à des classes de courbes. CourboTree permet d'obtenir conjointement une classification des courbes et les règles d'affectation du profil de courbe d'une classe à tout nouvel individu. **La méthode CourboTree se substitue à l'approche classique qui nécessite de combiner une classification puis une modélisation des classes**.

1. CourboTree

Soient en entrée :

- 1,..., n : individus de l'échantillon,
- p_1, ..., p_n : le système de poids de l'échantillon,
- X_1, ..., X_j, ..., X_p : les p variables explicatives où x_{ij} est la valeur de l'individu i pour X_j,
- Y_1, ..., Y_k, ..., Y_q : les q variables temporelles des courbes où y_{ik} est la valeur de l'individu i pour Y_k.

S'agissant de l'arbre, nous notons :

- t, un nœud de l'arbre divisé en deux nœuds fils t_G (resp. t_D),
- g_t, le centre de gravité des individus appartenant au nœud t. Ses coordonnées g_{t1}, …, g_{tq} sont les moyennes pondérées des individus du nœud t sur les q variables Y_1, …, Y_q,
- g, le centre de gravité de l'échantillon. Ses coordonnées sont les moyennes pondérées des individus de l'échantillon sur les q variables Y_1, …, Y_q.

2.1. Principe de l'algorithme

L'algorithme procède par divisions successives de l'échantillon pour produire un arbre binaire de régression. Contrairement à la méthode CART, le nombre de feuilles de l'arbre (c'est-à-dire de classes de courbes) est fixé a priori et c'est ce nombre qui est utilisé pour l'arrêt de l'algorithme. La construction de l'arbre nécessite de définir les éléments suivants :

- établissement pour chaque noeud de l'ensemble des divisions binaires admissibles ;
- choix du critère de coupure pour la sélection de la meilleure division binaire ;
- choix d'un nouveau noeud à diviser.

L'ensemble des divisions binaires admissibles d'un noeud est évalué sur l'espace d'observations des variables X_1, ..., X_p. La recherche de toutes les coupures est similaire à celle effectuée dans la méthode CART. Le critère de division optimale d'un noeud correspond au choix du noeud ayant l'inertie intra-groupe calculée sur les variables Y_1, ..., Y_q minimale. A chaque étape de l'algorithme, nous choisissons comme nouveau noeud à diviser celui qui possède la réduction d'inertie intra-groupe (issue de la division) la plus forte. Ce choix est différent de la méthode AID, où le noeud avec la plus forte inertie totale est retenu.

2.2. Critère de coupure à optimiser

Le critère de division à optimiser dans l'algorithme est la minimisation de l'inertie intra-groupe calculée sur les q composantes des courbes.
Soit t, un nœud terminal de l'étape en cours. L'inertie intra-groupe de t est égale à :

Soit $d=(X_j,V)$ une division admissible issue des observations de X_j sur le noeud t, où V est l'un des deux groupes de modalités de X_j obtenue par division binaire. Soient t_G et t_D la bi-partition des individus de t selon la division d telle que :

Les inerties intra-groupe des noeuds t_G et t_D sont donc égales à :

où g_{t_G} (resp g_{t_D}) est le centre de gravité du fils gauche (resp droit) de *t*.

A une étape donnée, on note *T* l'arbre de classification produit. L'inertie intra-groupe de l'arbre est égale à celle du nuage d'individus partitionnés selon les feuilles de *T* :

Pour une division admissible *d* sur le nœud *t* induisant une bi-partition $\{t_G, t_D\}$, le critère de division s'écrit comme la **réduction de l'inertie intra-groupe** obtenue :

A chaque étape de l'algorithme, l'inertie intra-groupe de la partition obtenue à partir des feuilles de *T* est minimisée. Ceci revient à partitionner le noeud dont la division admissible admet la plus forte réduction de l'inertie intra-groupe :

3. Application

3.1. Représentation XML de l'arbre de régression

CourboTree est une application implémentée en SAS qui fournit une sortie d'arbre dans le format XML. Ce fichier est construit à partir des sorties de la procédure SAS. La structure retenue est celle de tableaux emboîtés qui permettent une représentation très proche de l'arbre souhaité. Cette structure est définie par un fichier XSL. L'avantage du choix d'XML pour l'affichage des résultats est de s'affranchir de tout outil de visualisation spécifique : le fichier des sorties peut être directement visualisé avec un navigateur pour restituer l'arbre. Un exemple de sortie d'arbre est donné dans la figure 3.

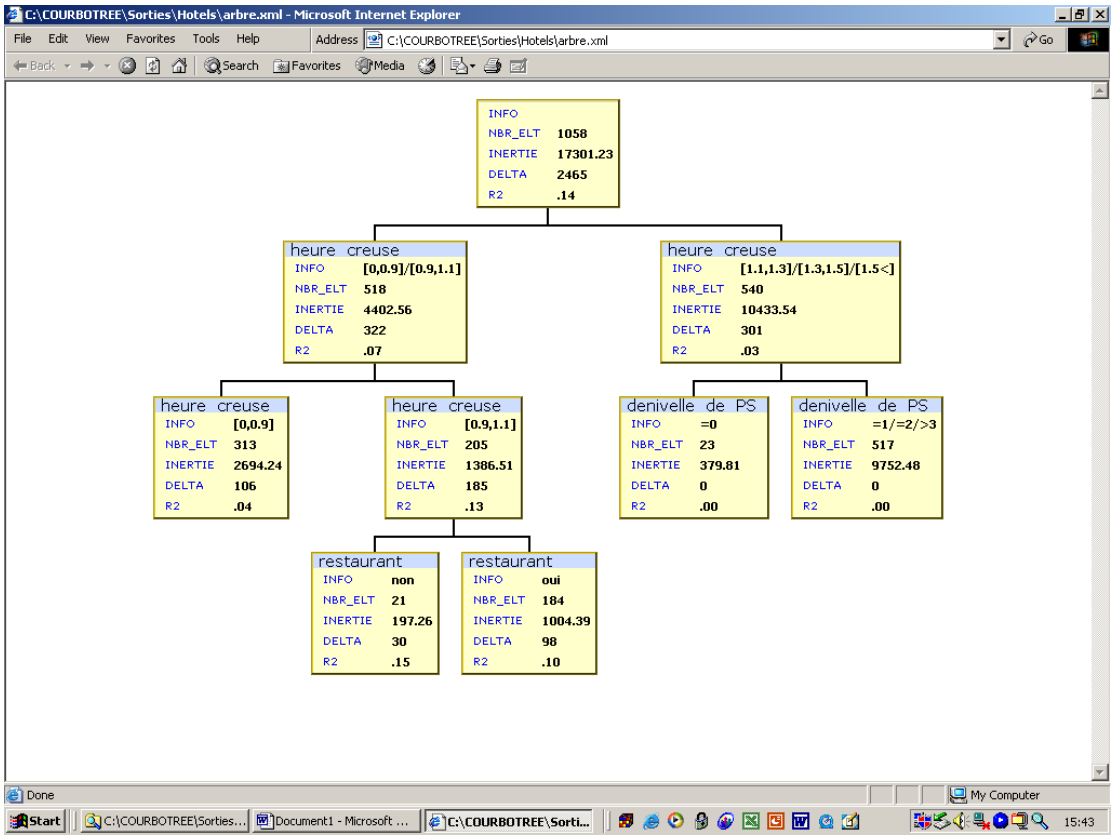

Figure 3 : Visualisation de l'arbre de régression par un fichier XML

3.2. Application aux données des hôtels

Courbotree a été utilisée sur un jeu de données de clients EDF composés d'hôtels (voir partie 1). Chaque hôtel est décrit par un ensemble de courbes de consommation journalière et par ses caractéristiques (activité, ..), ses usages (climatisation, ...) et son contrat (heure creuse, ...).
Une partie de l'arbre de régression en 20 classes a été présentée à la figure 1 du §1.1. Les deux exemples suivants fournissent deux classes de l'arbre et leur règle d'affectation :

- Si [heure_creuse ∈ {[0,0.9],[0.9,1.1]}] et [heure_creuse = [0,0.9]] alors Classe 1
- Si [heure_creuse ∈ {[1.1,1.3],[1.3,1.5],[1.5<]} et [deniv_puis > 0] et ... alors Classe 2.

Une variable explicative peut intervenir plusieurs fois pour une même classe (cf. Classe1). Dans l'exemple, la règle aurait pû être simplifiée comme : Si [heure_creuse = [0,0.9]] alors Classe 1
Les profils moyens de la classe 1 et de la classe 2 ainsi que l'intervalle de confiance autour de la moyenne sont fournis dans la figure 4.

Contrairement aux méthodes classiques, nous obtenons conjointement des classes de courbes homogènes et une interprétation à l'aide de variables auxiliaires. Sur l'exemple, nous pouvons ainsi en conclure que les hôtels de la classe 1 c'est-à-dire ayant un ratio heures creuses / heures pleines dont la valeur est inférieure à 0.9 ont un fort profil jour marqué (ce qui est appuyé par l'allure de la courbe).

Dans le cadre du load profiling, cette règle sera par la suite appliquée pour affecter la courbe moyenne de la classe 1 à un nouvel hôtel ayant une valeur de ratio comprise dans l'intervalle [0, 0.9].

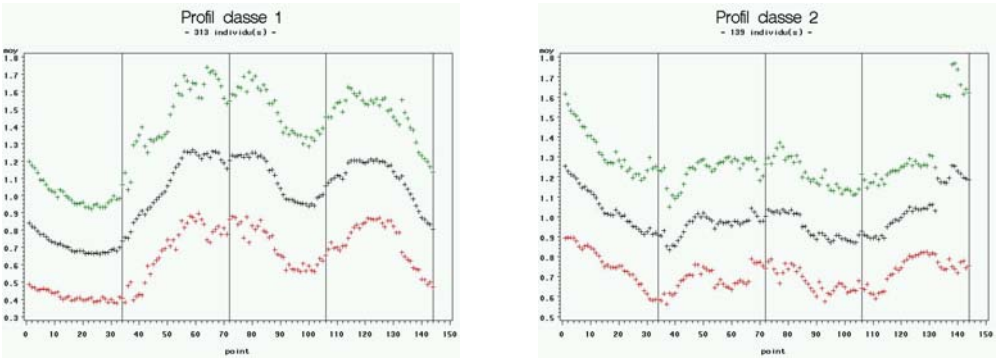

Figure 4 : Profil moyen journalier des classes 1 et 2

Parallèlement, nous avons réalisé l'approche classique consistant à effectuer une classification de l'échantillon des courbes puis à caractériser les classes à l'aide d'un arbre de segmentation. L'arbre de segmentation obtient des taux de bien classés assez médiocres du fait du nombre de classes retenu. De plus, si la première variable de coupure est la même que celle obtenue par CourboTree, certaines variables discriminantes sont masquées par cette approche.

L'utilisation de Courbotree dans une approche de load profiling est naturelle. Dans un premier temps, Courbotree est appliquée aux courbes de l'échantillon, avec une normalisation préalable pour s'affranchir des différences de niveaux de consommation. Les profils de courbes obtenus peuvent ensuite être extrapolés à l'ensemble des clients non télérelevés grâce aux règles d'affectation de l'arbre.

4. Conclusion et prespectives

CourboTree propose à la fois une classification de courbes directement interprétable et le modèle d'affectation d'une courbe moyenne à un nouvel individu en fonction de ses caractéristiques. L'interprétation des résultats est simple pour l'utilisateur puisque chaque classe est caractérisée par un profil de courbe et une règle d'affectation au profil. Cet aspect s'avère particulièrement intéressant dans les problématiques de profilage de courbes où l'on cherche à prédire la courbe d'un individu en fonction de ses caractéristiques, à partir de l'analyse d'un simple échantillon où l'on dispose à la fois des courbes et des variables explicatives.
En perspective, il serait intéressant d'introduire dans l'approche, la technique de validation proposée par Breiman et al. (1984). Une autre perspective est l'application d'une technique de réduction de dimension préalable de manière à rendre plus stable l'arbre de régression multivarié. Une telle réduction de dimension peut être réalisée au moyen d'une analyse en composantes principales. Un autre axe de recherche est une réduction de dimension par regroupement des composantes temporelles de la courbe en épisodes. Chaque nouvelle courbe est alors décrite par un ensemble d'épisodes qui peuvent représentés une consommation moyenne, ou un profil de consommation sur un intervalle de temps, par exemple. Courbotree pourrait alors être appliqué sur les nouvelles courbes symboliques obtenues.

Remerciements

L'auteur remercie Pierre Lé, Chercheur de la Division Recherche et Développement d'EDF, pour sa relecture attentive et ses remarques.

Bibliographie

[1] Morgan, J. N. and Sonquist, J.A. (1963) Problems in the analysis of survey data, and a proposal, *Journal of American Statistical Association*, 58:415-434.

[2] Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, C. J. (1984) *Classification and Regression Trees*, Wadsworth International Group, Belmont, California.

[3] Segal, M.R.(1992) Tree structured methods for longitudinal data, *Journal of American Statistical Association*, 87, 418:407-418.

[4] Zhang, H. (1998) Classification trees for multiple binary responses, *Journal of American Statistical Association*, 93, 441:180-193.

[5] Yu, Y. and Lambert, D. (2000) Fitting Trees to Functional Data, With an Application to Time of Day Patterns, *Journal of Computational and Graphical Statistic*, 8, 749-762.

[6] Debrégeas, A. and Hébrail, G. (1998) *Interactive Interpretation of Kohonen Maps Applied to Curves*, KDD 1998: 179-183.

[7] Chavent M. (1997) *Analyse des données symboliques : Une méthode divisive de classification*, Thèse de doctorat, Université Paris-IX Dauphine.