SEGMENTER LA CLIENTELE A PARTIR DES COURBES DE CHARGE

Anne Debregeas

EDF – Division Recherche et Développement
1 avenue du Général de Gaulle
92141 CLAMART Cedex

Cet article fait partie des Actes des 5èmes JOURNEES MODULAD qui ont eu lieu les 16 et 17 novembre 2000 à E.D.F. (Clamart).

1. Contexte

1.1. Les courbes, une source d'information riche mais complexe

Une courbe, ou série temporelle, représente un phénomène observé au cours du temps : le cours d'une action relevé tous les jours sur un an, la consommation d'électricité d'un client relevée toutes les 10 minutes sur une journée, etc. Elle intervient dans de très nombreux domaines : physique, médecine, finance, marketing, en fait tous les domaines manipulant des historiques de données).

Lorsque la courbe n'est pas trop longue (i.e. lorsqu'elle n'a pas été relevée sur beaucoup de points du temps), elle représente une donnée complexe mais " lisible " et est interprétable par l'oeil humain. Mais dès qu'il s'agit d'analyser un ensemble de courbes ou une courbe très longue, l'utilisation de techniques statistiques se révèle indispensable pour en extraire de l'information interprétable. La classification automatique, par exemple, constitue une aide précieuse : elle permet de dégager des formes-type, d'isoler des courbes atypiques. Cependant, l'expertise humaine reste nécessaire pour analyser et affiner les résultats obtenus par classification. La présentation des résultats ainsi que le caractère interactif du processus sont donc deux paramètres importants à respecter.

1.2. Un besoin croissant à EDF

Grâce à des appareils de télémesure, EDF enregistre la consommation d'électricité de ses plus grands clients toutes les 10 minutes, tout au long de l'année. Les consommations de panels de petits clients sont également relevées. Ces enregistrements de consommation constituent la « courbe de charge », qui représente une information très riche sur le comportement des clients. Elle peut être utilisée pour définir des segments de clientèle potentiellement intéressés par un service ou pour déterminer des comportements-type. A l'échelle d'un client, l'analyse de la courbe de charge constitue un guide précieux pour le choix du tarif.

Avec l'ouverture du marché de l'électricité, l'exploitation de l'information contenue dans les courbes de charge devient un enjeu majeur pour EDF, car elle lui permet de mieux connaître les comportements de ses clients et d'ajuster son offre tarifaire, par exemple. Or, le nombre de courbes de charge à disposition ne cesse de croître.

Depuis plusieurs années, la Division Recherche et Développement d'EDF étudie ces courbes de charge, notamment au moyen de techniques de classification (voir par exemple [CHA96] et [BOU97]). Cependant, l'ouverture du marché nous incite à passer d'une approche " étude ad hoc " à une approche " outil ", afin de donner aux services commerciaux les moyens d'effectuer eux-mêmes ces études. L'intérêt d'une telle démarche est double :

Elle permet d'effectuer des analyses à plus grande échelle, en multipliant le nombre d'analystes potentiels,

Elle tire meilleur parti de l'expertise des services commerciaux sur leurs données, en les intégrant dans l'analyse.

Or, il n'existe actuellement pas d'outil, à notre connaissance, permettant de combiner une analyse statistique interactive avec une visualisation conviviale des résultats et une aide à l'interprétation de ces résultats par l'expert. C'est dans ce contexte qu'a été conçu et développé le logiciel Courboscope.

2. Le Courboscope, un outil d'analyse exploratoire de courbes

2.1. Principe

Le logiciel Courboscope a été développé à la Division Recherche et Développement d'EDF pour permettre aux experts des données, non spécialistes en statistique, de mener une analyse

exploratoire sur leur ensemble de courbes, de manière interactive. Le Courboscope effectue une classification automatique sur les données, mais il permet également de prendre en compte l'expertise humaine dans une étape d'interprétation et d'affinement de la classification obtenue. Une attention particulière a été portée à l'ergonomie de l'outil et à la lisibilité des résultats de la classification, ce qui a motivé en partie le choix de l'algorithme de Kohonen.

L'outil a été initialement conçu pour l'analyse des courbes de charges des clients d'EDF, mais il s'adapte également aux problématiques de divers corps de métiers d'EDF (courbes de débit de réseau informatique, mesures électriques, etc)

2.2. Description générale du Courboscope

Ce logiciel a été développé en C++ pour tous les calculs numériques, et en TCL/TK pour l'interface utilisateur. Il fonctionne sur PC/Windows (95, 98, NT).

2.2.1. Les jeux de données

Les jeux de données se composent :

- d'un ensemble de courbes, décrites par une suite de valeurs.
- d'un ensemble de caractéristiques permettant d'interpréter ces courbes : les attributs

a) L'ensemble de courbes :

Les courbes du jeu de données doivent pouvoir être comparées entre elles, c'est-à-dire représenter la même mesure aux mêmes points du temps. Des pré-traitements peuvent être nécessaires pour améliorer les résultats de la classification : normalisation, lissage, resynchronisation, etc.

Lorsqu'on souhaite analyser une courbe très longue, on procède à un découpage de cette courbe en " épisodes " significatifs, puis on étudie ces épisodes à l'aide du Courboscope. Ainsi l'étude de la consommation d'un client peut se faire en découpant sa courbe de charges en périodes journalières, et en définissant quelques formes-types de consommation journalière. Il est également possible de découper cette consommation en période hebdomadaire ou mensuelle, en fonction du type d'analyse souhaité.

b) Les attributs :

Les attributs sont des informations qui n'interviendront pas dans la classification, mais qui peuvent aider l'utilisateur à interpréter les courbes. Il peut s'agir d'informations liées à la

période d'extraction (température, type de jour, saison, etc), ou à l'individu lui-même : pour des analyses multi-clients, le type d'activité, la région, le tarif, les usages de l'électricité peuvent par exemple être retenus.

Les attributs doivent être nominaux ou ordinaux (un module d'aide à la constitution des jeux de données permet de découper en intervalles les attributs continus).

2.2.2. Les fonctionnalités

Une analyse avec le Courboscope se fait en deux étapes :

- Une étape de classification automatique des courbes, basée sur l'algorithme de Kohonen,
- Une étape d'aide à l'interprétation de la classification obtenue, à l'aide notamment de tests d'indépendance sur les caractéristiques explicatives des courbes (attributs)

L'interprétation de la classification obtenue consiste :

- D'une part, à associer à chaque classe de Kohonen un libellé et éventuellement un commentaire, après une interprétation minutieuse du contenu de ces classes, tant en terme de forme des courbes qu'au regard des caractéristiques externes (attributs),

- D'autre part, à définir un deuxième niveau de classification, en regroupant les classes de Kohonen proches en termes de formes et/ou d'attributs en « super-classes ». Celles-ci pourront également se voir affecter un libellé et un commentaire par l'utilisateur.

Pour effectuer cette interprétation, le Courboscope met à disposition de l'utilisateur un ensemble de fonctionnalités, parmi lesquelles :

- un zoom sur une ou plusieurs classe(s), permettant de visualiser l'ensemble des courbes classées dans ce (ces) classe(s),

- l'élimination des courbes ou des classes atypiques perturbant l'analyse,

- une modification de la taille de la carte de Kohonen (i.e. du nombre de classes),

- la projection des classes de Kohonen sur le premier plan ACP, pour visualiser les distances entre classes, les classes atypiques, les défauts éventuels de classification (vrilles, retournements),

- la visualisation des corrélations entre les classes de Kohonen et les caractéristiques externes (attributs).

- le regroupement des classes de Kohonen en super-classes, au choix de l'utilisateur, avec possibilité d'initialisation automatique.

2.2.3. Les résultats

Le résultat d'une analyse avec le Courboscope est une carte interprétée : chaque " case " de la carte représente une classe de courbes homogènes, caractérisée par la courbe moyenne et les courbes des écarts-types, et par un libellé et un commentaire renseigné par l'utilisateur au terme de son analyse.

Cette carte interprétée, contenant l'expertise de l'analyste, peut ensuite être transférée à des utilisateurs plus " opérationnels ". Ceux-ci disposeront d'un module permettant de projeter leurs propres courbes sur la carte interprétée, afin d'en induire de l'information.

3. La classification automatique

3.1. Principe

La carte auto-organisatrice de Kohonen (ou algorithme de Kohonen, voir [KOH95]) est une méthode de classification non hiérarchique apparentée aux réseaux de neurones. Elle s'approche d'une méthode des K-Means avec contrainte de proximité. Le résultat est présenté sous forme de carte de forme et de dimension paramétrables, dans laquelle les classes produites sont organisées selon des critères de proximité. Ainsi, des données (ici, des courbes) de profils proches seront soit dans une même classe, soit dans des classes proches sur la carte.

Cette méthode de classification présente deux avantages majeurs pour le traitement des courbes :

- des temps de calcul très satisfaisants, grâce à une complexité de l'algorithme en n, alors que la classification hiérarchique est en n^2 (n étant le nombre de courbes) ;

- une présentation très lisible du résultat de la classification, grâce à la contrainte de proximité, même pour un grand nombre de classes. Cette lisibilité est essentielle pour une validation et une interprétation par l'utilisateur. Par ailleurs, elle permet de

représenter un nombre de classes beaucoup plus important que des classifications de type hiérarchique, ce qui favorise une analyse microscopique du jeu de données. Ainsi, une carte de Kohonen à 100 classes reste lisible alors qu'un algorithme sans propriété topologique rend une lecture des résultats difficile au-delà d'une dizaine de classes.

3.2. Algorithme

L'algorithme est itératif :

- Le nombre de classes et la forme de la carte sont fixés a priori (dans le Courboscope, la forme est rectangulaire et la taille est laissée au choix de l'utilisateur)

- On associe à chaque classe une courbe " neurone " choisie en théorie de manière aléatoire. Dans le Courboscope, on initialise les neurones des classes par Analyse en Composantes Principales (ACP)[1]

- On effectue N passages du jeu de données. A chaque passage :

- on choisit une observation au hasard (i.e. une courbe du jeu de données),

- on la compare à tous les neurones,

- on détermine la classe gagnante, c'est-à-dire celle dont le neurone est le plus proche au sens d'une distance donnée a priori (ici, la somme des carrés des distances point à point) ;

- on rapproche alors de l'observation le code de la classe gagnante et ceux des classes voisines par un gradient.

---

[1]L'initialisation par ACP (voir [G. Saporta, 1990]) permet :
- d'accélérer la convergence de l'algorithme (et donc les temps de traitement),
- d'obtenir une reproductibilité de la classification obtenue. En effet, l'algorithme de Kohonen est non déterministe : si on le relance plusieurs fois, on peut obtenir une carte différente (optima locaux, retournements), ce qui est troublant pour l'utilisateur. L'ACP remédie à ce problème.

4. Visualisation des courbes

4.1. Fenêtre principale : la carte de Kohonen

Le résultat de la classification de Kohonen se présente sous forme d'une carte de taille paramétrable (entre 4 et 100 classes). L'utilisateur a ainsi un aperçu immédiat de son jeu de données.

Dans l'exemple de la **figure 1**, la consommation d'électricité d'un client a été télérelevée toutes les 10 minutes pendant 1 an. Cette courbe annuelle a été découpée en 364 courbes journalières. L'utilisateur a choisi une classification en 16 classes. La carte de Kohonen fait apparaître les courbes représentant des jours de faible consommation sur la gauche, et les courbes de plus forte consommation sur la droite.

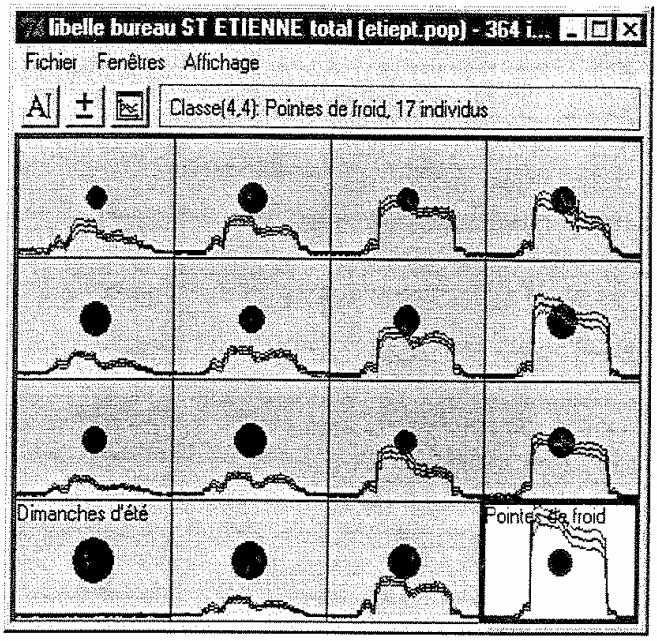

Figure 1

Pour chaque classe, les informations suivantes sont affichées :

- sa courbe moyenne
- les courbes représentant +/- 1,5 écart-types, indiquant la dispersion des courbes dans la classe,
- son effectif, représenté par un disque noir de taille proportionnelle.

Par ailleurs, l'utilisateur peut ajouter un libellé et un commentaire sur chaque cellule. Le libellé apparaît alors sur la carte (cf. cellule en bas à gauche, **figure 1** : « Dimanche d'été »)

4.2. La fenêtre zoom

Cette fenêtre affiche le détail (i.e. l'ensemble des courbes) de la sélection effectuée dans la fenêtre principale. La sélection comporte une ou plusieurs classes, ou une super-classe. Chaque courbe de la fenêtre zoom peut être sélectionnée à l'aide de la souris. Cela permet :

- de la faire apparaître en blanc, et donc de distinguer sa forme parmi l'ensemble des courbes de la sélection,

- de voir afficher son libellé dans le bandeau d'en-tête de la fenêtre,

- de la désactiver par simple clic de la souris. Elle apparaîtra dans la " corbeille " et sera ignorée lorsque l'utilisateur relancera la classification.

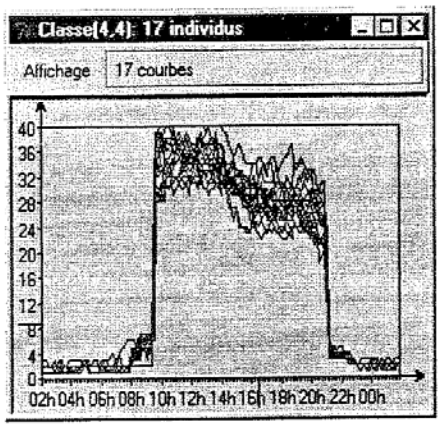

Figure 2

La fenêtre zoom permet également de mieux visualiser la dispersion des courbes de la sélection, d'obtenir la liste des individus de cette sélection, de tous les désactiver ou les réactiver.

4.3. La fenêtre des distances

Elle est affichée sur demande et permet de visualiser la topologie réelle de la carte réduite en deux dimensions (cf. **figure 3**).

La fenêtre des distances représente chaque classe de Kohonen par sa courbe prototype qu'elle projette sur le premier plan principal de l'ACP. Cette fenêtre est liée avec la carte de Kohonen. Ainsi, la sélection d'une classe sur la carte de Kohonen a pour effet d'afficher en blanc le point la représentant dans la fenêtre des distances. Les classes voisines de la classe sélectionnée apparaissent également en rouge. Inversement, la sélection d'un point dans la fenêtre des distances entraîne son affichage en blanc sur la carte de Kohonen.

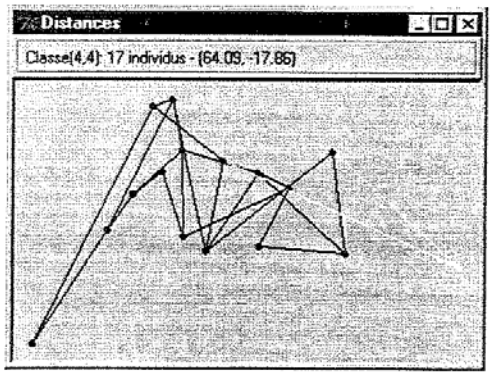

Figure 3
la classe en bas à droite a été sélectionnée dans la **figure 1**

5. Interprétation de la carte à l'aide des caractéristiques externes

Outre l'analyse visuelle à partir de la forme des courbes décrite au paragraphe précédent, le Courboscope permet à l'utilisateur d'interpréter les classes de Kohonen à l'aide de caractéristiques externes n'ayant pas pris part à la classification (les attributs). Ainsi, dans l'exemple de la **figure 1**, la classe en bas à gauche présentant des courbes plates se révèle correspondre aux courbes de charge des dimanches. De même, l'analyse de l'attribut « température » met en évidence l'effet température sur les consommations du client analysé.

5.1. Liste des attributs caractéristiques

Pour chaque attribut du jeu de données, le Courboscope effectue un test statistique (test du Chi2) pour déterminer s'il semble jouer un rôle dans le regroupement des individus en classes. Puis il associe à chaque attribut un signe (de rien à ++++) en fonction de son " pouvoir explicatif " potentiel. L'utilisateur y a accès par menu, par ordre décroissant de pouvoir explicatif.

Le principe consiste à déterminer, pour chaque attribut, si la répartition de ses modalités dans chaque classe est significativement différente d'une répartition aléatoire (voir [DER96], [AGR89] et [MOR84]. On utilise pour cela un test de comparaison d'échantillons (les J classes de la carte de Kohonen) pour chaque variable (attribut) à I modalités :

On calcule l'effectif théorique de chaque modalité dans chaque classe, qui représente une distribution sous l'hypothèse que les échantillons (i.e. les classes) proviennent de la même population;

On calcule, pour chaque classe et chaque modalité, l'écart entre effectif théorique et effectif réel

Connaissant la loi de distribution théorique de cet écart, on peut définir le pourcentage de chance qu'il soit dû à des classes significativement différentes pour la variable considérée.

5.2. Analyse des distributions de chaque attribut

Pour chaque attribut sélectionné par l'utilisateur, le Courboscope affiche un diagramme à bâtons lié à la carte de Kohonen (cf. **figure 4**).

Quand un utilisateur sélectionne une (ou des) classe(s) sur la carte de Kohonen, chaque diagramme à bâton représente la distribution de l'ensemble du jeu de données, en blanc, en comparaison à la distribution de la (des) classe(s) sélectionnée(s), en rouge. Ainsi, dans l'exemple de la **figure 4**, considérant l'attribut " température ", on observe que 41% des courbes de la classe sélectionnée (classe en bas à droite de la **figure 1**) appartiennent à la tranche " -2 à 0° ", alors que cette tranche de température ne représente que 5% de l'ensemble des courbes.

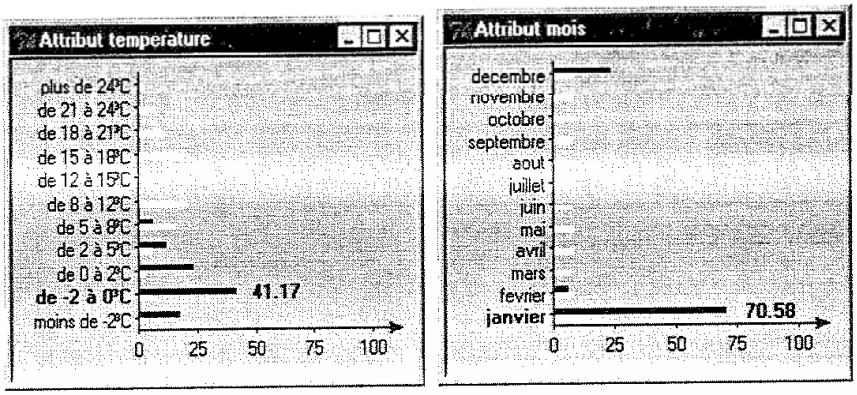

Figure 4

Inversement, l'utilisateur peut sélectionner une (ou plusieurs) modalité(s) sur un diagramme à bâtons :

- Dans les diagrammes à bâtons des autres attributs, les modalités liées apparaissent en vert. Elles correspondent à la distribution de la sous-population possédant la modalité sélectionnée dans le diagramme à bâtons actif ;

- Sur la carte de Kohonen, les classes caractéristiques et anti-caractéristiques de cette modalité sont représentées par un rectangle rouge (resp. bleu) dont la taille est proportionnelle au niveau de signification du test statistique utilisé, pour la (les) modalité(s) dans la classe.

Ainsi, dans l'exemple de la **figure 5**, on constate que les classes de droite sur la carte de Kohonen sont caractéristiques des températures froides, alors que les classes de gauche sont anti-caractéristiques de ces températures.

On constate également qu'elles correspondent à des niveaux de puissance élevées (cf. attribut « niveau pui »).

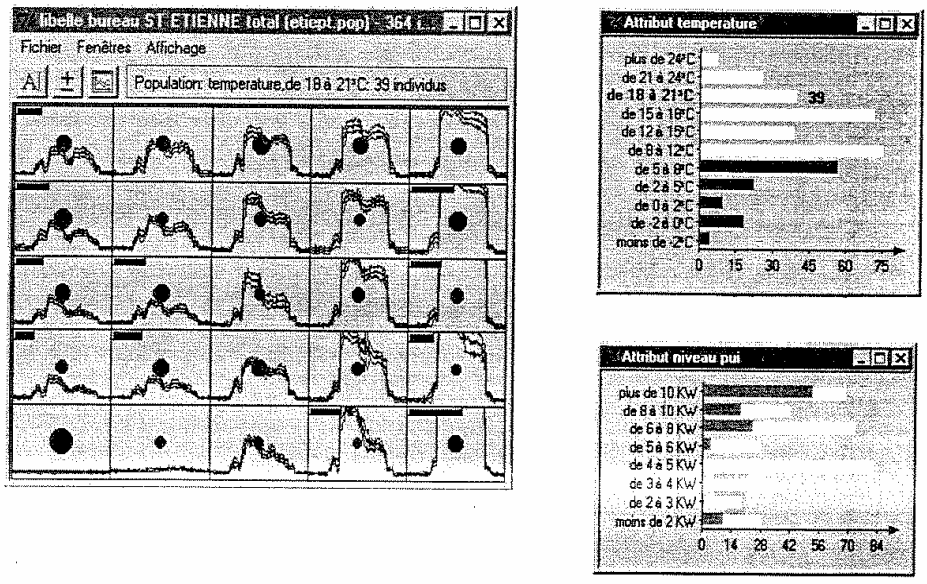

Figure 5

En se positionnant sur une classe de la carte de Kohonen, l'utilisateur peut également accéder à la liste ordonnée de toutes les modalités caractéristiques et anti-caractéristiques de la classe, chaque modalité étant associée à un signe indiquant son niveau de signification.

Une modalité est dite caractéristique (resp. anti-caractéristique) d'une classe si elle est représentée de manière anormalement élevée (resp. rare) dans la classe, au sens statistique. Le principe consiste, pour chaque modalité de chaque variable et pour chaque classe, à comparer à l'aide d'un test statistique la proportion de la modalité *j* dans la classe *i* et la proportion de la modalité *j* dans l'ensemble de la population (voir [DER96], [AGR89] et [MOR84].

6. Les super-classes : un regroupement interactif des classes

Le logiciel offre à l'utilisateur la possibilité de regrouper de manière interactive les classes de Kohonen dans des " super-classes " disjointes, visualisables sur la carte de Kohonen par des couleurs.

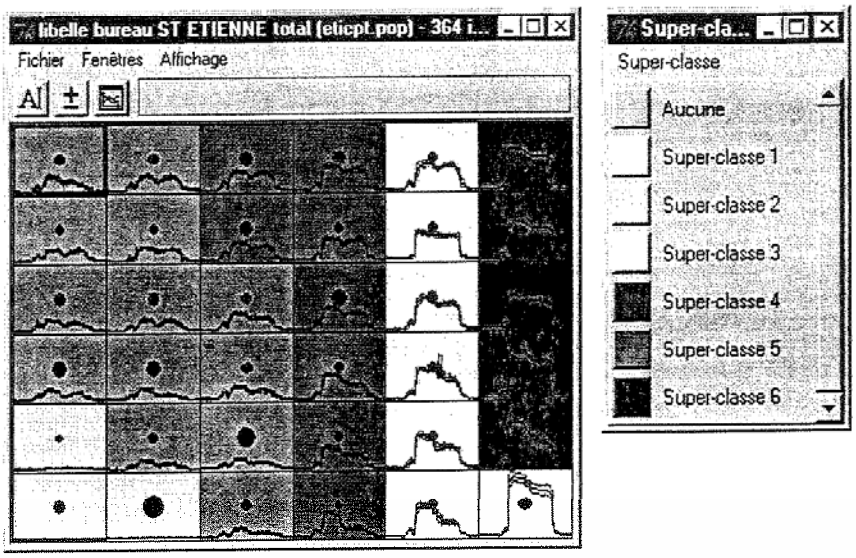

Figure 6
la carte de Kohonen comporte 36 classes. L'utilisateur a demandé la création de 6 super-classes selon la forme des courbes.

Cette classification de deuxième niveau peut être effectuée manuellement par l'utilisateur, qui crée les super-classes, leur attribue un libellé et une couleur, puis les construit en peignant certaines classes sur la carte de Kohonen avec la couleur de la super-classe (cf. **figure 6**).

L'utilisateur peut également demander l'initialisation automatique des super-classes, par deux méthodes alternatives :

Soit par une classification hiérarchique des neurones des classes (approximativement les courbes moyennes). Le logiciel conseille alors l'utilisateur dans le choix du nombre de super-classes (à l'aide du critère de Ward). Les attributs ne sont pas pris en considération.

Soit à partir de la distribution d'un attribut sélectionné par l'utilisateur. Le logiciel crée alors une super-classe par modalité de l'attribut sélectionné et lui associe toutes les classes pour lesquelles cette modalité (et uniquement celle-là) est caractéristique. Les classes n'ayant aucune modalité caractéristique pour l'attribut sélectionné ne sont pas affectées et restent grisées.

L'utilisateur peut ensuite modifier les super-classes ainsi initialisées (changement de libellé ou de couleur, réaffectation de classes).

Ces fonctionnalités permettent à l'utilisateur de construire rapidement des super-classes basées soit sur la forme des courbes, soit sur la valeur d'un attribut.

Les libellés associés aux super-classes peuvent ensuite remplacer ou s'ajouter aux libellés des classes, et apparaître sur la carte de Kohonen.

7. Le module de classement

Une version prototype d'un module de classement a également été développée. Ce module prend en entrée la carte de Kohonen interprétée, réalisée à l'aide du module précédemment décrit, et propose les fonctionnalités suivantes :

- une visualisation de la carte de Kohonen importée, chaque classe étant représentée par sa courbe moyenne,

une projection, sur cette carte, d'un ensemble de nouvelles courbes. Chacune de ces nouvelles courbes est affectée à la classe dont la courbe moyenne est la plus proche.

- L'effectif des nouvelles courbes affectées à chaque classe apparaît sous forme d'un disque de taille proportionnelle,

- une projection, sur chaque classe de la carte, d'une courbe particulière, avec indication par des couleurs des 5 classes les plus proches. Cela permet à l'utilisateur de valider visuellement l'affectation de la courbe aux classes, au regard de sa connaissance des courbes.

Ce module s'adresse à des utilisateurs opérationnels, à la différence du module précédemment décrit, plutôt destiné à des profils d'analystes. Il pourrait être utilisé par un négociateur commercial, par exemple, pour projeter les courbes de ses clients sur une carte associant à chaque classe un conseil tarifaire (carte qui serait fournie par les services marketing).

8. Conclusion et perspectives

Le Courboscope a été utilisé à la Division R&D pour diverses études. Appliqué à l'analyse des courbes de charges, il a été utilisé soit pour l'étude des courbes journalières d'un client, soit pour l'étude des différents sites d'un grand client, soit pour l'étude d'un portefeuille de clients. Citons par exemple les études suivantes :

- définition de comportements-type pour l'aide à la tarification d'un client éligible,

- définition des profils-type hebdomadaires pour un client multi-sites, puis reconstitution des courbes de charges d'un site particulier à partir de ces profils-type et de caractéristiques du client (données facturaires par exemple),

- visualisation des quelques courbes-type journalières d'un client, pour le négociateur commercial.

Le Courboscope a permis également d'étudier le comportement d'un réseau informatique, en analysant des mesures telles que les taux de charge, les taux d'erreur, de collision en regard d'informations sur les équipements et l'organisation des équipes par exemple.

Il a facilité l'exploitation d'une base de mesures électriques temporelles résultant de la simulation d'incidents sur le réseau électrique, en permettant de sélectionner et d'interpréter quelques mesures représentatives.

Aux vues de ces premiers retours d'expérience, les fonctionnalités proposées (analyse exploratoire et interactive des courbes) semblent répondre à un réel besoin. Une version industrielle a été élaborée pour une diffusion plus large, tant au sein de la Division R&D que dans les services commerciaux, mais également à l'externe d'EDF.

Des études autour de l'analyse symbolique sont menées en amont du Courboscope pour aider l'utilisateur à découper les courbes en " épisodes " optimaux, et pour classifier des courbes très longues.

REFERENCES

[AGR89] A. Agresti, 1989, Ed. Wiley & Sons, « Categorical Data Analysis »,.

[BOU97] E. Boudaillier, et G. Hébrail, 1997. Interactive Interpretation of Hierarchical Clustering. *Proceedings of PKDD'97, Note de lecture dans* Artificial Intelligence, Principles of Data Mining and Knowledge Discovery, 288-298, Springer.

[CHA96] D. Chantelou, G. Hébrail, and C. Muller, 1996. Visualizing 2,665 Electric Power Load Curves on a Single A4 Sheet of Paper. *Proceedings of the ISAP'96 Conference*, Orlando (Florida).

[DER96] C. Derquenne, 1996, *EDF-DER*, « Caractérisation statistique d'une typologie : la nouvelle macro QUALICLS » ,

[KOH95] T. Kohonen, 1995. *Self-organizing Maps*. Berlin: Springer.

[MOR84] A. Morineau, 1984. Note sur la Caractérisation Statistique d'une Classe et les Valeurs-test. Bulletin du CESIA, Vol.2, N°1-2, Paris.

[SAP90] G. Saporta, 1990 « Probabilités, analyse de données et statistique », Ed. Technip.