Une approche de classification non supervisée basée sur la détection de singularités et la corrélation de séries temporelles pour la recherche d'états: application à un bioprocédé fed-batch

Sébastien Régis

G.R.I.M.A.A.G.
Université Antilles-Guyane
Campus de Fouillole 97159 Pointe-à-Pitre Guadeloupe, France
sregis@univ-ag.fr
http://grimaag.univ-ag.fr/

Résumé. Nous proposons dans cet article une méthode de clustering qui combine l'analyse dynamique et l'analyse statistique pour caractériser des états. Il s'agit d'une méthode de fouille de données qui travaille sur des ensembles de séries temporelles pour détecter des états; ces états représentent les informations les plus significatives du système. L'objectif de cette méthode non supervisée est d'extraire de la connaissance à partir de l'analyse des séries temporelles multiples. Elle s'appuie sur la détection de singularités dans les séries temporelles et sur l'analyse des corrélations des séries entre les intervalles définis par ces singularités. Pour l'application présentée, les séries temporelles sont des signaux biochimiques mesurés durant un bioprocédé. Cette approche est donc utilisée pour confirmer et enrichir la connaissance des experts du domaine des bioprocédés sans utiliser la connaissance *a priori* de ces experts. Elle est appliquée à la recherche d'états physiologiques dans un bioprocédé de type fed-batch.

1 Introduction

L'extraction d'informations à partir de séries temporelles est un domaine de plus en plus important dans la fouille de données. En effet de nombreuses applications utilisant des signaux réels nécessitent l'utilisation des outils de ce champ d'étude. Ainsi, la fouille de données dans des ensembles de séries temporelles utilise divers outils pour extraire des informations intrinsèques ou d'interdépendance entre ces séries temporelles : recherche de similarités entre séries Marteau et al. (2006), corrélation entre séries Pelletier (2005), ou classification Nunez et al. (2002) par exemple. Les domaines d'application pour ces outils sont très variés : finance Takada et Bass (1998), données météorologiques Harms et Deogun (2004), données médicales Summa et al. (2006), biotechnologies, etc.
Par ailleurs les biotechnologies et plus particulièrement les bioprocédés offrent aujourd'hui des défis importants concernant l'extraction et la gestion des connaissances. L'analyse des états physiologiques survenant durant ces bioprocédés est un point essentiel pour leur contrôle et

leur optimisation. Ces bioprocédés ont longtemps utilisé des approches à base de modèles mathématiques. En effet les bioprocédés forment un système complexe de réactions biologiques qui peuvent être décrites par un système d'équations dynamiques non-linéaires. Mais bien que ces modèles aient montré leur utilité, ils ne peuvent gérer des situations inattendues, car ils sont basés sur des simulations et ne tiennent pas toujours compte de tous les variables réelles mesurées lors des bioprocédés. Les méthodes qui ne sont pas basées sur des modèles sont de plus en plus utilisées. Parmi ces méthodes, beaucoup s'appuient sur l'analyse de signaux biochimiques (qui sont des séries temporelles) mesurés durant le bioprocédé. Ces méthodes se basent notamment sur des techniques de classification Regis et al. (2003), d'apprentissage ou de règles d'expert de type "if-then" Steyer (1991), Steyer et al. (1991). Cependant, si ces approches fournissent de bons résultats dans le cas où les états du bioprocédé sont bien connus (comme par exemple pour des bioprocédés de type "batch" [1], voir Régis (2004), Roels (1983)), elles sont moins performantes dans le cas des bioprocédés pour lesquels on ne connaît pas parfaitement tous les états qui surviennent (c'est le cas des bioprocédés de type "fed-batch" voir Régis (2004), Roels (1983)). L'approche que nous proposons cherche à répondre à la problématique suivante :

Peut-on caractériser les états d'un système en s'appuyant sur l'analyse dynamique et statistique de séries temporelles ?
Dans le cadre de l'application sur les bioprocédés, est-il possible d'extraire de l'information concernant les états physiologiques d'un bioprocédé de type fed-batch, à partir des séries temporelles mesurées pendant l'expérience, et en utilisant peu ou pas de connaissances *a priori* des experts du domaine ?

Nous proposons d'utiliser une méthode de clustering qui combine l'analyse des propriétés dynamiques et des propriétés statistiques pour détecter et caractériser les états physiologiques d'un bioprocédé fed-batch. L'analyse dynamique consiste à détecter et sélectionner les singularités présentes dans les séries temporelles en utilisant la méthode du maximum du module de la transformée en ondelettes Mallat et Zhong (1992); Mallat et Hwang (1992) et l'évaluation du coefficient de Hölder (aussi appelé exposant de Hölder ou exposant de Lipschitz). Ces singularités correspondent aux frontières des différents états. L'analyse statistique consiste à calculer les corrélations des différentes séries temporelles entre le début et la fin d'un état afin de caractériser chaque état. On rappelle que ces séries temporelles représentent pour cette application des variables biochimiques mesurées durant l'expérience.
Il faut noter que l'approche que nous proposons est suffisamment générique pour être utilisée sur n'importe quel ensemble de séries temporelles ou n'importe quel système de flux de données, quelque soit le domaine d'application concerné. Cependant, pour une utilisation pertinente de cette approche, des corrélations implicites ou explicites doivent exister entre les séries temporelles, faute de quoi les résultats obtenus auront peu de sens et ne seront pas interprétables.
Le plan de cet article est le suivant.

[1] le bioprocédé fermentaire batch est un bioprocédé dans lequel le substrat est placé au début du procédé. Aucune intervention extérieure n'est faite jusqu'à la fin de l'expérience. Les états physiologiques apparaissant durant ce type de procédé sont parfaitement connus des microbiologistes. Par contre dans le bioprocédé de type fed-batch, des interventions sont réalisées pendant toute l'expérience. En particulier, les micro-organismes sont alimentés en substrat pendant toute l'expérience. Les états physiologiques ne sont pas tous connus dans ce type de procédé.

Nous présentons, dans le paragraphe 2 l'application biotechnologique, et dans le paragraphe 3, les travaux existants liés à cette application qui ont motivé la mise en place de la méthode proposée. Dans la section 4, nous présentons en détails l'approche que nous proposons. Enfin nous présentons des résultats expérimentaux dans le paragraphe 5 avant de conclure dans le paragraphe 6.

2 Les bioprocédés

Les bioprocédés sont des procédés industriels ou expérimentaux utilisant des micro-organismes dans le but de fabriquer des produits biochimiques (produits pharmaceutiques, biocarburants) ou de produire de la biomasse. Durant ces bioprocédés, de nombreuses variables biochimiques sont mesurées. Certaines peuvent être contrôlées tandis que d'autres expriment la biologie du système. Ces variables représentent essentiellement des gaz (dioxygène, dioxyde de carbone, etc.), des éléments rajoutés ou produits dans le milieu (substrat, base, biomasse, etc.) ou des variables physiques régulées (par exemple la variable agitation : c'est la vitesse de rotation des pales d'un moteur assurant l'homogénéité du milieu). Pour des détails sur la fonction biochimique des variables on peut se référer à Régis (2004) ou dans une moindre mesure à Régis et al. (2004a). Les mesures sont effectuées de manière régulière avec des capteurs en ligne ayant la même fréquence pour tous les variables. Ces variables biochimiques sont représentées sous forme de séries temporelles qui évoluent au cours du temps et expriment la dynamique du système pendant le procédé. Ces séries temporelles sont cruciales car elles permettent de comprendre les phénomènes biochimiques et physico-chimiques et par conséquent d'optimiser le procédé. Ainsi l'utilisation d'outils de classification et d'extraction d'informations permet de caractériser de façon automatique les états du système. La classification consiste à segmenter les séries temporelles de telle sorte qu'une classe corresponde à un état physiologique donné.
Pour effectuer l'analyse de ces séries temporelles, il convient de faire quelques remarques sur leurs propriétés (voir Régis (2004)). Ainsi il faut noter que :

- ce ne sont pas des signaux périodiques. Ces signaux traduisent des phénomènes biologiques faisant intervenir des micro-organismes. Ces micro-organismes n'ont pas de comportement physiologique périodique.
- ce ne sont pas des signaux pairs ou impairs. Il n'y pas de symétrie ou d'antisymétrie dans ces phénomènes biologiques.
- ce sont des signaux d'énergie finie. Ils n'ont pas de pics qui tendent vers l'infini. Ce sont des séries temporelles basse fréquence, elles n'ont pas des singularités oscillantes indéfiniment.
- ce ne sont pas -pour la plupart en tout cas- des signaux déterministes. Certains signaux physiques régulés peuvent néanmoins être décrits par des formules mathématiques mais la plupart dépend de la biologie et il est impossible à l'heure actuelle de trouver une formule mathématique décrivant parfaitement l'activité liée à ces variables biologiques.
- ce ne sont pas des signaux stationnaires.
- aucune hypothèse n'est faite concernant un quelconque bruit présent dans ces signaux. En effet, on suppose que les signaux ne sont pas bruités ou le sont très peu, en raison du dispositif matériel utilisé pour les capteurs.

Un exemple de ces variables biochimiques est donné sur la figure 1.

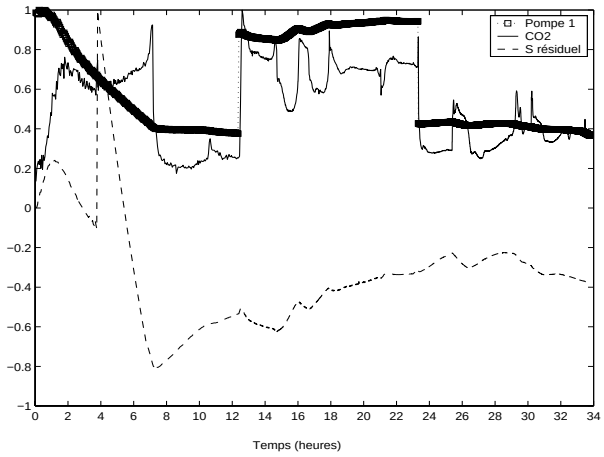

FIG. 1 – *Quelques-unes des variables biochimiques. L'axe des abscisses représente le temps en heure, l'axe des ordonnées représente les amplitudes des signaux. Les signaux ont été normalisés sur la figure pour une meilleure visualisation. La variable CO2 représente le dioxyde de carbone mesuré durant le bioprocédé, S résiduel est le substrat résiduel et la variable pompe 1 est une variable régulée correspondant à une pompe injectant un élement chimique dans le milieu.*

3 Travaux existants et motivation

Plusieurs travaux sur la détection des états ont montré que les singularités des signaux mesurés durant un procédé (que ce procédé soit biochimique ou physico-chimique) correspondent au début et à la fin d'un état du système. Ainsi plusieurs auteurs utilisant des méthodes très différentes les unes des autres sont arrivés à la conclusion que ces singularités représentaient les limites des états : c'est le cas de Steyer et al. (1991) (en utilisant la logique floue et un système expert), Bakshi et Stephanopoulos (1994) (en utilisant les ondelettes et un système expert) et Doncescu et al. (2002) (en utilisant la logique inductive). De plus, Jiang et al. (2003) se basent aussi sur cette assertion pour détecter le début et la fin d'un état.
Cette hypothèse n'est pas seulement valable pour les procédés chimiques ou biochimiques. En effet dans la plupart des applications réelles dans lesquelles on utilise des séries temporelles décrivant des phénomènes non stationnaires, les singularités représentent des informations significatives. Ainsi par exemple, Struzick montre que les singularités sont significatives dans des séries temporelles issues de la finance Struzik (2000) ou de données médicales Struzik (2003).

Par ailleurs, la détection des états peut être suivie d'une phase de caractérisation automatique de ces états. Une des particularités des bioprocédés de type fed-batch est qu'un état peut apparaître plusieurs fois et à des moments différents de l'expérience. Il est donc nécessaire de caractériser ces états pour savoir s'ils réapparaissent au cours du temps. Plusieurs méthodes statistiques ont été proposées pour caractériser ces états. Par exemple, des méthodes de clas-

sification basées sur l'Analyse en Composantes Principales (ACP) Ruiz et al. (2004), l'ACP adaptative Lennox et Rosen (2002), ou la kernel ACP Lee et al. (2004) permettent de distinguer et de caractériser les différents états d'un procédé.
Nous proposons d'étudier la corrélation des variables biochimiques entre les intervalles temporels définis par les singularités, afin de caractériser les états du système. On tient compte ainsi de l'évolution dynamique des corrélations à l'image de ce qui est proposé dans d'autres approches Nunez et al. (2002), Pelletier (2005) utilisées pour d'autres applications du data mining.
La méthode du maximum du module de la transformée en ondelettes Mallat et Zhong (1992); Mallat et Hwang (1992) est utilisée dans cet article pour détecter les singularités des signaux (afin de définir les limites d'un état) et l'étude des coefficients de corrélation entre ces signaux permet de caractériser ces différents états.

4 Description de la méthode proposée

La méthode que nous présentons se compose de deux étapes présentées dans les sous paragraphes suivants :

1. la détection et la sélection des singularités des séries temporelles, par les ondelettes et le coefficient de Hölder. Ces singularités correspondent aux frontières des différents états.
2. la caractérisation et la classification des états par la corrélation. Les corrélations sont calculées sur chaque intervalle temporel défini par les singularités.

4.1 Détection et sélection des singularités par les ondelettes et l'exposant de Hölder

Les singularités de séries temporelles (ici des signaux biochimiques) correspondent aux limites d'un état. Plusieurs méthodes utilisent les ondelettes pour trouver ces singularités afin de détecter les états : par exemple, Bakshi et Stephanopoulos (1994) et plus récemment Jiang et al. (2003). Ces singularités correspondent à des maxima des coefficients de la transformée en ondelettes des signaux biochimiques. Jiang et al. Jiang et al. (2003) proposent de sélectionner les maxima en utilisant un seuil. Cependant le choix de ce seuil reste empirique.
Nous proposons donc d'utiliser l'exposant de Hölder pour sélectionner ces maxima. Une singularité au point x_0 est caractérisée par son coefficient de Hölder. Ce coefficient est définie comme étant le plus grand exposant α vérifiant l'inégalité suivante :

$$|f(x) - P_n(x - x_0)| \leq C|x - x_0|^\alpha \tag{1}$$

où $P_n(x-x_0)$ est le développement de Taylor de f au voisinage de x_0 et où $n \leq \alpha < n+1$. La valeur de cet exposant de Hölder change en fonction de la nature de la singularité (voir Mallat et Hwang (1992)) et il est possible d'évaluer numériquement la valeur du coefficient de Hölder d'une singularité à partir des coefficients en ondelettes correspondants Jaffard (1989, 1997). Le coefficient de Hölder permet de caractériser le type de singularité comme cela est illustré sur les exemples de la figure 2 et dans le tableau 1.

Ainsi il est possible de sélectionner les singularités les plus significatives à partir de leur exposant de Hölder. Plusieurs méthodes de calcul du coefficient de Hölder existent : ces méthodes sont soit basées sur des régressions linéaires Mallat et Hwang (1992), soit basées sur

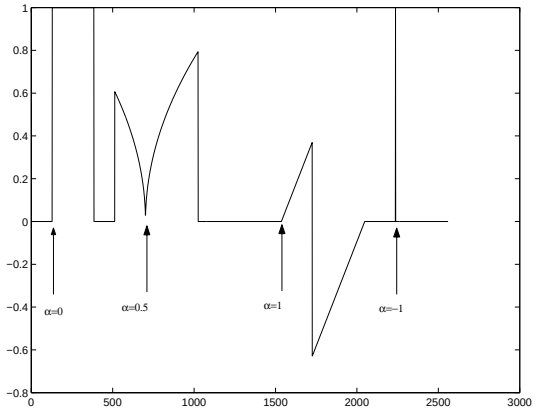

FIG. 2 – *Sur ce signal, les valeurs des coefficients de Hölder théoriques α sont donnés en fonction du type de singularités. Les singularités représentées sont respectivement (de gauche à droite) un palier, une singularité dite de type "cups", une rampe et un dirac.*

type de singularités	coefficient de Hölder
dirac	-1
palier	0
"cups"	0,5
rampe	1

TAB. 1 – *Tableau des valeurs du coefficient de Hölder en fonction du type de singularité.*

l'optimisation d'une fonction de coût Mallat et Zhong (1992). Une méthode utilisant les algorithmes génétiques a été proposée pour optimiser la fonction de coût, et semble fournir des résultats plus précis que les méthodes classiques Manyri et al. (2003).
La sélection des singularités est réalisée en fonction des attentes des experts en microbiologie. Ainsi pour les bioprocédés de type fed-batch, les experts privilégient les singularités brusques (pic, saut, palier, etc.) qui correspondent à des singularités dont les coefficients de Hölder sont inférieurs à 1. La détection des singularités par les ondelettes et l'évaluation du coefficient de Hölder ont déjà été testées dans un bioprocédé de type fed-batch pour une application proche de celle présentée dans cet article (voir Régis et al. (2004b)).

4.2 Caractérisation et classification des états par corrélation

Les singularités détectées et sélectionnées définissent les bornes d'intervalles temporels : on obtient ainsi une segmentation temporelle du bio-procédé en plusieurs intervalles de temps (un intervalle de temps représentant *a priori* un état). Nous proposons de caractériser chaque intervalle par les signes des différents coefficients de corrélations linéaires calculés deux à deux entre tous les variables.
En effet, les règles de la logique floue du type "if-then" décrivent les relations entre les va-

riables biochimiques du point de vue de l'expert en microbiologie Steyer et al. (1991); Steyer (1991). Nous avons déjà signalé que ces règles s'appliquaient parfaitement aux bioprocédés de type batch mais qu'elles rencontraient certaines difficultés face aux bioprocédés de type fed-batch. Nous faisons l'hypothèse dans l'approche que nous proposons, que les règles de type "if-then" peuvent être implicitement remplacées par l'analyse des corrélations entre les variables biochimiques. En fait, les corrélations décrivent les relations entre les variables mais d'un point de vue statistique. De plus ces corrélations décrivent les relations entre variables en fonction du contexte et de manière plus exhaustive que des règles d'expert. Par ailleurs cette approche permet de tenir compte implicitement de l'évolution dynamique des corrélations en fonction du temps, car ces corrélations sont modifiées du fait des réactions biochimiques complexes réalisées et régulées par les micro-organismes au cours du temps.
Cette hypothèse qui consiste à utiliser l'évolution des corrélations de séries temporelles multiples au cours du temps comme élément discriminant de classification n'est pas propre aux séries temporelles de cette application biotechnologique ; elle peut être étendue à d'autres applications utilisant des séries temporelles multiples. Ainsi par exemple dans Pelletier (2005) il est indiqué que l'évolution des corrélations entre séries temporelles financières caractérisent les différentes phases du système ; l'auteur de cet article s'appuie donc sur une hypothèse similaire à celle que nous proposons dans cette approche.
Les états physiologiques sont donc caractérisés par l'analyse des corrélations entre les signaux biochimiques. Sur chaque intervalle temporel défini à partir des singularités, le coefficient de corrélation est calculé entre les signaux deux à deux. Ce coefficient de corrélation (aussi appelé coefficient de Bravais-Pearson voir Saporta (1990)) est donné par l'équation suivante :

$$\frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sigma_x \sigma_y} \tag{2}$$

où x_i représente les valeurs de la première variable biochimique (sur un intervalle temporel donné), y_i les valeurs de la deuxième variable (sur le même intervalle temporel), n le nombre d'éléments, $\overline{x}$ la valeur moyenne des éléments x_i, $\overline{y}$ la valeur moyenne des éléments y_i, et σ_x et σ_y les écarts type pour chacune des deux variables. Les valeurs mesurées x_i et y_i des deux variables sont les mesures comprises entre les singularités détectées par les ondelettes (on rappelle que ces singularités représentent les bornes de l'intervalle temporel analysé).
Ce coefficient de corrélation est en fait équivalent au cosinus du produit scalaire de deux variables projetées sur le cercle de corrélation pour une Analyse en Composantes Principales (ACP) effectuée entre ces deux variables.
Sur chaque intervalle on garde le signe des coefficients de corrélation entre deux signaux. Chaque intervalle est ainsi caractérisé par un ensemble de signes positifs et négatifs. Les intervalles ayant la même série de signes sont regroupés dans la même classe et représente donc le même état (voir figure 3).

Ruiz et al. (2004) proposent également une méthode basée sur l'ACP pour une application voisine, concernant le traitement des eaux usées : la méthode consiste à classer les données projetées préalablement dans l'espace défini par les deux premières composantes principales. Cette méthode réduit la dimension de l'espace analysé mais l'ACP ne tient pas compte du temps : l'évolution des signaux n'est pas prise en compte. Pour pallier ce problème, Ruiz et al. proposent d'utiliser une fenêtre d'analyse de taille fixe contenant des données consécutives dans le temps. Cependant du fait que la taille de la fenêtre d'analyse est fixe, cette méthode

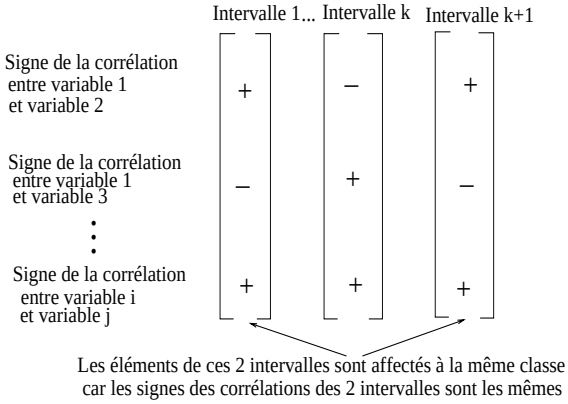

FIG. 3 – *Principe de la méthode de clustering basée sur la segmentation fournie par les ondelettes et la recherche de corrélation.*

ne tient pas réellement compte des changements survenant dans le procédé. Ainsi, la méthode basée sur la segmentation temporelle à partir de l'exposant de Hölder des singularités, semble mieux adaptée si l'on veut tenir compte de la dynamique du système.
Bien qu'elle repose sur une idée relativement simple (corrélation sur des intervalles temporels), l'approche que nous proposons peut être utilisée dans d'autres applications pour caractériser les divers états, car il existe plusieurs domaines dans lesquels les corrélations entre séries temporelles évoluent dynamiquement. On citera le cas de l'économétrie pour laquelle il existe des modèles qui tiennent compte de cette évolution dynamique des corrélations Pelletier (2005).

5 Résultats expérimentaux

Les tests ont été effectués sur un bioprocédé fermentaire de type fed-batch. Ce bioprocédé utilisant des micro-organismes (levures) appelés *Saccharomyces Cerevisiae* a duré environs 34 heures. 11 signaux biochimiques ont été mesurés durant l'expérience et ont chacun 2448 points de mesure. Les 11 séries temporelles ont été utilisées pour la classification. Ces signaux ont été normalisés (voir Régis (2004)). Pour la transformée en ondelettes, l'échelle maximum utilisée qui était égale à 2^{10}, a été choisi empiriquement après discussion avec un expert en microbiologie. La méthode a permis de détecter et de caractériser une action externe réalisée durant l'expérience. La classification est composée au total de 24 classes mais c'est surtout la classe numéro 8 qui nous intéressait (voir figure 4). En effet cette classe 8 correspond à l'ajout d'un acide dans le milieu. Toutes les apparitions de la classe 8 correspondent exactement à l'ajout de cet acide. Autant que nous le sachions, c'est la première fois qu'une méthode (qui n'est pas basée sur un modèle) permet de trouver automatiquement l'addition d'un acide dans un bioprocédé fed-batch. Les résultats sont donc encourageants et une analyse biologique plus approfondie doit être réalisée.
De même cette approche pourrait mettre en lumière des phénomènes récurrents dans d'autres applications utilisant des ensembles de séries temporelles. L'approche est générique pour les séries temporelles mulitples de quelque origine que ce soit, car :

– la méthode n'utilise pas de modèle, ni de connaissance spécifique à la microbiologie. Elle s'appuie uniquement sur les propriétés analytiques et statistiques des séries temporelles. La classification dépend donc uniquement du contexte et non de connaissance exogène.
– même sur des séries temporelles bruitées (comme celles issues des domaines économique ou physique par exemple) il est possible de détecter les singularités les plus significatives (qui ont souvent une amplitude supérieure à celle du bruit) en choisissant le maximum de l'échelle pour la transformée en ondelettes de façon adéquate (en ce qui concerne la détection des singularités par les ondelettes) et en sélectionnant les singularités en fonction de leur coefficient de Hölder (en ce qui concerne la caractérisation de ces singularités)

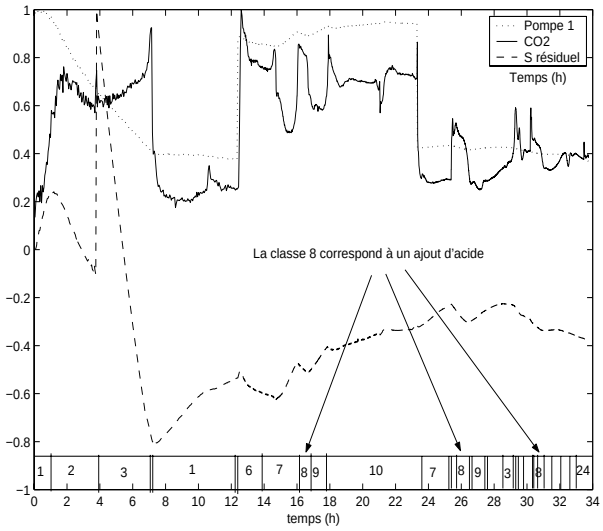

FIG. 4 – *Résultat de la classification basée sur la segmentation et les corrélations des signaux. L'axe des abscisses représente le temps en heure, l'axe des ordonnées représente les amplitudes des signaux. Les signaux ont été normalisés sur la figure pour une meilleure visualisation. Au dessus de l'axe des abscisses quelques-unes des classes les plus significatives (dont la classe 8) ont été représentées.*

6 Conclusion

Nous avons présenté une méthode de classification non supervisée basée sur l'analyse statistique et dynamique des variables mesurées pendant l'expérience. Elle s'appuie sur la détection de singularités par le maximum du module de la transformée en ondelettes et des valeurs de leurs différents coefficients de Hölder pour tenir compte de la dynamique du système. Elle utilise les corrélations entre variables pour analyser les propriétés statistiques du système et caractériser les états de ce système. La méthode a été testée sur une expérience réelle et les

résultats sont prometteurs. Elle a permis de détecter des informations pertinentes sans utiliser les informations *a priori* des experts de la microbiologie. Les perspectives et les approfondissement sont nombreux.
D'une part, les résultats obtenus ont été validés par l'expertise humaine, mais ils peuvent être évalués avec des critères d'évaluation numérique ou être comparés aux résultats d'autres méthodes de classification. D'autre part, des travaux supplémentaires consisteront à utiliser les classes trouvées dans une méthode supervisée à tester en temps réel. Une voie à explorer est d'utiliser des plages de valeurs définies avec l'expert au lieu des signes de corrélations afin d'obtenir une certaine souplesse dans la caractérisation des états. Une autre possibilité serait de regrouper les vecteurs de signe des corrélations légèrement différents les uns des autres mais suffisement semblables dans une même classe.
Cette méthode pourrait aussi être utilisée dans d'autres applications faisant intervenir des ensembles de séries temporelles. On rappelle en effet que l'approche n'est pas spécifique aux bioprocédés mais elle est utilisable sur tout système de séries temporelles dans lequel il existe un lien explicite ou implicite entre les séries. Ainsi tout système de flux de données pourrait être analysé par cette approche.

Références

Bakshi, B. et G. Stephanopoulos (1994). Representation of process trends-III. multiscale extraction of trends from process data. *Computer and Chemical Engineering 18*(4), 267–302.

Doncescu, A., J. Waissman, G. Richard, et G. Roux (2002). Characterization of bio-chemical signals by inductive logic programming. *Knowledge-Based Systems 15*(1-2), 129–137.

Harms, S. et J. Deogun (2004). Sequential association rule mining with time lags. *Journal of Intelligent Information Systems 22*(1), 7–22.

Jaffard, S. (1989). Exposants de Hölder en des points donnés et coefficients d'ondelettes. *Notes au compte-rendu de l'Académie des Sciences 308*(1), 79–81.

Jaffard, S. (1997). Multifractal formalism for functions part 1 and 2. *SIAM J. of Math. Analysis 28*(4), 944–998.

Jiang, T., B. Chen, X. He, et P. Stuart (2003). Application of steady-state detection method based on wavelet transform. *Computer and Chemical Engineering 27*(4), 569–578.

Lee, J.-M., C. Yoo, I.-B. Lee, et P. Vanrolleghem (2004). Multivariate statistical monitoring of nonlinear biological processes using kernel PCA. In *IFAC CAB'9*, Nancy, France.

Lennox, J. et C. Rosen (2002). Adaptative multiscale principal components analysis for online monitoring of wastewater treatment. *Water Science and Technology 45*(4-5), 227–235.

Mallat, S. et W.-L. Hwang (1992). Singularity detection and processing with wavelets. *IEEE Trans. on Information Theory 38*(2), 617–643.

Mallat, S. et S. Zhong (1992). Characterization of signals from multiscale edges. *IEEE Trans. on PAMI 14*(7), 710–732.

Manyri, L., S. Regis, A. Doncescu, J. Desachy, et J. Urribelarea (2003). Holder coefficient estimation by differential evolutionary algorithms for saccharomyces cerivisiae physiological states characterisation. In *ICPP-HPSECA*, Kaohsiung, Taiwan.

Marteau, P.-F., M. Fuad, et G. Ménier (2006). Echantillonnage adaptatif et classification supervisée de séries temporelles. In *EGC 06*, Lille.

Nunez, M., R. Morales, et F. Triguero (2002). Automatic discovery of rules for predicting network management events. *IEEE Journal of Selected Areas in Communications 20*(4), 736–745.

Pelletier, D. (2005). Regime switching for dynamic correlations. *Journal of Econometrics 131*, 445–573.

Régis, S. (2004). *Segmentation, classification, et fusion d'informations de séries temporelles multi-sources : application à des signaux dans un bioprocédé*. Thèse de Doctorat, Université des Antilles et de la Guyane.

Regis, S., J. Desachy, A. Doncescu, et J. Aguilar-Martin (2003). Comparaison de classification non supervisées de données biotechnologiques. In *Xe Rencontre de la Société Francophone de Classification*, Neuchâtel, Suisse.

Régis, S., A. Doncescu, J.-P. Asselin de Beauville, et J. Desachy (2004a). Evaluation de la pertinence de paramètres biochimiques et classification pour la caractérisation des états physiologiques dans un bioprocédé par la théorie de l'évidence. *Revue des Nouvelles Technologies de l'Information C-1*, 137–151.

Régis, S., L. Faure, A. Doncescu, J.-L. Uribelarrea, L. Manyri, et J. Aguilar-Martin (2004b). Adaptive physiological states classification in fed-batch fermentation process. In *IFAC CAB'9*, Nancy, France.

Roels, J. (1983). *Energetics and kinetics in biotechnology*, Chapter Macroscopic theory and microbial growth and product formation. Elsevier Biomedical Press.

Ruiz, G., M. Castellano, W. Gonzàlez, E. Roca, et J. Lema (2004). Algorithm for steady states detection of multivariate process : application to wastewater anaerobic digestion process. In *AutMoNet 2004*, pp. 181–188.

Saporta, G. (1990). *Probabilités, et Analyse des données et Statistique*. Technip.

Steyer, J. (1991). *Sur une approche qualitative des systèmes physiques : aide en temps rèel à la conduite des procédés fermentaires*. Thèse de Doctorat, Université Paul Sabatier, Toulouse France.

Steyer, J., J. Pourciel, D. Simoes, et J. Uribelarrea (1991). Qualitative knowledge modeling used in a real time expert system for biotechnological process control. In *IMACS International Workshop "Decision Support Systems and Qualitative Reasoning"*.

Struzik, Z. R. (2000). Wavelet methods in (financial) time-series processing. In *INS-R0023 Report*.

Struzik, Z. R. (2003). Taming surprises. In *Intelligent Information Systems : New Trends in Intelligent Information and Web Mining. Advances in Soft Computing*. Springer-Verlag.

Summa, M. G., F. Vautrain, L. Schwartz, M. Barrault, J.-M. Steyaert, et N. Hafner (2006). Multiple time series : New approches and new tools in data mining applications to cancer epidemiology. *Modulad 34*, 37–46.

Takada, H. et F. Bass (1998). Multiple time series analysis of competitive marketing behavior. *Journal of Business Research 43*, 97–107.

Clustering dynamique et statistique sur des séries temporelles.

Remerciements

Nous remercions l'équipe fermentation du LBB INSA pour avoir mis à notre diposition les données expérimentales et pour leur aide à la validation des résultats ainsi que M. Doncescu et M. Desachy pour leur remarques pertinentes sur la classification et l'analyse des séries. Nous tenons à remercier également les relecteurs anonymes pour leurs critiques et remarques constructives.

Summary

We present a clustering method using dynamical and statistical analysis. It is a data mining tool which works on multiple time series to detect meaningfull states. The aim of this unsupervised method is to find relevant information within the time series. It is based on the detection of singularities of the time series and on the analysis of their correlation factors in intervals defined by these singularities. For the application the time series are biochemical signals measured during a bioprocess. This approach is used to confirm the knowledge of experts without using the *a priori* knowledge of these experts. It is used for the search of physiological states in a fed-batch process.