Contrôle non paramétrique de procédés par lots basé sur STATIS et la classification

Non parametric on line control of batch processes based on STATIS and clustering

Ndèye Niang[1], Gilbert Saporta[1], Flavio S. Fogliatto[2]

[1] Chaire de Statistique Appliquée & CEDRIC CNAM
292, rue Saint Martin, 75141 Paris Cedex 03, France,
ndeye.niang_keita@cnam.fr
gilbert.saporta@cnam.fr

[2] ffogliatto@producao.ufrgs.br

Résumé
Nous proposons une nouvelle approche du contrôle de qualité des procédés par lots basée sur la méthode STATIS et des cartes de contrôles non paramétriques à partir d'enveloppes convexes. Cette approche générale peut être utilisée pour le contrôle en fin de fabrication des procédés par lots ainsi que pour le contrôle en cours de fabrication après une étape de classification. La méthode proposée est illustrée sur des données réelles.
Mots-clés : Procédés par lots, Classification, Contrôle de qualité multivarié, STATIS.

Abstract
We propose a new non parametric quality control strategy for monitoring batch processes based on the three way method STATIS and convex hull peeling. This general approach allows off line monitoring of batch processes as well as on line one after a clustering step. A real example illustrates the proposed method.
Keywords : Batch process, Clustering, Multivariate quality control, STATIS

1. Introduction

Les procédés par lots sont largement utilisés dans le secteur industriel notamment dans l'industrie agroalimentaire, chimique ou pharmaceutique. Dans ces procédés, les matières premières sont introduites dans un ordre spécifique et subissent une série de transformations pendant une durée qui peut être fixe ou variable donnant alors lieu à des procédés à temps fixe ou à temps variable. Le produit final obtenu est ensuite analysé pour vérifier s'il correspond à des standards de qualité désirés. Le suivi du procédé s'effectue à travers un ensemble de variables caractéristiques du procédé prélevées par un échantillonnage en ligne au fur et à mesure de son déroulement. Les données se présentent sous la forme d'un tableau à trois entrées ou « cube » de données. Due à la nature multidimensionnelle des données issues de tels procédés, les cartes de contrôle multivariées sont alors les seules adéquates pour le contrôle de leur qualité.

La carte multivariée la plus fréquemment utilisée est la carte T^2 de Hotelling (Lowry& Montgomery (1995)). En général, ces cartes de contrôle sont basées sur l'hypothèse d'indépendance des observations et de multinormalité des caractéristiques du procédé. Mais dans la pratique ces hypothèses ne sont pas toujours vérifiées. De plus les cartes de contrôle classiques ne permettent pas un contrôle efficace lorsque les standards de qualité sont décrits par des profils ou courbes. Dans le cas de tels procédés, le contrôle s'effectue à travers des cartes multivariées basées sur une analyse en composantes principales particulière (multiway principal component analysis) Nomikos & MacGregor (1995). Ces cartes seront notées MPCA-CCs dans la suite.
L'application des MPCA-CCs pour le monitoring des procédés par lots a été initialement proposé par Jackson & Mudhokar (1979), et largement étudiée par la suite par Nomikos & MacGregor (1995), Kourti & MacGregor (1996) et MacGregor (1997). Elle suppose, en plus de la normalité des variables, que tous les lots aient la même durée et ne peut donc pas être directement utilisée pour le contrôle des procédés à temps variable, ni pour le contrôle de procédés par lots en cours de fabrication. De nombreuses méthodes ont été proposées pour adapter les MPCA-CCs aux cas cités ci-dessus: Nomikos (1995), Kassidas *et al*. (1998), Doan & Srinivasan (2008). Elles peuvent être globalement considérées comme des méthodes de prétraitement dont le but est de donner la même longueur à tous les lots afin d'appliquer ensuite les MPCA-CCs classiques. Cependant ces méthodes présentent toutes quelques limitations (Niang et al. 2009).
Une approche générale basée sur la méthode STATIS a été proposée (Niang et al. 2009) permettant le contrôle en fin de fabrication des procédés par lots à temps fixe (avec STATIS) et à temps variable (avec STATIS DUAL) sans aucun traitement préalable des données ni hypothèse sur la distribution des variables. Elle consiste d'abord à utiliser la méthode STATIS pour réduire la dimension des données puis à construire des cartes de contrôles non paramétriques à partir des enveloppes convexes obtenues directement sur les plans factoriels issus de l'application de la méthode STATIS.
Nous nous intéressons au contrôle en cours de fabrication qui consiste à suivre le procédé au fur et mesure de son déroulement pour détecter le plus tôt possible une sortie des limites de contrôle plutôt que d'attendre la fin du lot. Il s'agit donc de vérifier le comportement du procédé à chaque instant noté t. L'application de l'approche décrite précédemment permet d'établir une distribution de référence pour le comportement du procédé jusqu'à l'instant t ou de manière équivalente une carte de contrôle sur des tableaux partiels obtenus en sélectionnant les t premières lignes des tableaux de données. En principe il faudrait autant de cartes de contrôle que d'instants de mesures. La dernière carte est identique à la carte pour le contrôle off line. Mais en pratique ne sont intéressantes que celles qui correspondent à des instants de changement important dans l'évolution du procédé, ces instants définissent une partition de l'ensemble des instants de mesures.
Dans cet article, nous proposons une approche basée sur la classification sous contrainte de contiguïté pour déterminer ces instants. Après un rappel sur les cartes de contrôle non paramétriques basées sur STATIS, nous présentons dans la section 3 notre proposition pour le contrôle on line. La méthode proposée est ensuite illustrée sur des données réelles d'un procédé à temps fixe.

2. Cartes de contrôle non paramétriques basées sur STATIS

On dispose d'un historique de *N* lots de référence c'est à dire des lots ayant donné un produit de bonne qualité définissant ainsi une distribution de référence représentant le bon fonctionnement du procédé. Dans le cas des procédés à temps fixe, les données se présentent sous la forme de *N* tableaux à *p* variables prélevées à *T* instants (figure 1). On est donc en présence de plusieurs tableaux décrivant un ensemble d'individus sur *p* variables. Il est alors possible de les analyser directement sans aucun traitement préalable en utilisant la méthode STATIS. Nous expliquons plus en détail notre méthode de contrôle de qualité après avoir rappelé brièvement la méthode STATIS.

	LOT 1	*LOT 2*		*LOT N*	
X =	*VARIABLES* $x_1,x_2,\ldots,x_p$ **X_1**	*VARIABLES* $x_1,x_2,\ldots,x_p$ **X_2**	……..	*VARIABLES* $x_1,x_2,\ldots,x_p$ **X_N**	*instant 1* *instant 2* . . *instant T*

Figure 1- Matrice des données

2.1 STATIS

STATIS est une méthode d'analyse exploratoire simultanée de plusieurs tableaux de données recueillies à différentes occasions Escoufier(2006). A notre connaissance son utilisation en contrôle de qualité se limite aux travaux de Scepi (2002).
L'idée essentielle est la recherche d'une structure commune aux tableaux pour voir si les distances entre individus sont stables d'un tableau à l'autre. Elle fonctionne en trois étapes. D'abord on effectue une analyse globale dans laquelle on cherche à comparer la structure des tableaux sans pouvoir donner une explication fine des éventuelles différences entre tableaux. Cette étape est appelée *interstructure*.
L'étude fine s'effectue dans la deuxième analyse appelée *intrastructure*. Elle repose sur la détermination d'un résumé global des tableaux appelé compromis qui permet de trouver un espace commun de représentation. L'étude de l'évolution de chacun des individus des tableaux sur cet espace de représentation permet d'expliquer au niveau individuel les écarts mis en évidence par l'interstructure.
Plus précisément, on dispose de $\mathbf{X}_i$ (i = 1,…, N) matrices contenant T observations de p variables. Préalablement à l'analyse, les données sont centrées réduites. STATIS associe à chaque $\mathbf{X}_i$ la matrice ($T \times T$) des produits scalaires entre individus $\mathbf{W}_i = \mathbf{X}_i\mathbf{X}_i^{'}$, où $\mathbf{X}_i^{'}$ est la matrice transposée de $\mathbf{X}_i$. C'est un objet représentatif de $\mathbf{X}_i$; il contient tous les liens inter-individus et ses vecteurs propres sont les composantes principales de $\mathbf{X}_i$. Pour comparer deux tableaux $\mathbf{X}_i$ et $\mathbf{X}_{i'}$, on utilise le coefficient *RV* de corrélation vectorielle, Escoufier (2006) défini par:

$$RV_{ii'} = \mathrm{trace}(\mathbf{W}_i\mathbf{W}_{i'}) / \sqrt{\mathrm{trace}(\mathbf{W}_i)^2\,\mathrm{trace}(\mathbf{W}_{i'})^2} \qquad (1)$$

RV varie entre 0 et 1; plus il est proche de 1, plus les deux matrices $\mathbf{W}_i$ et $\mathbf{W}_{i'}$ sont similaires.

2.1.1 Interstructure

L'interstructure consiste à étudier graphiquement les ressemblances globales entre tableaux. Comme en ACP, les deux premiers vecteurs propres de la matrice S contenant les coefficients *RV* entre $\mathbf{W}_i$ et $\mathbf{W}_{i'}$ (i ,i'= 1,…, N) définissent le premier plan principal ce qui permet de visualiser les proximités

entre tableaux en y projetant les objets $\mathbf{W}_i$: la coordonnée du tableau $\mathbf{W}_i$ sur le *kième* axe factoriel est donnée par $c_i^k = \sqrt{\lambda_k} u_i^k$ où λ_k est la valeur propre associée au *kième* vecteur propre $\mathbf{u}^k$. Les coefficients *RV* étant positifs, $\mathbf{u}^1$ a ses composantes toutes de même signe, elles seront prises positives.

2.1.2 Intrastructure

L'étude de l'intrastructure consiste d'abord à rechercher le compromis qui résume au mieux l'ensemble de tableaux. La solution est une moyenne pondérée des objets $\mathbf{W}_i$ les coefficients α_i^1 étant les composantes du premier vecteur propre $\mathbf{u}^1$ normalisé:

$$\mathbf{W} = \sum_{i=1}^{N} \alpha_i^1 \mathbf{W}_i \qquad (2)$$

Les poids α_i^1 représentent alors le niveau d'accord entre les tableaux et le compromis. Cette définition du compromis confère à STATIS une propriété de robustesse vis à vis des valeurs aberrantes: plus un tableau est différent des autres, moins il a d'influence sur le compromis. Cette propriété est particulièrement intéressante en contrôle de qualité dont le but est la détection de valeurs anormales.

Les vecteurs propres de la matrice compromis $\mathbf{W}$ permettent ainsi d'obtenir l'espace de représentation commun à l'ensemble des tableaux. Il est alors possible de visualiser sur le premier plan principal des points artificiels B_t (t =1,…, T) appelés points compromis. Les coordonnées sur le *k-ième* axe factoriel sont les éléments du vecteur suivant:

$$\mathbf{z}_k = \sqrt{\delta_k}\,\mathbf{v}^k = (1/\sqrt{\delta_k})\mathbf{W}\mathbf{v}^k \qquad (3)$$

où δ_k est la valeur propre associée au *k-ième* vecteur propre $\mathbf{v}^k$.

De plus, il est possible de représenter les individus de tous les tableaux $\mathbf{W}_i$ en les projetant sur le plan compromis par la technique des points supplémentaires. Les différentes positions d'un individu selon les tableaux définissent sa trajectoire qui permet de mettre en évidence des écarts entre tableaux au niveau individuel. On peut donc avoir une représentation détaillée du comportement commun des lots à un instant donné.

2.2 Cartes de contrôle non paramétriques

Les cartes de contrôle non paramétriques que nous proposons sont basées sur des enveloppes convexes directement construites sur les plans principaux de l'interstructure de STATIS. Pour établir une région de contrôle de confiance $(1-\alpha)$ on utilise une proposition de Zani *et al.* (1998), comprenant les trois étapes suivantes :

- on détermine sur le plan de l'interstructure une région intérieure qui contient une proportion π^* des points. Elle est obtenue par lissage par une *B*-spline des contours de l'enveloppe convexe contenant les points, cette dernière étant obtenue par pelages successifs de l'enveloppe convexe contenant l'ensemble des points.

---

* π est égale à 50% des points dans Zani *et al*. (1998), mais on peut utiliser une plus grande proportion.

- Ensuite on détermine une estimation du centre de la région en prenant par exemple la moyenne arithmétique des observations dans la région.

- Finalement, la carte de contrôle est obtenue par dilatation de l'enveloppe lissée en multipliant la distance entre le centre et frontière de la région par un nombre *l* correspondant à la probabilité α de fausse alarme désirée.

Avec cette méthode, à partir des *N* lots de référence associés à *N* tableaux de dimension *T* x *p*, on obtient une carte de contrôle non paramétrique. Le but du contrôle en fin de fabrication est de vérifier la conformité des données d'un nouveau lot représenté par la matrice $\mathbf{X}_{N+1}$ avec des standards de qualité résumés, représentés par la carte de contrôle issue des lots de référence. Le contrôle du nouveau lot s'effectue alors en projetant la matrice $\mathbf{X}_{N+1}$ sur la carte. Le lot sera déclaré sous contrôle si la matrice se projette à l'intérieur de la région de contrôle. Dans le cas contraire, le lot sera hors contrôle. La section 4 présente les résultats de l'application de cette méthode à des données issues d'un procédé de polymérisation.

3- Contrôle on line

Rappelons qu'il consiste à suivre le procédé au fur et mesure de son déroulement pour détecter le plus tôt possible une sortie des limites de contrôle plutôt que d'attendre la fin du lot. La démarche que nous proposons consiste à établir, selon la méthode décrite en section 2, de manière séquentielle des cartes de contrôle non paramétriques basées sur STATIS appliquée à des tableaux partiels issus des tableaux de référence. Elle comporte donc une étape préalable de détermination de la longueur des séquences ou de la taille des tableaux partiels que nous proposons de réaliser à partir d'un partitionnement de l'ensemble des instants de mesures.

Nous disposons de *N* matrices $\mathbf{X}_i$ contenant *T* observations de *p* variables. Elles représentent le comportement de référence du procédé produisant des lots de bonne qualité. Le problème est donc de trouver une partition *P* des *T* instants de mesure commune à l'ensemble des lots avec une contrainte de conservation de la chronologie.

L'application d'une classification ascendante hiérarchique sous contrainte de contiguïté temporelle Murtagh (1985) sur chaque lot permet d'obtenir un ensemble de *N* partitions des *T* instants en *K* classes. La variabilité des lots de référence peut entraîner une variabilité dans les tailles des classes: la classe 1 d'une partition P_i peut contenir les instants de 1 à t alors que la classe 1 de la partition P_j contient les instants de 1 à t-1. Plus formellement, soit n_{ik} la taille de la classe k de la partition *Pi* associé au lot i, les instants de contrôle associés au lot i et notés t_{ik}, sont définis par $t_{ki} = \sum_{l=1}^{k} n_{il}$ avec k variant de 1 à K.

En considérant l'ensemble des N lots de référence, on obtient pour chaque k un ensemble de *N* valeurs t_{ki} (i= 1,…, *N*) qui peut être assimilé à une période « critique » pendant laquelle il faudrait surveiller le procédé. Nous proposons de choisir comme instants pour le contrôle on line les valeurs $t_k = \sup_i t_{ik}$.

Cela revient à effectuer le contrôle au dernier instant de la période critique augmentant ainsi la probabilité de détection des lots hors contrôle. En effet, il est usuel en contrôle de qualité de supposer que si une cause assignable produit un déréglage des caractéristiques du procédé, ce dernier persiste. Effectuer le contrôle à l'instant t_k permet donc de détecter un plus grand nombre de lots qui ont pu être déréglés antérieurement à t_k . Lorsque les instants de mesure sont proches les uns des autres, cela

n'affectera pas beaucoup la période opérationnelle moyenne. Dans le cas contraire d'autres stratégies prenant en compte l'écart entre les instants de mesures devraient être considérées.
En appliquant la méthode proposée en 2.2 aux K ensembles de *N* tableaux obtenus en sélectionnant successivement les t_k premiers instants des tableaux de référence, on construit alors K cartes de contrôle. Plus précisément pour chaque instant de contrôle, on applique STATIS aux N tableaux à t_k individus et on obtient la carte de contrôle non paramétrique à partir du plan factoriel représentant l'interstructure.
Le contrôle d'un nouveau lot en cours de fabrication consiste ensuite à projeter les tableaux de taille t_k associés au lot en cours de fabrication sur les cartes correspondantes.

4- Application

Nous illustrons notre proposition sur des données réelles utilisées dans la littérature des procédés par lots par Nomikos, Mc Gregor ou Eriksson et al. par exemple. Les données sont issues d'un procédé de polymérisation et sont composées de 18 lots de référence sélectionnés comme représentant le comportement normal souhaité du procédé. Pour chaque lot, 10 variables ont été prélevées à 100 instants. Les variables x_1 x_2 x_3 x_6 et x_7 sont des mesures de température, x_4 x_8 et x_9 sont des mesures de pression et les variables x_5 et x_{10} représentent des vitesses d'écoulement de matières ajoutées au réacteur. On dispose de plus d'un ensemble de 11 lots supplémentaires pour tester les performances des méthodes. Il contient 4 lots de bonne qualité et 7 mauvais lots. Nous avons appliqué STATIS et les régions de contrôle avec un niveau de confiance de 99% sont construites sur les plans factoriels de l'interstructure.
La figure 2 montre les résultats du contrôle en fin de fabrication. Tous les 7 mauvais lots ont été signalés hors contrôle avec cependant un plus fort signal pour 6 d'entre eux (fig.2.a). Le mauvais lot proche de la limite a été diagnostiqué comme ayant un comportement différent des 6 autres et n'est en général pas détecté comme étant hors contrôle (Eriksson *et al*). La carte de contrôle (fig.2.b) montre les résultats pour les bons lots. 3 bons lots parmi les 4 sont signalés sous contrôle. On constate cependant une fausse alerte comme dans Eriksson et al.

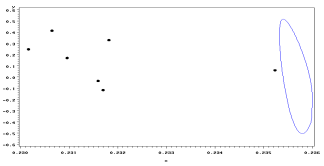

Fig. 2.a. Mauvais lots

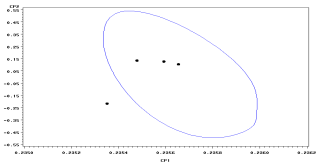

Fig. 2.b. Bons lots

Conclusion

Nous avons proposé une méthode pour le contrôle de qualité des procédés par lots en fin et en cours de fabrication basée sur la méthode STATIS. Le suivi du procédé est effectué à travers des cartes de contrôles non paramétriques utilisant toutes les observations disponibles pour le contrôle en fin de fabrication, et une partie des observations séquentiellement pour le contrôle en cours de fabrication. La méthode est illustrée sur des données réelles.
Des évaluations plus formelles des performances de la méthode sont en cours ainsi que des études comparatives avec d'autres méthodes proposées dans la littérature.

Bibliographie

Doan, X-T., Srinivasan, R. (2008) Online monitoring of multi-phase batch processes using phase-*based* multivariate statistical control. *Computers and Chemical Engineering*, 32: 230-243

Eriksson, L., Johansson, E., Kettaneh-Wold, N., and Wold, S.(2001).*Multi- and Megavariate Data Analysis.* Umetrics

Escoufier, Y. 2006. Operator related to a data matrix: a survey. *Proceedings in Computational Statistics* Rizzi A. *et al*. (eds), 285-297 Physica-Verlag.

Jackson, J.E. and Mudholkar, G.S. (1979) Control Procedures for Residuals Associated with Principal Component Analysis. *Technometrics*, 21 (3), 341–34.

Kassidas, A., MacGregor, J.F. and Taylor, P.A. (1998) Synchronization of batch trajectories using dynamic time warping. *AIChE Journal*, 44, 864–875.

Kourti, T. and MacGregor, J.F. (1996) Multivariate SPC Methods for Process and Product Monitoring. *Journal of Quality Technology*, 28 (4), 409–428.

Lowry, C.A. and Montgomery, D.C. (1995) A review of multivariate control charts. *IIEE Transactions*, 27 (6), 800–810.

MacGregor, J.F. (1997) Using on-line process data to improve quality: challenges for statisticians. *International Statistical Review*. 65 (3), 309–323.

Murtagh, F. (1985) A Survey of Algorithm for Contiguity-constrained clustering and Related problems. The computer journal, 28(1), 82-88

Niang, N., Fogliatto F. and Saporta, G. (2009) Batch Process Monitoring by Three-way Data Analysis Approach, *ASMDA'09*, Vilnius, July 2009, pp.294-298

Nomikos, P. and MacGregor, J.F. (1995) Multivariate SPC charts for monitoring batch processes. *Technometrics*, 37 (1), p.41–59.

Scepi, G. (2002) Parametric and non parametric multivariate quality control charts. In *Multivariate Total Quality Control*, Physica-Verlag , Lauro C. *et al*. (eds), 163–189.

Zani, S., Riani, M. and Corbellini, A. (1998) Robust bivariate boxplots and multiple outlier detection. *Computational Statistics & Data Analysis*, 28, 257-270.