
Amélioration des algorithmes d'étalonnage d'un spectromètre UV-visible pour le mesurage *in situ* en continu des polluants dans les effluents urbains

A. Torres — J.-L. Bertrand-Krajewski

URGC Hydrologie Urbaine, INSA de Lyon,
34 avenue des Arts, 69621 Villeurbanne cedex, France

Emails : andres.torres@insa-lyon.fr, jean-luc.bertrand-krajewski@insa-lyon.fr

RESUME. Les spectromètres UV-visible permettent d'estimer en continu les concentrations en polluants dans les effluents transportés par les réseaux d'assainissement. La régression PLS (Partial Least Squares) est présentée en tant qu'outil mathématique pour traiter les spectres d'absorption fournis par les spectromètres UV-visible afin d'estimer des concentrations équivalentes en polluants. Une méthode spécifique est proposée pour déterminer les variables indépendantes et le nombre de vecteurs latents. La méthode proposée est appliquée pour l'estimation des concentrations en MES (matières en suspension) et DCO (demande chimique en oxygène) à l'entrée d'un bassin de retenue-décantation sur un réseau séparatif pluvial. La précision des résultats fournis par ce type de capteur peut être améliorée en utilisant des techniques d'analyse multivariée comme la régression PLS.

ABSTRACT. UV-visible spectrometers may be used for on-line measurements of pollutants concentrations in wastewater transported in sewer systems. The PLS (Partial Least Squares) regression is presented as a mathematical tool to process absorption spectra provided by UV-visible spectrometers in order to estimate equivalent concentrations of pollutants. A specific method is proposed to determine the independent variables and the number of latent vectors. The proposed method is then applied to estimate TSS (total suspended solids) and COD (chemical oxygen demand) concentrations at the inlet of a storage-settling tank in a stormwater separate sewer system. The accuracy of the equivalent concentrations estimated by means of this type of sensors can be improved by using multivariate analysis techniques like the PLS regression.

MOTS-CLES : spectrométrie UV-visible, régression PLS, polluants, effluents urbains

KEYWORDS: UV-visible spectrometry, PLS regression, pollutants, urban wastewater

1. Introduction

Le mesurage des concentrations et des flux de polluants transportés dans et rejetés par les systèmes d'assainissement fournit une information essentielle pour la gestion, le contrôle et le dimensionnement de ces systèmes. Depuis quelques années, les progrès et innovations technologiques ont permis de développer des capteurs en continu pour le suivi de la qualité de l'eau *in situ* et en temps réel. Parmi ces capteurs, les spectromètres UV-visible permettent d'estimer des concentrations en polluants à partir d'un traitement mathématique des spectres d'absorption. Ce traitement mathématique peut être plus ou moins complexe en fonction de la nature de l'eau étudiée et des paramètres de qualité auxquels on s'intéresse. Il fait généralement appel à des méthodes telles que la régression multiple ou la régression PLS (Partial Least Squares). Dans cet article, nous présentons un programme PLS développé sous MatLab et son application à l'estimation des concentrations en MES (matières en suspension) et DCO (demande chimique en oxygène) en entrée d'un bassin de retenue-décantation à partir des spectres fournis par un spectromètre UV-visible.

2. Matériels et méthodes

Dans ce chapitre, nous présentons successivement les principes de la spectrométrie UV-visible, le spectromètre s::can[®] utilisé, le site expérimental, les principes de la régression PLS, et le programme de régression PLS développé sous MatLab.

2.1. Spectrométrie UV-visible

Les spectromètres UV-visible effectuent un mesurage de l'absorption de la lumière par les particules dissoutes ou en suspension dans des gammes de longueurs d'onde allant de l'ultraviolet au visible. Le spectromètre commercialisé par la société s::can[®], appelé spectro::lyser, est un capteur submersible de 60 cm de longueur et 44 mm de diamètre qui mesure l'atténuation de la lumière entre 200 nm et 750 nm (Langergraber *et al.*, 2003 ; Hochedlinger, 2005). Cet appareil est livré avec un étalonnage par défaut intégré dans le logiciel ana::pro, appelé « global calibration », établi par le constructeur. Pour une estimation précise des concentrations en polluants, cet étalonnage global peut ne pas être suffisant, et un étalonnage spécifique local, dit « local calibration », doit être effectué (Langergraber *et al.*, 2003). Cet étalonnage local est établi pour chaque polluant en corrélant d'une part des résultats de mesure *in situ* (spectres) fournis par le spectromètre et d'autre part des analyses en laboratoire (concentrations) effectuées sur des échantillons prélevés à proximité de la fente de mesure du spectromètre. A partir d'une base de données de spectres d'absorption et d'analyses faites en laboratoire sur les

échantillons correspondants, on peut établir un étalonnage local résultant de la construction d'une relation mathématique linéaire ou polynomiale entre d'une part les valeurs de concentrations équivalentes fournies par le capteur en mode étalonnage global et d'autre part les valeurs de concentration des échantillons fournies par le laboratoire d'analyse (de Bénédictis et Bertrand-Krajewski, 2005).

2.2. Site expérimental

Le capteur spectro::lyser a été installé en entrée du bassin de retenue-infiltration Django Reinhardt à Chassieu. Ce bassin a été construit en 1975 pour recueillir les eaux pluviales de la zone industrielle de Chassieu (185 ha). Il est composé de deux sous-bassins reliés par une canalisation de 60 cm de diamètre : un sous-bassin de retenue-décantation, suivi d'un sous-bassin d'infiltration (Bardin et Barraud, 2004). Quelques industries sont autorisées à rejeter dans le réseau d'eaux pluviales des eaux de refroidissement. Par temps sec, il y a donc un faible débit permanent arrivant dans les ouvrages. Les effluents entrent dans le sous-bassin de retenue-décantation (volume : 32200 m³, surface au sol : 11302 m²) par deux collecteurs circulaires de 1.6 m de diamètre. La vidange vers le sous-bassin d'infiltration est faite gravitairement avec un débit maximum de 700 L/s contrôlé par un régulateur, et par une surverse directe en partie supérieure en cas d'événement pluvieux exceptionnel. Une station de mesure en continu installée à l'entrée du sous-bassin de retenue-décantation permet d'acquérir des données par temps sec et par temps de pluie. Le débit entrant est estimé par mesurage simultané de la hauteur d'eau et de la vitesse dans la section d'écoulement. Les autres grandeurs sont mesurées hors du réseau dans un canal expérimental installé dans un bungalow et alimenté par une pompe péristaltique (débit 1 L/s et vitesse d'aspiration 1 m/s). Différents capteurs disposés dans le canal expérimental permettent de mesurer, avec un pas de temps de 2 minutes, la conductivité, la température, le pH et la turbidité. Un préleveur automatique réfrigéré de 24 flacons de 1 L permet de réaliser des prélèvements d'effluents dans le canal expérimental, prélèvements dont les concentrations en polluants sont déterminées ultérieurement en laboratoire.

2.3. Régression PLS (Partial Least Squares)

Pour étalonner le capteur spectro::lyser et exploiter les données, nous appliquons la régression PLS. Cette méthode est particulièrement utile quand le nombre de descripteurs (variables indépendantes) est comparable ou plus grand que le nombre d'observations et/ou quand il existe des corrélations entre les variables (VCCL, 2005).

L'objectif de la régression PLS est de prédire une variable Y à partir d'une matrice de variables X et de décrire leur structure commune. L'avantage de la régression PLS par rapport à d'autres méthodes réside dans sa capacité à identifier

les composantes de la matrice X qui sont également pertinentes pour expliquer Y . Plus particulièrement, la régression PLS cherche un jeu de composantes (appelées vecteurs latents) qui fournit une décomposition simultanée de X et de Y sous la contrainte que ces composantes expliquent le mieux possible la covariance entre X et Y . Il y a ensuite une étape de régression où la décomposition de X est utilisée pour prédire Y . Les détails mathématiques de la méthode sont expliqués dans par exemple dans Abdi (2003) et Tenenhaus (1998).

Dans la PLS, la phase de détermination des variables latentes constitue un point central dans l'estimation du modèle structurel, du fait que l'estimation des paramètres du modèle dépend fortement des vecteurs latents obtenus (Hanafi, 2004). En plus du nombre optimum de vecteurs latents, il existe aussi un nombre optimum de variables indépendantes à considérer dans le modèle. A partir d'expériences réalisées avec un spectromètre UV-visible :scan, Hochedlinger (2005) a montré que la qualité prédictive est meilleure si, au lieu d'utiliser la totalité des 214 longueurs d'onde composant les spectres, on ne choisit que 50 longueurs d'onde entre 240 et 500 nm pour prédire la DCO, et 50 longueurs d'onde entre 380 et 750 nm pour prédire les MES.

2.4 Programme de régression PLS

Nous avons développé un programme sous MatLab, appelé OPP (OTHU PLS Programme), en utilisant l'algorithme NIPALS (Nonlinear estimation by Iterative Partial Least Squares) originalement conçu par Wold (1966) (cité dans Tenenhaus, 1998) et présenté par Abdi (2003).

Le nombre optimal de vecteurs latents a été déterminé en utilisant une validation croisée de type Jackknife (Aji *et al.*, 2003). Comme indicateur de la qualité prédictive des modèles, nous avons utilisé la valeur RMSEP (Root Mean-Square Error of Prediction) pour le jeu de données destiné à la validation (Dantas Filho *et al.*, 2005) :

$$RMSEP = \sqrt{\frac{PRESS}{M}} \text{ avec } PRESS = \sum_{m=1}^M (y_m - \hat{y}_m)^2 \quad \text{Eq. 1}$$

avec y_m la valeur de la concentration de référence pour l'objet de validation numéro m , \hat{y}_m la valeur de la concentration prédite par la PLS pour l'objet de validation numéro m , et M le nombre total d'objets de validation.

Pour déterminer le nombre optimal de variables indépendantes, nous proposons une méthode consistant à classer les variables indépendantes par ordre de coefficient de corrélation croissant entre chaque variable indépendante et la variable dépendante, pour le jeu de données destiné à l'étalonnage, pour ensuite prendre les deux premières variables indépendantes (selon le classement) et tester la prédictibilité ($PRESS$) du modèle pour 1 et 2 vecteurs latents. La procédure est

réitérée jusqu'à atteindre le nombre total de variables indépendantes H et obtenir la matrice indiquée Figure 1. La cellule de cette matrice où la valeur $PRESS_{i,j}$ est minimum indique les variables indépendantes à prendre en compte et le nombre de vecteurs latents à utiliser pour la régression PLS.

		nombre de variables indépendantes						
		1	2	3	...	i	...	H
nombre de vecteurs latents	1	$PRESS_{1,1}$	$PRESS_{2,1}$	$PRESS_{3,1}$...	$PRESS_{i,1}$...	$PRESS_{H,1}$
	2		$PRESS_{2,2}$	$PRESS_{3,2}$...	$PRESS_{i,2}$...	$PRESS_{H,2}$
	3			$PRESS_{3,3}$...	$PRESS_{i,3}$...	$PRESS_{H,3}$
	⋮				⋮	⋮	...	⋮
	j					$PRESS_{i,j}$...	$PRESS_{H,j}$
	⋮						⋮	⋮
	h							$PRESS_{H,h}$

Figure 1. Schéma de la matrice fournie par le programme OPP, utilisée pour déterminer le nombre optimal de variables indépendantes et de vecteurs latents

3. Application au site expérimental Django Reinhardt

Pour étalonner le spectromètre UV-visible spectro::lyser avec la méthode proposée, nous avons réalisé des campagnes de mesures à l'entrée du bassin de retenue-décantation de Chassieu. L'objectif est d'établir un étalonnage spécifique pour ce site, de manière à estimer les flux de MES et de DCO par temps sec et par temps de pluie. Les données utilisées pour cet étalonnage sont : (i) par temps de pluie : huit échantillons (28 novembre 2005) ; (ii) par temps sec : huit échantillons (22 novembre et 15 décembre 2005).

Le choix des échantillons destinés à l'étalonnage et au test a été réalisé de manière aléatoire. Ces sous-jeux de données (étalonnage, test) ont été utilisés pour les différentes méthodes d'étalonnage utilisées. Le Tableau 1 présente une synthèse des résultats obtenus en laboratoire. Le capteur spectro::lyser fournit un spectre spécifique à chaque échantillon : on observe sur la Figure 2 que les absorbances sont plus élevées pour les spectres de temps de pluie que pour les spectres de temps sec, et ceci pour la plupart des longueurs d'onde. Un étalonnage spécifique pour chacun de ces deux contextes s'avère donc nécessaire si l'on veut obtenir une estimation des concentrations la plus satisfaisante possible.

De manière à comparer différentes méthodes d'étalonnage, nous avons utilisé le logiciel d'exploitation ana::pro fourni avec le capteur (global calibration, local calibration) et le programme OPP, pour corréler d'une part les spectres obtenus avec le capteur spectro::lyser et d'autre part les concentrations en MES et DCO mesurées sur échantillons en laboratoire. Par ailleurs, à partir des données du Tableau 1, nous avons établi des relations linéaires $MES = f(\text{turbidité})$ et $DCO = f(\text{turbidité})$.

No.	Pluie/Sec	Turbidité	MES (mg/L)	DCO (mg/L)	Etalonnage/Test
1	Pluie	162	152	230	Etalonnage
2	Pluie	347	345	364	Test
3	Pluie	431	390	342	Etalonnage
4	Pluie	348	271	291	Etalonnage
5	Pluie	248	159	194	Etalonnage
6	Pluie	239	166	123	Etalonnage
7	Pluie	95	41	31	Test
8	Pluie	31	13	30	Etalonnage
9	Sec	8	2	24	Test
10	Sec	23	13	54	Etalonnage
11	Sec	28	11	24	Etalonnage
12	Sec	31	11	17	Test
13	Sec	28	31	6	Etalonnage
14	Sec	18	2	4	Etalonnage
15	Sec	14	5	51	Test
16	Sec	23	6	70	Etalonnage

Tableau 1. Résultats obtenus en laboratoire sur 8 échantillons de temps de pluie et 8 échantillons de temps sec à l'entrée du bassin de retenue-décantation Django Reinhardt à Chassieu

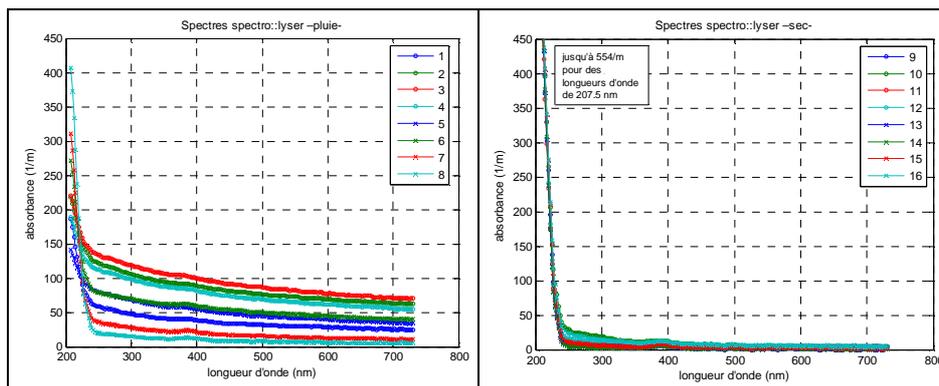


Figure 2. Spectres fournis par le capteur spectro::lyser sur 8 échantillons de temps de pluie (à gauche) et 8 échantillons de temps sec (à droite) à l'entrée du bassin de retenue-décantation Django Reinhardt à Chassieu

Le Tableau 2 montre la qualité des ajustements des modèles obtenus avec les quatre méthodes d'étalonnage utilisées (relation avec la turbidité, global calibration, local calibration et programme OPP). Les figures relatives aux résultats obtenus avec le programme OPP sont fournies Figure 3. On observe que le capteur spectro::lyser présente une meilleure qualité prédictive (test) que le turbidimètre, quel que soit la méthode d'étalonnage utilisée : pour les MES, par exemple, on passe de $r^2 = 0.981$ et $RMSEP = 31$ mg/L pour les résultats fournis par la relation $MES = f(\text{turbidité})$ à r^2 entre 0.981 et 0.998 et $RMSEP$ entre 11 et 23 mg/L pour les résultats fournis par le spectro::lyser. Les résultats obtenus confirment qu'il est préférable d'utiliser un étalonnage local, soit avec le logiciel ana::pro du spectro::lyser, soit

avec une méthode de régression multivariée comme la PLS. Au niveau des résidus moyens (*RMSEP*), le programme OPP montre une nette amélioration de la précision du modèle, notamment en ce qui concerne la DCO.

METHODE	MES				DCO			
	Etalonnage		Test		Etalonnage		Test	
	r^2	<i>RMSEP</i>	r^2	<i>RMSEP</i>	r^2	<i>RMSEP</i>	r^2	<i>RMSEP</i>
relations avec la turbidité	0.972	21	0.981	31	0.879	40	0.931	49
global calibration (ana::pro)	-	-	0.981	23	-	-	0.911	40
local calibration (ana::pro)	0.975	20	0.998	13	0.902	36	0.940	39
OPP *	0.996	8	0.996	11	0.978	17	0.991	14

* MES : variables longueur d'onde utilisées : 290 à 730 nm ; nombre de vecteurs latents : 3
 DCO : variables longueur d'onde utilisées : 207.5 à 730 nm ; nombre de vecteurs latents : 4

Tableau 2. Qualité des ajustements des modèles obtenus pour les quatre méthodes d'étalonnage

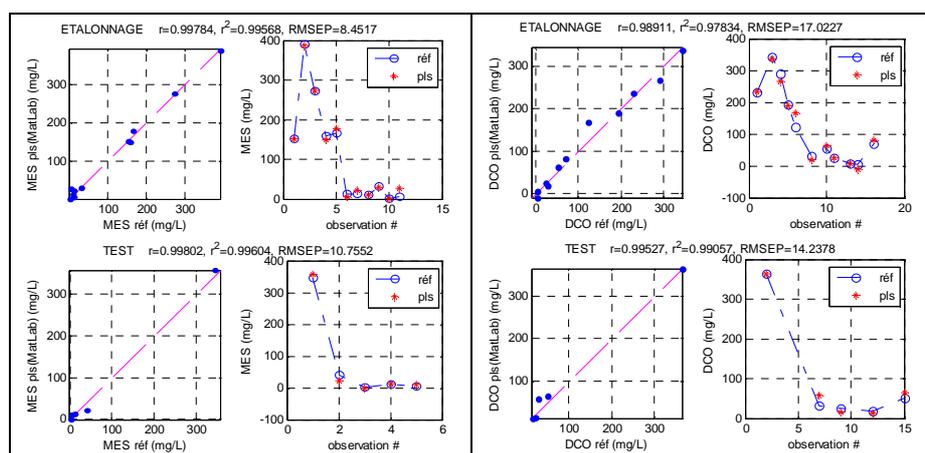


Figure 3. Représentation de l'ajustement des résultats de MES et DCO fournis par le capteur spectro:lyser en utilisant le programme OPP aux valeurs de MES et DCO de référence

4. Conclusions

Cette étude confirme l'intérêt de l'utilisation des spectromètres UV-visible pour le suivi de la qualité des effluents en réseau d'assainissement, en temps réel ou différé. Un étalonnage local est indispensable pour améliorer la précision des résultats et pour prendre en compte la variabilité et la spécificité de la composition locale des effluents. Le programme OPP fournit des modèles un peu plus précis que ceux établis avec le logiciel ana::pro fourni avec le capteur. En ce qui concerne le site expérimental choisi (bassin de retenue-décantation Django Reinhardt à Chassieu), d'autres campagnes de mesure en entrée et sortie de cet ouvrage doivent

être réalisées en 2006-2007 pour différents événements pluvieux, afin d'évaluer la variabilité éventuelle des relations établies par PLS avec le programme OPP. Après étalonnage en entrée et sortie du bassin, cette technique permettra de réaliser un suivi en continu de l'efficacité épuratoire de l'ouvrage, et d'envisager des règles nouvelles en vue d'une gestion améliorée. Des aspects non mentionnés dans cette communication, comme la mise en place et l'entretien de ce type de capteur, sont de première importance et doivent être pris en compte (Gruber *et al.*, 2005).

Remerciements

Ce travail a été réalisé avec le financement du programme Alban, programme de bourses de haut niveau de l'Union Européenne pour Amérique Latine, bourse n° E04D034837CO.

6. Références

- Abdi H. (2003). Partial least squares (PLS) regression. In: *Encyclopedia of social sciences research methods*. M. Lweis-Beck, A. Bryman, T. Futing (eds). Thousand Oaks, CA (USA): Sage, 792-795.
- Aji S., Tavoraro S., Lantz F., Faraj A. (2003). Apport du bootstrap à la régression PLS : application à la prédiction de la qualité des gazoles. *Oil & Gas Science and Technology*, 58(5), 599-608.
- Bardin J.-P., Barraud S. (2004). *Aide au diagnostic et à la restructuration du bassin de rétention de Chassieu*. Villeurbanne (France) : INSA de Lyon – URGC, rapport pour la Direction de l'Eau du Grand Lyon, juin 2004, 62 p.
- Dantas Filho H. A., de Souza E. S., Visani V., de Barros S. R., Saldanha T. C., Araújo M. C., Galvao R. K. (2005). Simultaneous Spectrometric Determination of Cu²⁺, Mn²⁺ and Zn²⁺ in Polivitaminic/Polimineral Drug Using SPA and GA Algorithms for Variable Selection. *J. Braz. Chem. Soc.*, 16(1), 58-61.
- de Bénédictis J., Bertrand-Krajewski J.-L. (2005). Mesurage de la concentration en polluants dans les eaux usées par spectrométrie UV-visible. *Actes de la conférence SHF-ASTEE-GRAlE « Autosurveillance, diagnostic permanent et modélisation des flux polluants en réseaux d'assainissement urbains »*, Marne-la-Vallée (France), 5-6 juin 2005, 201-210.
- Gruber G., Bertrand-Krajewski J.-L., de Bénédictis J., Hochedlinger M., Lettl W. (2005). Practical aspects, experiences and strategies by using UV/VIS sensors for long-term sewer monitoring. *Proceedings of the 10th International Conference on Urban Drainage*, Copenhagen (Denmark), 21-26 August 2005, 8 p.
- Hanafi M. (2004). Approche PLS : une hiérarchie des stratégies pour la détermination des variables latentes. *Actes des XXXVIèmes Journées de Statistique*, Montpellier (France), 24-28 mai 2004, 5 p.
- Hochedlinger M. (2005). *Assessment of combined sewer overflow emissions*. PhD thesis : Faculty of Civil Engineering, University of Technology Graz (Austria), June 2005, 174 p. + annexes.
- Langergraber G., Fleischmann N., Hofstädter F. (2003). A multivariate calibration procedure for UV/VIS spectrometric quantification of organic matter and nitrate in wastewater. *Water Science and Technology*, 47(2), 63-71.
- Tenenhaus M. (1998). *La régression PLS, théorie et pratique*. Paris (France) : Technip, 254 p.
- VCCL (2005). *Partial Least Squares Regression (PLSR)*. Virtual Computational Chemistry Laboratory, <http://146.107.217.178/lab/pls/>