

N° d'ordre : 06/2015-M/MT

République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
Université des Sciences et de la Technologie Houari Boumediene
Faculté de Mathématiques



MÉMOIRE

Présenté pour l'obtention du diplôme de Magister

En : Mathématiques

Spécialité : Statistiques Mathématiques et Probabilités

Par Amel SAHRAOUI

Thème

Etude des Propriétés des estimateurs de la régression PLS

Soutenu publiquement, le 22/10/2015 , devant le jury composé de :

Mme H. SAGGOU	Maître de Conférence/A	à l'U.S.T.H.B	Présidente.
Mme K. DJABALLAH	Maître de Conférence/A	à l'U.S.T.H.B	Directrice de mémoire.
Mme D. SEDDIKI	Maître de Conférence/A	à l'U.S.T.H.B	Examinatrice.

Table des matières

Table des matières	1
Table des figures	5
<i>Abréviations et symboles</i>	13
<i>Résumé</i>	14
<i>Chapitre 1 : Notion de la quasi-colinéarité</i>	1
1 Multicolinéarité	2
2 Quasi-colinéarité	4
3 Comment mesurer la quasi-colinéarité	6
4 Quasi-colinéarité et qualité du modèle	8
5 Régression linéaire multiple et difficultés d'interprétation	10
<i>Chapitre 2 : La régression PLS univariée</i>	12
1 Étapes de calcul de la régression PLS1	12
1.1 Principe de l'algorithme et passage à l'écriture matricielle	13
1.1.1 Étape 1	13
1.1.2 Étape 2	14
1.2 Algorithme	17
1.3 Calcul des coefficients de régression	18
1.4 Propriétés des composantes PLS	18

1.4.1	Optimalité des composantes	18
1.4.2	Indépendance (Non corrélation) entre composantes	20
1.5	Choix du nombre de composantes par validation croisée	21
1.5.1	Principe	21
1.5.2	Calcul du RSS	23
 Chapitre 3 : La regression PLS multivariée		26
1	Introduction	26
2	L'algorithme de la régression PLS2	27
3	Recherche et propriétés des composantes PLS	28
3.1	Recherche de t_1 et u_1	28
3.2	Recherche de t_2 et u_2	30
4	Propriété de la régression	32
4.1	Relations cycliques	32
4.2	Relations d'orthogonalité	33
5	Formules de décomposition	35
6	Équations de régressions PLS	38
7	Résultats de la régression PLS	39
7.1	Calcul des composantes t_h, u_h	39
7.2	Valeurs explicatives des composantes PLS	40
7.3	Interprétation des composantes PLS	40
8	Conclusion	41
 Chapitre 4 : Application		42

1	Problématique de la conduite d'un four rotatif de cimenterie	42
1.1	Introduction	43
1.2	Analyse en composante principale (ACP)	43
1.2.1	Tableau de données	44
1.2.2	Choix de l'origine	45
1.2.3	Matrice à diagonaliser	47
1.2.4	Qualité de représentation	49
1.2.5	Choix du nombre de composantes	50
1.3	Interprétation de l'ACP	50
1.3.1	Corrélation entre les variables et les facteurs	54
1.3.2	Nuage des points variables	54
2	Application de la régression PLS univariée	61
2.1	Interprétation de la régression PLS1 pour la variable Y_1	61
2.1.1	Variables importante dans la projection (VIP)	70
2.1.2	Equation du modèle	72
2.1.3	Coefficients d'ajustement de la Variable Y_1	73
2.2	Interprétation de la régression PLS1 pour la variable Y_2	80
2.2.1	Equation du modèle	89
3	Application de la régression PLS multiples à 2 variables	95
3.0.2	Equation de régression de Y_k sur t_1 et t_2	101
3.0.3	Règles d'interprétation	101
3.0.4	Variable importante dans la projection (VIP)	101
3.0.5	Equation du modèle	104
3.1	Conclusion	115

Annexe A **i****1 Programmes** **iii****Annexe B** **i****1 Application de la régression PLS pour diagnostiquer l'état d'un four** **i**1.1 Introduction i**2 Contexte de l'étude** **iv**2.1 Introduction iv2.2 Atelier cuisson v2.2.1 Tour échangeur vi2.2.2 Four rotatif vii2.2.3 Refroidisseur viii2.3 Choix des signaux utilisés xi**3 Problématique de la conduite d'un four rotatif de cimenterie** **xi****Annexe C** **i****1 Identification des variables** **i**1.1 Le tableau X : variables explicatives i1.2 Le tableau Y : variables à expliquer iv**2 Les sorties de l' ACP** **v**2.1 Matrice de corrélation de l'ACP v2.2 Valeurs propres et vecteurs xiii**3 Les Tableaux des sorties PLS1** **xv**

4 Les Tableaux des sorties PLS2

xxiv

Références

116

Table des figures

2.1	Evolution de la valeur du déterminant de $X_\alpha^t X_\alpha$ en fonction de α	5
2.2	Evolution de la valeur des coefficients estimés en fonction de α	5
4.1	Relation cyclique	33
1.1	Nuage des points variables de X et Y	55
1.2	Observation sur les axes F_1 et F_2	57
1.3	Les trajectoires des Coordonnées d'observations de F_1 , F_2 et F_3	58
1.4	La trajectoire de F_1	58
1.5	La trajectoire de F_2	59
1.6	La trajectoire de F_3	60
1.7	Biplot des axes F_1 et F_2	61
2.1	Ecart type	62
2.2	Qualité du modèle par nombre de composantes	63
2.3	Matrice des corrélations de PLS1	64
2.4	carte superposant les variables dépendantes sur c et les variables explicatives sur w^*	65
2.5	Les trajectoires de t_1 , t_2 et t_3	66
2.6	Observations sur les axes t_1 et t_2	67
2.7	Corrélations sur les axes t_1 et t_2	68
2.8	Les trajectoires de u_1 , u_2 et u_3	69
2.9	Observations sur $\tilde{u}_1$ et $\tilde{u}_2$	69
2.10	graphique des VIP à 1 composante	70
2.11	Graphique des VIP à 2 composantes	71
2.12	Graphique des VIP à 3 composantes	72

2.13 Y_1 /coefficient normalisé	74
2.14 le signal Y_1 /Résidus normalisés	75
2.15 prédiction du signal Y_1 en fonction du résidus normailès	76
2.16 Les observations en fonction des résidusnormalisés de la variable Y_1 (actives)	77
2.17 Les observations en fonction des résidus normalisés de la variable Y_1 (validation)	78
2.18 Graphique des DModX normalisées	79
2.19 Geaphique des DModY normalisées	80
2.20 Ecart type	80
2.21 Qualité du modèle par nombre de composantes	81
2.22 Regression PLS t_1, t_2	82
2.23 Carte des vecteurs (w^*, c_1) et (w^*, c_2)	83
2.24 Trajectoires de t_1, t_2, t_3	84
2.25 Carte des deux premières composantes PLS (les observations)	84
2.26 Plan des commposantes t_1 et t_2 PLS	85
2.27 trajectoires de u_1, u_2 et u_3	86
2.28 Carte des observations sur $\tilde{u}_1$ et $\tilde{u}_2$	87
2.29 Carte des VIP à 1composante (variable Y_2)	88
2.30 Carte des VIP à 2 composante (variable Y_2)	88
2.31 Carte des VIP à 3 composante (variable Y_2)	89
2.32 Coefficients normalisées en fonction des variables	91
2.33 Résidus normalisés de la variable Y_2	91
2.34 résidus normalisé et prédiction de la variable Y_2	92
2.35 Prédiction de la variable Y_2	92
2.36 Les observations en fonction des Résidus normalisés de la variable Y_2 Active	92

2.37	Les observations en fonction des Résidus normalisés de la variable Y2	
	(38-355_11/TE) Validation	93
2.38	Graphique des DModX,N	94
2.39	Graphique des DModY,N	94
3.1	Qualité du modèle par nombre de composantes	95
3.2	Plan des composantes t_1 et t_2	96
3.3	Graphique des écarts- types des X	98
3.4	Régression PLS. Carte des variables w^*/c_1 et w^*/c_2	99
3.5	Trajectoire de t_1 et t_2	99
3.6	Trajectoire de u_1 et u_2	100
3.7	Observations sur les axes t_1 et t_2	101
3.8	Graphique des VIP (1composante)	102
3.9	Graphique des VIP (2 composante)	103
3.10	Régression PLS à 2 composantes .coefficients normalisés de régression	
	PLS	106
3.11	Résidus normalisés de la variable Y_1 (Actives ,Validation)	107
3.12	Prédiction de la variable Y_1	107
3.13	Résidus normalisés de la variable Y_1 (active)	108
3.14	Résidus normalisés de la variable Y_1 (validation)	109
3.15	Coefficients normalisés Température en C° de l'air secondaire (Y_2) . .	110
3.16	Prédictions et résidus normalisés	110
3.17	Résidus normalisés en fonction de la variables Y_2	111
3.18	Prédiction de la variable Y_2	111
3.19	L'histogramme des résidus centrés réduits	112
3.20	Résidus normalisés en fonction de la variable Y_2	113
3.21	Graphique des DModX,N	114
3.22	Graphique des DModY,N	115
2.1	Tour échangeur type Humboldt	vii

2.2	Four et Profil deTempératures	viii
2.3	Profile de température dans le refroidisseur	ix
2.4	Profile de température, zones et phases d'après K.S. Stadler [1]	x
2.5	Tour-Four-Refroidisseur	xi

Etude des Propriétés des estimateurs de la régression PLS

Amel SAHRAOUI

Département de Probabilités et Statistiques

Faculté des Mathématiques

Université des Sciences et de la Technologie HOUARI Boumediene

U. S. T. H. B

21 novembre 2015

“ A ma très chère mère ”

Remerciements

Je remercie Dieu de m'avoir donné le courage et m'avoir mis sur le chemin du savoir, d'abord je remercie Mme K. DJABALLAH pour m'avoir proposé le sujet et apporté le soutien nécessaire. De part ses qualités pédagogiques, ses précieux conseils et ces qualités humaines.

Mes remerciements vont à Mme H. SAGGOU de l'honneur qu'elle m'a fait en acceptant de présider le jury et de l'intérêt qu'elle a porté à ce travail.

Bien évidemment, je remercie Mme D. SEDDIKI de l'honneur qu'elle m'a fait d'être l'examinatrice de ce mémoire, du temps et du soin qu'elle a apporté à la rédaction de son rapport.

Un grand merci à tous ceux qui m'ont aidé pour que ce travail voit le jour, Mr N.Kouadri, Mr A.Bensmail, Mr M.Hanafi, Mlle A.Baouta, Mr A.Kessira, Melle D.Sedkaoui. Je remercie aussi ma sœur Dalel et mes amies Mlle M.Zeghidi, Mme N.Atoum, Mme Kahina, Mme Fadila, Mme T.Ouelmane, Melle Zhore, Mr R.Chergui, Melle Nawel, Mr Oualid et tous ceux qui m'ont soutenu durant mon travail mes parents mes soeurs et frères et leurs enfants.

Je tiens à exprimer ma reconnaissance à toute ma famille, mes amis et tous ceux qui m'ont aidé de près ou de loin.

Abréviations et symboles

<i>PLS</i>	La régression des moindres carrés partiels (Partial Least Squares).
<i>PLS1</i>	La régression des moindres carrés partiels à une seule variable.
<i>PLS2</i>	La régression des moindres carrés partiels à plusieurs variables.
<i>PCR</i>	la Régression sur Composantes Principales.
<i>ACP</i>	Analyse en composantes principales.
<i>RESS</i>	Pouvoir explicatif.
<i>PRESS</i>	Pouvoir prédictif.
<i>MCO</i>	Moindre carrée ordinaire.
<i>DDL</i>	Degré de liberté.
<i>MCE</i>	Mean squared error (Moyen des carrés des erreurs).
<i>RMCE</i>	Root mean squared error (Racine du Moyen des carrés des erreurs).

Résumé :

Il s'agit d'étudier et de construire des modèles de prédictions d'un groupe de variables à partir d'un autre groupe de variables. Quand on a un problème de quasi colinéarité, la régression linéaire multiple n'est pas adaptée et par conséquent on fait appel à d'autres méthodes ; (régression pas à pas, régression ridge, régression orthogonale, et la régression PLS). Nous avons choisi d'étudier la régression PLS. Pour ce faire on a détaillé l'algorithme de cette méthode dans le cas où le groupe de variables à prédire est univarié ou multivarié, la théorie de la régression PLS univarié et la régression PLS multivarié. En ce qui concerne l'application, des données ont été traitées par différentes stratégies : ACP des variables explicatives, régression PLS univarié à chacune des variables à expliquer et régression PLS2, chaque application est bien interprétée. En ce qui concerne l'application des données issues d'un processus de fabrication de ciment. Ces données impliquent un groupe de variables explicatives de 50 variables et un groupe de variable à expliquer constitué de deux variables. Plusieurs stratégies d'analyse ont été élaborées : ACP du tableau des données explicatives, régression PLS univarié de chacune des variables à expliquer sur les données explicatives et finalement une régression PLS2 sur les données explicatives, plusieurs sorties graphiques et l'algorithme du logiciel (MATLAB, XLSTAT) sont répertoriés en annexe.

Introduction générale

Il est d'usage courant dans beaucoup de domaines scientifiques de vouloir décrire, expliquer ou prédire le comportement d'une ou plusieurs variables (les variables expliquées, encore appelées réponses ou variables endogènes) par des variables plus faciles à mesurer ou à contrôler (les variables explicatives, encore appelées facteurs ou respectivement variables exogènes).

Lorsque ces variables explicatives sont peu nombreuses, qu'elles ne sont pas trop redondantes (colinéaires), la régression linéaire multiple est la technique appropriée pour ajuster un modèle aux données. Cependant, si cette condition n'est pas remplie, la régression multiple peut s'avérer peu efficace. C'est souvent le cas, par exemple, en chimie où l'on essaie souvent d'estimer des paramètres biochimiques (par exemple teneur en protéines des aliments) par des mesures fournies par la spectroscopie infrarouge qui présentent souvent une grande colinéarité.

La régression des moindres carrés partiels (Partial Least Squares (PLS)) est une méthode statistique permettant de construire des modèles prédictifs lorsque les variables explicatives sont nombreuses et très corrélées. L'accent est mis dans cette méthode aussi bien sur la prédiction que sur l'exploration. Cette dernière problématique se partage les mêmes objectifs aussi des techniques descriptives d'analyse factorielle.

On peut en principe inclure de nombreuses variables explicatives dans un modèle de régression multiple. Cependant, si on laisse le nombre de variables explicatives

croître sans contrôle voire dépasser le nombre d'observations, le risque est d'obtenir un modèle qui s'ajuste parfaitement aux données mais dont les estimations obtenues par combinaisons des variables explicatives s'avèrent incapables de prédire correctement les valeurs des variables expliquées pour de nouvelles observations souvent dans ces situations de saturation du modèle, bien que les variables explicatives a priori soient nombreuses, seul un petit nombre de variables latentes (combinaisons linéaires des variables explicatives) permet de rendre compte de la variabilité des variables expliquées.

L'objectif de la régression PLS est d'extraire des variables latentes responsables de la variation des variables explicatives, qui modélisent au mieux le comportement des variables expliquées. Pour prédire les variables expliquées à partir des variables explicatives, la régression PLS procède alors de manière indirecte en extrayant des composantes T à partir des variables explicatives X pour estimer des composantes U qui permettront de calculer les réponses Y .

Si le nombre de composantes extraites est supérieur ou égal à la dimension de l'espace des facteurs, la régression PLS fournit alors la solution des moindres carrés ordinaires (MCO). Cependant, la méthode PLS permet en général de se limiter à un nombre restreint de composantes. Le choix du nombre de composantes à extraire est effectué sur la base de validation croisée. Il s'agit de construire le modèle PLS avec un nombre fixé de variables explicatives pour un sous ensemble particulier des données, puis à tester le modèle obtenu sur un autre sous ensemble des données en choisissant le nombre de composantes qui minimise l'erreur d'estimation.

La régression PLS a été initialement développée par Herman Wold dans les années 60, puis fut popularisée auprès des chimistes par son propre fils, Svante Wold dans ce cadre, elle a été appliquée au contrôle de processus industriels.

A partir de ces travaux un certain nombre de modifications ou d'extensions de la régression PLS ont été proposés. Par exemple, l'algorithme SIMPLS5, proposée par De Jong en 1993, donne toujours des résultats très proches du modèle originel des

moindres carrés partiels, tout en facilitant considérablement les calculs en particulier si le nombre de variables explicatives est élevé. Le modèle de régression continuement paramétrable, (continuum régression), proposé par [Stone et Brooks, 1990], permet de faire varier continument le model de régression suivant les valeurs d'un paramètre α , $0 \leq \alpha \leq 1$, du modèle de la régression multiple ($\alpha = 0$) au modèle de la régression sur composantes principales ($\alpha = 1$) en passant par celui des moindres carrés partiels ($\alpha = 0,5$). En 1992, De Jong et Kiers ont proposé une technique apparentée, la régression sur covariables principales (principal covariates regression).

En dernière analyse, la technique PLS peut constituer, en chimiométrie comme dans bien d'autre domaines, un outil privilégié d'analyse dans la mesure où il s'avère possible d'interpréter les composantes en termes de facteurs sous-jacents aux variables explicatives.

La régression PLS demeure une technique de modélisation statistique en constante évolution. On pourra consulter [Geladi et Kowalski, 1986] pour une introduction classique à l'application des moindres carrés partiels en chimométrie parmi les références bibliographiques citées, [Naes et Martens, 1985] et [de Jong, 1993] constituent des présentations techniques plus détaillées de la méthodologie PLS une synthèse récente, couvrant un très large champ, est disponible désormais en français [Tenenhaus, 1998].

Le présent mémoire propose une présentation détaillée de la méthode PLS tant sur le plan théorique que pratique. Il est organisé en quatre chapitres. Les annexes qui accompagnent ce mémoire fournissent les codes des simulations présentées dans le mémoire les programmes ont été fait par le logiciel MATLAB ainsi qu'une présentation de la cimenterie. La mise en œuvre de la PLS sur des données réelles a été réalisée par logiciel XLSTAT.

Chapitre

1

Notion de la quasi-colinéarité

Nous allons nous intéresser dans ce chapitre aux modèles de prédiction à partir d'un groupe de variables et à leur construction. Nous ne reviendrons pas sur la régression linéaire multiple qui est l'outil statistique le plus communément utilisé pour résoudre ce type de problématique. Par contre, nous présenterons plus en détail la question de la quasi-colinéarité en régression linéaire multiple, car elle justifie l'introduction de la régression PLS1 dite méthodes de régression sur variables latentes.

De nombreux problèmes dans la pratique peuvent être présentés sous la forme d'un problème de régression. L'objectif peut alors consister être double :

- (i) Décrire les relations entre une variable ou plusieurs variables, appelée variables expliquées notée Y , et p variables explicatives notées X .
- (ii) Prédire pour des nouvelles observations prises sur les variables X les valeurs de ces observations pour les variables Y .

Le modèle linéaire de la régression linéaire multiple est défini par l'équation matricielle ci-dessous :

$$y = X\beta + \varepsilon \tag{1}$$

Dans laquelle :

- y est le vecteur $(y_1, y_2, \dots, y_n)^t$ des n valeurs observées de la variable expliquée Y .

- X est la matrice des données à n lignes et p colonnes, la colonne j ($j = 1, \dots, p$) étant définie par le vecteur $(x_1^j, x_2^j, \dots, x_n^j)^t$.
- $\beta = (\beta_1, \dots, \beta_p)^t$ est le vecteur des coefficients de régression.
- ε est le vecteur $(\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n)^t$ de la supposée variable résiduelle ε de variance σ^2 . On note R la matrice de corrélation entre les p variables $X_1, \dots, X_p$. On la suppose inversible (de rang p) : elle possède donc p valeurs propres strictement positives. La méthode classique d'estimation des paramètres est fondée sur le critère des moindres carrés. L'estimateur de la régression linéaire multiple est donné par la formule ci-dessous :

$$\hat{\beta} = \left(\frac{X^t X}{n} \right)^{-1} X^t Y \quad (2)$$

En notant r le vecteur des coefficients de corrélation observés $(r_1, r_2, \dots, r_p)^t$ entre les variables explicatives X_j et la variable expliquée Y . Le coefficient de détermination est par définition le carré du coefficient de corrélation linéaire de la variable expliquée Y et de $X\hat{\beta}$, il est égal à :

$$R^2 = r^t R^{-1} r \quad (3)$$

L'estimation d'un modèle par la régression linéaire multiple a l'avantage d'être relativement simple à comprendre, car il possède des propriétés géométriques, et sa mise en œuvre est facile seulement en présence de données fortement corrélées entre elles, la régression linéaire multiple peut présenter quelques limitations que nous proposons de détailler dans ce chapitre.

1 Multicolinéarité

On parle de multicolinéarité entre plusieurs variables lorsque l'une d'elles s'exprime comme combinaison linéaire des autres. On dira qu'il existe une relation li-

néaire "exacte" entre les variables explicatives.

L'exemple didactique ci-dessous permet d'illustrer ce propos :

Y	-1	14	14	14	6	-12	7	16
X_1	5	18	-22	8	3	-13	-4	3
X_2	35	27	-13	30	7	0	7	-2
X_3	40	45	-35	38	10	-13	3	1

Tableau 1 : Exemple didactique, les variables sont en lignes, les individus en colonnes

Une relation linéaire exacte existe entre les 3 variables X_1 , X_2 et X_3 , en effet :

$$X_3 = X_1 + X_2$$

Soient X la matrice $[X_1, X_2, X_3]$ et X^t sa transposée. Comme $\hat{\beta} = \frac{1}{n}(X^t X)^{-1} X^t Y$, le calcul de $\hat{\beta}$ nécessite l'inversion de $X^t X$. Or, si au moins deux variables explicatives sont colinéaires, le déterminant de $X^t X$ est nul, et de ce fait, on ne peut pas calculer l'inverse de $X^t X$.

Pour remédier à ce problème d'inversion, on fait appel à l'inverse généralisée de Moore Penrose. Dans ce cas, le calcul de $\hat{\beta}$ n'est pas unique, un ensemble de $\hat{\beta}$ sont alors possibles parmi lesquels on choisira celui dont la norme est minimale.

Rappelons que l'inverse généralisée de Moore Penrose, ou pseudo inverse, de $X^t X$ peut se définir ainsi :

Définition :

Soit la décomposition en valeurs propres de X :

$X^t X = U \Delta U^t$ où U est une matrice orthogonale et $\Delta = (\Delta_{ij})$ une matrice diagonale.

L'inverse généralisée de Moore Penrose de $X^t X$ est la matrice $(X^t X)^{-1}$ définit de la manière suivante :

$$(X^t X)^{-1} = U D^{-1} U^t$$

$$\text{avec : } d_{ii}^{-1} = \begin{cases} \frac{1}{\Delta_{ii}} & \text{si } \Delta_{ii} \neq 0 \\ 0 & \text{si } \Delta_{ii} = 0 \end{cases} \quad \text{et } d_{ij}^{-1} = 0 \text{ pour } i \neq j$$

En l'occurrence, dans l'exemple présenté précédemment, on trouve

$$\hat{\beta} = \begin{pmatrix} 0,5638 \\ -0,4738 \\ 0,0900 \end{pmatrix}$$

Ainsi, le problème de l'inversion de $X^t X$ engendré par la multicolinéarité de variables explicatives peut être résolu. Il suffit d'ajouter une contrainte implicite norme minimale de $\hat{\beta}$, prise en compte par le calcul de l'inverse généralisée de $X^t X$.

2 Quasi-colinéarité

Le plus souvent les données réelles présentent par nature une relation " presque " linéaire entre plusieurs variables dues à des corrélations fortes entre celles-ci : on parlera alors de quasi-colinéarité. En situation de quasi-colinéarité, si on tolère une perturbation des variables X par une matrice relativement petite, on aboutit à des variables parfaitement colinéaires au sens de la définition ci-dessus.

En situation de quasi-colinéarité, la matrice $X^t X$ est inversible : l'inversion de $X^t X$ ne pose cette fois pas de problème. En revanche, cette situation induit une instabilité des coefficients.

Afin d'illustrer cette situation, reprenons l'exemple précédent (tableau 1) et faisons varier cette fois la valeur prise par la 8ème observation de la variable X_3 : $X_3(8) \leftarrow x = X_3(8) + \alpha$, où α est un paramètre qui varie entre 0 et 1. Lorsque $\alpha = 0$, x prend la valeur 28,6 et, comme on l'a déjà vu, X_1 , X_2 et X_3 sont multicolinéaires. Lorsque α est proche de 0, on est en présence de quasi-colinéarité.

Voyons maintenant ce qui se passe lorsqu'on fait varier avec un pas de 0,01 ci-dessous les déterminants successifs de $X_\alpha^t X_\alpha$ en fonction de α :

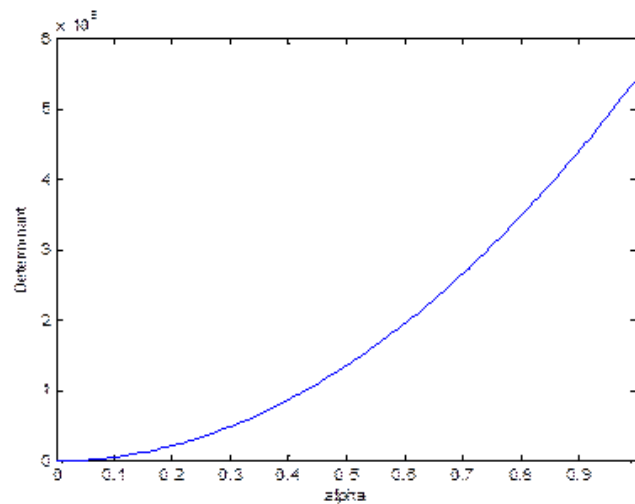


FIG. 2.1: Evolution de la valeur du déterminant de $X_{\alpha}^t X_{\alpha}$ en fonction de α

On constate que, plus la valeur de α est proche de 0, plus celle du déterminant tend vers 0 : on se rapproche de la multicollinéarité parfaite. On est donc en présence d'une quasi-colinéarité de plus en plus forte.

On effectue une régression pour chaque valeur de α . On représente graphiquement les estimations de chacun des coefficients :

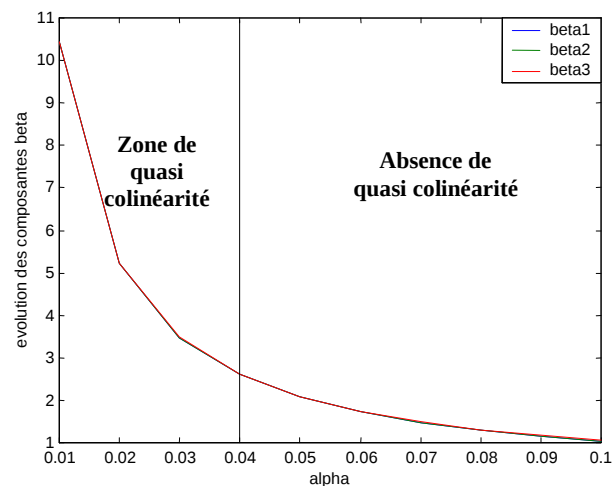


FIG. 2.2: Evolution de la valeur des coefficients estimés en fonction de α

On voit que les coefficients varient énormément : on passe de valeurs proches de 11 à environ 0 pour une faible variation de α . Cela signifie qu'une faible erreur de

mesure (par exemple 1,02 au lieu de 1,05 pour $X_3(8)$) nous fournira des modèles très différents.

Sur le graphique 2, on a placé une "limite de quasi-colinéarité" à $\alpha = 0,04$ car au delà de ce seuil les coefficients estimés varient nettement moins. Toutefois, ce qui apparaît être arbitraire peut se justifier par la suite.

3 Comment mesurer la quasi-colinéarité

On s'est jusqu'à présent intuitivement basés sur la valeur du déterminant pour déterminer si on était en présence de quasi-colinéarité ou non. Cependant, l'une des difficultés lorsqu'on travaille avec le déterminant est qu'il est très sensible aux changements d'échelles. Cela vient de la relation : $\det(\lambda A) = \lambda^p \det(A)$ pour une matrice A de dimension $p \times p$.

Ainsi, par exemple, si on dispose de données concernant 10 poids différents mesurés sur plusieurs individus et exprimés en livres. On calcul le déterminant de $X^t X$ et on trouve 0,01, ce qui est relativement proche de 0, on pourrait donc penser que les données sont quasi-colinéaires. Maintenant, si au lieu de mesurer les poids en livres on les prend en kilogrammes ; la relation entre les deux est la suivante : $K = L \times 2,2046$.

On a donc : $\det(X_k^t X_k) = 2,2046^{10} \det(X_k^t X_k) = 27,11$, ce qui est suffisamment éloigné de 0 pour qu'on ne songe cette fois pas à la quasi colinéarité. Pourtant on a simplement multiplié par environ 2 toutes les données, donc si on est très proche de la multicollinéarité lorsque les mesures sont effectuées en livres, on ne devrait pas en être si éloigné lorsqu'on passe aux kilogrammes (le déterminant est environ 2711 plus grand).

Le critère du déterminant de $X^t X$, s'il s'avère utile pour déterminer la présence d'une multicollinéarité parfaite entre les variables explicatives, n'est donc pas très pertinent lorsqu'il s'agit de détecter une quasi-colinéarité.

Il apparaît donc nécessaire de trouver d'autres indices permettant de détecter la présence d'une quasi-colinéarité et d'en mesurer la force.

– Les valeurs propres et le conditionnement

En Analyse sur Composantes Principales (ACP), on construit de nouveaux facteurs qui sont des combinaisons linéaires des variables dont on dispose déjà et restituent toute l'information qu'elles contiennent. Les valeurs propres de la matrice de variance-covariance de X correspondent à la variance de chacune des composantes principales et donc plus une valeur propre est grande, plus la composante principale qui lui est associée explique une grande part de l'information contenue dans X . Les valeurs propres sont généralement classées par ordre décroissant : $\lambda_1 \geq \lambda_2 \dots \geq \lambda_k$ (avec k le rang de X).

Pour chaque valeur propre λ_j on définit

$$CI_j = \frac{\lambda_1}{\lambda_j}$$

La première valeur propre aura donc toujours un CI égal à 1. Le plus grand CI_j est celui associé à la plus petite valeur propre (λ_k), il s'agit du conditionnement :

$$CN = \frac{\lambda_1}{\lambda_k}$$

De grandes valeurs de CI_j ou CN implique donc que la part d'information restituée par la $j^{\text{ème}}$ composante principale (voire la $k^{\text{ème}}$ dans le cas de CN) est faible comparée à celle restituée par la première : elle n'apporte presque pas d'information supplémentaire ce qui laisse à supposer qu'au moins deux combinaisons linéaires des X_i sont quasiment égales, on est par conséquent en présence de quasi-colinéarité.

Belsley et *al* (1980) préconisent de considérer qu'un conditionnement supérieur ou égal à 30 est synonyme d'une quasi-colinéarité.

- Le facteur d'inflation de la variance (*VIF*)

Le coefficient de corrélation multiple de X_i avec les autres régresseurs est noté

$$R^2(X_i | X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_k)$$

ou encore R_i^2 et correspond au R^2 résultat de la régression de X_i sur l'ensemble des autres régresseurs, autrement dit :

$$R_i^2 = \frac{\text{variance de } X_i \text{ expliquée par les autres régresseurs}}{\text{variance totale de } X_i} = \frac{\sum_{j=1, j \neq i}^k (\hat{X}_{ij} - \bar{X}_i)^2}{\sum_j (X_{ij} - \bar{X}_i)^2}$$

Ainsi, plus R_i^2 est élevé et proche de 1, plus la variance de X_i est expliquée par les autres régresseurs donc plus la quasi-colinéarité de X_i avec l'ensemble des autres variables explicatives est forte.

C'est à partir de cette constatation qu'on définit le facteur d'inflation de la variance ou variance inflation factor (*VIF*) :

$$VIF_i = \frac{1}{1 - R_i^2}$$

Comme on l'a vu, si on est en présence d'une forte quasi-colinéarité, R_i^2 sera proche de 1 donc $1 - R_i^2$ sera proche de 0 et VIF_i tendra vers l'infini : une grande valeur de VIF_i indique une quasi-colinéarité de X_i avec d'autres régresseurs.

4 Quasi-colinéarité et qualité du modèle

L'exemple auquel nous nous sommes intéressés jusqu'à présent pour illustrer la notion de quasi-colinéarité et les problèmes qu'elle implique est basé sur des données simulées et destinées à bien mettre en évidence le problème de casicolinéarité ; voyons maintenant ce qu'il en est dans le cas d'une étude sur des données réelles.

Nous considérons tout d'abord l'exemple proposé par Cornell [5], également ana-

lysé par Tenenhaus [22] : l'indice d'octane moteur Y de 12 mélanges différents a été enregistré afin de déterminer l'influence des composants

Le conditionnement de la matrice X des variables explicatives (X_1 à X_7) est égal à 864,6123, ce qui est très important, on se trouve donc en présence d'une quasi-colinéarité des variables explicatives.

Afin d'illustrer le phénomène d'instabilité des coefficients estimés par régression linéaire multiple, lié à la quasi-colinéarité des variables, nous sélectionnons deux individus très proches dans nos données et nous construisons un modèle de prédiction sans l'un de ces individus, puis sans l'autre. Les données étant similaires, qu'on retire l'un ou l'autre des individus, on s'attend à trouver deux modèles de prédiction semblables.

Lorsqu'on observe les distances entre individus, on constate que les individus 8 et 9 sont extrêmement proches ($d(8, 9) = 0,332$).

Les matrices de variance-covariance de $X_{(-8)}$ (X privée de l'individu 8) et de $X_{(-9)}$ (X privée de l'individu 9) ont quasiment la même structure. Les modèles de prédiction estimés à partir de $X_{(-8)}$ et $X_{(-9)}$ devraient donc être très similaires, voyons si tel est le cas :

	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	$\hat{\beta}_6$	$\hat{\beta}_7$
$\hat{\beta}_{(-8)}$	-26,376	-3,1375	1,6478	-12,198	-1,6023	11,1121	30,5539
$\hat{\beta}_{(-6)}$	-79,0269	-8,1116	80,8098	-16,9141	-6,4697	6,2541	23,4585

Les deux estimations sont en fait très différentes alors que les matrices de variance-covariances étaient presque identiques, c'est la manifestation d'une quasi-colinéarité. Cela signifie qu'en modifiant très peu les données (en prenant l'individu 8 au lieu du 9 ou inversement), on obtient, par la régression linéaire multiple, un modèle de prédiction totalement différent. On retrouve ici la notion d'instabilité des coefficients liée à la quasi-colinéarité.

Ceci illustre les conséquences néfastes de la présence d'une quasi-colinéarité sur le modèle de prédiction par régression linéaire multiple.

5 Régression linéaire multiple et difficultés d'interprétation

Outre l'instabilité des coefficients déjà évoquée, d'autres problèmes d'interprétation peuvent survenir en régression linéaire multiple.

On s'intéresse au calcul de $\hat{\beta}$ dans le cas où l'on a deux régresseurs : $X = [x_1 x_2]$. Pour plus de clarté, on suppose toutes les variables centrées et réduites.

$$\hat{\beta} = (X^t X)^{-1} X^t Y$$

Or, comme les variables sont centrées et réduites :

$$(X^t X)^{-1} = \frac{1}{1 - r_{12}^2} \begin{bmatrix} 1 & -r_{12} \\ -r_{12} & 1 \end{bmatrix}$$

Où r_{12} est le coefficient de corrélation entre x_1 et x_2 . Et

$$X^t Y = \begin{bmatrix} r_{1Y} \\ r_{2Y} \end{bmatrix}$$

On a donc

$$\hat{\beta} = \begin{bmatrix} \frac{r_{1Y} - r_{12}r_{2Y}}{1 - r_{12}^2} \\ \frac{r_{2Y} - r_{12}r_{1Y}}{1 - r_{12}^2} \end{bmatrix}$$

Le signe de $\hat{\beta}_i$ va donc dépendre non seulement de celui de $r_{iY} - r_{ij}r_{jY}$: plus la variable est fortement corrélée avec Y , plus c'est le signe de cette corrélation qui va l'emporter ; par contre si l'autre variable est plus fortement corrélée avec Y ou que les variables explicatives sont très corrélées entre elles le signe peut être différent.

On voit donc bien ici que plusieurs facteurs rentrent en compte dans la détermination du signe de $\hat{\beta}$: les corrélations entre Y et les variables explicatives et les corrélations entre les variables entre elles. Ceci ne nous permet pas de conclure que

la quasi-colinéarité joue vraiment un rôle dans le problème de l'interprétation des signes car, même si une forte corrélation entre les variables aura une influence et que cette forte corrélation implique une quasi-colinéarité, l'inverse est faux. Comme le problème d'incohérence des signes ne vient pas systématiquement d'une forte corrélation entre les variables, on peut très bien rencontrer cette incohérence sans qu'il y ait de quasi-colinéarité.

Une chose à noter ici est que, même s'il peut sembler intéressant d'interpréter le signe des $\hat{\beta}_i$ pour savoir comment les variables explicatives peuvent influencer Y , il faut rester très prudent quant à cette interprétation.

Plusieurs éléments ressortent à l'issue de ce premier paragraphe. Tout d'abord, la multicollinéarité parfaite pose un problème pour le calcul des coefficients estimés en régression linéaire multiple ; toutefois, l'utilisation de l'inverse généralisée permet d'y remédier. D'autre part, la quasi-colinéarité, elle n'engendre généralement pas de problème particulier pour l'estimation du modèle, c'est en revanche la qualité de ce modèle qui est affectée : les estimations des coefficients sont instables et l'on peut rencontrer des difficultés d'interprétation. Enfin, dans tous les cas, il est préférable d'interprétation le signe des coefficients estimés avec beaucoup de prudence.

Comme on l'a vu, la quasi-colinéarité des variables explicatives pose un problème dans le cadre de la régression linéaire multiple. Pourtant lorsqu'on dispose de données réelles, ce problème est très fréquent. Des méthodes alternatives telles que la Ridge Régression, la Régression sur Composantes Principales ou *PCR* (Principal Component Régression), ou encore la régression *PLS* (Partial Least Square) ont été proposées. Ces méthodes alternatives sont basées sur un principe commun, pour construire le modèle, elles utilisent toutes des variables latentes

Chapitre

2

La régression PLS univariée

Etant donné les limitations évoquées dans le chapitre 1 de la régression linéaire multiple, la régression PLS présente plusieurs avantages :

- Dans le cas où une des variables explicatives serait une stricte combinaison linéaire des autres, la régression linéaire ne peut avoir lieu sans enlever au moins une variable explicative de l'analyse. La régression PLS ne présente pas cet inconvénient.

- La régression PLS peut traiter des cas où les individus seraient moins nombreux que les variables explicatives. La régression linéaire ne peut le faire.

- Lorsque les variables explicatives sont fortement corrélées entre-elles, la régression linéaire devient très peu pertinente, au sens où les coefficients qui en ressortent deviennent très instables lorsque l'on « bruite » les données (on fait varier, de manière aléatoire et très légère, les données de l'échantillon). La régression PLS, basée sur des critères de covariance, est considérée comme étant plus robuste. Les coefficients demeurent stables et gardent une certaine significativité, même en présence de corrélations fortes entre les variables.

Ce chapitre est dédié aux mécanismes d'estimation de la régression PLS1.

1 Étapes de calcul de la régression PLS1

On considère que les vecteurs x_j des variables explicatives et que le vecteur y de la variable à expliquer sont centrés. Les variables peuvent être centrées-réduites,

mais seul le centrage des variables est utilisé lors du passage à l'écriture matricielle. On note X la matrice des variables explicatives de dimension $n \times p$. On va donc chercher le vecteur $b_h = (b_{h1}, \dots, b_{hp})$ des coefficients de régressions du modèle à h composantes.

1.1 Principe de l'algorithme et passage à l'écriture matricielle

La stratégie de calcul va consister à chercher des composantes « variables latentes » $(t_1, t_2, \dots, t_h)$ qui s'expriment en fonction des variables explicatives $(x_1, \dots, x_p)$ en trouvant une série de coefficients (pour chaque composante : un coefficient associé à chacune des variables) pour chaque composante, à la manière d'une régression linéaire, à la différence près que les coefficients sont calculés sur la base d'un critère de covariance.

1.1.1 Étape 1

On construit la première composante t_1 comme une combinaison linéaire des p variables explicatives x_j . Les coefficients $w_1 = (w_{11}, \dots, w_{1j}, \dots, w_{1p})$ de cette combinaison linéaire cherchent à "résumer" au mieux les variables explicatives x_j et à "expliquer" au mieux la variable y :

$$t_1 = w_{11}x_1 + \dots + w_{1p}x_p$$

$$w_{1j} = \frac{\text{cov}(x_j, y)}{\sqrt{\sum_{k=1}^p \text{cov}^2(x_k, y)}}$$

On effectue ensuite une régression simple de y sur t_1

$$y = c_1 t_1 + y_1$$

Où y_1 est le vecteur des résidus et c_1 est le coefficient de régression de y sur t_1 :

$$c_1 = \frac{\text{cov}(y, t_1)}{\sigma^2(t_1)}$$

$\sigma^2(t_1)$ désigne la variance de t_1 . On en déduit une première équation de régression :

$$y = \overbrace{c_1 w_{11}}^{b_{11}} x_1 + \cdots + \overbrace{c_1 w_{1p}}^{b_{1p}} x_p + y_1$$

Le passage à l'écriture matricielle sachant que les vecteurs x_j et y sont centrés :

$$\text{cov}(x_j, y) = \sum_{i=1}^n x_{ij} y_i = x_j^t y$$

$$\text{cov}(y, t_1) = \sum_{i=1}^n y_i t_{1i} = y^t t_1$$

$$\sigma_{t_1}^2 = \sum_{i=1}^n t_{1i} t_{1i} = t_1^t t_1$$

D'où :

$$w_1 = \frac{X^t y}{||X^t y||}$$

$$t_1 = X w_1$$

$$c_1 = \frac{y^t t_1}{t_1^t t_1}$$

$$b_1 = c_1 w_1$$

1.1.2 Étape 2

On construit une deuxième composante t_2 , non corrélée à t_1 et expliquant bien le résidu y_1 . Cette composante t_2 est combinaison linéaire des résidus x_{1j} des régressions simples des variables x_j sur t_1 :

$$t_2 = w_{21} x_{11} + \cdots + w_{2j} x_{1j} + \cdots + w_{2p} x_{1p}$$

$$w_{2j} = \frac{\text{cov}(x_{1j}, y_1)}{\sqrt{\sum_{j=1}^p \text{cov}^2(x_{1j}, y_1)}}$$

Pour calculer les résidus x_{1j} , on réalise une régression linéaire de toutes les variables x_j sur t_1 :

$$x_j = p_{1j}t_1 + x_{1j}$$

où x_{1j} est le vecteur de résidus et p_{1j} est le coefficient de régression :

$$p_{1j} = \frac{\text{cov}(x_j, t_1)}{\sigma_{t_1}^2}$$

D'où

$$x_{1j} = x_j - p_{1j}t_1$$

On effectue ensuite une régression de y sur t_1 et t_2 :

$$y = c_1t_1 + c_2t_2 + y_2$$

Où c_1 est le coefficient de régression de la première étape, c_2 est le coefficient de la régression simple de y_1 sur t_2 et y_2 le vecteur des résidus de cette régression :

$$y_1 = c_2t_2 + y_2$$

D'où

$$c_2 = \frac{\text{cov}(y_1, t_2)}{\sigma_{t_2}^2}$$

Cette équation, à l'étape 2, paraît déjà fort complexe sous forme de formule. Néanmoins, dans le cas d'un exemple concret, elle est écrite de manière tout à fait similaire à une régression linéaire multiple (lorsque les coefficients sont connus numériquement). Seuls les coefficients affectés à chaque variable varient.

Nous verrons dans la section 3 de ce chapitre comment calculer le vecteur b_{2p} des

coefficients de l'équation de régression :

$$y = b_{21}x_1 + \cdots + b_{2p}x_p + y_2$$

Le passage à l'écriture matricielle sachant que les x_{1j} et y_1 sont centrés et en notant $X_1 = (x_{11}, \dots, x_{1p})$ la matrice des résidus x_{1j} :

$$p_1 = \frac{X_1^t t_1}{t_1^t t_1}$$

$$X_1 = X - t_1 p_1$$

$$y_1 = y - c_1 t_1$$

$$w_2 = \frac{X_1^t y_1}{\|X_1^t y_1\|}$$

$$t_2 = X_1 w_2$$

$$c_2 = \frac{y_1^t t_2}{t_2^t t_2}$$

Étapes suivantes :

Naturellement, on peut encore ajouter des étapes supplémentaires pour affiner la qualité de la régression, selon le même principe. Cela ne compliquera pas vraiment l'étude du modèle définitif car il sera toujours aussi facile à analyser (un seul coefficient définitif pour chaque variable explicative, même si ce coefficient s'obtient par un calcul de plus en plus long au fur et à mesure que l'on ajoute des étapes). Cette procédure itérative peut se poursuivre en utilisant de la même manière les résidus y_2 et $x_{21}, \dots, x_{2p}$. Le nombre de composantes $t_1, \dots, t_H$ à retenir est habituellement déterminé par validation croisée.

1.2 Algorithme

Cette version de l'algorithme PLS1 ne traite pas les données manquantes. Il s'agit donc d'une version simplifiée de celui p. 99 dans [Tenenhaus, 1998].

Étape 1 : $X_0 = X$ et $y_0 = y$

Étape 2 : Pour $h = 1, \dots, H$:

Étape 2.1 : Calcul du vecteur $w_h = (w_{h1}, \dots, w_{hj}, \dots, w_{hp})$

$$w_h = \frac{X_{h-1}^t Y_{h-1}}{\|X_{h-1}^t Y_{h-1}\|} \quad (1.1)$$

Étape 2.2 : Calcul de la composante t_h

$$t_h = X_{h-1} w_h \quad (1.2)$$

Étape 2.3 : Calcul du coefficient de régression c_h de y_{h-1} sur t_h

$$c_h = \frac{y_{h-1}^t t_h}{t_h^t t_h} \quad (1.3)$$

Étape 2.4 : Calcul du vecteur y_h des résidus de la régression de y_{h-1} sur t_h

$$y_h = y_{h-1} - c_h t_h \quad (1.4)$$

Étape 2.5 : Calcul du vecteurs p_h des coefficients des régressions de x_{hj} sur t_h

$$p_h = \frac{X_{h-1}^t t_h}{t_h^t t_h} \quad (1.5)$$

Étape 2.6 : Calcul de la matrice X_h des vecteurs des résidus des régressions de x_{hj} sur t_h

$$X_h = X_{h-1} - t_h p_h^t \quad (1.6)$$

1.3 Calcul des coefficients de régression

$$\begin{aligned}
 (X - t_2 p_2^t) w_3 &= (X - X w_2 p_2^t) w_3 \\
 &= \underbrace{(X (w_3 - w_2 p_2^t w_3))}_{w_3^*}
 \end{aligned}$$

- D'où en généralisant on retrouve (1.6) :

$$\begin{aligned}
 w_h^* &= w_h - \overbrace{w_1^* p_1^t}^{w_1^*} w_h - w_2^* p_2^t w_h - \dots - w_{h-1}^* p_{h-1}^t w_h \\
 &= w_h - \sum_{k=1}^{h-1} w_k^* p_k^t w_h
 \end{aligned}$$

1.4 Propriétés des composantes PLS

1.4.1 Optimalité des composantes

Chaque étape de la régression PLS peut se résumer comme étant le produit de la maximisation de la covariance (au carré, car ce qui importe, c'est la covariance en valeur absolue) entre la variable à expliquer (ou les résidus de la projection de celle-ci sur le modèle tel qu'il était à l'étape précédente) et la nouvelle composante, avec pour contrainte que celle-ci soit formée linéairement à partir des variables explicatives (ou des résidus de leur projection sur le modèle tel qu'il était à l'étape précédente), la somme des coefficients au carré (associés à chaque variable explicative) étant égale à 1.

Cette covariance au carré, peut s'écrire, à la première étape « $Cov^2(Y, t_i)$ » (Y étant la variable expliquée). Il s'agit en réalité du produit du coefficient de détermination entre Y et t_i ($R^2(Y, t_i)$) et de la variance de t_i . Dit autrement :

$$Cov^2(Y, t_i) = R^2(Y, t_i) \times Var(t_i)$$

Il s'agit donc de maximiser ce produit, en jouant sur les deux termes, qui sont les deux critères que nous allons retenir dans le cadre du choix du nombre de composantes.

Le premier terme, $R^2(Y, t_i)$, n'est autre que le critère normal que l'on cherche à maximiser lorsque l'on effectue une régression linéaire multiple au sens des MCO. Il est donc important que les composantes retenues apportent toutes une part significative en termes de corrélation par rapport à la variable Y , sous peine de risquer de n'avoir aucune vertu explicative.

Le second terme, $Var(t_i)$, représente, en quelques sortes, la fraction de la structure interne de l'ensemble formé par les variables explicatives (ensemble X) expliquée par la composante t_i , il s'agit donc de la prise en compte de la structure propre à l'ensemble X , et en lui-même ce terme n'a pas vocation à expliquer la variable Y .

Chaque composante répond donc au compromis d'expliquer au mieux la fraction de la variance de Y non prise en compte par les composantes précédentes, tout en rendant au mieux compte de ce qui est propre à l'ensemble des variables explicatives et qui n'a pas encore été pris en compte par les composantes précédentes.

Ces deux critères doivent être les plus élevés possibles s'agissant des composantes retenues. L'idéal est que pour toute composante retenue, soient suffisantes. Mais parfois, on peut justifier la conservation d'une étape supplémentaire sur base d'un seul des deux critères. Le tout est de comparer ce qu'apporte chaque composante supplémentaire par rapport à ce qu'il reste à expliquer, sachant que chaque critère, pris individuellement, peut être plus élevé pour des étapes pourtant précédées par d'autres étapes pour lequel le critère est moins élevé.

Il se peut donc même, parfois, qu'il soit nécessaire de retenir une étape intermédiaire dans le but d'être en mesure de retenir les étapes qui la suivent, même si cette étape, en elle-même, au regard des critères pris isolément, n'est pas suffisamment significative que pour justifier sa retenue. Par exemple, la 3^{ème} étape peut sembler non-significative, mais si la 4^{ème} étape l'est davantage au regard de l'un des deux

critères, il est préférable de se poser la question de retenir les 4 premières étapes, et dans ce cas on ne pourra pas passer outre la 3ème. Mais dans ce cas, il faut être sûr que l'importance de la 4ème étape est capitale au regard de ce qu'il reste à expliquer.

1.4.2 Indépendance (Non corrélation) entre composantes

Une des propriétés primordiales de la régression PLS est l'indépendance des composantes $t_1, \dots, t_i, \dots, t_H$ formées à partir des variables explicatives. En effet, la première composante t_1 est formée à partir des variables explicatives, en leur donnant certains coefficients sur base de leur covariance avec la variable expliquée « y » (ou de leur coefficient de corrélation avec la variable y si les variables sont centrées réduites). Pour ce faire, la variable t_1 sera représentative d'une partie de la variance des variables explicatives. Bien entendu, si y n'est pas une combinaison linéaire des variables explicatives, et qu'il y a plus d'une variable explicative dans l'analyse (et qu'aucune de ces variables n'est combinaison linéaire des autres), la variable t_1 sera insuffisante pour expliquer toute la variance de y , de même qu'elle sera insuffisante pour expliquer toute la variance des variables explicatives, et toute la covariance des variables explicatives avec y .

Il en demeurera un résidu. La variance de y ne sera pas totalement expliquée par la variance de t_1 . Il y a moyen d'améliorer le pouvoir explicatif du modèle. Pour cela, on s'intéresse aux résidus, qui ont été « oubliés » par la première composante. Cette première composante est indépendante des résidus. Or, on se sert de ces résidus pour construire la seconde composante t_2 , qui sera par la même occasion indépendante de t_1 . La composante t_i s'intéressera donc à la variance de y qui n'est pas expliquée par t_1 . Les résidus qui en résulteront, qui sont donc indépendants de t_i , et indépendants de t_1 (ils sont le résultat d'une régression sur des résidus qui sont déjà indépendants de t_1), serviront à la création de t_3 . t_3 sera donc indépendante de t_1 et t_i . Il en sera de même pour toutes les composantes, qui seront toutes indépendantes entre elles.

Cette indépendance peut se démontrer assez facilement d'un point de vue mathématique. Voici la démonstration telle qu'elle est présentée dans [22], avec quelques précisions supplémentaires :

1.5 Choix du nombre de composantes par validation croisée

Dans cette section, nous allons présenter en détails comment calculer les critères $PRESS$, RSS et Q_h^2 [22].

1.5.1 Principe

La procédure de validation croisée pour le choix du nombre de composantes (pp.77 et 83, [22]) est la suivante. A chaque étape h et donc pour chaque nouvelle composante t_h on calcule le critère :

$$Q_h^2 = 1 - \frac{PRESS_h}{RSS_{h-1}}$$

Pour $h = 1$, on a

$$RSS_0 = \sum_{i=1}^n (y_i - \bar{y})^2 = n - 1$$

lorsque la variable est centrée réduite en utilisant la division par $n - 1$ pour le calcul de la variance. Une nouvelle composante t_h est significative et donc conservée si $Q_{2h} \geq 0.0975$. Il faut donc définir les critères RSS_h et $PRESS_h$. Classiquement et c'est la présentation adoptée dans [22], on utilise le modèle de régression de y sur les h composantes suivantes :

$$y = \overbrace{(c_1 t_1 + \dots + c_h t_h)}^{\hat{y}_h} + y_h \quad (1.7)$$

pour calculer la prédiction $\hat{y}_h = y - y_h$. En fait, pour chaque observation i :

- on calcule la prédictions $\hat{y}_{hi}$ de y_i à l'aide du modèle (1.7) obtenu en utilisant toutes les observations.

• on calcule la prédiction $\hat{y}_{h(-i)}$ de y_i à l'aide du modèle (1.7) obtenu sans utiliser l'observation i .

Définition :

Les critères RSS_h (Residual Sum of Square) et $PRESS_h$ (Prediction Error Sum Of Squares) sont alors définis par :

$$RSS_h = \sum_{i=1}^n (y_i - \hat{y}_{hi})^2 \quad (1.8)$$

Et

$$PRESS_h = \sum_{i=1}^n (y_i - \hat{y}_{hi})^2 \quad (1.9)$$

L'idée qui nous a permis de retrouver ces résultats numériques est la suivante. A chaque étape h , c'est-à-dire pour chaque nouvelle composante t_h , les critères RSS_h et $PRESS_h$ sont calculés à partir du vecteur des résidus $y(h-1)$ de l'étape précédente et non plus à partir de y :

Etape 1 : calcul du $PRESS$ et du RSS sur y .

Etape 2 : calcul du $PRESS$ et du RSS sur y_1 vecteur des résidus de la régression de y sur t_1 .

À partir de là, à chaque étape h , trois modèles de régression peuvent être utilisés pour calculer la prédiction $\tilde{y}_h = y_{(h-1)} - y_h$ selon que l'on considère la régression du vecteur des résidus y_{h-1} sur t_h , sur la matrice des résidus $X_{(h-1)}$ ou sur la matrice initiale X . En effet, sachant que

$$t_h = X_{(h-1)} w_h$$

et

$$t_h = X w_h^*$$

Nous avons les trois modèles suivant :

$$y_{(h-1)} = \overbrace{c_h t_h}^{\tilde{y}_h} y_h \quad (1.10)$$

$$y_{(h-1)} = \overbrace{c_h X_{(h-1)} w_h}^{\tilde{y}_h} + y_h \quad (1.11)$$

$$y_{(h-1)} = \overbrace{c_h X w_h^*}^{\tilde{y}_h} + y_h \quad (1.12)$$

Pour une observation i , la prédiction $\tilde{y}_{hi}$ de $y_{(h-1)i}$ est la même avec les trois modèles obtenus en utilisant toutes les observations. En revanche, la prédiction $\tilde{y}_{h(-i)}$ de $y_{(h-1)i}$ est différente avec les trois modèles obtenus en retirant l'observation i . En conséquence, le choix du modèle n'intervient pas dans le calcul du RSS mais uniquement dans le calcul du $PRESS$. Ces deux critères sont maintenant définis par :

$$RSS_h = \sum_{i=1}^n (y_{(h-1)i} - \tilde{y}_{hi}) \quad (1.13)$$

Et

$$PRESS_h = \sum_{i=1}^n (y_{(h-1)i} - \tilde{y}_{h(-i)})^2 \quad (1.14)$$

1.5.2 Calcul du RSS

Nous venons de voir qu'il est équivalent de calculer le RSS_h à partir de la formule (1.13) et de l'un des trois modèles (1.10), (1.11) ou (1.12). De plus, il est équivalent de calculer RSS_h :

- sur y avec la formule (1.8) et le modèle (1.7) :

$$y = \overbrace{c_1 t_1 + \dots + c_h t_h}^{\hat{y}_h} + y_h$$

D'où

$$RSS_h = \sum_{i=1}^n (y_i - \hat{y}_{h(i)})^2 = \|y - \hat{y}_h\|^2 = \|y_h\|^2$$

- sur le vecteur des résidus $y_{(h-1)}$ avec la formule (1.13) et le modèle (1.10) :

$$y_{h-1} = \underbrace{c_h t_h}_{\tilde{y}_{h(i)}} + y_h$$

D'où

$$RSS_h = \|y_{h-1} - \tilde{y}_h\|^2 = \|y_{h-1} - c_h t_h\|^2 = \|y_h\|^2$$

Finalement, on calcule :

$$\begin{cases} RSS_1 = \|y_1\|^2 \\ RSS_2 = \|y_2\|^2 \\ \vdots \\ RSS_h = \|y_h\|^2 \end{cases}$$

Nous venons de voir qu'il y a trois manières différentes de calculer la prédiction $\tilde{y}_h$ selon que l'on considère la régression de y_{h-1} sur t_h , sur la matrice des résidus X_{h-1} ou sur la matrice initiale X c'est à dire selon que l'on considère le modèle (1.10), (1.11) ou (1.12). On a donc à chaque étape h et pour chaque observation i trois manières de calculer la prédiction $\tilde{y}_{hi}$:

$$\begin{aligned} \tilde{y}_{hi} &= c_h t_{hi} \\ &= c_h X_{(h-1)i} w_h \\ &= c_h X_i w_h^* \end{aligned} \tag{1.15}$$

Où

$X_{(h-1)i}$ est la i -ème ligne de la matrice des résidus $X_{(h-1)}$ calculée à l'étape 2.6 de l'algorithme (à l'étape $h-1$), X_i est la i -ème ligne de la matrice initiale X et t_{hi} est la i -ème valeur de la composante t_h calculée à l'étape 2.2 de l'algorithme.

Le coefficient c_h , les vecteurs w_h et w_h^* sont calculés respectivement aux étapes 2.3, 2.1 et 2.7 de l'algorithme.

Il y a donc à chaque étape h et pour chaque observation i , trois manières non

équivalentes cette fois, de calculer la prédiction $\tilde{y}_{h(-i)}$. En fait, pour retrouver les résultats numériques nous avons utilisé l'égalité (1.15) c'est-à-dire la régression du vecteur des résidus y_{h-1} sur la matrice des résidus X_h . Cette égalité s'écrit alors

$$\tilde{y}_{h(-i)} = c_{h(-i)} X_{(h-1)i} w_{h(-i)} \quad (1.16)$$

- $X_{(h-1)i}$ est la i -ème ligne de la matrice des résidus $X_{(h-1)}$
- $c_{h(-i)}$ et $w_{h(-i)}$ sont calculés sur $n - 1$ individus, l'individu i étant retiré. Plus précisément, ils sont calculés à partir des formules (1.3) et (1.1) utilisées aux étapes 2.3 et 2.1 de l'algorithme, mais ces formules sont appliquées à la matrice des résidus $X_{(h-1)}$ privée de la ligne i et au vecteur des résidus $y_{(h-1)}$ privé de sa i -ème valeur.

Finalement, on note $X_{(h-1)(-i)}$ la matrice des résidus $X_{(h-1)}$ privée de sa i -ème ligne et $y_{(h-1)(-i)}$ le vecteur y_{h-1} privé de sa i -ème valeur. Ces notations sont un peu lourdes, mais le calcul du *PRESS* est simple et peut s'inclure à l'algorithme à la suite de l'étape 2.3 :

Étape (2.3)' : Calcul de $PRESS_h$

- Pour $i = 1, \dots, n$ en utilisant les formules (1.1), (1.2) et (1.3) on calcule :

$$w_{h(-i)} = \frac{X'_{(h-1)(-i)} y_{(h-1)(-i)}}{\|X'_{(h-1)(-i)} y_{(h-1)(-i)}\|}$$

$$t_{h(-i)} = X_{(h-1)(-i)} w_{h(-i)}$$

$$c_{h(-i)} = \frac{X'_{(h-1)(-i)} y_{(h-1)(-i)}}{\|X'_{(h-1)(-i)} y_{(h-1)(-i)}\|}$$

$$\tilde{y}_{h(-i)} = c_{h(-i)} X_{(h-1)i} w_{h(-i)}$$

- Puis on calcule :

$$PRESS_h = \sum_{i=1}^n (y_{(h-1)i} - \tilde{y}_{h(-i)})^2$$

où $y_{(h-1)i}$ est la i -ème valeur du vecteur des résidus $y_{(h-1)}$

Chapitre

3

La régression PLS multivariée

La régression PLS permet de relier un ensemble de variables dépendantes $Y = \{Y_1, \dots, Y_p\}$ à un ensemble de variables indépendantes $X = \{X_1, \dots, X_M\}$ lorsque le nombre de variables indépendantes et/ou dépendantes est élevé. La régression PLS consiste à effectuer une analyse en composantes principales de l'ensemble de variables X sous la contrainte que les (pseudo-) composantes principales des X_j soient aussi explicatives que possible de l'ensemble de variables Y . Il est alors possible de prédire les Y_k à partir des X_j en séparant mieux le signal (ce qui est synthétique, structuré, commun aux données) du bruit (ce qui est plus spécifique à l'échantillon de données étudié). Nous présentons dans ce chapitre la régression PLS2, l'application de la régression PLS2 sera traitée sur un exemple réel dans le chapitre 4.

1 Introduction

Après avoir détaillé la théorie, le présent chapitre détaille les aspects théoriques de la méthode.

Données et notations

Les notations de la régression PLS sont relativement bien standardisées dans la littérature, les voici :

N : Nombre d'individus

M : Nombre de variables indépendantes X_j

P : Nombre de variables dépendantes Y_k

A : Nombre de composantes retenues

X : Matrice des données $(N \times M)$ pour les variables indépendantes

Y : Matrice des données $(N \times P)$ pour les variables dépendantes

E_0 : Matrice des variables X_j centrées-réduites

F_0 : Matrice des variables Y_k centrées-réduites

E_h : Matrice des résidus de la décomposition de E_0 en utilisant h composantes

E_{hj} : j -ième colonne de E_h

F_h : Matrice des résidus de la décomposition de F_0 en utilisant h composantes

F_{nk} : k -ième colonne de F_h

Il s'agit d'étudier les liaisons entre les variables dépendantes Y_k et les variables indépendantes X_j , en compte les relations internes entre les variables de chaque groupe.

L'algorithme de la régression PLS est itératif. Nous allons en décrire le principe puis étudier plus en détail les propriétés de la solution.

2 L'algorithme de la régression PLS2

Etape 0 On part des tableaux E_0 et F_0

Etape 1 On construit une combinaison linéaire u_1 des colonnes de F_0 et une combinaison linéaire t_1 des colonnes de E_0 maximisant $cov(u_1, t_1) = cor(u_1, t_1) \cdot \sqrt{(var(u_1) \cdot var(t_1))}$. On obtient donc deux variables u_1 et t_1 aussi corrélées que possible et résumant au mieux les tableaux F_0 et E_0 . On construit ensuite les régressions :

$$E_0 = t_1 p_1^t + E_1$$

$$F_0 = t_1 r_1^t + F_1$$

Etape 2 On reprend l'étape 1 en remplaçant les tableaux E_0 et F_0 par E_1 et F_1 . On obtient donc deux nouvelles composantes : u_2 , combinaison linéaire des colonnes de F_1 , t_2 , combinaison linéaire des colonnes de E_1 . D'où les décompositions obtenues par régression :

$$E_0 = t_1 p_1^t + t_2 p_2^t + E_2$$

$$F_0 = t_1 r_1^t + t_2 r_2^t + F_2$$

On itère la procédure jusqu'à ce que les composantes $t_1, \dots, t_A$ expliquent suffisamment F_0 . Les composantes t_h sont des combinaisons linéaires des colonnes de E_0 , et non corrélées entre elles. De la décomposition $F_0 = t_1 r_1^t + \dots + t_h r_h^t + F_h$ on peut donc déduire les équations de régression PLS :

$$Y_k = \beta_{k0} + \beta_{k1} X_1 + \dots + \beta_{kM} X_M + F_{hk}$$

3 Recherche et propriétés des composantes PLS

Nous allons décrire dans cette section la recherche des deux premières composantes PLS et donner ensuite des résultats généraux.

3.1 Recherche de t_1 et u_1

On recherche une composante $t_1 = E_0 w_1$ et une composante $u_1 = F_0 c_1$, avec $\|w_1\| = \|c_1\| = 1$, telles que la covariance entre t_1 et u_1 soit maximum. On adopte ici la norme usuelle. Comme

$$\text{cov}(t_1, u_1) = \sqrt{(\text{var}(t_1) \text{var}(u_1))} \cdot \text{cor}(t_1, u_1)$$

on essaie ainsi de maximiser simultanément la variance expliquée par t_1 , la variance expliquée par u_1 et la corrélation entre ces deux composantes. On cherche donc des vecteurs normés w_1 et c_1 maximisant $\langle t_1, u_1 \rangle = \|t_1\| \cdot \|u_1\| \cdot \text{Cor}(t_1, u_1)$.

Utilisons la méthode des multiplicateurs de Lagrange.

On obtient les relations :

on pose $s = w_1' E_0^t F_0 c_1 - \lambda_1 (w_1' w_1 - 1) - \lambda_2 (c_1' c_1 - 1)$

On annule les dérivées partielles :

$$\frac{\partial s}{\partial \lambda_1} = -(w_1' w_1 - 1) = 0 \quad (3.1)$$

$$\frac{\partial s}{\partial \lambda_2} = -(c_1' c_1 - 1) = 0 \quad (3.2)$$

$$\frac{\partial s}{\partial w_1} = E_0^t F_0 c_1 - 2\lambda_1 w_1 \quad (3.3)$$

$$\frac{\partial s}{\partial c_1} = F_0^t E_0 w_1 - 2\lambda_2 c_1 \quad (3.4)$$

Des égalités (3.1) à (3.4) on déduit :

$$2\lambda_1 = 2\lambda_2 = \theta_1 = w_1' E_0^t F_0 c_1 = \langle t_1, u_1 \rangle$$

$$E_0^t F_0 c_1 = \theta_1 w_1 \quad (3.5)$$

$$F_0^t E_0 w_1 = \theta_1 c_1 \quad (3.6)$$

$$E_0^t F_0 F_0^t E_0 w_1 = \theta_1^2 w_1 \quad (3.7)$$

$$F_0^t E_0 E_0^t F_0 c_1 = \theta_1^2 c_1 \quad (3.8)$$

Ainsi w_1 est vecteur propre de $E_0^t F_0 F_0^t E_0$ associé à la plus grande valeur propre θ_1^2 , et c_1 est vecteur propre de $F_0^t E_0 E_0^t F_0$ associé à la plus grande valeur propre θ_1^2 .

On effectue ensuite deux régressions de E_0 sur t_1 , puis de F_0 sur t_1 :

$$E_0 = t_1 p_1^t + E_1 \quad (3.9)$$

Où

$$P_1 = E_0^t t_1 / \|t_1\|^2 \quad (3.10)$$

Est le vecteur des coefficients de régression des E_{0j} sur t_1

$$F_0 = t_1 r_1^t + F_1 \quad (3.11)$$

Où

$$r_1 = F_0^t t_1 / \|t_1\|^2 \quad (3.12)$$

est le vecteur des coefficients de régression des F_{0k} sur t_1 .

On a les propriétés suivantes :

$$\text{a) } P_1^t w_1 = 1$$

$$\text{b) } r_1 = b_1 c_1 \text{ ou } b_1 \text{ est le coefficient de régression de } u_1 \text{ sur } t_1.$$

Les vecteurs r_1 et c_1 sont donc colinéaires.

$$\text{Montrons (a) : } p_1^t w_1 = t_1^t E_0 w_1 / \|t_1\|^2 = t_1^t t_1 / \|t_1\|^2 = 1$$

$$\begin{aligned} \text{Montrons (b) : } r_1 &= F_0^t t_1 / \|t_1\|^2 = F_0^t E_0 w_1 / \|t_1\|^2 \\ &= \theta_1 c_1 / \|t_1\|^2 = b_1 c_1 \end{aligned}$$

Où $b_1 = \theta_1 / \|t_1\|^2 = u_1^t t_1 / \|t_1\|^2$ est bien le coefficient de régression de u_1 sur t_1 .

3.2 Recherche de t_2 et u_2

On remplace dans la section précédente E_0 et F_0 par les matrices résiduelles E_1 et F_1 D'où :

$$t_2 = E_1 w_2$$

$$u_2 = F_1 c_2$$

$$\theta_2 = u_2^t t_2$$

$$p_2 = E_1^t t_2 / \|t_2\|^2$$

$$r_2 = F_1^t t_2 / \|t_2\|^2 = \theta_2 c_2 / \|t_2\|^2 = b_2 c_2$$

où $b_2 = \frac{u_2^t t_2}{||t_2||^2}$ est le coefficient de régression de u_2 sur t_2 .

$$E_1 = t_2 p_2^t + E_2$$

$$F_1 = t_2 r_2^t + F_2$$

$$E_0 = t_1 p_1^t + t_2 p_2^t + E_2$$

$$F_0 = t_1 r_1^t + t_2 r_2^t + F_2$$

Et ainsi de suite pour les autres étapes.

Nous allons présenter dans cette section les propriétés des composantes PLS. Ces résultats sont nécessaires à une compréhension approfondie de la méthode.

On obtient les résultats suivants :

A l'étape h les équations (3.5) à (3.8) deviennent :

$$E_{h-1}^t F_{h-1} c_h = \theta_h w_h \quad (3.13)$$

$$F_{h-1}^t E_{h-1} w_h = \theta_h c_h \quad (3.14)$$

$$E_{h-1}^t F_{h-1} F_{h-1}^t E_{h-1} w_h = \theta_h^2 w_h \quad (3.15)$$

$$F_{h-1}^t E_{h-1} E_{h-1}^t F_{h-1} c_h = \theta_h^2 c_h \quad (3.16)$$

Les équations (3.13) et (3.16) ou (3.14) et (3.15) permettent d'obtenir w_h et c_h et les composantes PLS $u_h = F_{h-1} c_h$ et $t_h = E_{h-1} w_h$, avec $\theta_h = u_h^t t_h$

Les équations (3.9) à (3.12) deviennent

$$E_{h-1} = t_h p_h^t + E_h \quad (3.17)$$

où

$$p_h = \frac{E_{h-1}^t t_h}{||t_h^2||} \quad (3.18)$$

$$F_{h-1} = t_h r_h^t + F_h \quad (3.19)$$

où

$$r_h = \frac{F_{h-1}^t t_h}{||t_h||^2} \quad (3.20)$$

$$r_h = b_h c_h \quad (3.21)$$

avec b_h coefficient de régression de u_h sur t_h . On a aussi :

$$w_h^t p_h = 1 \quad (3.22)$$

et

$$t_h^t E_h = 0 \quad (3.23)$$

4 Propriété de la régression

4.1 Relations cycliques

Partant de w_h on obtient la suite de relations suivantes :

$$t_h = E_{h-1} w_h$$

$$c_h = 1/\theta_h F_{h-1}^t t_h \quad (4.1)$$

$$u_h = F_{h-1} c_h$$

$$w_h = 1/\theta_h E_{h-1}^t u_h \quad (4.2)$$

la relation (4.1) provient de (3.14) et (4.2) de (3.13).

On peut schématiser

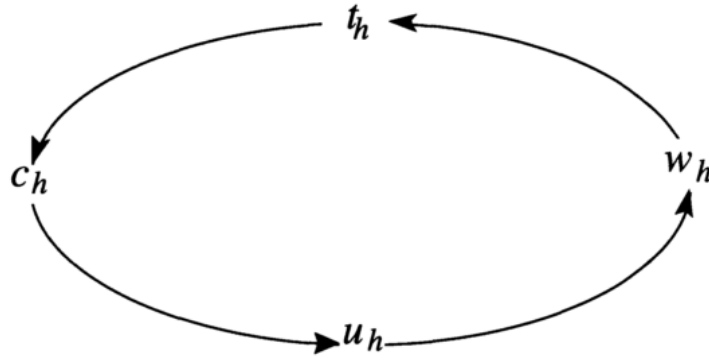


FIG. 4.1: Relation cyclique

4.2 Relations d'orthogonalité

Il y a de nombreuses relations d'orthogonalité au niveau de l'analyse des variables X_j . Ceci n'est plus vrai au niveau de l'analyse des Y_k .

On a les relations d'orthogonalité suivantes :

(a) $w_h^t w_l = 0, l > h$

(b) $t_h^t t_l = 0, l > h$

(c) $w_h^t p_l = 0, l > h$

(d) $w_h^t E_l^t = 0, l \geq h$

(e) $t_h^t E_l = 0, l \geq h$

1) Démonstration de $t_h^t t_l = 0, l > h$

La preuve se fait par récurrence.

- $t_1^t t_2 = t_1^t E_1 w_2 = 0$, puisque $t_1 E_1 = 0$.

- Supposons $t_1, \dots, t_h$ orthogonaux, alors : $t_1, \dots, t_{h+1}$ sont orthogonaux.

Montrons que t_{h+1} est orthogonal aux vecteurs $t_1, \dots, t_h$:

$$t_h^t t_{h+1} = t_h^t E_h w_{h+1} = 0, \text{ puisque } t_h^t E_h = 0.$$

$$\begin{aligned} t_{h-1}^t t_{h+1} &= t_{h-1}^t E_h w_{h+1} = 0 \\ &= t_{h-1}^t [E_{h-1} - t_h p_h^t] w_{h+1} \\ &= [t_{h-1}^t E_{h-1} - t_{h-1}^t t_h p_h^t] w_{h+1} \\ &= 0 \end{aligned}$$

Puisque $t_{h-1}^t E_{h-1} = 0$ et $t_{h-1}^t t_h = 0$.

$$\begin{aligned} t_{h-2}^t t_{h+1} &= t_{h-2}^t E_h w_{h+1} \\ &= t_{h-2}^t [E_{h-2} - t_{h-1}^t p_{h-1}^t - t_h p_h^t] w_{h+1} \\ &= 0 \end{aligned}$$

Et ainsi de suite, d'où le résultat.

2) Démonstration de $w_h^t E_l^t = 0, l \geq h$

On a :

$$\begin{aligned} w_h^t E_h^t &= w_h^t [E_{h-1} - t_h p_h^t]^t \\ &= t_h^t - w_h^t p_h^t t_h^t = 0 \end{aligned}$$

Puisque $w_h^t p_h = 1$.

Montrons maintenant que $w_h^t E_l^t = 0, l \geq h$, implique $w_h^t E_{l+1}^t = 0$. On a :

$$\begin{aligned} w_h^t E_{l+1}^t &= w_h^t [E_l - t_{l+1} p_{l+1}^t]^t \\ &= w_h^t E_l^t - (w_h^t E_l^t t_{l+1}) / (||t_{l+1}||^2) \cdot t_{l+1}^t \\ &= 0, \end{aligned}$$

D'où le résultat.

3) Démonstration de $w_h^t p_l = 0, l > h$

On a :

$$\begin{aligned} w_h^t p_l &= w_h^t E_{l-1}^t t_l / ||t_l||^2 \\ &= 0, \text{ pour } l > h \end{aligned}$$

Puisque $w_h^t E_{l-1}^t = 0$, pour $l-1 > h$, soit $l > h$.

4) Démonstration de $w_h^t w_l = 0, l > h$

on a :

$$\begin{aligned} w_h^t w_l &= w_h^t \cdot \frac{1}{\theta_l} E_{l-1}^t u_l \\ &= 0, l > h \end{aligned}$$

puisque $w_h^t E_{l-1}^t = 0$ pour $l-1 \geq h$, soit $l \geq h$

5) Démonstration de $t_h^t E_{l=0}^t, l \geq h$

L'égalité est vraie pour $l = h$.

Si $l > h$ on a :

$$\begin{aligned}
t_h^t E_l &= t_h^t (E_{l-1} - t_l p_l^t) \\
&= t_h^t E_{l-1} \\
&= t_h^t (E_{l-2} - t_{l-1} p_{l-1}^t) \\
&= t_h^t E_{l-2} = \dots = t_h^t E_h = 0
\end{aligned}$$

5 Formules de décomposition

Les matrices E_0 et F_0 peuvent se décomposer par régression sur les composantes $t_1, \dots, t_A$, où A est le rang de E_0 :

$$E_0 = t_1 p_1^t + t_2 p_2^t + \dots + t_A p_A^t \quad (5.1)$$

$$F_0 = t_1 r_1^t + t_2 r_2^t + \dots + t_A r_A^t + F_A \quad (5.2)$$

Les vecteurs t_h étant orthogonaux entre eux on retrouve bien les mêmes coefficients de régression dans les équations (26) et (27) que par la méthode itérative décrite plus haut.

La matrice des résidus E_h à l'étape h peut s'exprimer en fonction de E_0 :

$$E_h = E_0 (I - w_1 p_1^t) (I - w_2 p_2^t) \dots (I - w_h p_h^t) \quad (5.3)$$

Montrons ce résultat.

$$\text{On a } E_1 = E_0 - t_1 p_1^t = E_0 - E_0 w_1 p_1^t = E_0 (I - w_1 p_1^t)$$

Supposons la décomposition (28) vraie pour $h = j$. Montrons qu'elle est vraie pour $h = j + 1$:

$$\begin{aligned}
E_{j+1} &= E_j - t_{j+1} p_{j+1}^t = E_j - E_j w_{j+1} p_{j+1}^t \\
&= E_j (I - w_{j+1} p_{j+1}^t) \\
&= E_0 \Pi_{h=1}^j (I - w_h p_h^t) (I - w_{j+1} p_{j+1}^t)
\end{aligned}$$

D'où le résultat.

Ces décompositions entraînent trois résultats importants au niveau de l'interprétation. Notons

$$||E_0||^2 = ||t_1||^2 \cdot ||p_1||^2 + ||t_2||^2 \cdot ||p_2||^2 + \dots + ||t_A||^2 \cdot ||p_A||^2 \quad (5.4)$$

Cette formule permet de mesurer le pouvoir de chaque t_h (ou des h premiers) pour résumer E_0 . On utilise

$$||t_h||^2 \cdot ||p_h||^2 / ||E_0||^2 \text{ ou } \sum_{l \leq h} ||t_l||^2 \cdot ||p_l||^2 / ||E_0||^2$$

Montrons (29) :

Les $t_1, \dots, t_A$ étant orthogonaux on obtient pour chaque colonne E_{0j} . en utilisant (26) :

$$||E_{0j}||^2 = \sum_{h=1}^A ||t_h||^2 \cdot p_{hj}^2$$

Et par conséquent

$$||E_0||^2 = \sum_{(h=1)}^A ||t_h||^2 \cdot ||p_h||^2$$

On aura de même la décomposition :

$$||F_0||^2 = ||t_1||^2 \cdot ||r_1||^2 + \dots + ||t_A||^2 \cdot ||r_A||^2 + ||F_A||^2 \quad (5.5)$$

On mesure le pouvoir de chaque t_h (ou des h premiers) pour expliquer F_0 à l'aide de $||t_h||^2 \cdot ||r_h||^2 / ||F_0||^2$ ou $\sum_{l \leq h} ||t_l||^2 \cdot ||r_l||^2 / ||F_0||^2$

Enfin il est important de remarquer que la composante t_h est combinaison linéaire des colonnes de E_0 :

$$\begin{aligned} t_h &= E_{h-1} w_h \\ &= E_0 \prod_{(l=1)}^j (I - w_l p_l^t) w_h = E_0 \tilde{w}_h \end{aligned}$$

Le vecteur $E_0 \tilde{w}_h = \prod_{(l=1)}^j (I - w_l p_l^t) w_h$ des coefficients des variables E_{0j} n'est cependant pas normé à 1.

Les composantes PLS $t_1, \dots, t_h$ sont donc des combinaisons linéaires des colonnes

de E_0 , non corrélées entre elles, résumant au mieux E_0 tout en expliquant autant que possible F_0 . Ces composantes PLS sont donc analogues à des composantes principales des $X_1, \dots, X_M$ expliquant au mieux les variables $Y_1, \dots, Y_p$.

Les composantes PLS $u_1, \dots, u_h$ n'ont pas cette propriété. La composante u_1 est bien combinaison linéaire des colonnes de F_0 , mais ce n'est plus le cas des autres composantes. Les composantes $u_2, \dots, u_h$ sont des variables intermédiaires plus difficiles à interpréter.

Le plan (t_1, t_2) permet de visualiser la liaison entre les Y_k et les X_j détectée par ces premières composantes PLS.

Pour interpréter les composantes PLS t_h en fonction des X_j et des Y_k il est naturel de calculer les corrélations entre X_j et t_h , puis entre Y_k et t_h . On obtient :

$$Cor(X_j, t_h) = \frac{\frac{1}{(N-1)t_h^t E_{0j}}}{\sqrt{\frac{1}{(N-1)t_h^t t_h}}} = \sqrt{(var(t_h))} p_{hj}$$

Bien que la division par $N - 1$ plutôt que N ne s'impose pas, nous l'avons utilisée pour être cohérent avec les sorties du logiciel SIMCA.

Notons λ_h la variance de t_h on obtient :

$$cor(X_j, t_h) = \sqrt{(\lambda_h)} \cdot p_{hj}$$

Et de même

$$Cor(Y_k, t_h) = \sqrt{\lambda_h} \cdot r_{hk}$$

Pour interpréter t_1 et t_2 , on construit donc le cercle des corrélation en visualisant dans un plan les points

$$A_j = (\sqrt{\lambda_1} \cdot p_{1j}, \sqrt{\lambda_2} \cdot p_{2j})$$

Et

$$B_k = (\sqrt{\lambda_1} \cdot r_{1k}, \sqrt{\lambda_2} \cdot r_{2k})$$

représentant les corrélations de X_j et Y_k avec t_j et Y_k avec t_1 et t_2 .

Les produits scalaires $\langle A_j, A_j^t \rangle$, $\langle B_k, B_k^t \rangle$ et $\langle A_j, B_k \rangle$ représentent des approximations à l'ordre 2 des $\text{cor}(X_j, X_j^t)$, $\text{cor}(Y_k, Y_k^t)$ et $\text{cor}(X_j, Y_k)$ respectivement.

En effet :

$$\begin{aligned} \text{Cor}(X_j, X_j^t) &= 1/(N-1) E_{0j}^t E_{0j}^t \\ &= \lambda_1 p_{1j} p_{1j}^t + \lambda_2 p_{2j} p_{2j}^t + \dots + \lambda_A p_{Aj} p_{Aj}^t \\ &\approx \langle A_j, A_j^t \rangle \end{aligned}$$

On démontre de même les autres approximations.

La norme de A_j représente la corrélation multiple entre X_j et (t_1, t_2) et celle de B_k la corrélation multiple entre Y_k et (t_1, t_2) . On a en effet

$$R^2(X_j; t_1, t_2) = \text{cor}^2(X_j, t_1) + \text{cor}^2(X_j, t_2)$$

Puisque les variables t_1 et t_2 sont non corrélées. D'où le résultat. Et de même pour Y_k .

6 Équations de régressions PLS

De la décomposition de F_0 sur $t_1, \dots, t_h$ on déduit la régression PLS de chaque variable Y_k sur $X_1, \dots, X_M$. On part de

$$\begin{aligned} F_{0k} &= \frac{y_k - \bar{y}_k}{s_{yk}} = \sum_{l=1}^h r_{lk} t_l + F_{hk} \\ &= \sum_{l=1}^h r_{lk} E_0 \tilde{w}_l + F_{hk} \\ &= \sum_{l=1}^h r_{lk} \sum_{j=1}^M \tilde{w}_{lj} \frac{X_j - \bar{x}_j}{s_{xj}} + F_{hk} \end{aligned}$$

Posons $\beta_j = \sum_{l=1}^h r_{lk} \tilde{w}_{lj}$.

On trouve dans les programmes de régression PLS trois types de résultats :

1) Les variables Y_k et les X_j sont centrées-réduites.

D'où :

$$F_{0k} \approx \sum_{j=1}^M \beta_j E_0$$

2) On utilise les variables réduites $Y_k^t = Y_k/s_{yk}$ et $X_j^t = X_j/s_{xj}$

D'où :

$$Y_k^t \approx \beta_0 + \sum_{j=1}^M \beta_j X_j^t$$

$$\text{Ou } \beta_0 = \frac{\bar{y}_k}{s_{y_k}} - \sum_{j=1}^M \beta_j \bar{x}_j / s_{x_j}$$

Les coefficients de régression β_j sont comparables au niveau de l'explication de Y_k par les X_j .

3) On utilise les variables d'origine (non centré et non réduite).

D'où :

$$Y_k \approx \beta_0^t + \sum_{j=1}^M \beta_j^t X_j$$

$$\text{Ou } \beta_0^t = \beta_0 \cdot s_{y_k} \text{ et } \beta_j^t = \beta_j \cdot \frac{s_{y_k}}{s_{x_j}}.$$

Chaque composante t_h étant combinaison linéaire des $X_j^* = E_{0j}$, on peut écrire l'équation de régression PLS

$$Y^* = \beta_1 X_1^* + \dots + \beta_M X_M^* F_h$$

7 Résultats de la régression PLS

7.1 Calcul des composantes t_h, u_h

Le logiciel permet d'obtenir les vecteurs propres w_h, c_h et les composantes PLS t_h, u_h .

Dans le logiciel chaque composantes PLS u_h est normalisée de manière à ce que le coefficient de régression de u_h sur t_h soit égal à 1.

Les composantes PLS u_h sont définies par :

$$u_h = \frac{1}{b_h \cdot F_{h-1} c_h}$$

Ou $b_h = \frac{t_h^t F_{h-1} c_h}{t_h^t t_h}$ est le coefficient de régression de $F_{h-1} c_h$ sur t_h .

Les résultats de régressions

$$E_0 = t_1 p_1^t + t_2 p_2^t + t_3 p_3^t$$

$$F_0 = t_1 r_1^t + t_2 r_2^t + t_3 r_3^t + F_3$$

Coefficient des régressions de E_0 et F_0 sur t_1, t_2, t_3 .

$$\text{De } r_h = \frac{F_{h-1}^t t_h}{t_h^t t_h} = \frac{F_{h-1}^t E_{h-1} w_h}{t_h^t t_h} = \frac{t_h^t F_{h-1} c_h}{t_h^t t_h} c_h = b_h \cdot c_h$$

On déduit

$$||r_h||^2 = b_h^2$$

Et

$$u_h = 1/(||r_h||^2) F_{h-1} r_h$$

Les composantes PLS u_h pouvant s'exprimer directement en fonction des r_h , le logiciel ne fournit pas les vecteurs propres c_h .

7.2 Valeurs explicatives des composantes PLS

Les formules (29) et (30) s'écrivent :

$$||E_0||^2 = ||t_1||^2 \cdot ||p_1||^2 + ||t_2||^2 \cdot ||p_2||^2 + ||t_3||^2 \cdot ||p_3||^2$$

Et

$$||F_0||^2 = ||t_1||^2 \cdot ||r_1||^2 + ||t_2||^2 \cdot ||r_2||^2 + ||t_3||^2 \cdot ||r_3||^2 + ||F_3||^2.$$

On a :

$$||E_0||^2 = ||F_0||^2 = 3(N-1)$$

D'où les variances expliquées par les composantes t_h .

Du pouvoir explicatif des t_h on constate que les données étudiées sont essentiellement unidimensionnelles, au niveau de la relation entre Y et X .

7.3 Interprétation des composantes PLS

Les composantes PLS t_h peuvent s'écrire en fonction des X_j centrées-réduites :

$$t_1 = E_0 w_1$$

$$t_2 = E_1 w_2 = E_0 (I - w_1 p_1^t) w_2 = E_0 \tilde{w}_2$$

$$t_3 = E_2 w_3 = E_0 (I - w_1 p_1^t) (I - w_2 p_2^t) w_3 = E_0 \tilde{w}_3$$

Corrélations entre les variables X_j et les composantes PLS t_h , on déduit les corrélations.

La composante PLS $u_1 = F_0 r_1 / \|r_1\|^2$

La composante u_1 On a $\text{cor}(u_1, t_1)$.

La composante u_2

Où F_{1k} est le vecteur des résidus de la régression de F_{0k} sur t_1 . Cette composante u_2 ne peut s'exprimer en fonction des variables y_k et u_2 . La composante u_2 est redondante avec u_1 . On a $\text{cor}(u_1, u_2)$. IL en va de même pour la composante u_3 :

$\text{cor}(u_1, u_3)$ et $\text{cor}(u_2, u_3)$

8 Conclusion

La régression PLS permet de rapprocher le statisticien du chercheur dans les problèmes de modélisation. En général ce dernier souhaite en effet conserver dans son modèle toutes les variables importantes tout en obtenant des équations de régression cohérentes. Lorsque, en régression multiple il y a multicollinéarité et/ou un nombre important de variables explicatives par rapport au nombre d'observations, la solution la plus courante consiste à exclure des variables explicatives par des méthodes de pas-à-pas. La régression PLS permet dans ces situations de conserver toutes les variables explicatives tout en obtenant une équation de régression cohérente.

Lorsqu'il s'agit de relier un ensemble de Y_k à un ensemble de X_j , on peut essayer l'analyse canonique. L'intérêt pratique de cette méthode est souvent faible car elle ne fait le plus souvent qu'isoler des couples (X_j, Y_k) à forte corrélation.

Ainsi les variables canoniques expliquant une part qui peut être faible des variances des groupes de variables X et Y .

La régression PLS apparaît comme un compromis entre l'analyse canonique des groupes de variables X et Y et des analyses en composantes principales de chacun de ces groupes de variables.

Chapitre

4

Application

1 Problématique de la conduite d'un four rotatif de cimenterie

Le four sujet de cette étude est actuellement conduit d'une manière manuelle centralisée, à partir d'un centre de commandement (salle de contrôle) où toute l'information des capteurs installés sur le procédé est disponible et toutes les commandes vers la partie opératoire peuvent être envoyées, des opérateurs humains estiment à chaque instant ces commandes en se basant sur les valeurs de mesures observées et leurs propres expériences.

Les signaux observés sont regardés séparément, un seuillage de type Shewhart est appliqué sur chaque signal seul pour en déduire s'il est en état normal (entre les deux seuils maximal et minimal). Une alarme est émise par le système de supervision en cas de dépassement de l'un des deux seuils pour le signal. Ce type de supervision est dit classique. Le nombre de signaux est très grand pour qu'il soit suivi par un seul opérateur, en réalité une centaine de signaux de mesures sont émises par l'équipement four seul, une cinquantaine sont des mesures reflétant directement l'état de son fonctionnement, dont chaque divergence peut avoir une conséquence néfaste sur la qualité du produit final du four et par la suite un coût économique négatif sur la trésorerie de l'entreprise. Les autres sont des mesures nécessaires pour

sa propre sécurité tel que les températures des parties tournantes, de graissage, des huiles et autres.

Un développement de cette supervision est donc plus que nécessaire pour lui confier un aspect à la fois intelligent et ergonomique. Au lieu que chaque signal soit observé seul, il est préférable de résumer l'essentiel de ces signaux est de le fournir à l'opérateur sous une forme simple et efficace. Des techniques statistiques multi-variées ont cette capacité. L'opérateur ne sera plus contrarié par le grand nombre d'alarmes émises par chaque signal en cas de situation anormale, mais une seule alarme ou deux uniquement peuvent résumer la situation. Les opérateurs non encore qualifiés peuvent être énormément aidés par l'implémentation d'un aspect intelligent dans la supervision actuelle, qui peut comparer l'état du four normal appris auparavant avec chaque future situation du four et en déduire d'une manière automatique si celui-ci est normal ou non.

1.1 Introduction

Pour répondre à cette problématique nous avons utilisé une méthode de prédiction qui est la régression des carrées partiel PLS (Partial Least Square) ; nous avons fait une applications sur des données réelles brutes issus de l'analyse d'un four de cimenterie de Aine el kbira à sitif, nous avons 52 variables (50 variables explicatives qui sont désigne par le tableau X et 2 variables à expliquer qui sont désigné par le tableau Y) les individus sont la variation du temps par minute(256 individus, individu 1 = 23h 39mn ; individu 256 = 3h 54mn soit 5h 03mn), l'identification des variables se trouve en annexe.

L'application est faite avec le Logiciel Excel stat PLS.

1.2 Analyse en composante principale (ACP)

Interprétation de L'ACP sur les variables du tableau X et des variables supplémentaires du tableau Y

1.2.1 Tableau de données

Les données sont les mesures effectuées sur n unités $\{u_1, u_2, \dots, u_i, \dots, u_n\}$. Les p variables quantitatives qui représentent ces mesures sont $\{v_1, v_2, \dots, v_i, \dots, v_n\}$.

Le tableau des données brutes à partir duquel on va faire l'analyse est noté X et a la forme[1] suivante :

$$X = \begin{bmatrix} & v_1 & v_2 & . & v_i & . & v_p \\ u_1 & X_{11} & X_{12} & & X_{i1} & & X_{1p} \\ u_2 & X_{21} & X_{22} & & X_{i2} & & X_{2p} \\ . & & & & & & \\ u_i & X_{i1} & X_{i2} & & X_{ij} & & X_{ip} \\ . & & & & & & \\ u_n & X_{n1} & X_{n2} & & X_{nj} & & X_{np} \end{bmatrix}$$

On peut représenter chaque unité par le vecteur de ses mesures sur les p variables :

$$U_i^t = [x_{i1} x_{i2} \dots x_{ij} \dots x_{ip}] \text{ ce qui donne } U_i = \begin{bmatrix} X_{i1} \\ X_{i2} \\ \\ X_{ij} \\ \\ X_{ip} \end{bmatrix}$$

Alors U_i est un vecteur de R^p .

De façon analogue, on peut représenter chaque variable par un vecteur de R^n dont les composantes sont les valeurs de la variable pour les n unités :

$$V_i = \begin{bmatrix} X_{1j} \\ X_{2j} \\ \\ X_{ij} \\ \\ X_{np} \end{bmatrix}$$

Pour avoir une image de l'ensemble des unités, on se place dans un espace affine

en choisissant comme origine un vecteur particulier de R^p , par exemple le vecteur dont toutes les coordonnées sont nulles. Alors, chaque unité sera représentée par un point dans cet espace. L'ensemble des points qui représentent les unités est appelé traditionnellement “nuage des individus”.

En faisant de même dans R^n , chaque variable pourra être représentée par un point de l'espace affine correspondant. L'ensemble des points qui représentent les variables est appelé “nuage des variables”.

On constate, que ces espaces étant de dimension supérieure en général à 2 et même 3, on ne peut visualiser ces représentations. L'idée générale des méthodes factorielles est de trouver un système d'axes et de plans tels que les projections de ces nuages de points sur ces axes et ces plans permettent de reconstituer les positions des points les uns par rapport aux autres, c'est-à-dire avoir des images les moins déformées possible.

1.2.2 Choix de l'origine

Le point 0 correspondant au vecteur de coordonnées toutes nulles n'est pas forcément une origine satisfaisante, car si les coordonnées des points du nuage des individus sont grandes, le nuage est éloigné de cette origine. Il apparaît plus judicieux de choisir une origine liée au nuage lui-même : le centre de gravité du nuage. Pour définir ce centre de gravité, il faut choisir un système de pondération des unités :

$$\forall i = 1, \dots, n; p_i = \text{poids de l'unité } u_i \text{ tel que } \sum_{i=1}^n p_i = 1$$

Définition :

Par définition le centre de gravité est défini comme le point tel que :

$$\sum_{i=1}^n p_i \vec{Gu}_i = \vec{0}$$

Pour l'ACP on choisit de donner le même poids $1/n$ pour tous les individus.

Le centre de gravité G du nuage des individus est alors le point dont les coor-

données sont les valeurs moyennes des variables :

$$1/n \sum_{i=1}^n X_{i1}$$

$$G = \begin{bmatrix} 1/n \sum_{i=1}^n X_{i1} \\ \vdots \\ 1/n \sum_{i=1}^n X_{ij} \\ \vdots \\ 1/n \sum_{i=1}^n X_{ip} \end{bmatrix} = \begin{bmatrix} X_{.1} \\ \vdots \\ X_{.j} \\ \vdots \\ X_{.p} \end{bmatrix}$$

Prendre G comme origine, conformément à la figure suivante, revient alors à travailler sur le tableau des données centrées :

$$X_c = \begin{bmatrix} X_{11} - X_{.1} & \cdot & X_{1h} - X_{.j} & \cdot & X_{1p} - X_{.p} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ X_{i1} - X_{.1} & \cdot & X_{ij} - X_{.j} & \cdot & X_{ip} - X_{.p} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ X_{n1} - X_{.1} & \cdot & X_{nj} - X_{.j} & \cdot & X_{np} - X_{.p} \end{bmatrix}$$

Et le vecteur des coordonnées centrées de l'unité u_i est : $U_{ci} = \begin{bmatrix} X_{i1} - X_{.1} \\ X_{i2} - X_{.2} \\ \cdot \\ X_{ij} - X_{.j} \\ \cdot \\ X_{ip} - X_{.p} \end{bmatrix}$

Celui des coordonnées centrées de la variable v_j est : $v_{ci} =$

$$\begin{bmatrix} X_{1j} - X_{.j} \\ X_{2j} - X_{.j} \\ \vdots \\ X_{ij} - X_{.j} \\ \vdots \\ X_{nj} - X_{.j} \end{bmatrix}$$

Donc le nouveau tableau de données X^* (centré-réduit) est formé des valeurs :

$$X_{ij}^* = \frac{X_{ij} - \bar{X}}{\sigma_j}$$

Toutes les variables ainsi transformées sont " comparables " et ont même dispersion : $\sigma_j = 1$.

1.2.3 Matrice à diagonaliser

L'analyse du tableau transformé X^* nous conduit à diagonaliser la matrice $C = X^{*'} X^*$. Le terme général c'_{jj} de cette matrice s'écrit :

$$c'_{jj} = \sum_{i=1}^n x_{ij}^* x_{ij'}^* = 1/n \sum_{i=1}^n (x_{ij} - \bar{X}_j)(x_{ij'} - \bar{X}_{j'}) / S_j S_{j'} \triangleq \text{cor}(j, j')$$

c'_{jj} est le coefficient de corrélation empirique entre les variables j et j' .

La matrice à diagonaliser est donc la matrice de corrélation C .

Les principaux axes et composantes principales

Premier axe principal : On cherche à faire passer une droite Δ_1 le mieux possible au milieu du nuage de points N^* . On mesure la dispersion du nuage N^* autour d'une droite Δ à l'aide de l'inertie $I(N^*, \Delta)$ du nuage N^* par rapport à la droite Δ .

Définition :

d^2 : distance euclidienne

$$I(N^*, \Delta) = 1/n \sum_{i=1}^n d^2(x_i^*, y_i)$$

Où y_i est la projection orthogonale $p_{\Delta}(x_i^*)$ du point x_i^* sur la droite Δ .

La droite Δ_1 cherchée minimise $I(N^*, \Delta_1)$, et s'appelle le premier axe principal du nuage N^* .

On peut montrer que la droite Δ_1 passe par l'origine O , centre de gravité du nuage N^* des données centrées réduites. C est la matrice des corrélations entre les variables X_j . On peut alors montrer que l'axe principal Δ_1 est engendré par le vecteur unitaire u_1 , vecteur propre normé de la matrice C associé à la plus grande valeur propre λ_1 .

Première Composante : La première composante principale est une nouvelle variable définie pour chaque individu i par la longueur algébrique $\overline{O_{y1i}}$, coordonnée de la projection y_{1i} du point x_i^* sur l'axe Δ_1 . La valeur $Y_1(i)$ est donc égale au produit scalaire entre les vecteurs x_i^* et u_1 :

$$Y_1(i) = \langle x_i^*, u_1 \rangle = \sum_{j=1}^p u_1 x_{ji}^* = \sum_{j=1}^p u_{1j} (x_{ij} - \overline{X_j}) / s_j$$

Globalement, la première composante principale Y_1 s'écrit donc : $Y_1 = \sum_{i=1}^p u_1 X_i^*$

Deuxième axe principal : On cherche un axe Δ_2 orthogonal à Δ_1 et minimisant l'inertie $I(N^*, \Delta_2)$. Ce deuxième axe principal passe par l'origine O et est engendré par le vecteur propre normé u_2 de la matrice des corrélations C associée à la deuxième plus grande valeur propre λ_2 .

Deuxième composante principale : La deuxième composante principale Y_2 est une nouvelle variable définie pour chaque individus i par la longueur algébrique $\overline{O_{y2i}}$ coordonné de la projection y_{2i} du point x_i^* sur l'axe Δ_2 . la valeur $Y_2(i)$ est donc égale au produit scalaire entre les vecteurs x_i^* et u_2

$$Y_2(i) = \langle x_i^*, u_2 \rangle = \sum_{j=1}^p u_2 x_{ji}^* = \sum_{j=1}^p u_{2j} ((x_{ij} - \overline{X_j}) / s_j)$$

La composante principale Y_2 s'écrit globalement : $Y_2 = \sum_{i=1}^p u_2 X_i^*$

Inertie L'inertie du nuage $U_1, U_2, \dots, U_n$ est donnée par :

$$I_g = \sum_{i=1}^n p_i \|U_i^t - g\|^2$$

Où g centre de gravité

Le pourcentage d'inertie exprimé par un axe principal (factoriel) α est donnée par :

$$I_g = \lambda_\alpha / \left(\sum_{\alpha=1}^p \lambda_\alpha \right) * 100$$

Le pourcentage d'inertie exprimé par le plan principal (factoriel) est donnée par :

$$I_{\alpha, \alpha'} = (\lambda_\alpha + \lambda'_{\alpha'}) / \sum_{\alpha=1}^p \lambda_\alpha \times 100$$

1.2.4 Qualité de représentation

La qualité de projection d'un individu sur l'axe α est donnée par le carré du cosinus de l'angle θ^α :

$$\cos^2 \theta_i^\alpha = \Psi_{\alpha i}^2 / \|U_i^t - g\|^2$$

La qualité de projection d'un individu sur un sous espace principal (α') est additive :

$$\cos^2 \theta_i(\alpha, \alpha') = \cos^2 \theta_i^\alpha + \cos^2 \theta_i^{\alpha'}, \sum_{\alpha=1}^p \cos^2 \theta_i^\alpha = 1$$

Plus le $\cos^2 \theta_i^\alpha$ est proche de 1 plus l'individu est bien représenté.

Définition :

puisque tous les individus ont le même poids. Un individu contribuera d'autant plus à la confection d'un axe, que sa projection sur cet axe sera éloignée du centre de gravité du nuage. Inversement, un individu dont la projection sur un axe sera proche du centre de gravité contribuera faiblement à l'inertie portée par cette axe. On se sert de ces contributions pour interpréter les nouveaux axes de l'ACP en fonction des individus.

La contribution à l'inertie d'un individu portée par un axe α est :

$$CTR_{\alpha}(U_i) = (P_i \Psi_{\alpha i}^2 / \lambda_{\alpha}), \sum_{\alpha=1}^p CTR_{\alpha}(U_i) = 1$$

On considère que l'individu i a une contribution importante si elle dépasse son poids.

	CTR faible	CTR forte
$\cos^2$ faible	Elément peu contributif quasi indépendant de l'axe	Elément très contributif mais peu illustratif de l'axe.
$\cos^2$ fort	Elément peu contributif mais bien illustratif de l'axe	Elément particulièrement caractéristique de l'axe.

1.2.5 Choix du nombre de composantes

On considère qu'une composante principale Y_h fournit une synthèse globale au tableau des données si sa variance λ_h est supérieure à 1.

En général, on conserve au moins deux composantes principales afin d'avoir accès aux graphiques des individus et des variables.

1.3 Interprétation de l'ACP

L'analyse en composantes principale sera faite sur le tableau des données de X ($X_1, \dots, X_{50}$), et Y (Y_1, Y_2) l'identification des variables se trouve en annexe pages[91 – 94].

Les variables sont homogènes, nous pouvons effectuer une ACP non normée, mais nous avons opté pour une ACP normée. Les résultats ne changent pas, et cela nous permet d'avoir en sortie la matrice de corrélation qui est plus simple à interpréter.

D'après la matrice des corrélations, on constate les corrélations entre variables :

Variables corrélées positivement sont :

- La Température en C° des gaz en sortie du cyclone deux, tour II X_{24} , et la Température en C° des gaz en sortie de la tour échangeur, tour II X_{22} , Sont

très fortement corrélée positivement avec la Température en C° des gaz en sortie de la tour échangeur (sortie cyclone un), tour $I X_6$, respectivement avec les valeurs 0,801, 0,796.

La Vitesse, en *coups/min* de la grille dynamique X_{47} , et la Température en C° des gaz en sortie du cyclone trois, tour $II X_{26}$, sont fortement corrélée positivement avec la Température en C° des gaz en sortie du cyclone trois, tour $I X_{10}$ respectivement avec la valeur de 0,758, 0,822.

- La Dépression en mbar des gaz en entrée du cyclone quatre, tour $I X_{14}$, est très corrélée positivement avec la Dépression en mbar des gaz en sortie du cyclone quatre, tour $I X_{11}$, avec la valeur de 0,844.
- La Température de la matière en entrée four, tour $II X_{32}$, la Température des gaz en sortie de la boîte à fumer, tour $II X_{30}$, et la Vitesse en *tour/min* du ventilateur de refroidissement $VI X_{47}$ sont très fortement corrélée positivement avec la Température des gaz en sortie boîte à fumer, tour $I X_{18}$, respectivement avec la valeurs 0,870, 0,879, 0,817.
- La Température en C° des gaz en sortie du cyclone deux, tour $II X_{24}$, et la Pression en mbar de l'air sous la chambre II de la grille statique X_{43} sont très fortement corrélées corrélée positivement avec la Température en C° des gaz en sortie de la tour échangeur, tour $II X_{22}$, avec les valeurs (0,889, 0,763).
- La Température en C° d'air excédentaire du refroidisseur X_{33} , et la Commande issue du régulateur de pression du capot de chauffe, pour les vitesses des ventilateurs de tirage du filtre refroidisseur, en *tour/min* X_{48} , sont très fortement corrélée positivement avec la Température en C° des gaz en sortie du cyclone quatre, tour $II X_{28}$ respectivement avec les valeurs 0,900, 0,789.
- X_{48} et (la Vitesse en *tour/min* du ventilateur de refroidissement VI) X_{46} , sont très fortement corrélée positivement avec la variable Température en C° d'air excédentaire du refroidisseur X_{33} respectivement avec les valeurs 0,812, 0,808.

Les variables les plus corrélées négativement :

De la matrice de corrélation on remarque que les variables les plus corrélées négativement sont :

- La Vitesse en *tour/min* du ventilateur de refroidissement *VI* X_{46} , est très fortement corrélée négativement avec la Température en C° des gaz en sortie de la tour échangeur (sortie cyclone un), tour *I*. X_6 avec la valeur $-0,858$.
- La Température en C° d'air excédentaire du refroidisseur X_{33} , est très fortement corrélée négativement avec la Température en C° des gaz en sortie de la tour échangeur, tour *II* X_{22} , avec la valeur $-0,907$.
- X_{48} , La Vitesse en *tour/min* du ventilateur de refroidissement *VI* X_{46} , La Température en C° d'air excédentaire du refroidisseur X_{33} , et la Température en C° des gaz en sortie du cyclone quatre, tour *II* X_{28} sont très corrélées négativement avec la variable Température en C° des gaz en sortie du cyclone deux, tour *II*. X_{24} respectivement avec les valeurs $(-0,812, -0,836, -0,961, -0,845)$.
- La Pression en mbar de l'air sous la chambre *II* de la grille statique. X_{43} est corrélée négativement avec la Température en C° d'air excédentaire du refroidisseur. X_{34} avec la valeur $-0,833$.

Variables corrélées négativement sont :

De la matrice de corrélation on remarque que les variables les plus corrélées négativement sont :

- La Vitesse en *tour/min* du ventilateur de refroidissement *VI* X_{46} , est très fortement corrélée négativement avec la Température en C° des gaz en sortie de la tour échangeur (sortie cyclone un), tour *I*. X_6 avec la valeur $-0,858$.
- La Température en C° d'air excédentaire du refroidisseur X_{34} , est très fortement corrélée négativement avec la Température en C° des gaz en sortie de la tour échangeur, tour *II* X_{22} , avec la valeur $-0,907$.
- La Commande issue du régulateur de pression du capot de chauffe, pour les

- vitesse des ventilateurs de tirage du filtre refroidisseur, en *tour/min* X_{48} , La Vitesse en *tour/min* du ventilateur de refroidissement *VI* X_{46} , La Température en C° d'air excédentaire du refroidisseur X_{34} , et la Température en C° des gaz en sortie du cyclone quatre, tour *II* X_{28} sont très corrélées négativement avec la variable Température en C° des gaz en sortie du cyclone deux, tour *II*. X_{24} respectivement avec les valeurs $(-0,812, -0,836, -0,961, -0,845)$.
- La Pression en mbar de l'air sous la chambre *II* de la grille statique. X_{43} est corrélée négativement avec la Température en C° d'air excédentaire du refroidisseur. X_{34} avec la valeur $-0,833$.

Test de sphéricité de Bartlett :

$Khi^2(Valeurobservée)$	$-Inf$
$Khi^2(Valeurcritique)$	1307,53717450748
DDL	1225
$p - value$	
α	0,05

Interprétation du test :

H_0 : Il n'y a pas de corrélation significativement différente de 0 entre les variables.

H_a : Au moins l'une des corrélations entre les variables est significativement différente de 0.

Etant donné que la p-value calculée est inférieure au niveau de signification $\alpha = 0,05$, on doit rejeter l'hypothèse nulle H_0 , et retenir l'hypothèse alternative H_a .

Le risque de rejeter l'hypothèse nulle H_0 alors qu'elle est vraie est inférieur à 0,01%.

Le tableau des valeurs propres se trouve en annex.

Tableau **partiel**, le tableau complet se trouve en annexe.

	F_1	F_2	F_3	F_4	F_5	F_6
<i>Valeur propre</i>	10,571	4,694	4,042	2,210	1,994	1,736
<i>Variabilité(%)</i>	22,492	9,986	8,600	4,701	4,242	3,693
<i>%cumulé</i>	22,492	32,478	41,078	45,780	50,021	53,714

On voit que la première valeur propre vaut 10.57 et représente 22.5% de la variabilité. Cela signifie que si l'on représente les données sur un seul axe, on aura alors 22.5% de la variabilité totale qui sera préservée.

On remarque que le premier plan (axe1 + axe2) représente 32,478% de l'inertie du tableau de données donc la représentation sur les deux premiers axes factoriels est d'assez bonne qualité mais on peut retenir les trois premiers axes pour une meilleure représentation.

D'après les valeurs propres et la règle du coude on prend les trois premières composantes F_1 , F_2 et F_3 avec 41,078% d'inertie expliquée.

1.3.1 Corrélation entre les variables et les facteurs

1.3.2 Nuage des points variables

L'axe F1 Cet axe explique 22.492% de l'inertie totale. C'est un axe d'opposition entre les variables :

{Température en C° des gaz en sortie de la tour échangeur (sortie cyclone un), tour I, Température en C° des gaz en sortie du cyclone trois, tour I, Température en C° des gaz en sortie du cyclone quatre, tour I., Température en C° des gaz en sortie de la tour échangeur, tour II, Température en C° des gaz en sortie du cyclone deux, tour II, Température en C° des gaz en sortie du cyclone trois, tour II, Pression en mbar de l'air sous la chambre II de la grille statique, Pression en mbar de l'air sous la chambre III, de la grille dynamique, Débit de combustible en $Nm^3/heure$, dans le brûleur four.} **et** {Température des gaz en sortie boîte à fumer, tour I, Température en C° des gaz en sortie du cyclone quatre, tour II, Température de la matière en

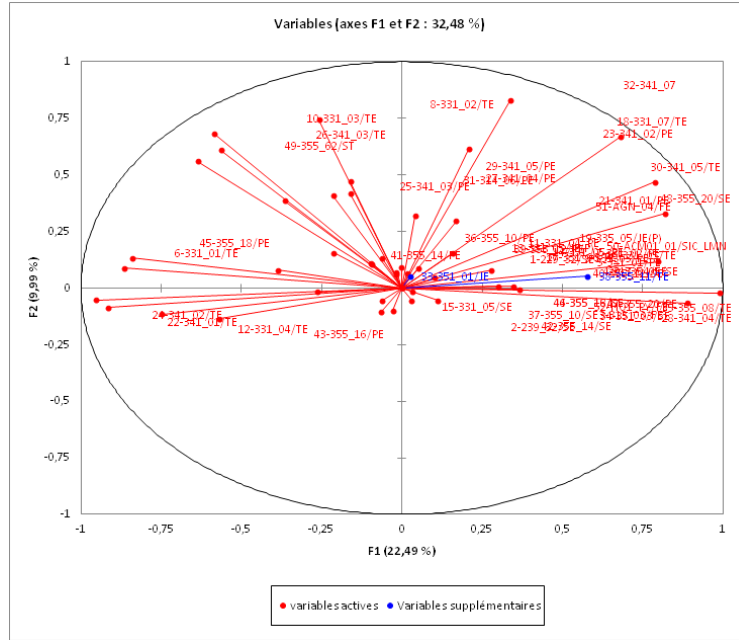


FIG. 1.1: Nuage des points variables de X et Y

entrée four, tour *II*, Pression en mbar de l'air sous la grille statique, (refoulement du ventilateur *I*), Commande issue du régulateur de pression du capot de chauffe, pour les vitesses des ventilateurs de tirage du filtre refroidisseur, en *tour/min*, Débit de combustible en $m^3/heure$, dans le brûleur de pré-calcination}.

Autrement dit les individus qui se projettent à gauche de l'axe1 ont de fortes valeurs de {6 10 12 22 24 26 43 45 49} et de faibles valeurs de {18 28 30 35 48 50} et inversement les individus qui se projettent à droite de l'axe1 ont de fortes valeurs de {18 28 30 35 48 50} et de faibles valeurs de {6 10 12 22 24 26 43 45 49}.

Si on regarde les coordonnées des individus sur le premier axe, à part quelques exceptions (100, 103 et 106), les premiers individus (jusqu'au 112) ont des coordonnées positives tandis que les derniers individus (à partir de l'individu 113) ont des coordonnées négatives.

Les individus sont la variation du temps par minute (256 individus et individu1 = 23h39mn , individu256 = 3h54mn soit 5h03mn).

Donc le premier axe oppose les premiers individus, autrement dit les premiers

instants de 23h39 à 01h29, caractérisés par de fortes valeurs de {18 28 30 35 48 50} et de faibles valeurs de {6 10 12 22 24 26 43 45 49} aux derniers individus, les instants entre 01h31 à 03h54, caractérisés par de fortes valeurs de {6 10 12 22 24 26 43 45 49} et faibles valeurs de {18 28 30 35 48 50}.

Axe2 Cet axe explique 9.987% de l'inertie totale ($I_T = 50$). Toutes les coordonnées sont positives ce qui fait que l'axe 2 est un axe d'échelle.

Les plus grandes coordonnées des variables sur l'axe 2 :

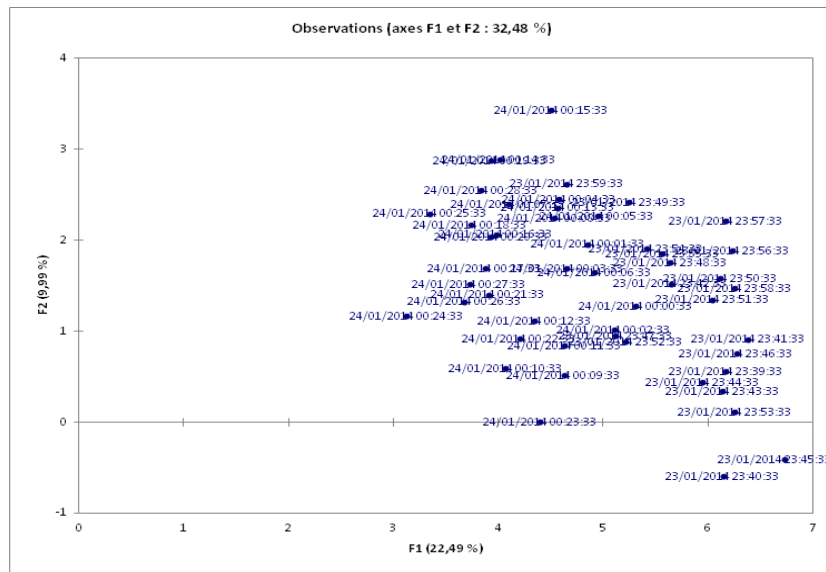
{Température en C° des gaz en sortie du cyclone deux, tour I, Température en C° des gaz en sortie du cyclone trois, tour I, Température des gaz en sortie boîte à fumer, tour I, Dépression en mbar des gaz en sortie du cyclone deux, tour II, Température en C° des gaz en sortie du cyclone trois, tour II, Température des gaz en sortie de la boîte à fumer, tour II, Débit de combustible en $Nm^3/heure$, dans le brûleur four}.

L'axe 2 est un axe d'échelle. Cela signifie que les projections des individus sur cet axe vont être ordonnées des plus fortes valeurs des variables citées ci-dessus aux plus faibles.

Les individus ayant les plus fortes valeurs de {8 10 18 23 26 32 49} sont : 11, 19, 26, 27, 29, 30, 35, 36, 37, 38, 40, 41, 42, 47, 50, 51, 52, 53, 56, 58, 219 à 237, 240 à 255.

Les individus ayant les plus faibles valeurs de {8 10 18 23 26 32 49} sont : 92, 99, 100 à 114, 117 à 153, 155, 158, 207.

On remarque que les derniers instants les température en C° des gaz en sortie du cyclone deux, tour I, température en C° des gaz en sortie du cyclone trois, tour I, température des gaz en sortie boîte à fumer, tour I, dépression en mbar des gaz en sortie du cyclone deux, tour II, température en C° des gaz en sortie du cyclone trois, tour II, température des gaz en sortie de la boîte à fumer, tour II, débit de combustible en $Nm^3/heure$, dans le brûleur four **sont faibles**

FIG. 1.2: Observation sur les axes F_1 et F_2

Axe3 Cet axe explique 8.6% de l'inertie totale et est expliqué par les variables 11 et 14, Axe d'échelle expliquant les Dépression en mbar des gaz en sortie du cyclone quatre, tour I . et Dépression en mbar des gaz en entrée du cyclone quatre, tour I .

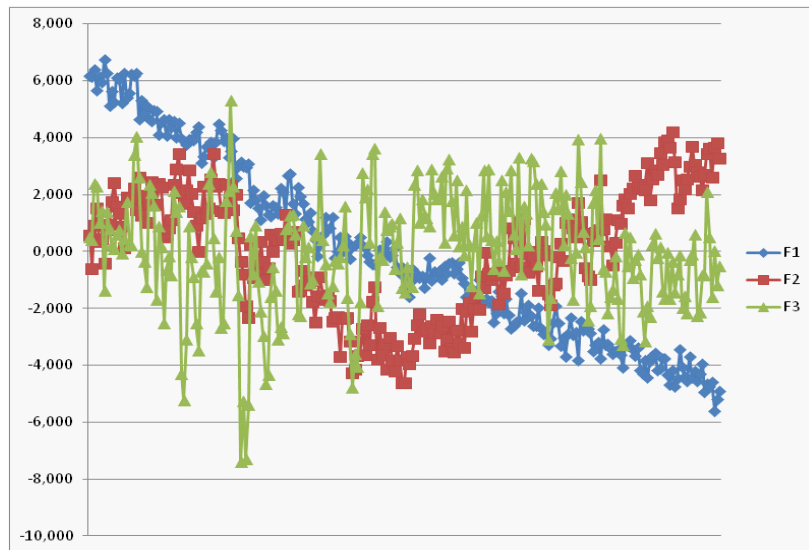


FIG. 1.3: Les trajectoires des Coordonnées d'observations de F_1 , F_2 et F_3

La figure représente les trajectoires des coordonnées d'observation F_1 , F_2 et F_3 en fonction du temps par minute

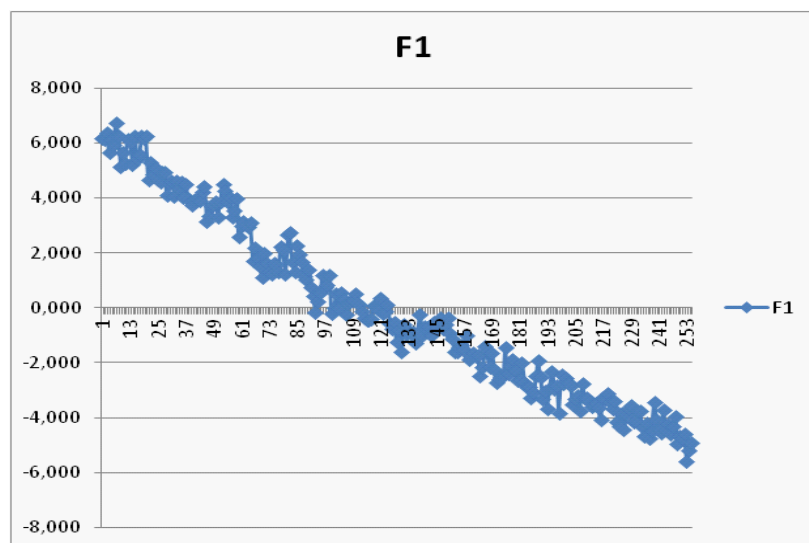


FIG. 1.4: La trajectoire de F_1

La trajectoire F_1 est une courbe décroissante avec le temps elle a l'allure d'une droite.

Elle commence avec une valeurs maximale à environ la valeur 6 à la première minute elle décroît avec le temps, elle prend la valeurs 0 après environ $1H37mn$; elle

décroit toujours avec le temps après $4h21mn$ environs elle atteint ça valeur minimale à la valeurs (-6) .

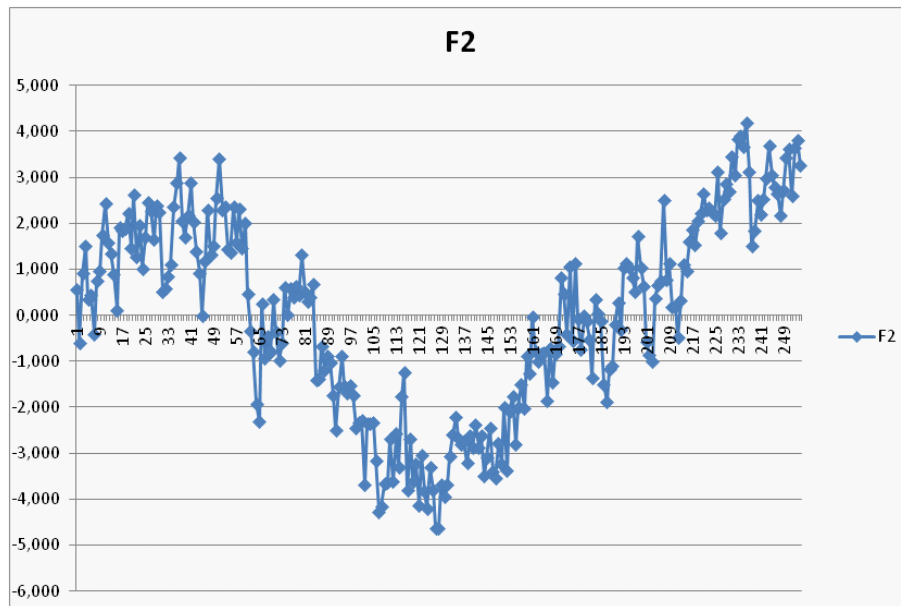
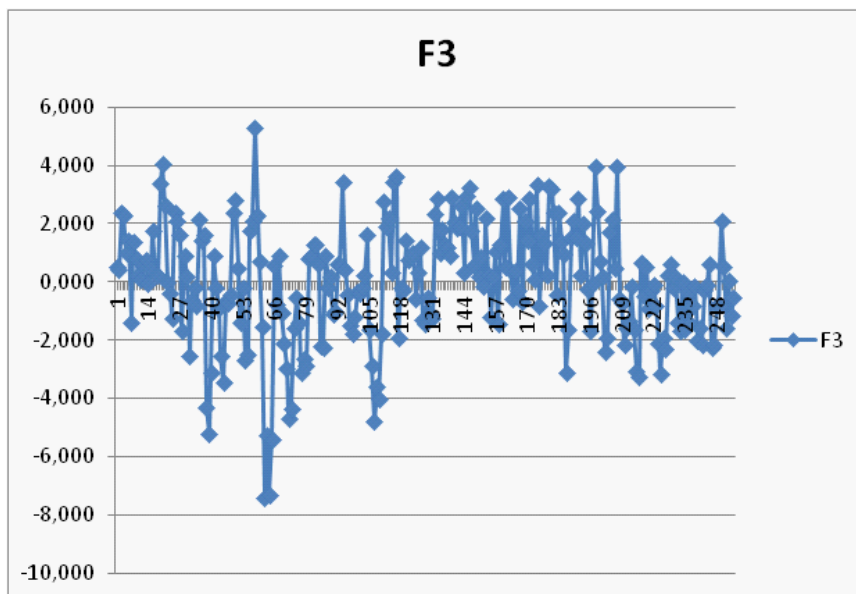


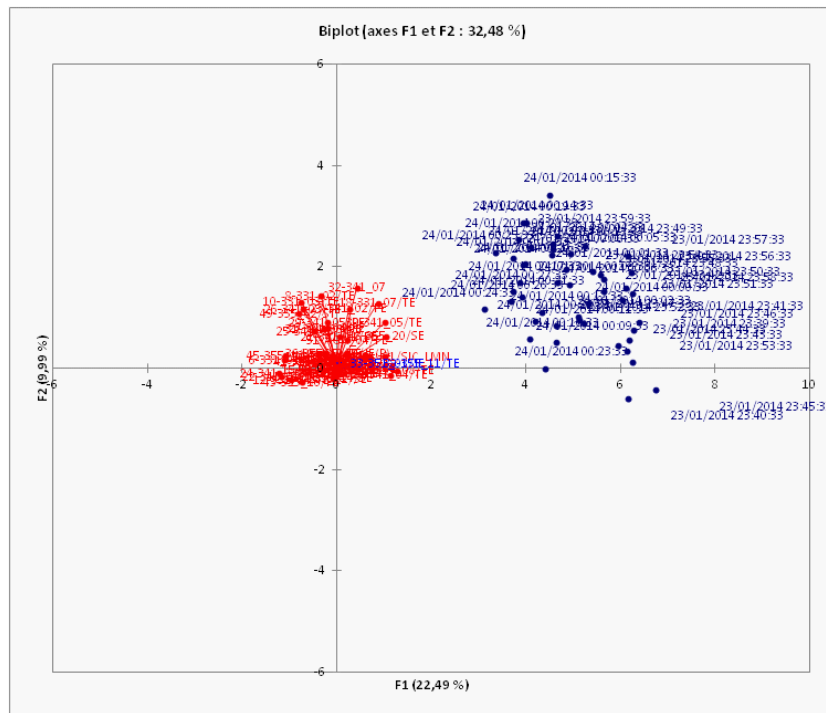
FIG. 1.5: La trajectoire de F_2

La trajectoire F_2 commence à la 1^{er} minute avec une valeur négative elle croit et décroît avec des intervalle de $5mn$ jusqu'à atteindre la valeur 3,42 à la $57mn$, puis elle décroît brutalement à la valeur $-4,622$ à $1h05mn$; elle croit jusqu'à la valeur 4,119 après $4h3mn$

*Globalement la trajectoire de F_2 a une allure croissante puis décroissante puis croissante

FIG. 1.6: La trajectoire de F_3

La trajectoire de F_3 a une allure presque stable sauf pour les deux piques dont le premier est à la $58mn$ avec la valeur $5,296$, le deuxième pique juste après le premier à $1h62mn$ c'est une valeur négative $-7,405$

FIG. 1.7: Biplot des axes F_1 et F_2

2 Application de la régression PLS univariée

Les variables Y_1 et Y_2 sur les quelles nous allons faire la régression PLS sont :

Y_1 : Somme des puissances en KW , des deux moteurs tournant le four.

Y_2 : Température en C° de l'air secondaire.

2.1 Interprétation de la régression PLS1 pour la variable Y_1

Les statistiques de base des différentes variables (les variables explicatives et les deux variables dépendantes ainsi que la matrice de corrélations correspondante sont données dans l'annexe. Deux points méritent d'être souligné :

(i) Le premier est l'hétérogénéité des variances des variables et suggère de centrer et réduire l'ensemble des variables.

(ii) Le second est la corrélation faible entre les variables dépendantes et suggère d'envisager deux régressions PLS, chacune sur une variable au lieu des deux simultanément.

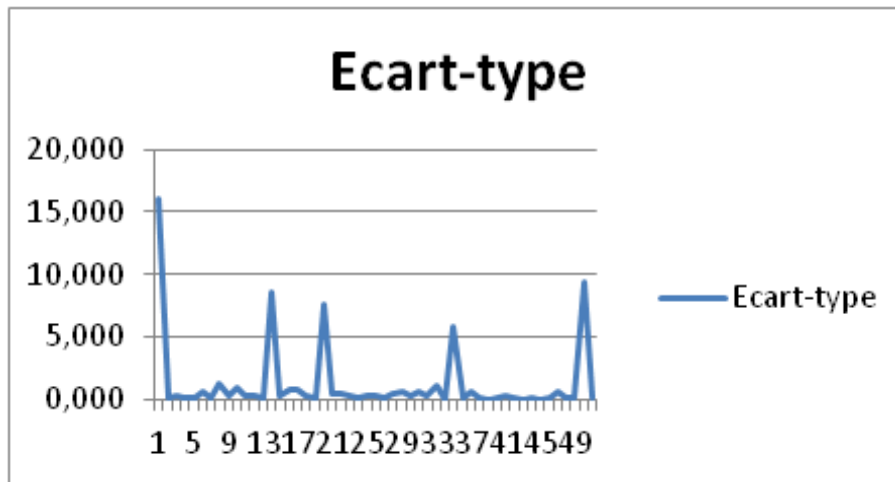


FIG. 2.1: Ecart type

En premier lieu, la procédure PLS1 produit un tableau permettant de vérifier comment chaque composante PLS extraite rend compte de la variabilité du tableau explicatif et la réponse Y au sein du modèle. Ainsi, on peut constater que les trois premières composantes PLS extraites représente près de 22.4% de la variabilité des signaux et expliquent plus de 35.3% de la variabilité du signal, Y_1 (Somme des puissances en KW , des deux moteurs tournant le four) Y_1 .

	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9	t_{10}
<i>variance</i>	2,222	4,364	4,670	0,906	2,107	1,139	0,780	0,800	0,582	0,654
$\%varexp(t_i)$	0,044	0,087	0,093	0,018	0,042	0,023	0,015	0,016	0,011	0,013

Le tableau et le graphique correspondant permettent de visualiser la qualité de la régression PLS en fonction du nombre de composantes retenues.

L'indice Q^2 cumulé est une mesure globale de la qualité de l'ajustement et de la qualité prédictive de 10 modèles. On voit que l'indice Q^2 reste stable à partir de la troisième composante. Les R_Y^2 cum et R_X^2 cum qui correspondent aux corrélations entre les composantes et les variables de départ sont proches de 0,4 dès la quatrième composante, et reste stable pour les autres dimensions ce qui indique que 4 composantes suffisent mais ces composantes sont relativement représentatives des X et des Y .

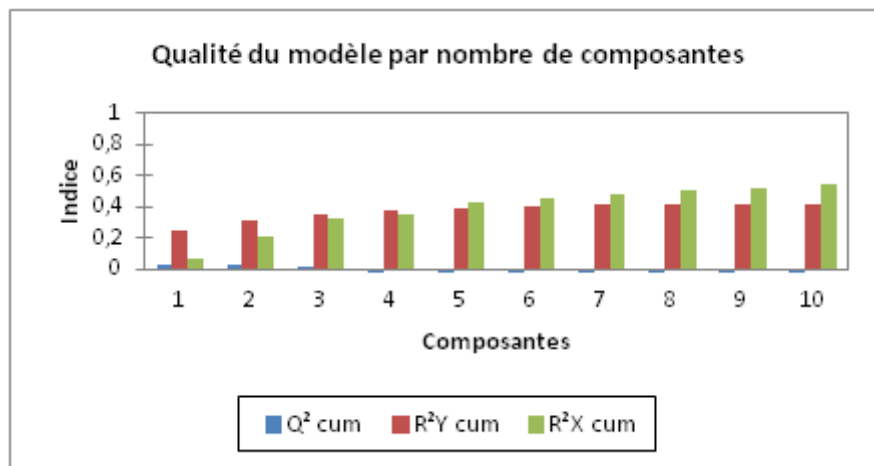


FIG. 2.2: Qualité du modèle par nombre de composantes

Le graphique des corrélations permet de visualiser sur les deux premières composantes générées par la régression PLS les corrélations entre X et les Y du modèle

Le tableau ci-dessous propose alors le tableau des pourcentages de variances expliquées par des modèles de 1 à 10 composantes [voir en annexe C page xiv].

En premier lieu, la procédure PLS1 produit un tableau permettant de vérifier comment chaque composante PLS extraite rend compte de la variabilité du tableau explicatif et la réponse Y au sein du modèle. Ainsi, on peut constater que les trois premières composantes PLS extraites représente près de 28% de la variabilité des signaux et expliquent plus de 55% de la variabilité du signal, Y_1 (Somme des puissances en KW , des deux moteurs tournant le four).

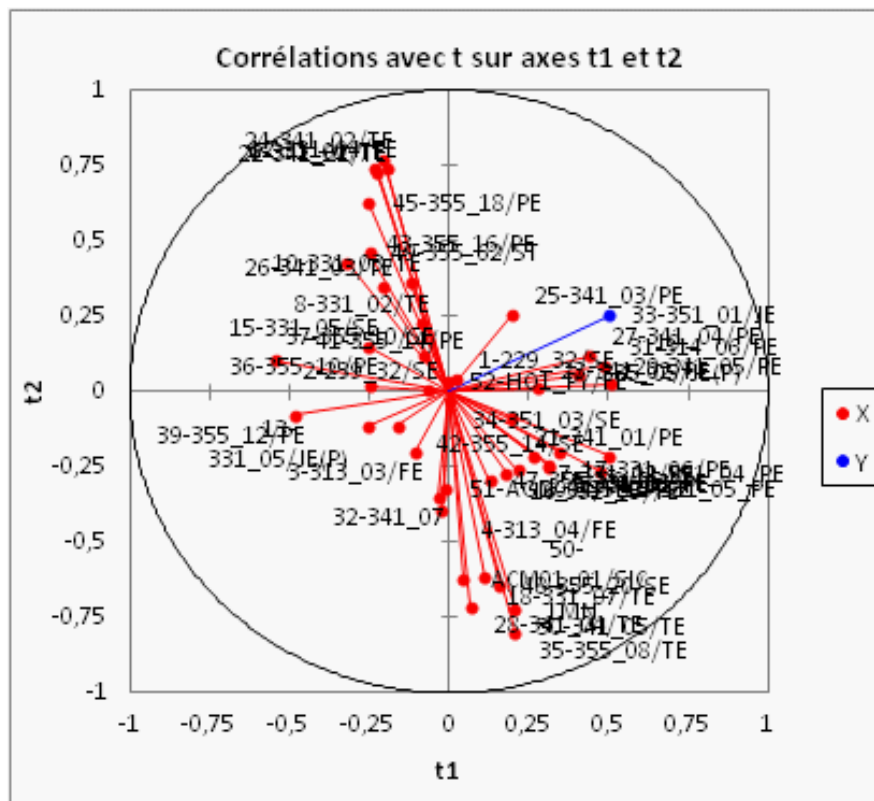


FIG. 2.3: Matrice des corrélations de PLS1

On note que parmi les variables explicatives, que seule X_6 est mal représentée sur le graphique. Cette dernière est donc globalement faiblement explicative de la variable à expliquer. La carte superposant les Variables dépendantes sur les vecteurs c et les variables explicatives sur les vecteurs w^* permet de visualiser la relation globale entre les variables, sachant que les w^* sont représentatifs du poids des variables dans les modèles.

Vecteurs c :

Variable	c_1	c_2	c_3	c_4	c_5
Y_1	0,335	0,120	0,091	0,171	0,068
Variable	c_6	c_7	c_8	c_9	c_{10}
Y_1	0,102	0,111	0,074	0,074	0,049

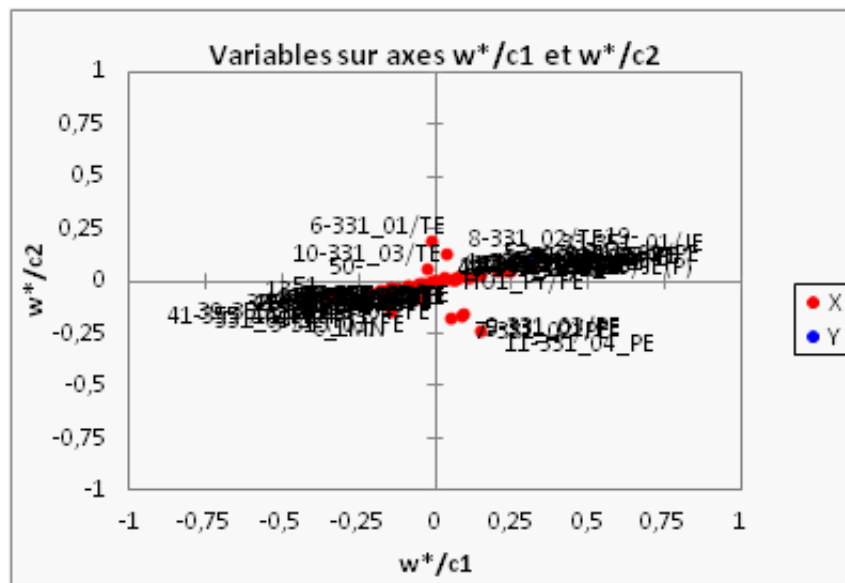


FIG. 2.4: carte superposant les variables dépendantes sur c et les variables explicatives sur w^*

Si l'on projette une variable explicative sur le vecteur d'une variable dépendante, on a une idée de son poids dans le modèle concernant cette même variable dépendante.

Coordonnées des observations sur t :

Un tableau donne les coordonnées (ou trajectoires) sur les composantes t . Le graphique correspondant sur les axes t_1 et t_2 est ensuite affiché. On notera que les trajectoires sont bien distinguées.

Les trajectoires sur t :

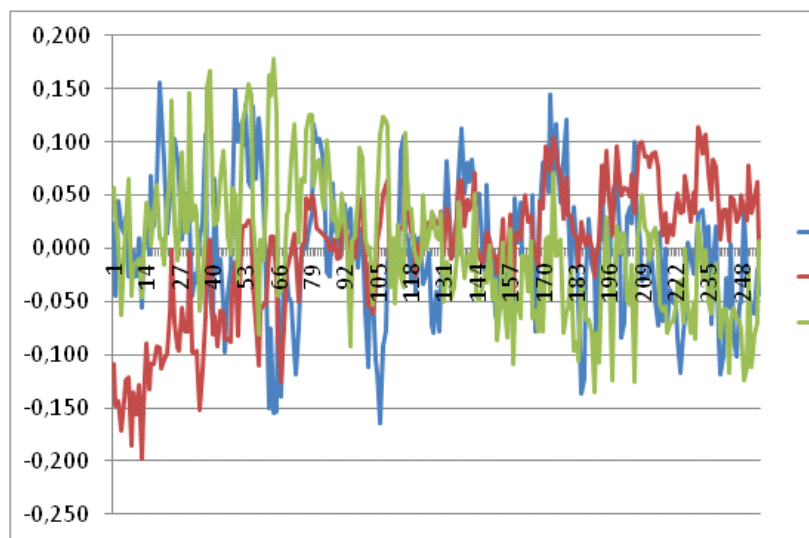
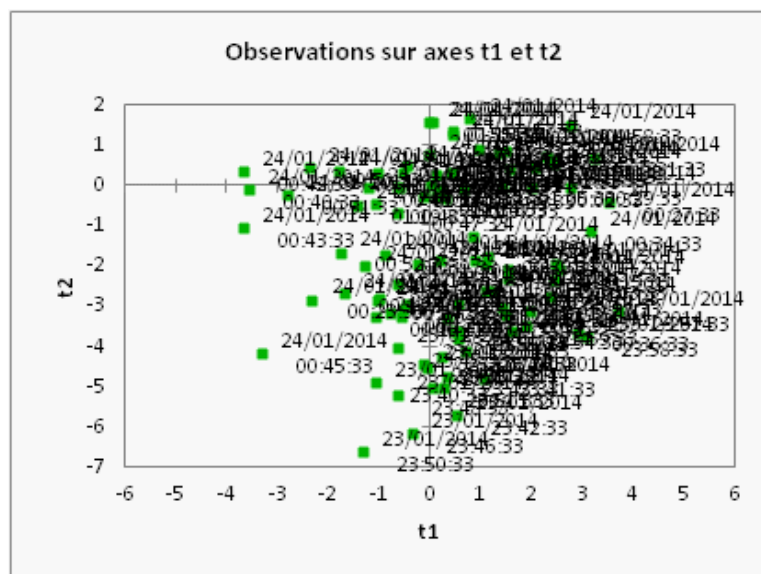
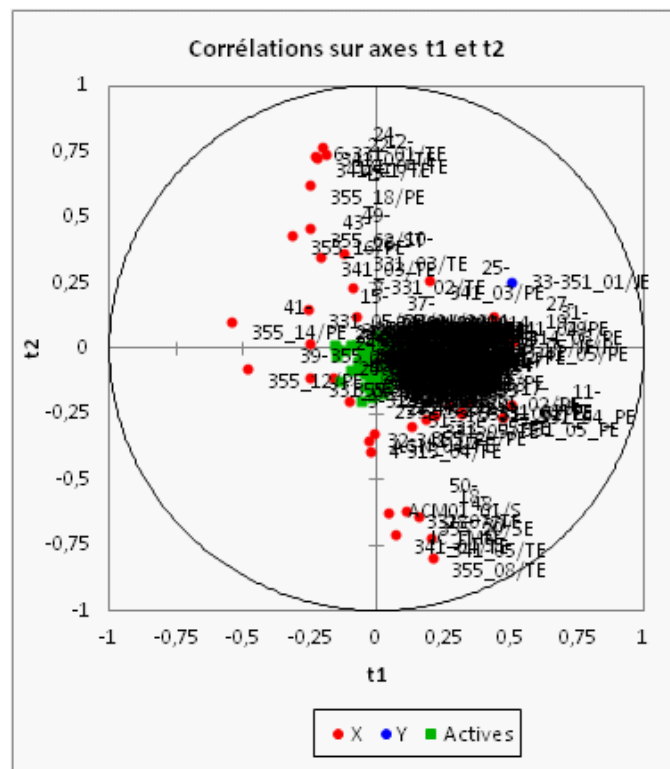


FIG. 2.5: Les trajectoires de t_1 , t_2 et t_3

FIG. 2.6: Observations sur les axes t_1 et t_2

FIG. 2.7: Corrélations sur les axes t_1 et t_2

Les trajectoires sur u :

Deux tableaux fournissant des résultats pour les composantes u . Un graphique permet de visualiser les observations (ici les trajectoires) dans l'espace des $\tilde{u}$.

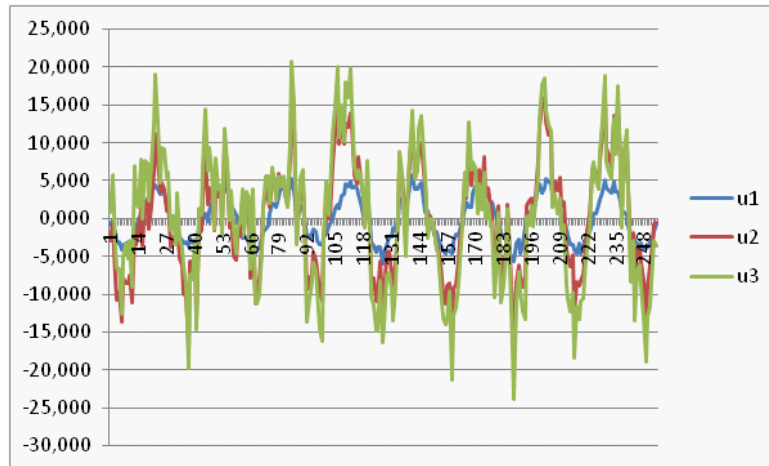


FIG. 2.8: Les trajectoires de u_1, u_2 et u_3

Les composantes u_1, u_2, u_3 ont la meme allure

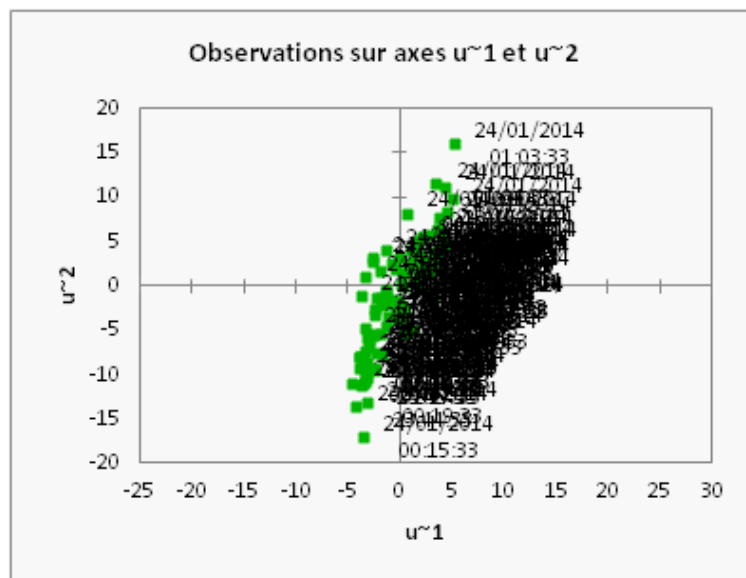


FIG. 2.9: Observations sur $\tilde{u}_1$ et $\tilde{u}_2$

2.1.1 Variables importante dans la projection (VIP)

Le tableau suivant présente les VIP (Variable Importance for the Projection) pour chacune des variables explicatives, sur chacun des modèles avec un nombre croissant de composantes. Cela permet d'identifier rapidement quelles sont les variables explicatives les plus importantes sur l'ensemble des modèles. Sur la première composante, la variable X_{42} , apparaissent comme étant peu influentes. On voit aussi que seuls les coefficients des variables X_{31} , X_{29} , X_{27} , X_{13} , X_{11} , X_{35} , X_{26} , X_{49} , X_{40} , X_3 , X_{25} , X_{15} sont significatifs.

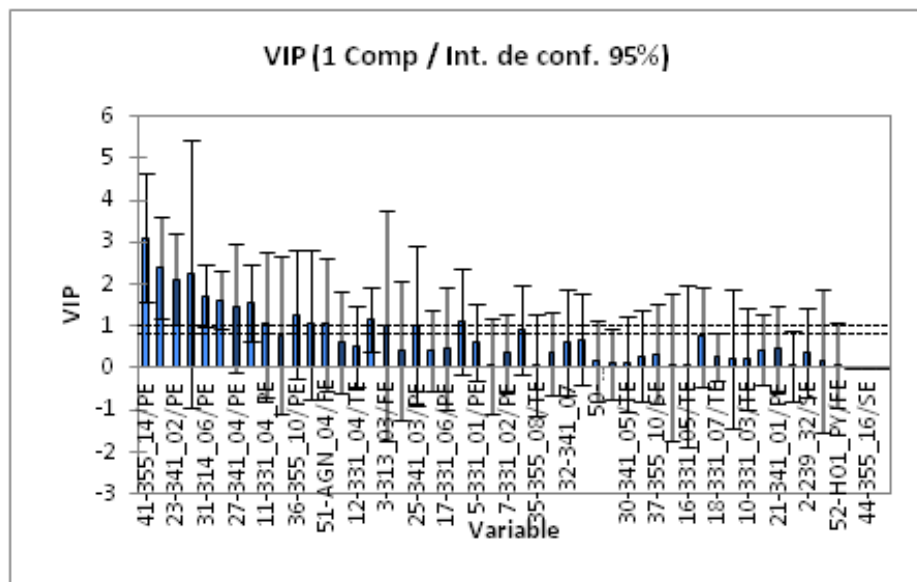
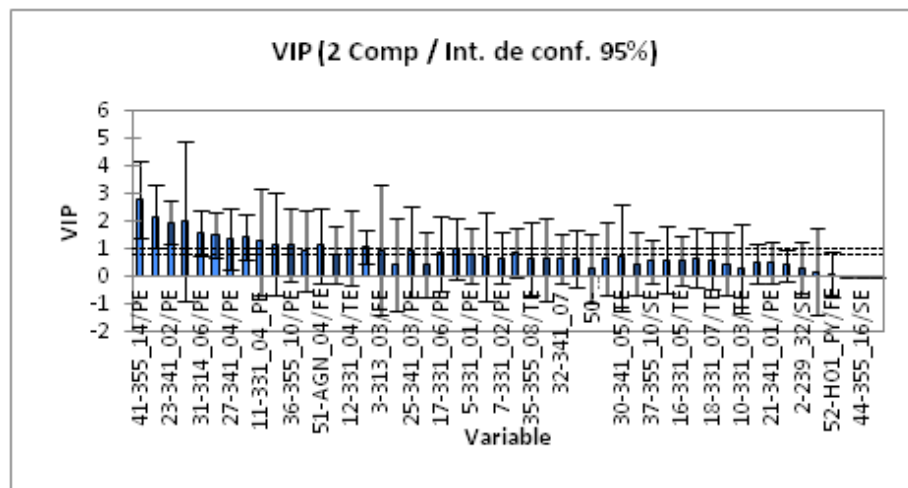


FIG. 2.10: graphique des VIP à 1 composante

FIG. 2.11: Graphique des *VIP* à 2 composantes

Même indication que pour la 1^{ère} composante

En ce qui concerne la deuxième composante on a presque la même chose, les variables X_{39} , X_{37} , X_{19} et X_{23} sont les plus importantes dans la construction de Y_1 .

On voit ici que seuls les coefficients des variables Pression en mbar de l'air sous la chambre *I* de la grille dynamique, (refoulement du ventilateur *IV*) X_{39} , Pression en mbar de l'air sous la grille statique, (refoulement du ventilateur *II*) X_{37} , Puissance en *KW* du moteur de la ventilateur exhauster, tour *II* X_{19} , sont significatifs. Le tableau des prédictions et résidus permet de vérifier que les résultats de la Somme des puissances en *KW*, des deux moteurs tournant le four Y_1 sont très bien reproduites par cette dernière.

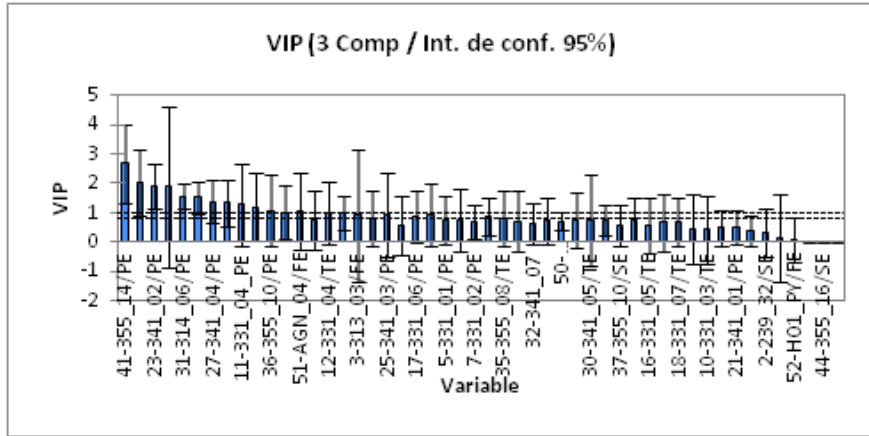


FIG. 2.12: Graphique des VIP à 3 composantes

Le tableau des paramètres des modèles correspondant à chacune des variables dépendantes est ensuite affiché (Annexe). Les équations sont ensuite affichées afin de faciliter une éventuelle utilisation ultérieure.

2.1.2 Equation du modèle

$$\begin{aligned}
 Y_1 = & 16872.359 - 1.55 * X_1 - 1.9356 * X_2 - 21.132 * X_3 + 6.425 * X_4 - 0.872 * X_5 - \\
 & 19.605 * X_6 - 6.879E - 02 * X_7 + 5.200 * X_8 + 1.254 * X_9 + 7.455 * X_{10} + 1.126 * X_{11} + \\
 & 7.606 * X_{12} - 0.144 * X_{13} + 0.623 * X_{14} - 1.121 * X_{15} - 0.179 * X_{16} + 1.520 * X_{17} - \\
 & 3.316 * X_{18} + 0.255 * X_{19} + 2.200 * X_{20} + 0.541 * X_{21} + 4.882 * X_{22} + 14.453 * X_{23} + \\
 & 1.034 * X_{24} + 1.707 * X_{25} - 14.353 * X_{26} + 1.785 * X_{27} + 0.221 * X_{28} + 6.444 * X_{29} - \\
 & 0.612 * X_{30} + 5.618 * X_{31} - 0.940 * X_{32} - 277.904 * X_{33} - 9.936 * X_{34} - 4.029 * X_{35} + \\
 & 0.580 * X_{36} - 23.924 * X_{37} - 41.931 * X_{39} - 3.570 * X_{40} - 4.965 * X_{41} + 3.912 * X_{43} - \\
 & 7.2687 * X_{45} + 2.423 * X_{46} - 9.543 * X_{47} + 10.597 * X_{48} - 0.119 * X_{49} - 2.545 * X_{50}
 \end{aligned}$$

Indice de qualité Q^2 :

<i>Composante</i>	Y_1	<i>Total</i>
<i>Comp1</i>	0,031	0,031
<i>Comp2</i>	0,001	0,001
<i>Comp3</i>	-0,027	-0,027
<i>Comp4</i>	-0,267	-0,267
<i>Comp5</i>	-0,195	-0,195
<i>Comp6</i>	-0,293	-0,293
<i>Comp7</i>	-0,336	-0,336
<i>Comp8</i>	-0,327	-0,327
<i>Comp9</i>	-0,323	-0,323
<i>Comp10</i>	-0,278	-0,278

En prend la 1ère composante pour Y_1

Si on ne garde que les variables importantes, cette équation devient :

$$Y_1 = 16872.359 - 21.132 * X_3 + 1.126 * X_{11} - 0.144 * X_{13} - 1.121 * X_{15} + 0.255 * X_{19} + 14.453 * X_{23} + 1.707 * X_{25} - 14.353 * X_{26} + 1.785 * X_{27} + 6.444 * X_{29} + 5.618 * X_{31} - 4.029 * X_{35} - 23.924 * X_{37} - 41.931 * X_{39} - 3.570 * X_{40} - 0.119 * X_{49}.$$

2.1.3 Coefficients d'ajustement de la Variable Y_1

Le tableau des coefficients d'ajustement, le tableau des coefficients normalisés (correspondant aux coefficients bêta de la régression linéaire classique) et enfin le tableau des prédictions et résidus. L'analyse du modèle correspondant à la variable Y_1 nous permet de conclure à une bonne qualité du modèle, le R^2 valant 0.88. Cependant, le nombre de degrés de liberté est faible ($DDL = 1$), et on risque un sur-ajustement du modèle. Cela tend à être confirmé par le fait que pour l'ensemble des coefficients normalisés, les intervalles de confiance comprennent la valeur 0. Etant donné que le Q^2 cumulé diminue dès la troisième composante (cf tableau des Q^2 , Annexe C page xv), il est fort probable qu'une qualité similaire aurait été atteinte

avec seulement deux composantes.

<i>Observations</i>	254,000
<i>Sommedespoids</i>	254,000
<i>DDL</i>	243,000
R^2	0,421
<i>Ecart – type</i>	12,428
<i>MCE</i>	147,763
<i>RMCE</i>	12,156

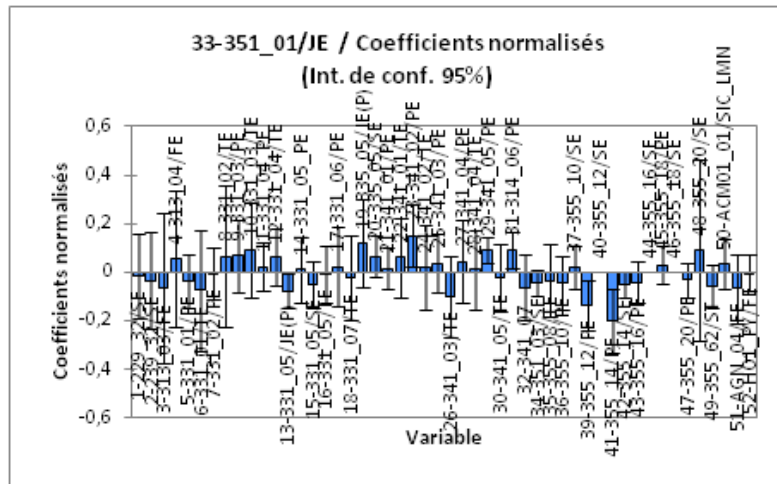


FIG. 2.13: Y_1 /coefficient normalisé

On voit ici que seuls les coefficients des variables Pression en *mbar* de l'air sous la chambre *I* de la grille dynamique, (refoulement du ventilateur *IV*) X_{39} ; Pression en *mbar* de l'air sous la grille statique, (refoulement du ventilateur *II*) X_{37} , Puissance en *KW* du moteur du ventilateur exhausteur, tour *II* X_{19} , sont significatifs. Le tableau des prédictions et résidus permet de vérifier que les résultats de la Somme des puissances en *KW*, des deux moteurs tournant le four Y_1 sont très bien reproduits par cette dernière.

Le diagramme des résidus normalisés permet de contrôler la présence de points atypiques. Ce diagramme des résidus normalisés, ou nomogramme, croise les résidus de l'estimation en Y réalisée par le modèle en ordonnée avec les scores normaux

empiriques de ces résidus en abscisse, selon le principe de la Droite de Henry, pour fournir un test graphique de la normalité des résidus.

La détection de points aberrants peut également se faire en calculant des indices de distance au modèle, soit dans l'espace des X et dans l'espace des Y :

Pour déterminer quels sont les facteurs à exclure du modèle, il est possible de consulter les coefficients de régression (matrice B) et les indices VIP (importance variables de la projection) pour chacun des facteurs. Les coefficients de régression représente l'importance de chacun des facteurs dans l'ajustement du modèle, à la fois pour les facteurs et pour les réponses. Si un facteur a un coefficient de régression relativement petit en valeur absolue et si son indice VIP a une valeur relativement faible (moins de 0.8) alors la variable explicative peut être exclue du modèle. On se limite à interpréter les deux premières composantes.

Prédictions et résidus de la Variable Y_1 :

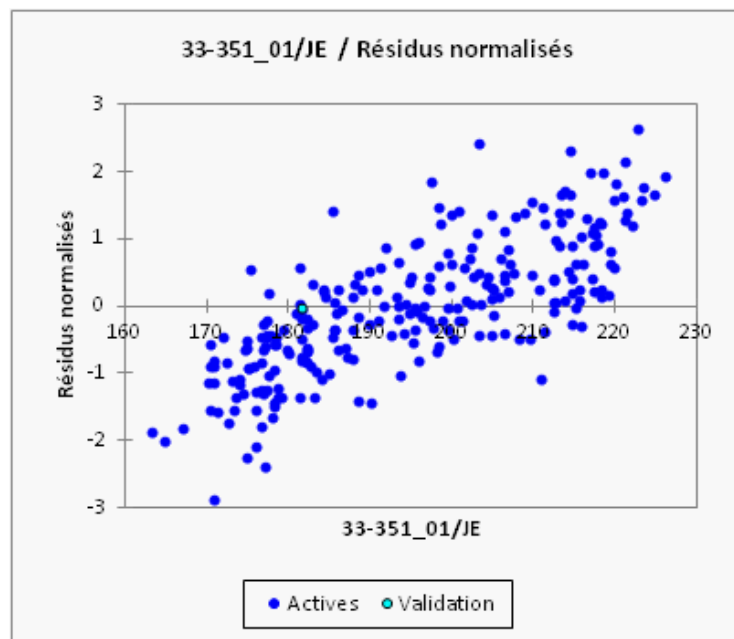
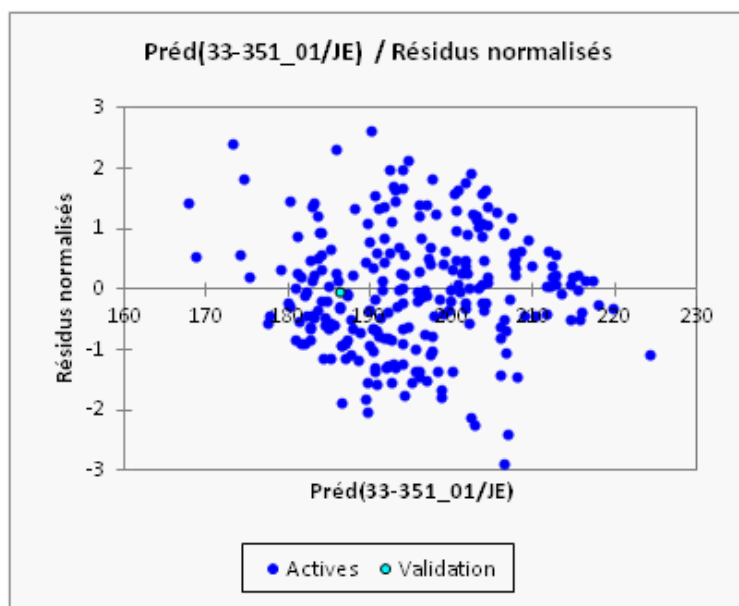
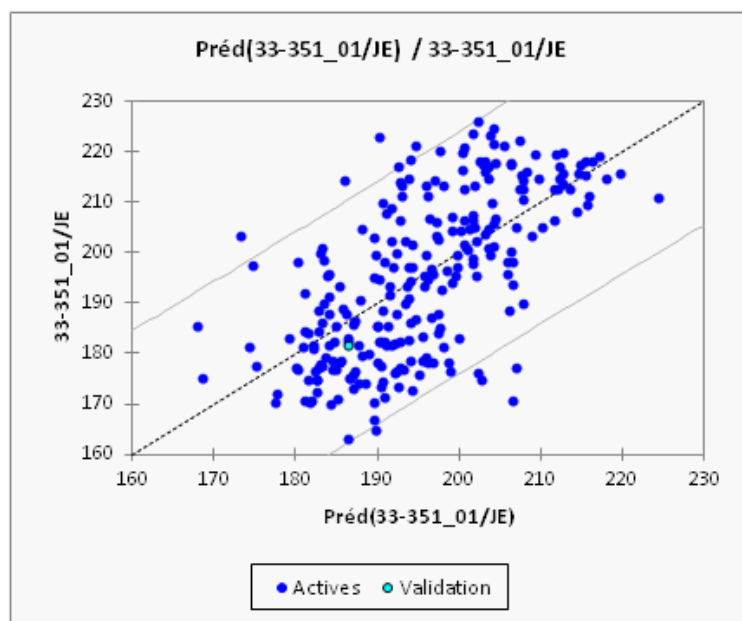


FIG. 2.14: le signal Y_1 /Résidus normalisés

Le diagramme des Y -résidus contre les Y -estimés permet de juger de la pertinence de la spécification du modèle :

FIG. 2.15: prédiction du signal Y_1 en fonction du résidus normailès

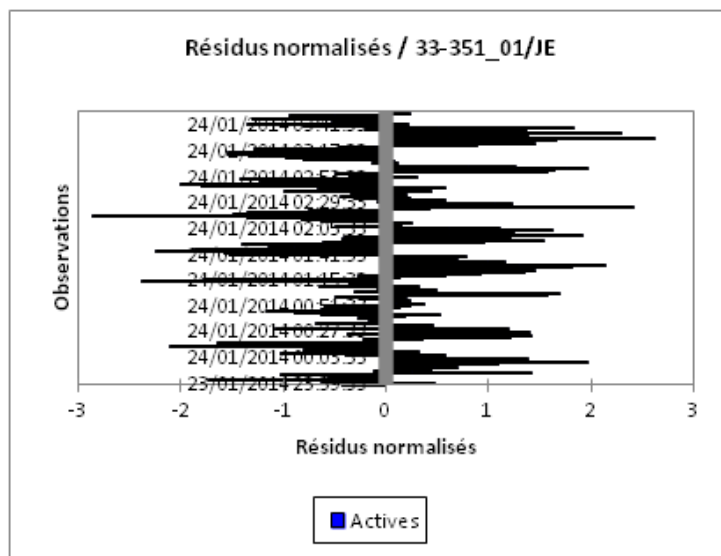


FIG. 2.16: Les observations en fonction des résidus normalisés de la variable Y_1 (actives)

<i>date</i>	<i>heure</i>	<i>valeurs</i>
24/01/2014	00 : 15 : 33	-2,102
24/01/2014	01 : 16 : 33	-2,388
24/01/2014	01 : 32 : 33	2,14
24/01/2014	01 : 45 : 33	-2,247
24/01/2014	02 : 19 : 33	-2,868
24/01/2014	02 : 27 : 33	2,415
24/01/2014	02 : 48 : 33	-2,010
24/01/2014	03 : 30 : 33	2,632
24/01/2014	03 : 36 : 33	2,301

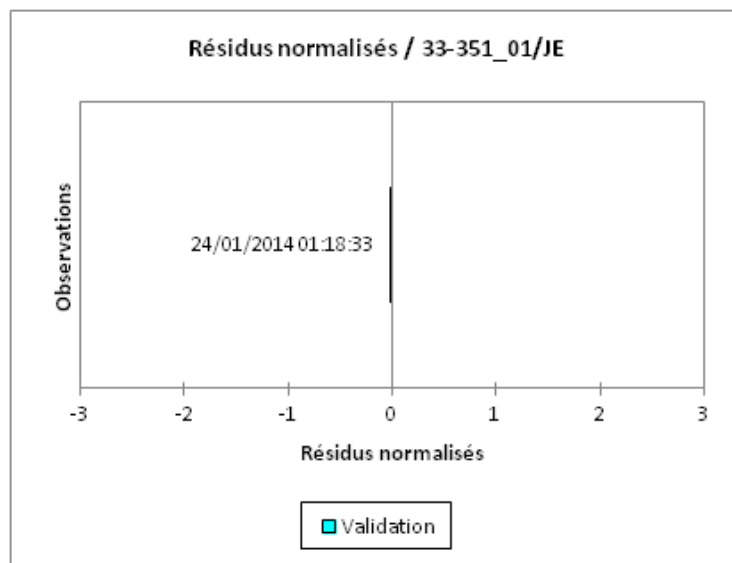


FIG. 2.17: Les observations en fonction des résidus normalisés de la variable Y_1 (validation)

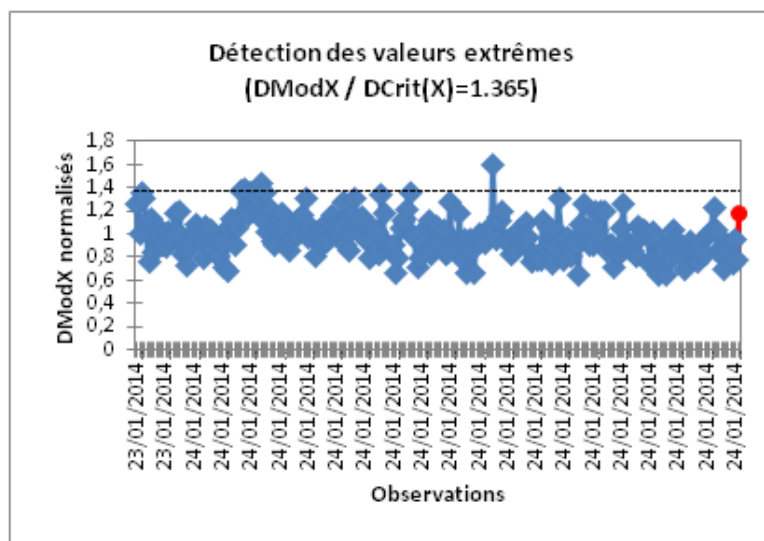
Détection des valeurs extrêmes :

FIG. 2.18: Graphique des DModX normalisées

$$DCrit(X)=1.365$$

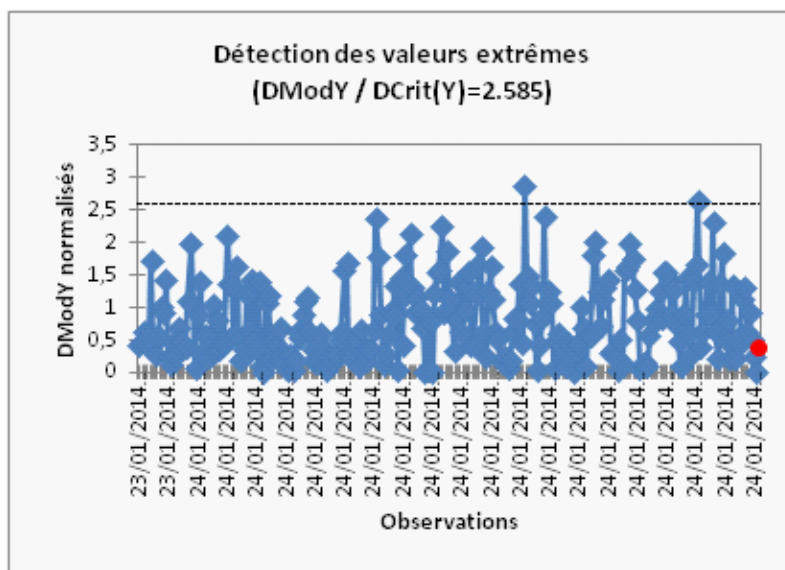


FIG. 2.19: Geaphique des DModY normalisées

$$DCrit(Y)=2.585$$

2.2 Interprétation de la régression PLS1 pour la variable Y_2

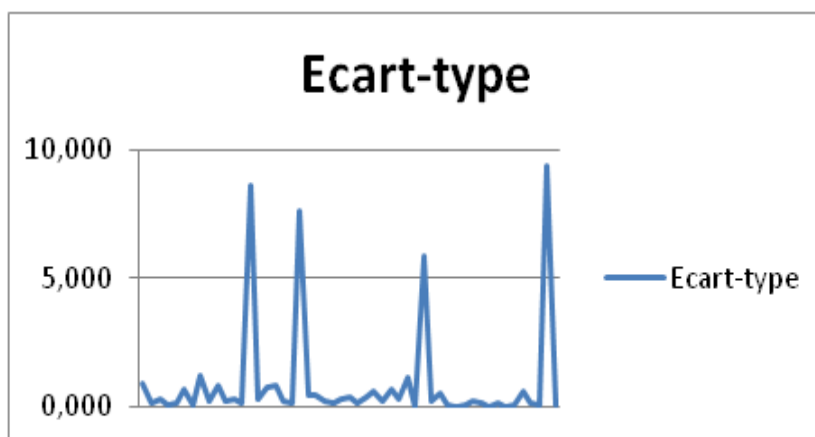


FIG. 2.20: Ecart type

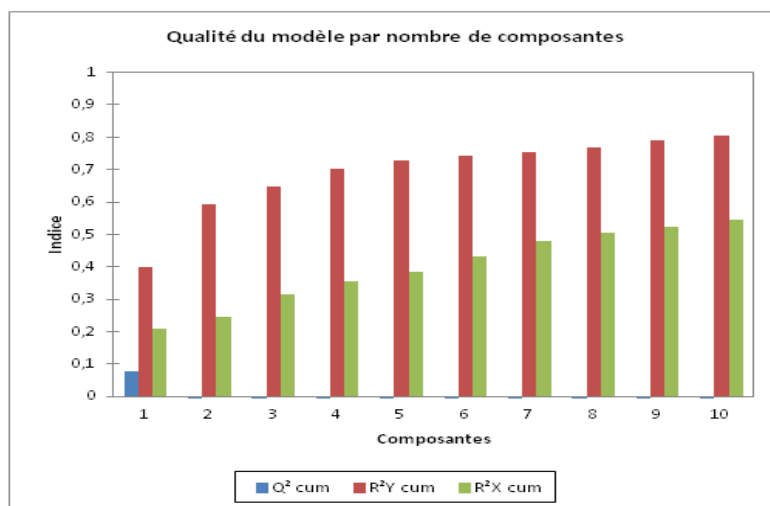
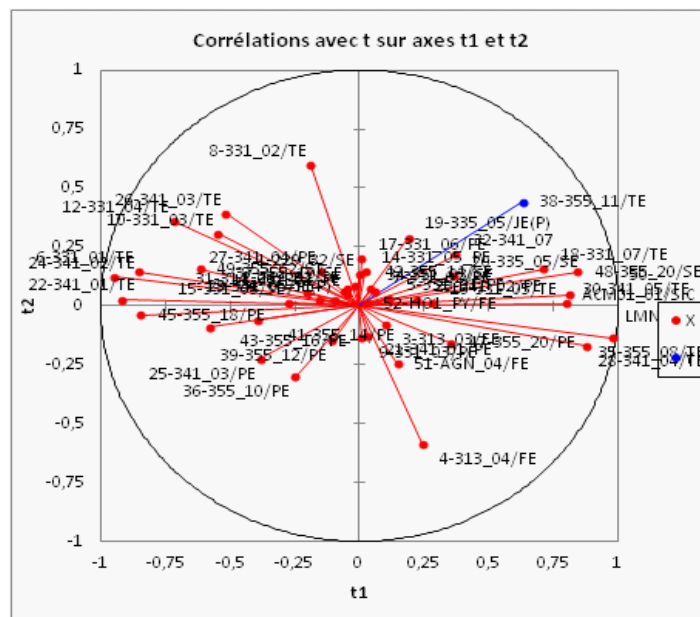
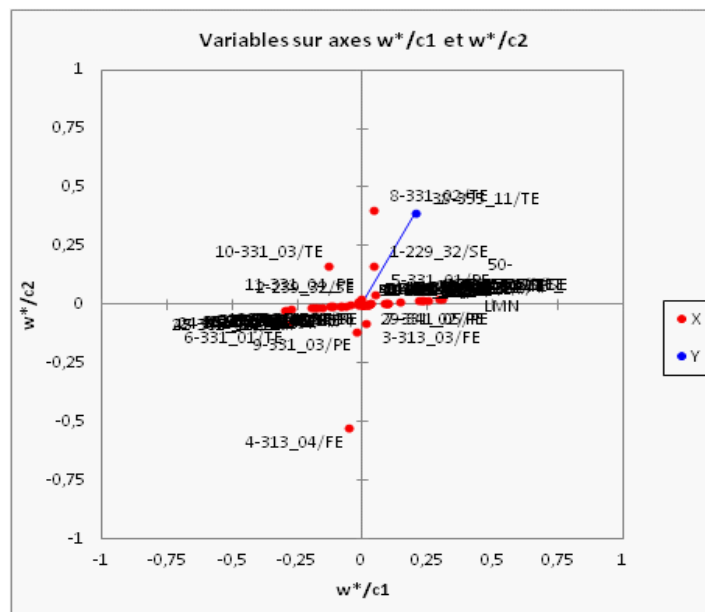
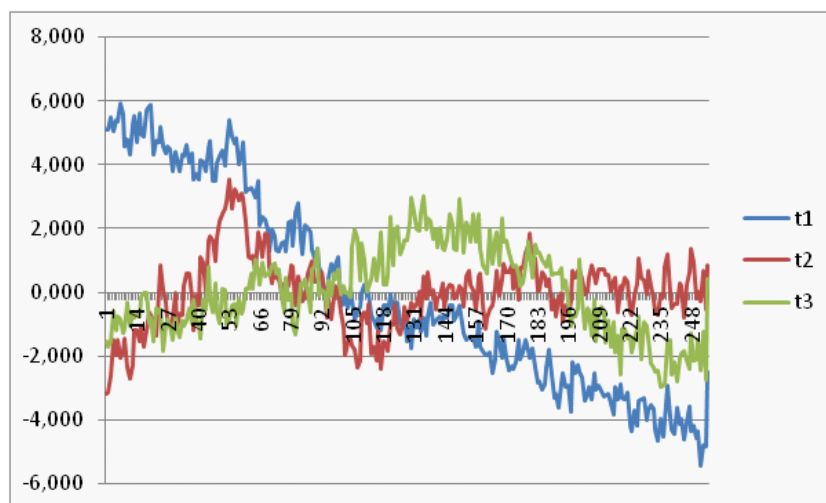


FIG. 2.21: Qualité du modèle par nombre de composantes

FIG. 2.22: Regression PLS t_1, t_2

FIG. 2.23: Carte des vecteurs (w^*, c_1) et (w^*, c_2)

Coordonnées des observations sur t FIG. 2.24: Trajectoires de t_1 , t_2 , t_3

La trajectoire t_1 à une tendance à la baisse

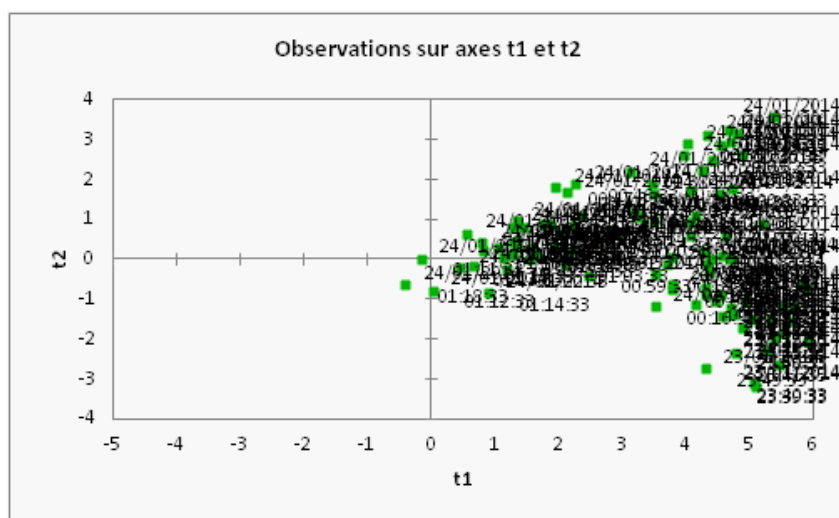
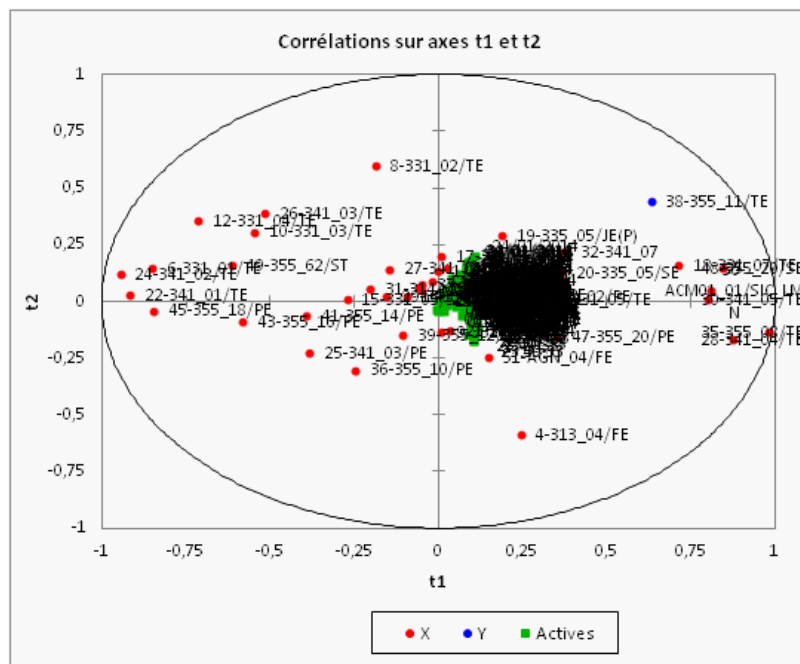


FIG. 2.25: Carte des deux premières composantes PLS (les observations)

FIG. 2.26: Plan des composantes t_1 et t_2 PLS

Coordonnées des observations sur u :

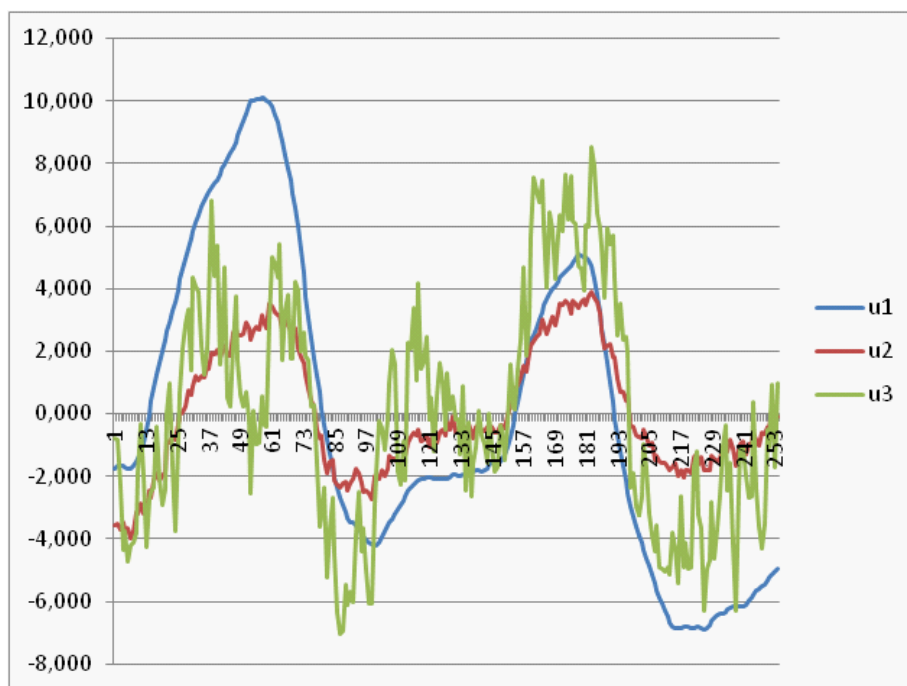


FIG. 2.27: trajectoires de u_1 , u_2 et u_3

En remarque dans le schéma des trajectoires que les courbes sont des série chronologiques

u_1 est une composante périodique

u_2 est une composante saisonnière

u_3 il y a une volatilité

Dans le tableau en dessous on a pris les valeurs extrême en fonction du temps

Le tableau de la courbe de u_1 :

<i>date</i>	<i>heure</i>	<i>valeurs</i>
24/01/2014	23 : 39 : 33	-1,743
24/01/2014	00 : 53 : 33	10,097
24/01/2014	00 : 58 : 33	0,134
24/01/2014	01 : 20 : 33	-4,116
24/01/2014	02 : 19 : 33	-2,203
24/01/2014	02 : 13 : 33	-0,003
24/01/2014	02 : 38 : 33	5,049
24/01/2014	02 : 52 : 33	-0,176
24/01/2014	03 : 15 : 33	-6,861
24/01/2014	03 : 53 : 33	-5,033

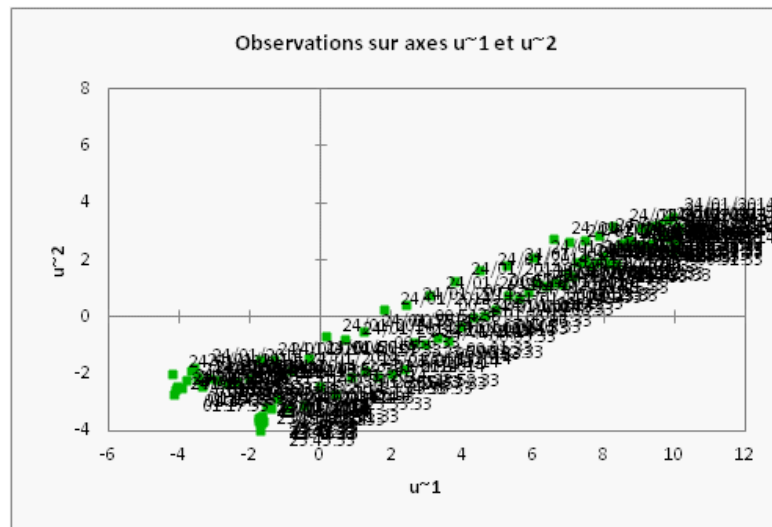


FIG. 2.28: Carte des observations sur $\tilde{u}_1$ et $\tilde{u}_2$

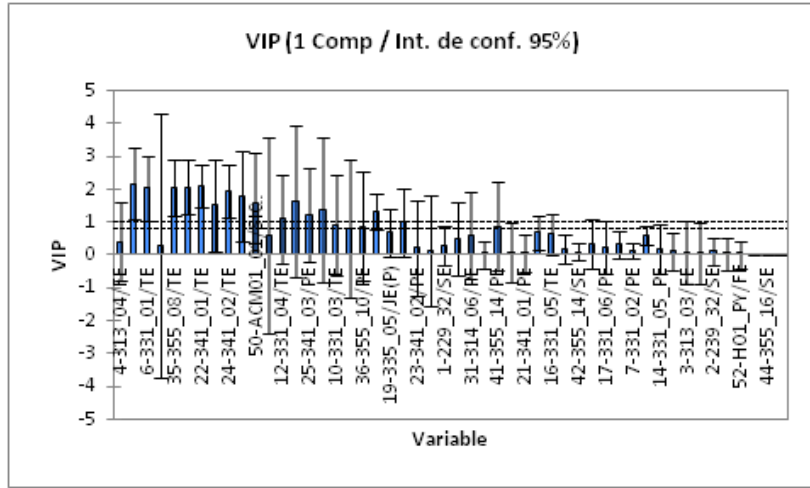


FIG. 2.29: Carte des VIP à 1composante (variableY2)

les VIP (variables importantes) sont : X_{46} , X_6 , X_{34} , X_{43} , X_{22} , X_{24} , X_{30} , X_{18} , X_{48} , X_{28} , X_{47} , X_{41} et X_{25} elles sont les plus importantes dans la construction de Y_2 (elles sont significatives) sur la première composante.

Les variables X_{38} , X_{42} et X_{44} n'ont aucun effet.

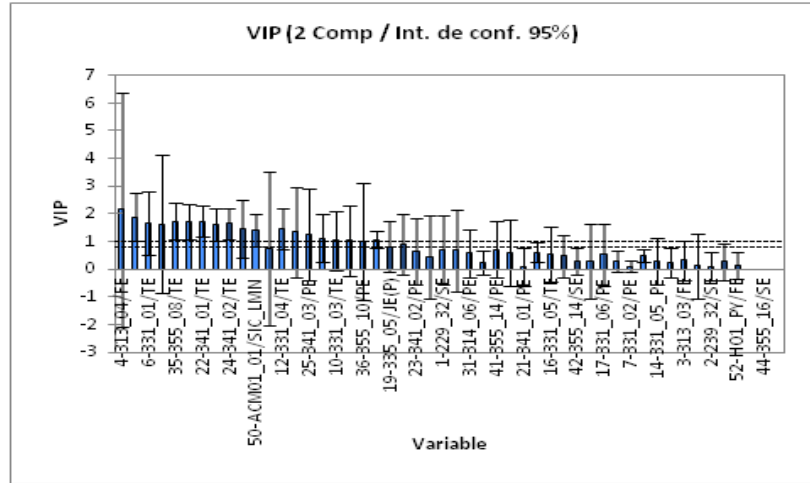
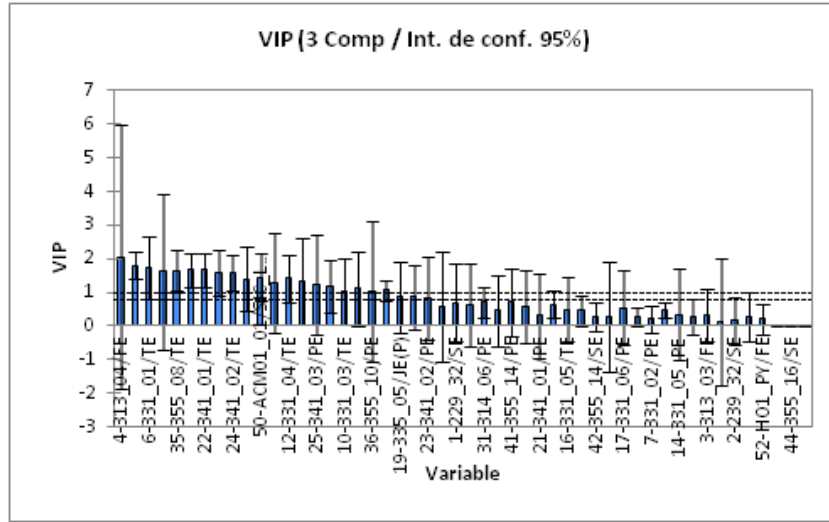


FIG. 2.30: Carte des VIP à 2 composante (variableY2)

En ce qui concerne la deuxième composante on a presque la même chose, les variables (X_4 , X_8), X_{46} , X_6 , X_{34} , X_{43} , X_{22} , X_{24} , X_{30} , X_{18} , X_{48} , X_{28} , X_{47} , X_{41} , X_{25} , (X_{10} , X_{26} , X_{35}), Sont les plus importantes dans la construction de Y_2 , les variables X_{38} , X_{42} , X_{44} on aucun effet.

FIG. 2.31: Carte des VIP à 3 composante (variable Y_2)

Les variables X_4 , X_{48} , X_6 , X_8 , X_{35} , X_{45} , X_{22} , X_{28} , X_{24} , X_{30} , X_{50} , X_{12} , X_{18} , X_{25} ,
Sont les plus importantes dans la construction de Y_2 alors que les variables X_{40} ,
 X_{44} , X_{46} elles ont aucun effet (les variables importante sont dans le tableau des
paramètres du modèle en annexe).

2.2.1 Equation du modèle

$$\begin{aligned}
 Y_2 = & 975.200 + 0.130 * X_1 - 0.022 * X_2 - 1.419 * X_3 - 1.221 * X_4 + 0.014 * X_5 - 0.686 * \\
 & X_6 - 0.001 * X_7 + 1.313 * X_8 - 0.017 * X_9 + 0.532 * X_{10} - 0.049 * X_{11} - 0.016 * X_{12} + 0.001 * \\
 & X_{13} + 0.036 * X_{14} - 0.048 * X_{15} + 0.014 * X_{16} - 0.010 * X_{17} + 0.667 * X_{18} - 0.000 * X_{19} + 0.015 * \\
 & X_{20} + 0.038 * X_{21} - 0.378 * X_{22} + 0.002 * X_{23} - 0.312 * X_{24} - 0.150 * X_{25} - 0.466 * X_{26} - 0.012 * \\
 & X_{27} - 0.024 * X_{28} + 0.105 * X_{29} + 0.231 * X_{30} - 0.232 * X_{31} - 0.201 * X_{32} + 0.630 * X_{33} + 0.025 * \\
 & X_{34} - 0.174 * X_{35} + 0.011 * X_{36} - 0.159 * X_{37} + 0.036 * X_{39} + 0.073 * X_{40} - 0.165 * X_{41} - \\
 & 1.032 * X_{43} + 0.597 * X_{45} + 0.186 * X_{46} - 0.397 * X_{47} - 0.145 * X_{48} + 0.0001 * X_{49} - 0.482 * X_{50}
 \end{aligned}$$

Indice de qualité Q^2 :

<i>Composante</i>	Y_2	<i>Total</i>
<i>Comp1</i>	0,076	0,076
<i>Comp2</i>	-0,571	-0,571
<i>Comp3</i>	-0,427	-0,427
<i>Comp4</i>	-0,393	-0,393
<i>Comp5</i>	-0,570	-0,570
<i>Comp6</i>	-0,595	-0,595
<i>Comp7</i>	-0,314	-0,314
<i>Comp8</i>	-0,652	-0,652
<i>Comp9</i>	-0,681	-0,681
<i>Comp10</i>	-0,510	-0,510

D'après les résultats de la validation croisée en prend la première composante

On ne considérant que les VIP cela donne :

$$Y_2 = 975.200 - 0.686 * X_6 + 0.667 * X_{18} - 0.378 * X_{22} - 0.312 * X_{24} - 0.150 * X_{25} - 0.024 * X_{28} + 0.231 * X_{30} + 0.025 * X_{34} - 0.165 * X_{41} - 1.032 * X_{43} + 0.186 * X_{46} - 0.397 * X_{47} - 0.145 * X_{48}.$$

Coefficients d'ajustement de la variable Y_2

<i>Observation</i>	254,000
<i>Sommedespoids</i>	254,000
<i>DDL</i>	243,000
R^2	0,804
<i>Ecart - type</i>	0,417
<i>MCE</i>	0,167
<i>RMCE</i>	0,408

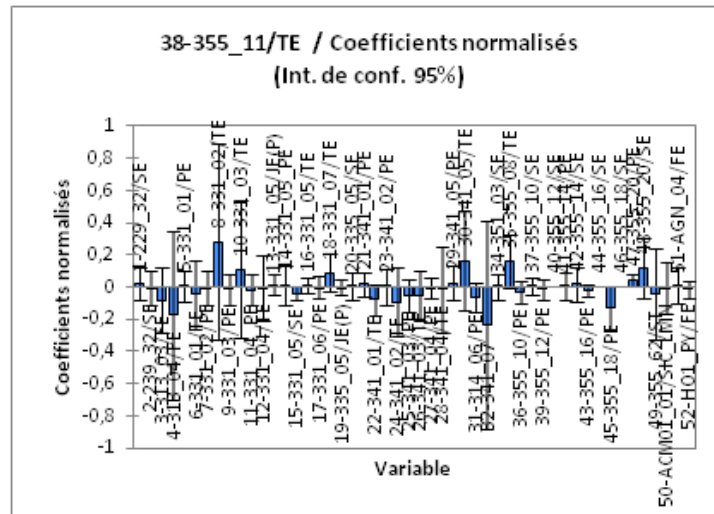
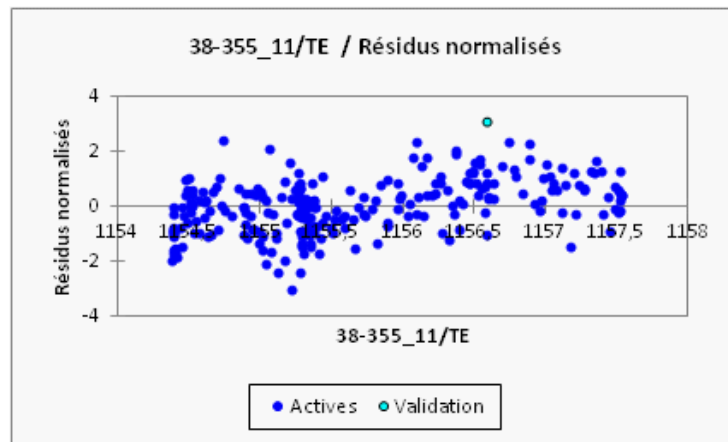
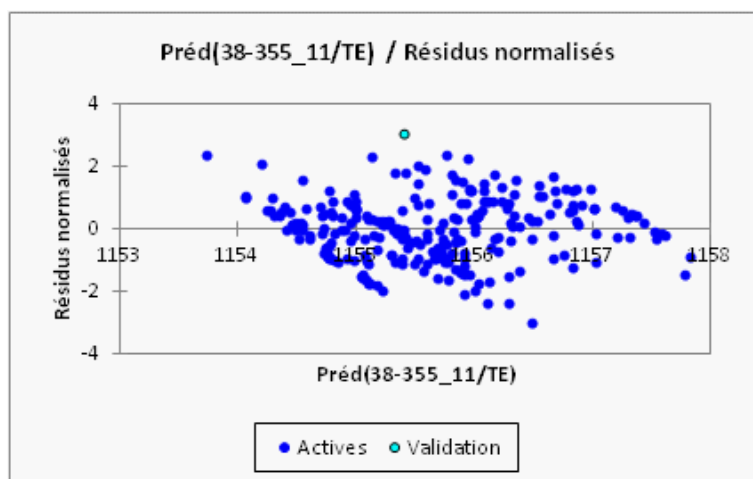
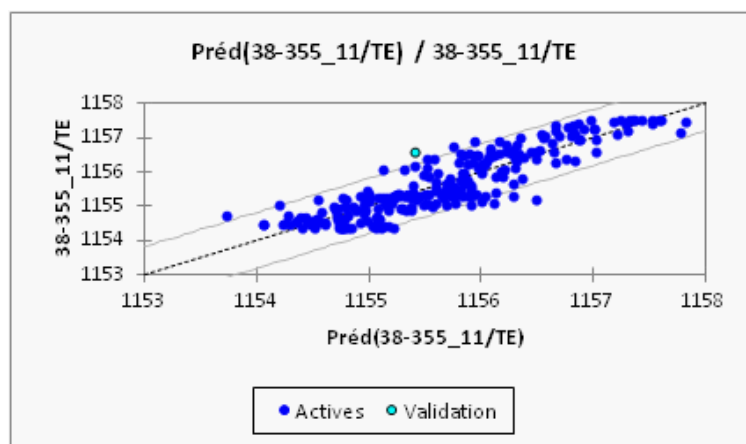
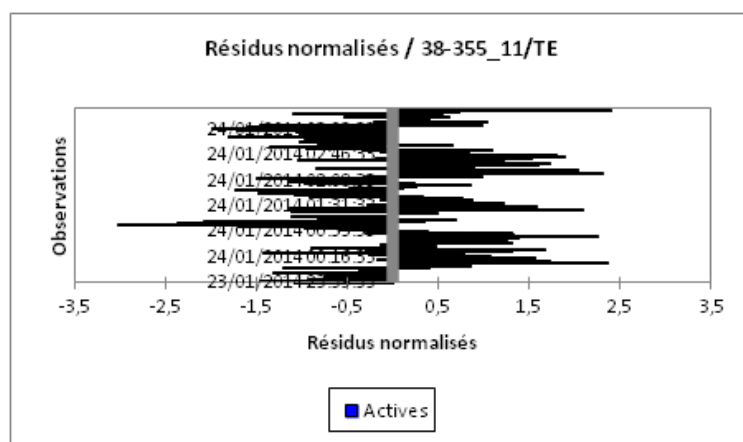
Coefficients normalisés de la variable Y_2 

FIG. 2.32: Coefficients normalisés en fonction des variables

Prédictions et résidus Variable Y_2 FIG. 2.33: Résidus normalisés de la variable Y_2

FIG. 2.34: résidus normalisé et prédiction de la variable Y_2 FIG. 2.35: Prédiction de la variable Y_2 FIG. 2.36: Les observations en fonction des Résidus normalisés de la variable Y_2 Active

<i>date</i>	<i>heure</i>	<i>valeurs</i>
24/01/2014	00 : 09 : 33	2,375
24/01/2014	01 : 48 : 33	2,257
24/01/2014	01 : 03 : 33	-2,406
24/01/2014	01 : 04 : 33	-2,379
24/01/2014	01 : 08 : 33	-2,079
24/01/2014	01 : 26 : 33	2,087
24/01/2014	02 : 19 : 33	2,320
24/01/2014	02 : 26 : 33	2,035
24/01/2014	03 : 54 : 33	2,403
24/01/2014	02 : 40 : 33	3,071

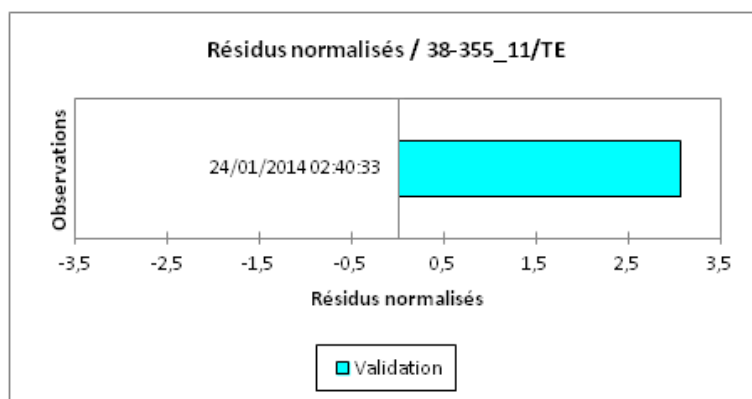


FIG. 2.37: Les observations en fonction des Résidus normalisés de la variable Y2 (38-355_11/TE) Validation

Détection des valeurs extrêmes

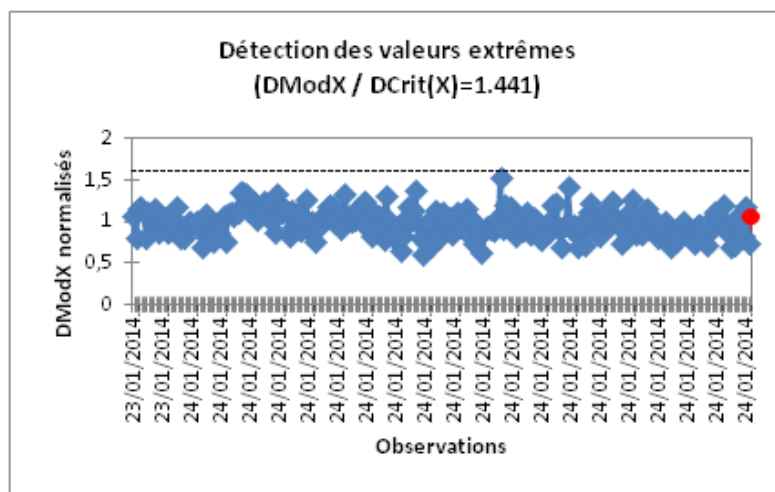


FIG. 2.38: Graphique des DModX,N

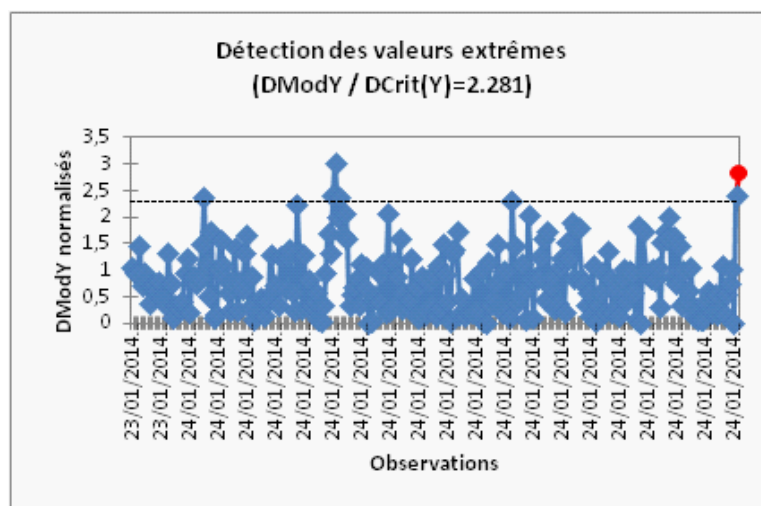


FIG. 2.39: Graphique des DModY,N

3 Application de la régression PLS multiples à 2 variables

33 – 351_01/ JE : (Y_1) : la Somme des puissances en KW , des deux moteurs tournant le four.

38 – 355_11/ TE : (Y_2) : Température en C° de l'air secondaire.

Qualité du modèle par nombre de composantes

Le premier tableau et le graphique correspondant permettent de visualiser la qualité de la régression PLS en fonction du nombre de composantes retenues.

<i>Indice</i>	<i>Comp1</i>	<i>Comp2</i>
Q^2_{cum}	0.029	0.049
R^2Y_{cum}	0.198	0.335
R^2X_{cum}	0.209	0.264

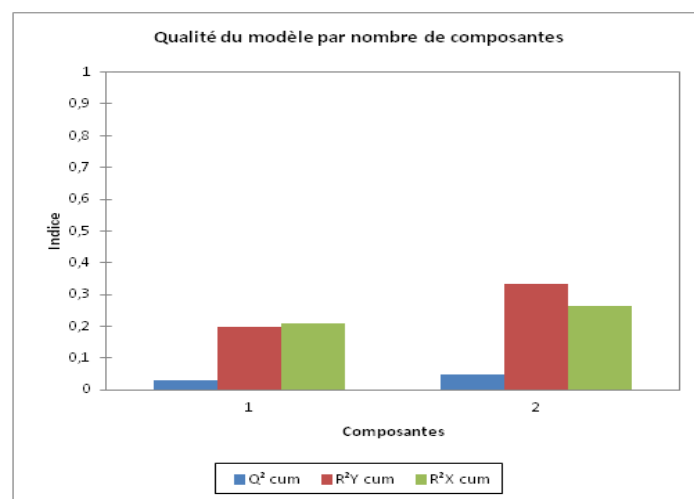


FIG. 3.1: Qualité du modèle par nombre de composantes

On remarque que R^2Y_{cum} prend la plus grande valeur, la 2eme composante contient plus d'information

L'indice Q^2 cumulé est une mesure globale de la qualité de l'ajustement et de la qualité prédictive des modèles. XLSTAT-PLS a automatiquement retenu 2 composantes. On voit que l'indice Q^2 reste faible. Cela suggère que la qualité de l'ajuste-

ment peut être très variable en fonction des variables X à expliquer.

Les R^2Y cum et R^2X cum qui correspondent aux corrélations entre les composantes et les variables de départ sont proches de 0,4 dès la 2^{ème} composante, ce qui indique que les composantes sont à la fois bien représentatives des X et des Y .

Matrice des corrélations :

La matrice des corrélations fournit des éléments de description entre variables, sa lecture donne une première idée sur l'existence de liaisons entre les variables.

Rappelons qu'une variable j est corrélée avec une variable j' si leurs coefficient de corrélation est proche de 1 en valeur absolu d'où :

Le premier graphique des corrélations permet de visualiser sur les deux premières composantes générées par la régression PLS les corrélations entre X et les Y du modèle.

-la variable X_{22} Température en C° des gaz en sortie de la tour échangeur, tour II }, est fortement corrélée, avec la variable X_{24} Température en C° des gaz en sortie du cyclone deux, tour II }.

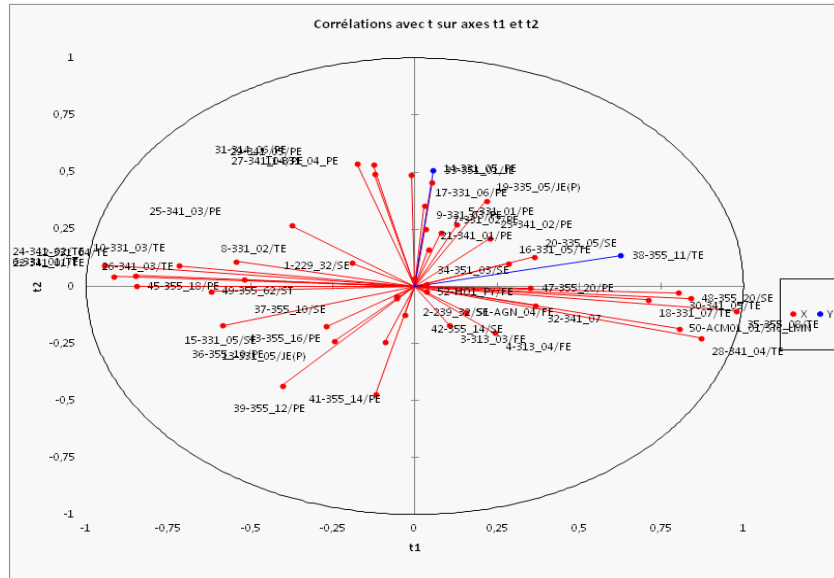


FIG. 3.2: Plan des composantes t_1 et t_2

On note que pour quelques variables au centre du graphique les corrélations sont

très faibles. En se reportant au tableau, on voit par exemple que la variable Y_2 Température en C° de l'air secondaire). n'est liée qu'à la composante t_2 , globalement peu corrélée avec les variables explicatives.

En ce qui concerne les variables explicatives, on note que seule la variable X_4 Débit de la matière première en *tonnes/heure*, entrant dans le four, tour *II*). est mal représentée sur le graphique.

*On remarque qu'il y a 3 groupes de variables ; le premier groupe contient une majorité des variables de X qui sont regroupés au tour du centre de gravité,

*Le 2^{eme} groupe de variables contient les variables ($X_{46}, X_{30}, X_{18}, X_{48}, X_{28}$) sont corrélée positivement entre eux sur l'axe t_2 .

*Le 3^{eme} groupe de variables ($X_{10}, X_{24}, X_{22}, X_{23}, X_{26}, X_{43}, X_{47}, X_{35}, X_{15}$) sont corrélée positivement entre eux Sur l'axe t_2 .

*Les variables du 3^{eme} groupe sont corrélées négativement avec les variables du 2^{eme} groupe.

*En fin on a le 4^{eme} groupe de variables il se trouve sur l'axe t_1 c'est les variables ($X_{31}, X_{29}, X_{27}, X_{11}, \dots, Y_1, X_{17}, X_9, X_5$) sont corrélées entre eux mais ce groupe de variables n'est pas corrélés avec les autres groupes de variables il est indépendant

*la variable Y_2 est liée à la 1^{er} dimension t_1

*la variable Y_1 est liée à la 2^{eme} dimension t_2

*la variable Y_2 est faiblement corrélée à l'ensemble des variables

*la variable Y_1 elle est corrélée à un certain nombre de variables

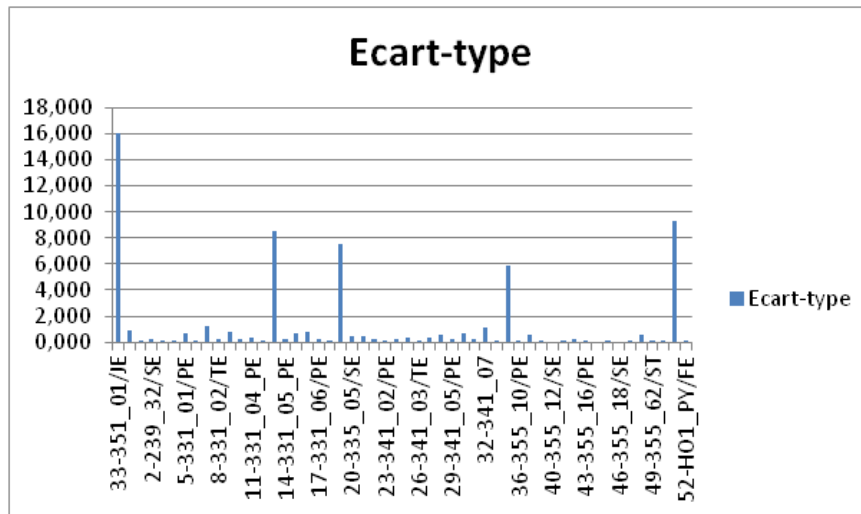
Ecart-type :

FIG. 3.3: Graphique des écarts- types des X

Il y a une grande variabilité des données de X . En remarque que la Somme des puissances en KW , des deux moteurs tournant le four Y_1 ; à le plus grand écart type et qui est égale à 16, équivalent à une variance égale à 256.

Vecteur c :

La carte superposant les variables dépendantes sur les vecteurs c et les variables explicatives sur les vecteurs w^* permet de visualiser la relation globale entre les variables, sachant que les w^* sont représentatifs du poids des variables dans les modèles.

<i>variable</i>	C_1	C_2
Y_1	0,018	0,352
Y_2	0,200	0,094

Variables sur axes w^*/c_1 et w^*/c_2 :

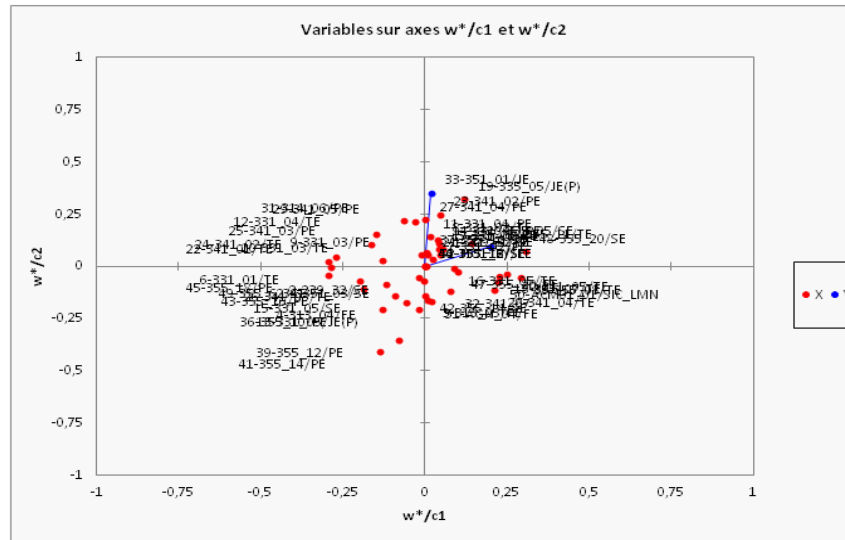


FIG. 3.4: Régression PLS. Carte des variables w^*/c_1 et w^*/c_2

Coordonnées des observations sur t :

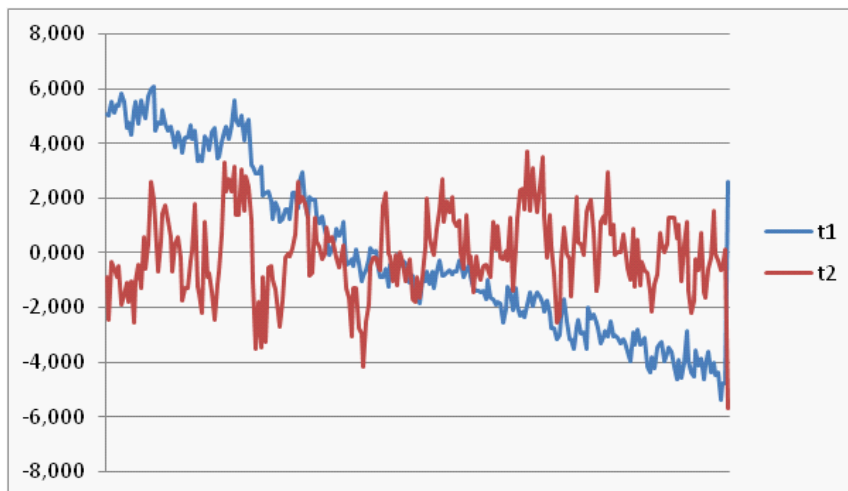
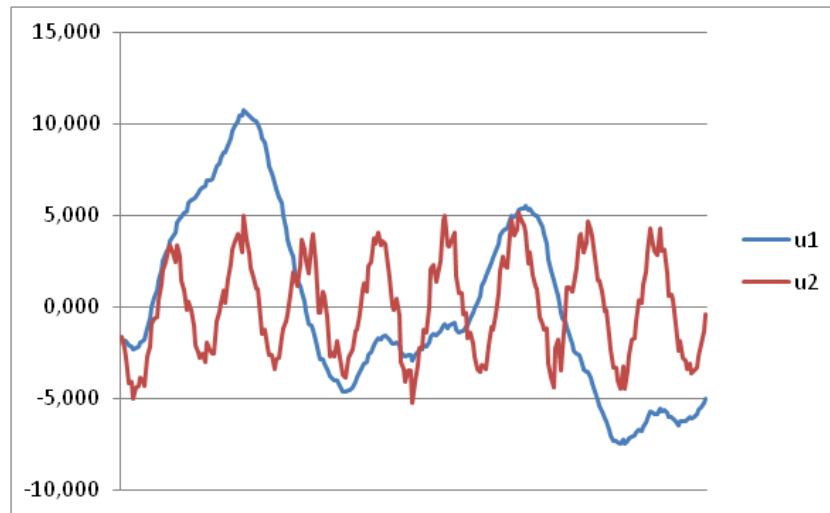


FIG. 3.5: Trajectoire de t_1 et t_2

Pour La trajectoire t_1 il y a une décroissance, pour celle de t_2 il y a des piques donc une variabilité

- La trajectoire de X la plus corrélée est la trajectoire t_1

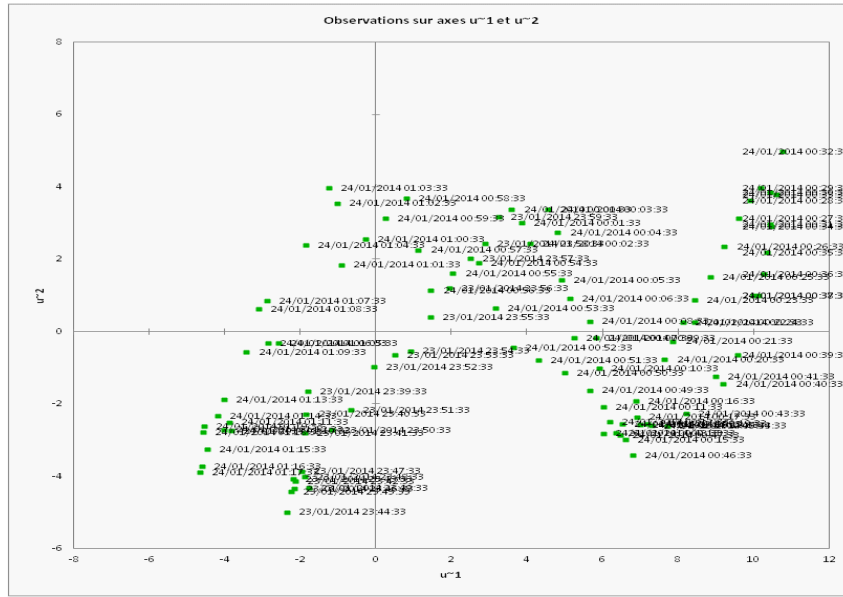
Coordonnées des observations sur u :FIG. 3.6: Trajectoire de u_1 et u_2

Les trajectoires ont un lien linéaire avec Y

Une trajectoire pour X

Une trajectoire pour Y

Un tableau donne les coordonnées des variables sur les composantes t . Le graphique correspondant sur les axes t_1 et t_2 est ensuite affiché. On notera que les variables sont bien distingués.

FIG. 3.7: Observations sur les axes t_1 et t_2

3.0.2 Equation de régression de Y_k sur t_1 et t_2

$$Y_k = C_1 t_1 + C_2 t_2 + \text{résidu}$$

3.0.3 Règles d'interprétation

Les projections des variables X sur Y reflètent le signe et l'ordre de grandeur des coefficients de régression PLS des Y sur X .

3.0.4 Variable importante dans la projection (VIP)

L'étude des Y

Pouvoir explicatif d'une variable X_j sur l'ensemble des Y

(VIP = Variable Importance in the projection)

Dans SIMCA-P le VIP_{hj} est défini par :

$$VIP_{hj} = \sqrt{\left(p / (Rd(Y; t_1, \dots, t_2)) \right) \sum_{l=1}^h Rd(Y; t_l) w_{lj}^2}$$

On a

$$\sum_{j=1}^p VIP_{hj}^2 = p$$

La contribution d'une variable x_j à la construction de la composante t_l est mesurée par le poids w_{lj}^2 . Pour chaque l la somme de ces poids sur l'ensemble des p variables x_j vaut 1. Pour mesurer la contribution de la variable x_j à la construction de Y à travers les composantes t_l , il faut prendre en compte le pouvoir explicatif de la composante t_l , mesuré par la redondance $Rd(Y; t_l)$. Un même poids w_{lj}^2 indique un pouvoir explicatif de la variable x_j sur l'ensemble des variables Y d'autant plus important que la redondance $Rd(Y; t_l)$ est élevée.

D'où la formule du VIP utilisée. Le VIP permet de classer les variables x_j en fonction de leur pouvoir explicatif de Y . Les variables ayant un fort $VIP(> 1)$ sont les plus importantes dans la construction de Y .

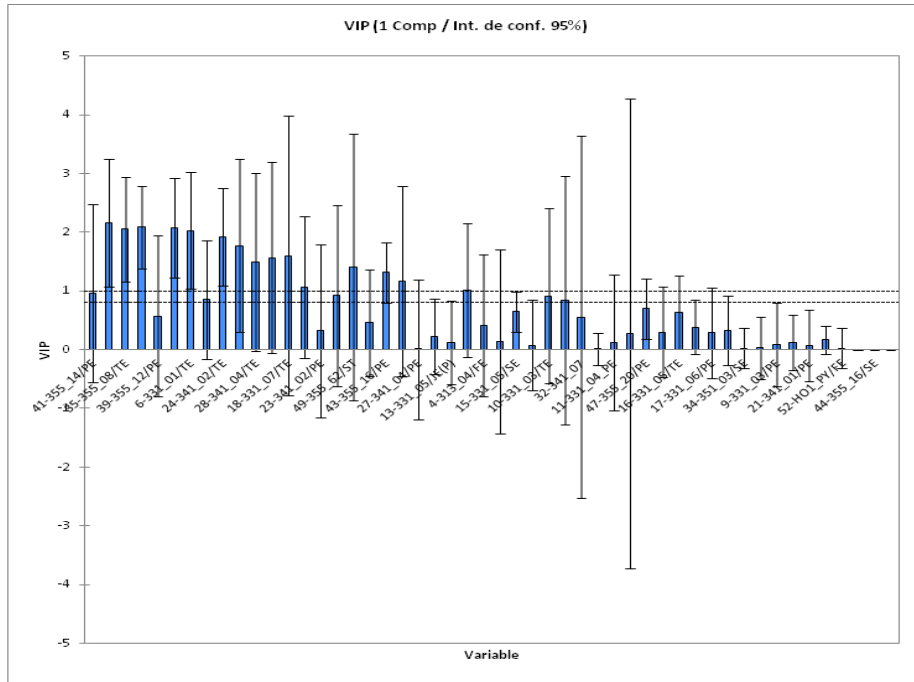


FIG. 3.8: Graphique des VIP (1composante)

les VIP (Variable Importance for the Projection) pour chacune des variables explicatives des modèles à une composante pour la prédiction de Y sont : X_{46} , X_{34} ,

$X_{22}, X_{43}, X_6, X_{24}, X_{30}, X_{28}, X_{48}, X_{18}, X_{12}, X_{47}, X_{41}, X_{25}, X_{20}$.

Sur la première composante les variables, Pression en mbar de l'air sous la grille statique, (refoulement du ventilateur II) X_{37} , Puissance en KW du moteur de la ventilateur exhausteur, tour II X_{19} ; Dépression en mbar des gaz en sortie du cyclone deux, tour II X_{23} ; Dépression des gaz en sortie de la boîte à fumer, tour II X_{31} ; Dépression en mbar des gaz en sortie du cyclone quatre, tour II X_{27} ; Dépression en entrée du cyclone quatre, tour II X_{29} ; Puissance en KW du moteur de la ventilateur exhausteur, tour I X_{13} ; Débit de la matière première en tonnes/heure, entrant dans le four, tour II X_4 ; Débit de combustible en $Nm^3/heure$, dans le brûleur four X_{49} ; Vitesse en $tour/min$ du ventilateur exhausteur, tour I X_{15} ; apparaissent comme étant peu influentes.

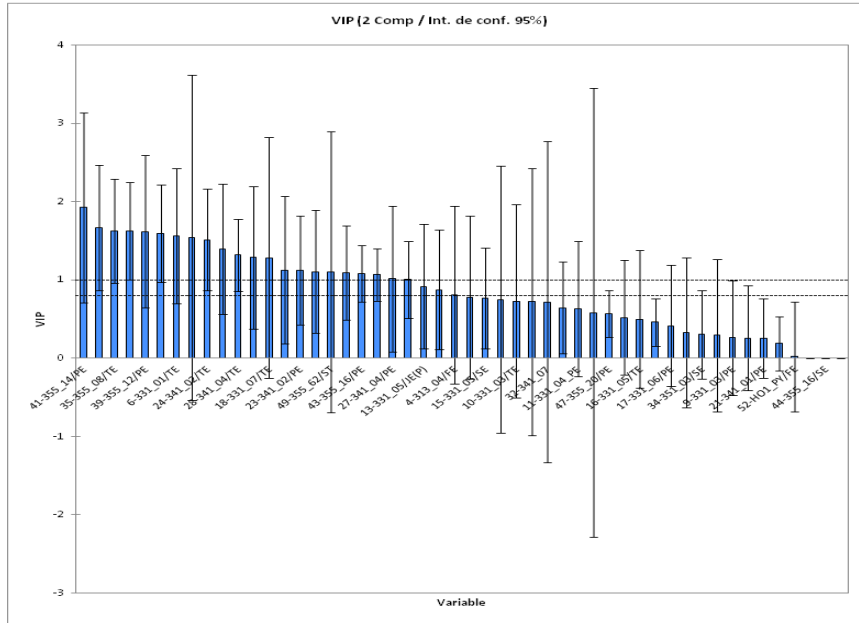


FIG. 3.9: Graphique des VIP (2 composante)

les VIP les variables importante pour chacune des variables explicatives des modèles à deux composante pour la prédiction de Y sont : $X_{46}, X_{34}, X_{22}, X_{39}, X_{37}, X_{43}, X_6, X_{19}, X_{24}, X_{30}, X_{28}, X_{48}, X_{18}, X_{12}, X_{23}, X_{35}, X_{47}, X_{31}, X_{41}, X_{25}, X_{27}, X_{29}$.

3.0.5 Equation du modèle

$$Y_1 = 16416.795 + 2.182 * X_1 - 1.088 * X_2 - 17.605 * X_3 - 7.394 * X_4 + 0.770 * X_5 - 1.872 * X_6 + 0.162 * X_7 + 3.622 * X_8 + 0.363 * X_9 + 0.556 * X_{10} + 2.643 * X_{11} + 6.274 * X_{12} - 0.135 * X_{13} + 2.397 * X_{14} - 1.140 * X_{15} - 0.052 * X_{16} + 2.010 * X_{17} - 1.748 * X_{18} + 0.248 * X_{19} + 1.376 * X_{20} + 0.752 * X_{21} + 0.185 * X_{22} + 8.642 * X_{23} + 0.641 * X_{24} + 1.669 * X_{25} - 4.613 * X_{26} + 3.312 * X_{27} - 0.965 * X_{28} + 5.452 * X_{29} - 0.192 * X_{30} + 4.97 * X_{31} - 0.581 * X_{32} - 149.666 * X_{33} - 0.036 * X_{34} - 6.829 * X_{35} + 0.697 * X_{36} - 22.273 * X_{37} - 30.134 * X_{39} - 3.341 * X_{40} - 4.784 * X_{41} - 2.799 * X_{43} - 1.674 * X_{45} + 0.815 * X_{46} - 4.274 * X_{47} - 6.937 * X_{48} - 0.101 * X_{49} + 0.473 * X_{50}$$

$$Y_2 = 1262.375 + 0.922 * X_1 - 0.027 * X_2 - 0.235 * X_3 - 0.191 * X_4 + 0.27 * X_5 - 0.864 * X_6 + 0.06 * X_7 + 0.091 * X_8 + 0.003 * X_9 - 0.112 * X_{10} + 0.050 * X_{11} - 0.111 * X_{12} - 0.002 * X_{13} + 0.066 * X_{14} - 0.040 * X_{15} + 0.019 * X_{16} + 0.062 * X_{17} + 0.303 * X_{18} + 0.007 * X_{19} + 0.076 * X_{20} + 0.016 * X_{21} - 0.273 * X_{22} + 0.182 * X_{23} - 0.170 * X_{24} - 0.063 * X_{25} - 0.261 * X_{26} + 0.051 * X_{27} + 0.047 * X_{28} + 0.058 * X_{29} + 0.065 * X_{30} + 0.029 * X_{31} + 0.04 * X_{32} - 2.479 * X_{33} + 0.008 * X_{34} - 0.238 * X_{35} + 0.012 * X_{36} - 0.500 * X_{37} - 0.774 * X_{39} - 0.049 * X_{40} - 0.315 * X_{41} - 0.491 * X_{43} + 0.223 * X_{45} + 0.104 * X_{46} - 0.414 * X_{47} + 0.685 * X_{47} - 0.001 * X_{48} + 0.025 * X_{50}$$

Indice de qualité Q^2 :

<i>Composante</i>	Y_1	Y_2	<i>Total</i>
<i>Comp1</i>	-0,006	0,064	0,029
<i>Comp2</i>	0,083	-0,083	0,020
<i>Comp3</i>	-0,052	-0,573	-0,283

En prend la 2^{ème} composante pour Y_1 :

$$Y_1 = 16416.795 - 1.872 * X_6 + 6.274 * X_{12} - 1.748 * X_{18} + 0.248 * X_{19} + 0.185 * X_{22} + 8.642 * X_{23} + 0.641 * X_{24} + 1.669 * X_{25} + 3.312 * X_{27} - 0.965 * X_{28} + 5.452 * X_{29} - 0.192 * X_{30} + 4.97 * X_{31} - 0.036 * X_{34} - 6.829 * X_{35} - 22.273 * X_{37} - 30.134 * X_{39} - 4.784 * X_{41} - 2.799 * X_{43} + 0.815 * X_{46} - 4.274 * X_{47} - 6.937 * X_{48}$$

En prend la 1ère composante pour Y_2 :

$$Y_2 = 1262.375 - 0.864 * X_6 - 0.111 * X_{12} + 0.303 * X_{18} + 0.076 * X_{20} - 0.273 * X_{22} - 0.063 * X_{25} + 0.047 * X_{28} + 0.065 * X_{30} + 0.008 * X_{34} - 0.315 * X_{41} - 0.491 * X_{43} + 0.104 * X_{46} - 0.414 * X_{47} + 0.685 * X_{47} - 0.001 * X_{48}$$

Coefficients d'ajustement de la Variable Y_1 :

<i>Observation</i>	254,000
<i>Sommedespoids</i>	254,000
<i>DDL</i>	251,000
<i>R²</i>	0,260
<i>Ecart – type</i>	13,823
<i>MCE</i>	188,812
<i>RMCE</i>	13,741

Coefficients normalisés de la variable Y_1 :

Nous avons donc réalisé une nouvelle régression *PLS*, en ne sélectionnant qu'une seule variable Somme des puissances en *KW*, des deux moteurs tournant le four Y_1 . Le graphique ci-dessous correspond aux coefficients normalisés pour le modèle avec 2 composantes.

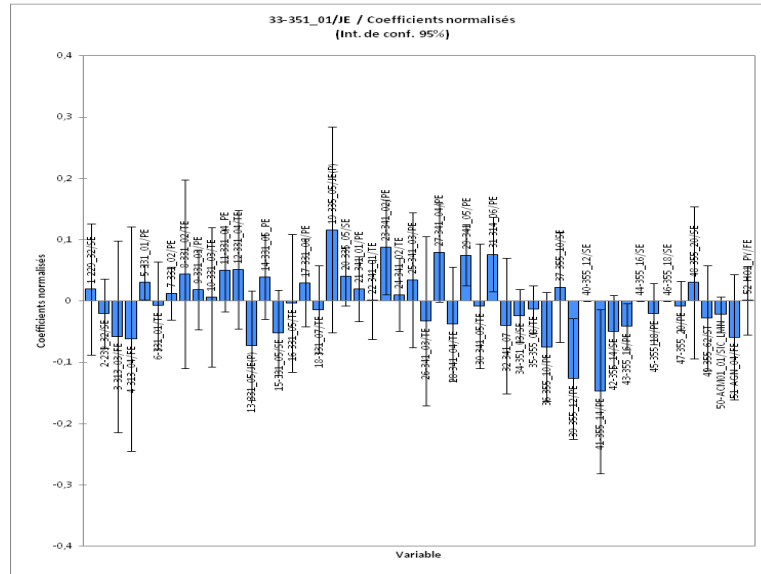


FIG. 3.10: Régression PLS à 2 composantes .coefficients normalisés de régression PLS

On voit ici que seuls les coefficients des variables Pression en mbar de l'air sous la chambre I de la grille dynamique, (refoulement du ventilateur *IV*) X_{39} ; Pression en mbar de l'air sous la grille statique, (refoulement du ventilateur *II*) X_{37} ; Puissance en *KW* du moteur du ventilateur exhausteur, tour *II* X_{19} ; sont significatifs. Le tableau des prédictions et résidus permet de vérifier que les résultats de la Somme des puissances en *KW*, des deux moteurs tournant le four Y_1 sont très bien reproduites par cette dernière.

Prédictions et résidus de la Variable Y_1 :

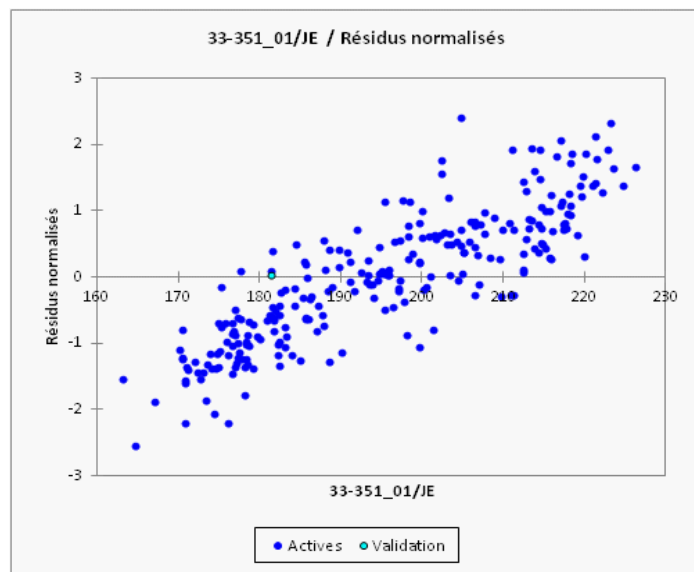


FIG. 3.11: Résidus normalisés de la variable Y_1 (Actives ,Validation)

Le nuage de points à une allure d'une droite

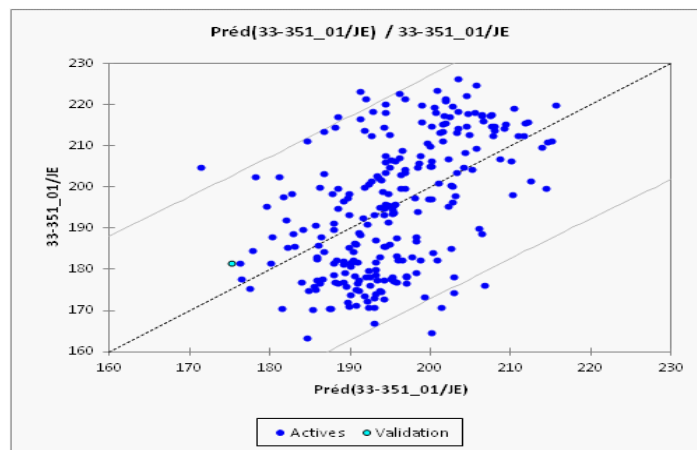


FIG. 3.12: Prédiction de la variable Y_1

Résidus normalisés de la variable Y_1 :

Le premier graphique permet de visualiser les résidus centrés réduits en fonction de la Somme des puissances en KW , des deux moteurs tournant le four Y_1 . Il semble indiquer que les résidus croissent en fonction de la variable Y_1 . L'historique des résidus centrés réduits permet de repérer rapidement et visuellement la présence de valeurs hors de l'intervalle $[-2, 2]$.

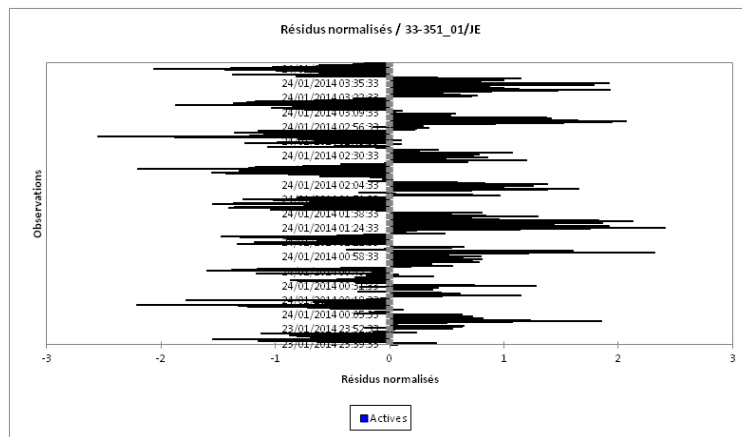
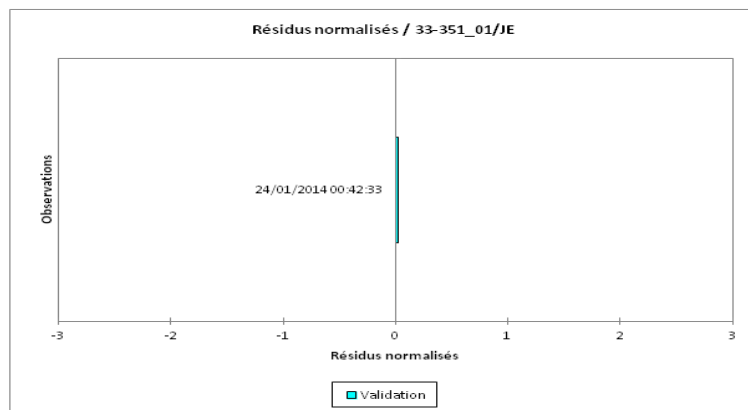


FIG. 3.13: Résidus normalisés de la variable Y_1 (active)

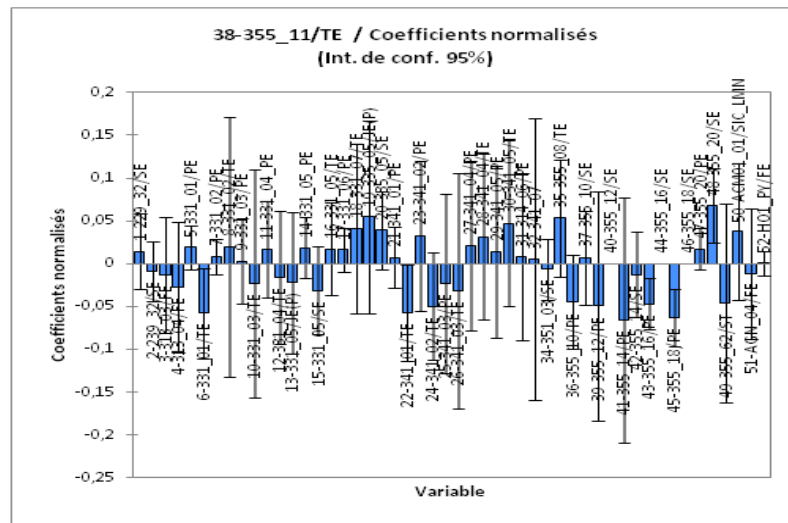
<i>date</i>	<i>heure</i>	<i>valeurs</i>
24/01/2014	00 : 15 : 33	-2,209
24/01/2014	01 : 03 : 33	2,318
24/01/2014	01 : 25 : 33	2,413
24/01/2014	01 : 32 : 33	2,129
24/01/2014	02 : 19 : 33	-2,203
24/01/2014	02 : 48 : 33	-2,556
24/01/2014	03 : 02 : 33	2,071
24/01/2014	03 : 49 : 33	-2,062

FIG. 3.14: Résidus normalisés de la variable Y_1 (validation)

le 24/01/2014 à 00 : 42 : 33 avec une valeur de 0,028

Coefficients d'ajustement de la Variable Y_2

<i>Observation</i>	254,000
<i>Somme des poids</i>	254,000
<i>DDL</i>	251,000
R^2	0,410
<i>Ecart – type</i>	0,709
<i>MCE</i>	0,496
<i>RMCE</i>	0,705

FIG. 3.15: Coefficients normalisés Température en C° de l'air secondaire (Y_2)

Prédictions et résidus normalisés de la variable Y_2 :

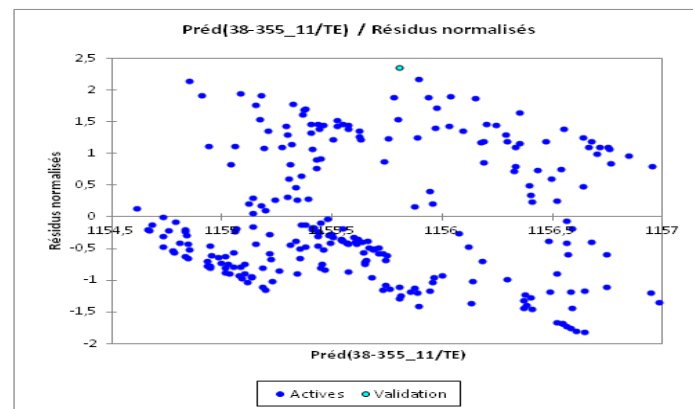
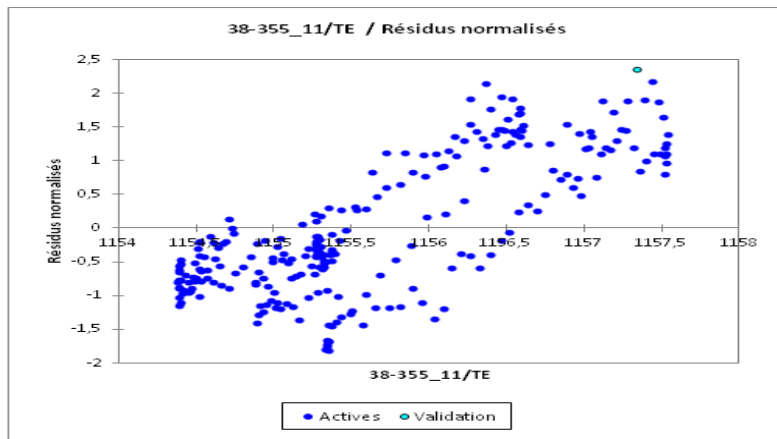
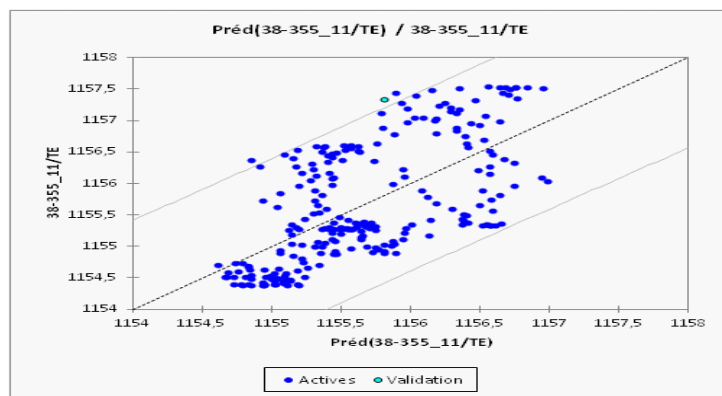


FIG. 3.16: Prédictions et résidus normalisés

FIG. 3.17: Résidus normalisés en fonction de la variables Y_2

Le graphique ci-dessous permet de comparer les prévisions et les valeurs observées. Les courbes de confiance permettent d'identifier les individus atypiques.

FIG. 3.18: Prédiction de la variable Y_2

Résidus normalisés de la variable Y_2 :

L'histogramme des résidus centrés réduits permet de repérer rapidement et visuellement la présence de valeurs hors de l'intervalle $[-2, 2]$.

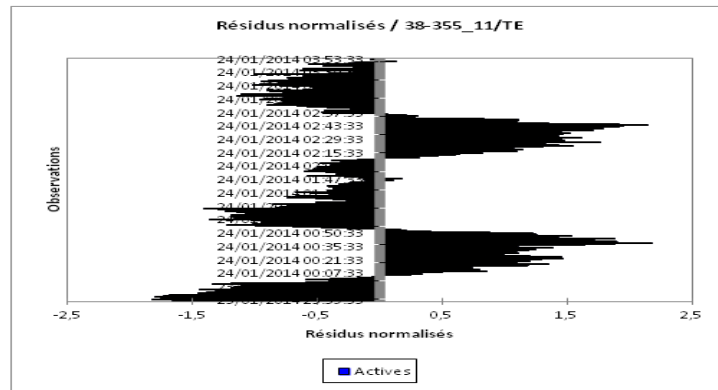
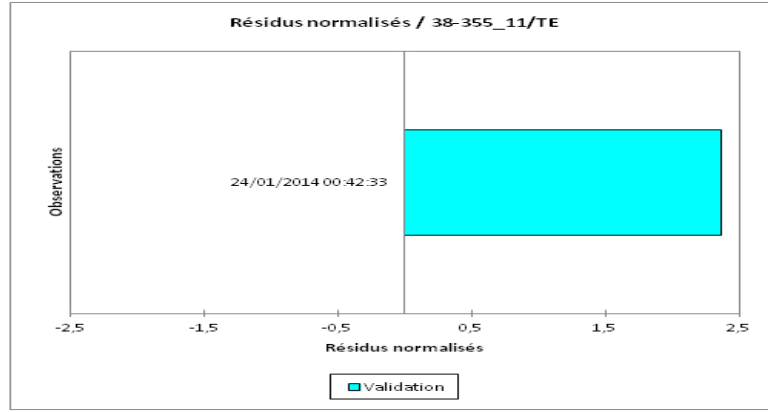


FIG. 3.19: L'histogramme des résidus centrés réduits

On remarque qu'il y a 4 point extrêmes, (le 24/01/2014 à 02 : 45 : 33 ça valeur est 2,183 et 24/01/2014 à 00 : 40 : 33 avec la valeur 2,141) sur le plan positive.

Sur le plan négative on a (le 23/01/2014 à 23 : 41 : 33 avec une valeur de $-1,820$ et le 24/01/2014 à 01 : 17 : 33 avec une valeur de $-1,412$)

<i>date</i>	<i>heure</i>	<i>valeurs</i>
23/01/2014	23 : 41 : 33	$-1,820$
24/01/2014	00 : 40 : 33	2,141
24/01/2014	01 : 17 : 33	$-1,412$
24/01/2014	02 : 45 : 33	2,183

FIG. 3.20: Résidus normalisés en fonction de la variable Y_2

La on remarque qu'il y a une seul date qui surgit qui est le 24/01/2014 à 00 : 42 : 33 elle prend la valeur de 2,364.

Détection des valeurs extrêmes :

$DModX$: distance au modèle de l'observation i dans l'espace des x_j

$DModY$: distance au modèle de l'observation i dans l'espace des y_j

Les valeurs de $DModX$ et $DModY$ supérieures aux valeurs critiques respectives sont affichées en gras ($DCrit(X) = 1.306/DCrit(Y) = 1.939$)

$DModX$:

Qualité de représentaton des données à l'aide des composantes PLS t_h ($DModX, DModY$)

Il s'agit d'étudier les résidus e_{ji} et f_{ki} des régressions de x_j et y_k sur les composantes $t_1, \dots, t_H$. Pour chaque individu i ayant participé à la construction du modèle on note :

$$DModX_i = \sqrt{\left(\sum_{j=1}^p e_{ji}^2\right)/(p-H) \times \sqrt{(n/(n-H-1))}}$$

et

$$DModY_i = \sqrt{\left(\sum_{k=1}^q f_{ki}^2\right)/(q-H)}$$

Ces quantités mesurent la distance au modèle l'individu i dans l'espace des X et des Y . Le terme correctif $\sqrt{(n/(n-H-1))}$ corrige en quelque sorte le trop bon ajustement de la régression de X sur des composantes formées elle-même à partir

des X .

On calcule ensuite des écarts-types de ces résidus :

$$S_X = \sqrt{\left(\left(\sum_{i=1}^n \sum_{j=1}^p e_{ji}^2 \right) / ((n - H - 1) \times (p - H)) \right)}$$

$$S_Y = \sqrt{\left(\left(\sum_{i=1}^n \sum_{j=1}^p f_{ji}^2 \right) / ((n - H - 1) \times (q - H)) \right)}$$

Puis on construit les distances normalisées

$$DModX, N_i = (DModX_i) / S_X$$

$$DModY, N_i = (DModY_i) / S_Y$$

Dans SIMCA-P la distance normalisée $DModX, N_i$ est considérée comme significative si elle dépasse le seuil $DCrit = \sqrt{F_{0.95}(k_1, k_2)}$

Ou les degrés de liberté k_1, k_2 sont définis comme en régression PLS1.

Détection des valeurs extrêmes(DModX / DCrit(X)=1.306) :

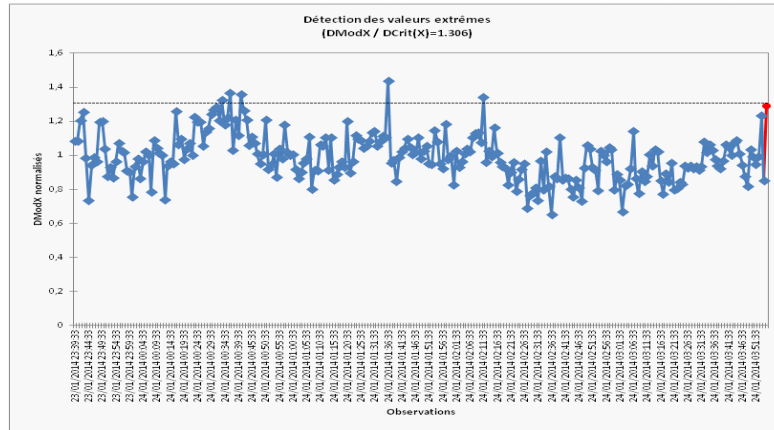


FIG. 3.21: Graphique des DModX,N

Le 24/01/2014 à 01 : 36 : 33 est une valeur extrême elle prend la valeur 1,439

Détection des valeurs extrêmes(DModY / DCrit(Y)=1.939) :

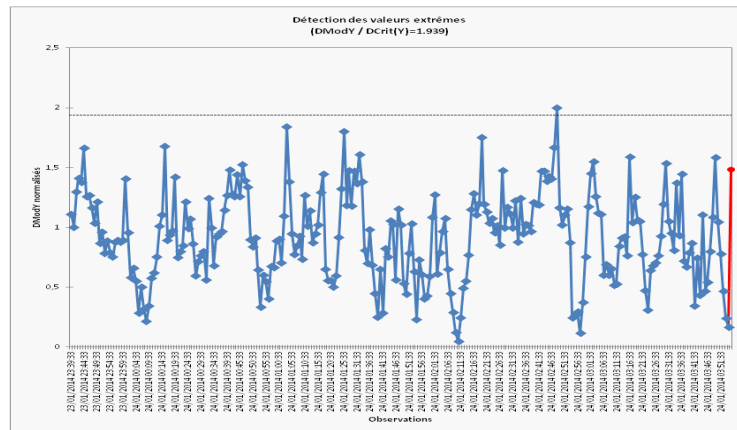


FIG. 3.22: Graphique des DModY,N

Le point ou l'observation 24/01/2014 02 : 48 : 33 qui prend la valeur 2,001 est un point extrême.

3.1 Conclusion

On conclut que la méthode régression PLS est une très bonne méthode de prédiction elle nous a permis d'analyser l'état du four durant son fonctionnement à travers les divers changements de mécanisme de fonctionnement (température, pression ...ect). Cela nous a permis de comprendre pourquoi et comment certains phénomènes tels que la température subissent des changements, et comment cela pose un impact sur la qualité de production qui est le ciment.

Annexes

Annexe A

Déroulement d'un programme avec Matlab sur un exemple de la casicolinéarité :

Programme avec MATLAB Simuler :

```
data =  $\begin{bmatrix} -1 & 5 & 35 & 40 \\ 14 & 18 & 27 & 45 \\ 14 & -22 & -13 & -35 \\ 14 & 8 & 30 & 38 \\ 6 & 3 & 7 & 10 \\ -12 & -13 & 0 & -13 \\ 7 & -4 & 7 & 3 \\ 16 & 3 & -2 & 1 \end{bmatrix}$ 
```

```
%ExtractiondeX
```

```
X = data(:,2:end)
```

```
X = X - ones(size(X,1),1) * mean(X);
```

```
%ExtractiondeY
```

```
Y = data(:,1);
```

```
Y = Y - ones(size(Y,1),1) * mean(Y);
```

```
% Calcul des  $X_{alpha}$  et du determinant et des coefficients de regression
```

```
%  $B_{alpha}$  stockés dans  $BB_{ta}$ 
```

```
BBta = zeros(101,3);
```

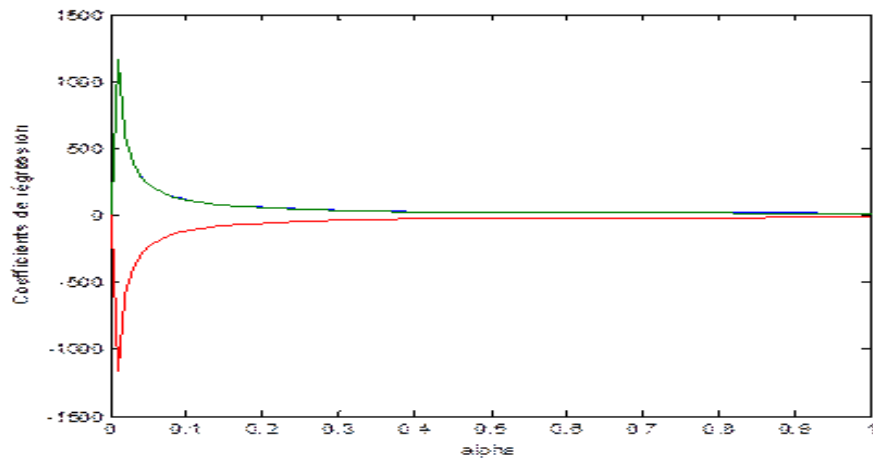
```
i = 0
```

```
for alpha = 0 : 0.01 : 1
```

```

i = i + 1
XX = [zeros(8, 2)[00000000alpha]I];
Xalpha = X + XX;
d(i) = det(Xalpha' * Xalpha);
beta = pinv(Xalpha' * Xalpha) * X' * Y;
BBta(i, :) = beta'
end
figure
plot([0 : 0.01 : 1]I, d)
xlabel('alpha')
ylabel('Matricedecovariance')
figure
plot([0 : 0.01 : 1]I, BBta)
xlabel('alpha')
ylabel('Coefficientsderégression')

```



1 Programmes

Programel :

```
function [VX, PX, CX, WY, UY] = PLS2_Hdimension(X, Y, H0, tol)
```

```
[nx, px] = size(X);
```

```
[ny, py] = size(Y);
```

```
Xh = X;
```

```
Yh = Y;
```

```
H0 = min([H0rank(X)]);
```

```
VX = zeros(px, H0);
```

```
CX = zeros(nx, H0);
```

```
WY = zeros(py, H0);
```

```
UY = zeros(ny, H0);
```

```
for h = 1 : H0
```

```
[vxh, cxh, wyh, uyh] = PLS2_Onedimension(Xh, Yh, tol);
```

```
VX(:, h) = vx(h);
```

```
CX(:, h) = cxh;
```

```
WY(:, h) = wyh;
```

```
UY(:, h) = uyh;
```

```
pxh = Xh' * cxh / (cxh' * cxh);
```

```
PX(:, h) = pxh;
```

```
Xh = Xh - cxh * pxh';
```

```
Yh = Yh - cxh * wyh';
```

```
end
```

Programe2 :

```
function[vx, cx, wy, uy] = PLS2_OneDimension(X, Y, tol)
```

```
[n, p] = size(X);
```

```
%1.Initialisatio
```

```
vx0 = rand(p, 1);
```

```
%2.Iteration
```

```
delta = 1;
```

```
iter = 0;
```

```
while(delta > tol)
```

```
    iter = iter + 1;
```

```
    [vx1] = PLS2_Oneiteration(X, Y, vx0);
```

```
    delta = norm(vx1/norm(vx1) - vx0/norm(vx0));
```

```
    vx0 = vx1;
```

```
    disp(['totalnumberofiterations' num2str(iter)]);
```

```
end
```

```
vx = vx1;
```

```
cx = X * vx / (vx' * vx);
```

```
wy = Y' * cx / (cx' * cx);
```

```
Wy = wy / norm(wy)
```

```
uy = Y * wy / (wy' * wy);
```

Programme3 :

```
function[vx1] = PLS2_Oneiteration(X,Y,vx0)
%vx0 est le vecteur d'initialition(Xloadings)
%X est la matrice des variables explications
%Y est la matrice des variables à expliquer
% =====
%2.Iteration
vx = vx0/norm(vx0);
cx = X * vx / (vx' * vx);
wy = Y' * cx / (cx' * cx);
wy = wy/norm(wy)
uy = Y * wy / (wy' * wy);
vx1 = X' * uy / (uy' * uy);
vx1 = vx1/norm(vx1);
```

Programme 4 :

```
function[B] = XLOADINGS_STAR(A,W)
maxdim = size(A,2);
p = size(A,1);
B = zeros(p,maxdim);
B(:,1) = A(:,1);
for dim = 2 : maxdim
%Transformation = eye(p) - Ah(:,1 : dim - 1) * Wh(:,1 : dim - 1)'
%b(h) = transformation * a(h)
B(:,dim) = A(:,dim) - A(:,1 : dim - 1) * (W(:,1 : dim - 1)' * A(:,dim));
end
```

Annexe B

1 Application de la régression PLS pour diagnostiquer l'état d'un four

1.1 Introduction

Pour diagnostiquer l'état d'un four cette étude vise à Analyser un ensemble de signaux qui proviennent de la cimenterie d'Ain-El-Kabira à l'est algérien. 52 signaux supposée pertinent ont était collecté durant un fonctionnement normal de l'installation. Une dizaine parmi eux peuvent être ajustés par les opérateurs de la salle de contrôle pour conserver un état de fonctionnement sain du four, les valeurs que peuvent prendre le reste des signaux sera conjointement le résultat de ces réglages effectué sur les consignes et les conditions actuelle dans les quelles le four travail.

Il est important de diagnostiquer à partir de ces signaux si le fonctionnement est normale ou pas. Il est d'une grande utilité pour le staff de l'unité de production à analyser et comprendre les relations et les corrélations cachées et apparentes qui peuvent existées entre ses variables du procédé, en particulier durant le fonctionnement normal et comment ils seront influées lors d'un fonctionnement instable ou en défaut.

Détection d'anomalie à partir des signaux :

Selon l'expérience des cimentiers, deux signaux d'intérêt, en particulier parmi eux sont d'une grande utilité à la surveillance de l'état réel du four. Ils sont riches

en information et leur interprétation correcte peut facilement expliquer et conclure sur l'état du fonctionnement du four, s'il est sain, perturbé ou carrément en défaut.

Malgré que les signaux collectés sont difficile à analyser Des études statistiques exploratoires de ces signaux sont alors nécessaires en vue de découvrir les secrets de ces données et par la suite du procédé lui-même et aider les opérateurs conduisant le four à mieux réagir à des situations de défaut vers les quelles le procédé peu diverger.

Des techniques statistiques telle que l'Analyse en Composantes principales (ACP) peuvent fournir des réponses quant aux dépendances et relations entres variables contenues dans le groupe de données.

En outre la compréhension du fonctionnement du procédé, un autre objective de l'étude peut être la prédiction des deux fameux signaux informatifs, en utilisant le reste des signaux et en construisant un modèle mathématique par un processus d'apprentissage convenable de l'état de fonctionnement normal du four. Le modèle peut servir par la suite à comparer l'état future du four observé à travers les deux signaux en provenance de capteurs physiques installés sur le four et la sortie du modèle, les deux signaux physiques et les deux prédits doivent converger l'un vers l'autre à tout instant, tant que le procédé est en fonctionnement normale. Chaque déviation importante entre l'une des prédictions et sa mesure peut être considérée comme synonyme d'un défaut naissant. A la lumière de cette détection, les opérateurs de la salle de contrôle peuvent anticiper des actions correctives en agissant sur les variables de consignes adéquates.

Un autre intérêt peut être aussi l'estimation à chaque instant des deux signaux en se basant sur les autres signaux au même instant et éventuellement leurs valeurs aux instants passées de manière à construire ce qui est communément connus comme capteur logiciel (soft sensor) et avoir par conséquence une redondance de ces deux mesures. En cas d'un dysfonctionnement sur le capteur physique (panne), les opérateurs de la salle de contrôle peuvent s'appuyer sur sa valeur estimée pour continuer à conduire le four sans interruptions, les agents de maintenance aussi auront un temps

suffisant pour la maintenance du capteur défectueux.

2 Contexte de l'étude

2.1 Introduction

Une cimenterie est constituée de plusieurs ateliers, depuis le gisement de la matière première jusqu'à l'expédition du produit fini. La matière première est en majorité du calcaire. Il est transporté de la carrière vers une première installation de concassage, où il est réduit sous forme d'agréats, de petite taille. Puis il est transporté vers un hall de stockage et de pré homogénéisation. Dans ce hall les autres matières secondaires tel que la marne, le gypse et les minerais de fer sont aussi stockés après avoir subi à leur tour une même opération de concassage.

Le deuxième atelier étant la zone crue où un produit intermédiaire nécessaire à la fabrication des ciments est produit, dit la farine. Ce dernier produit est préparé par le broyage sous une température moyenne de $90\text{ }^{\circ}\text{C}$ du calcaire, marne et minerais de fer, avec des proportions adéquates. La farine est stockée dans des silos d'homogénéisation et stockage, ces derniers assurent un bon mélange de cette matière et servent comme alimentateurs pour le four rotatif dans l'atelier suivant.

Le troisième atelier est le cœur d'une cimenterie, la zone de cuisson, ici, la farine obtenue précédemment est alimentée vers le four rotatif par le biais des doseurs pondéraux, après avoir traversé une zone de préchauffage ou de pré calcination selon l'installation existante. Le four rotatif est un grand dispositif cylindrique (quelques dizaines de mètres de longueur et quelques mètres de diamètre), protégé intérieurement par une couche de briques réfractaires. Dans le four, la température de matière s'élève à environ $1450\text{ }^{\circ}\text{C}$, et celle des gazes de 2000 à $3000\text{ }^{\circ}\text{C}$ juste à proximité du brûleur principale. Dans le four la matière passe par plusieurs phases et subit plusieurs réactions physicochimiques, pour obtenir en finalité un produit dit clinker. Celle-ci est refroidie brusquement (trompe) dans un équipement à la sortie du four dit refroidisseur. Le clinker est stocké à son tour dans des silos qui servent à l'alimentation de l'atelier suivant, proprement dit zone ciment.

Le quatrième atelier est celui où le proprement et communément appelé ciment est fabriqué, par le broyage du produit issu du four, i.e clinker, du gypse et quelques ajouts tel que pouzzolane ou laitier. Les ajouts sont nécessaires pour régler les caractéristiques mécaniques, selon le type du ciment, le gypse joue le rôle de régulateur de prise, lors de l'utilisation du ciment. Il est important de signaler que la finesse (de la matière crue et du ciment) lors du broyage est régulièrement contrôlé. Après le broyage le ciment est stocké à son tour dans des silos.

Le cinquième et dernier atelier est celui de l'expédition où le ciment est conditionné par des machines ensacheuses et envoyé vers des camions pour la vente. Deux formes son disponibles le ciment ensaché et vrac.

Dans la présente étude tout notre intérêt sera porté sur l'étude détaillée de la troisième zone qui est la cuisson, quand c'est nécessaire des éclaircissements sur les zones en amont seront donnés au moment opportun, mais pas pour les deux dernières zones, du moment où notre but est le contrôle de qualité du clinker qui doit répondre aux normes de qualité et dont la qualité du produit final (ciment) en découle.

2.2 Atelier cuisson

Dans cette zone le clinker est produit depuis la matière crue sous forme de farine ou patte. Elle est soumise à une haute température et traverse plusieurs étapes et changement dans les phases. L'équipement principale est le four, il tourne à une faible vitesse et placé a une position légèrement inclinée, en amont du four un équipement pour le séchage et la déshydratation de la matière, et en aval un autre pour la trompe et le refroidissement du clinker produit, les trois équipements travail en étroite relation. Du point de vue dynamique cet atelier est caractérisé par son fort couplage, retard important et des processus d'échange de chaleurs non linéaires.

2.2.1 Tour échangeur

L'échangeur suspendu est composé de plusieurs cyclones disposés l'un en dessus de l'autre en constituant une tour, la matière crue est acheminée vers le sommet de celle ci, ou elle est déversée par des quantités maîtrisables juste en dessous du dernier cyclone.

Les cyclones servent principalement à filtrer la matière (poussière) des gaz montants. Ils sont liés entre eux par deux canaux, qui sont la gaine gaz, reliant le centre supérieur du cyclone de l'étage en bas au coté supérieur de l'étage en haut, les gaz chaud transmettent leurs chaleur à 80% dans cette gaine, la deuxième, sert à récupérer la matière gagné par le filtre et la transmettre vers le cyclone en bas, plus précisément, elle est injecté dans la partie, la plus basse de la première gaine, pour augmenter le temps d'échange entre les contres courants, des gaz chauds en provenance du four qui sont tirés vers le haut par les ventilateurs exhausteurs, d'un coté et la matière descendante vers le bas par sa propre gravitation.

Une petite partie de l'échange de chaleur est faite dans les cyclones eux même. L'excès de chaleur est utilisé dans les opérations de séchage et de broyage cru. La déshydratation complète et un début de décarbonatation est entamé dans cet échangeur. Pour augmenter le rendement du four par transformation de la longueur utilisée initialement à la décarbonatation en une zone de cuisson (gain) ; une bonne partie de la décarbonatation devrait se faire dans le préchauffeur lui même, c'est dans ce sens que venait l'idée d'installer un bruleur secondaire aux premiers étages de l'échauffeur, dit précalcinateur . La productivité du four sera de ce fait augmentée. Cette technique a révolutionné le procédé des ciments, de nos jours, il est devenu normale de consommer jusqu'à 60% du total de combustible dans le précalcinateur.

Pour la combustion ce dernier utilise l'oxygène acheminé par l'air secondaire qui passe à travers le four, mais parfois aussi dans une conduite spéciale dite d'aire tertiaire, acheminé hors du cylindre four, cette dernière technique est préférable en particulier si la présence de cycles d'alkalin et chlorides est inévitable, dans ce cas un

tirage de ces nuisibles en état gazeux est possible sans perte de chaleur. La figure 6.1 présente une tour échangeur en suspension.

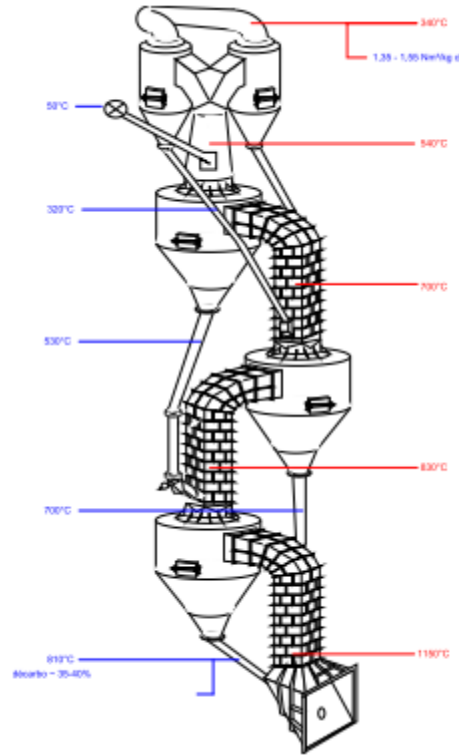


FIG. 2.1: Tour échangeur type Humboldt

2.2.2 Four rotatif

Le four rotatif est la pièce maitresse d'une unité de fabrication des ciments, là où on produit le clinker, matière principale constituant les ciments. Ce four est constitué d'un cylindre métallique creux, dont les dimensions sont parfois colossaux, à l'unité de Ain-El-Kebira (Est de l'Algérie) sa longueur est de $80m$ et son diamètres est de $5.4m$, le volume du four est directement liés à sa capacité de production de clinker par jour.

Deux principaux types de four existent, type long (longueur = 32 à 35 fois le diamètre du four) et type court(longueur = 15 à 17 fois le diamètre du four), le four en question donc appartient à la deuxième catégorie, qui est conservée aux

voies semi-sèche ou sèche. La figure 6.2 montre un four rotatif avec un profil de changement de température à titre indicatif.

La production de clinker nécessite un milieu oxydant avec une température de matière près de 1450°C . Des phénomènes physicochimiques extrêmement compliqués apparaissent durant cette étape, tel que la déshydratation des minéraux contenus dans l'argile, la décarbonatation des carbonates (dégagement du dioxyde de carbone : calcination), des réactions solides, d'autres introduisant la phase liquide (les fondants) et des processus de cristallisation. Ces phénomènes ne sont pas affectés uniquement par la composition chimique de la matière première mais aussi par sa composition minéralogique et d'autres facteurs physiques.

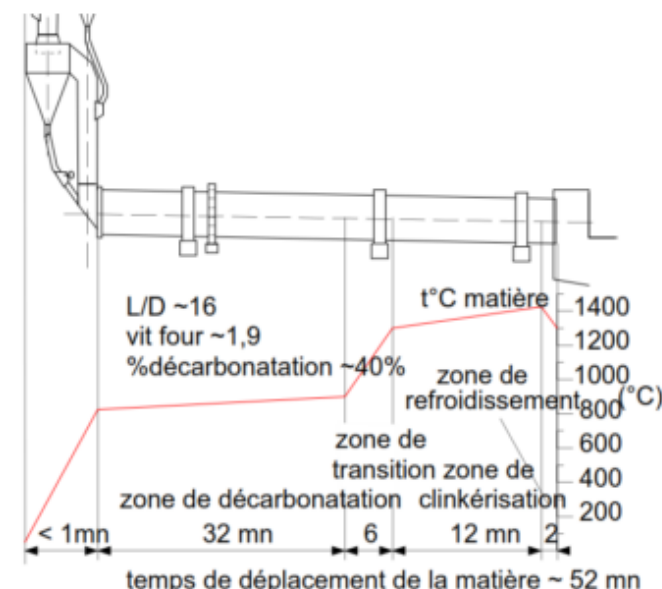


FIG. 2.2: Four et Profil deTempératures

2.2.3 Refroidisseur

Le refroidisseur à clinker est l'équipement installé en aval du four. Bien qu'il existe plusieurs technologies, le cas étudié est de type refroidisseur à grilles, c'est une grosse enceinte séparée horizontalement en deux parties par une grille, celle-ci est constituée de deux parties une statique, inclinée d'une certaine pente pour

permettre le mouvement du clinker, la deuxième est horizontale et dynamique, actionnée en mouvements de tamisage périodique. Le clinker passe sur cette grille. Sa partie inférieure est constituée de plusieurs chambres dont chacune souffle au moins un ventilateur l'air afin de refroidir le clinker. Après que le clinker subisse un léger refroidissement à la sortie du four (zone de refroidissement), celui-ci chute directement dans le refroidisseur où il subira une trompe rapide par l'air. Alors une vitrification du clinker est réalisée pour maintenir les phases déjà existantes en lui. Le flux d'air permet aussi la récupération des calories du clinker refroidie qui seront acheminées par le tirage vers le four et à travers ce dernier vers le préchauffeur pour l'air secondaire, et via un conduit séparé pour l'air tertiaire, en cas d'une tour échangeur dotée d'un précalcinateur. La figure 6.3 montre un profile de température lors de refroidissement du clinker.

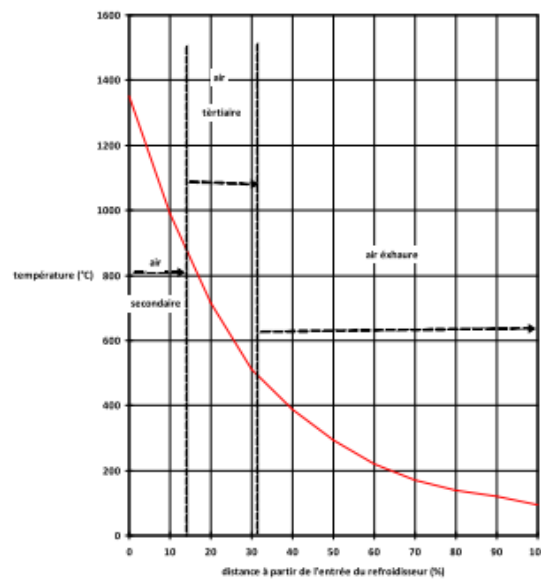


FIG. 2.3: Profile de température dans le refroidisseur

Cette courbe représente la chute dans la température de clinker en fonction de la longueur du refroidisseur en %. A partir de la sortie du four jusqu'à 15% de sa longueur se situe la zone de l'air secondaire, de là jusqu'à 32% vient la zone de l'aire tertiaire, le reste c'est pour l'air d'exhausteur qui sera tiré par les ventila-

teurs de tirage avale du procédé. En remarque que dans les deux premières zones la température de matière chute de 70%.

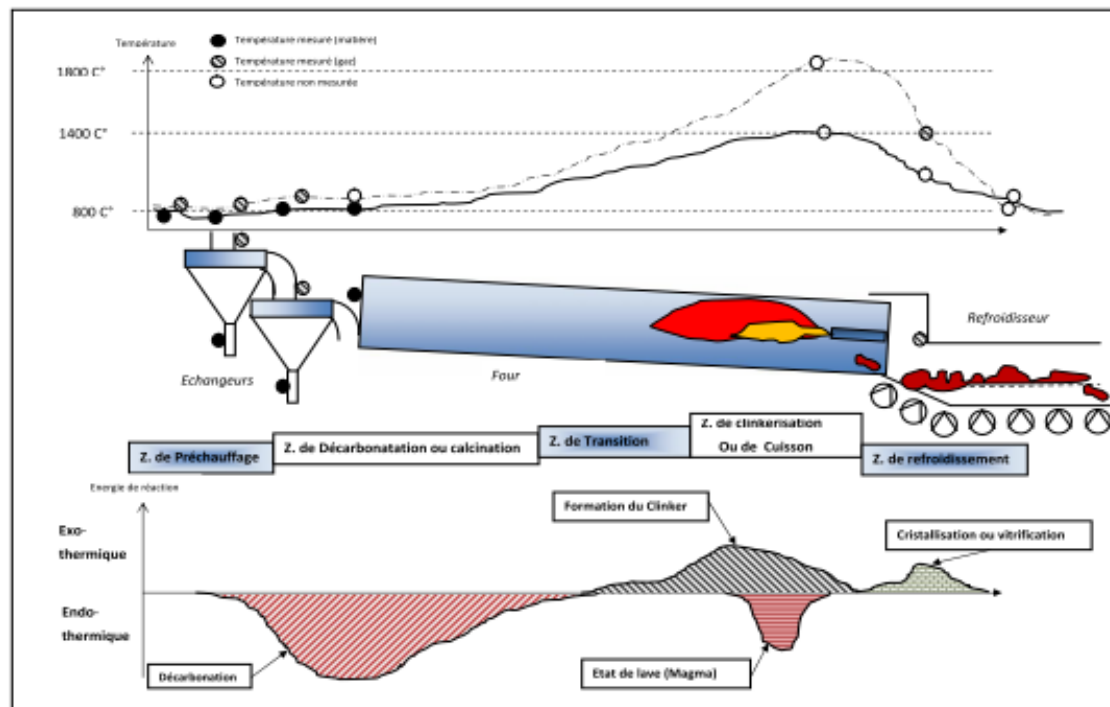


FIG. 2.4: Profile de température, zones et phases d'après K.S. Stadler [1]

2.3 Choix des signaux utilisés

La figure 6.5 montre un schématisation globale d'un four rotatif de cimenterie avec deux tours échangeurs, dont chaque signal utilisé est repéré par un numéro1, pour son identification. La liste en annexe explique la signification de chacun des signaux utilisés.

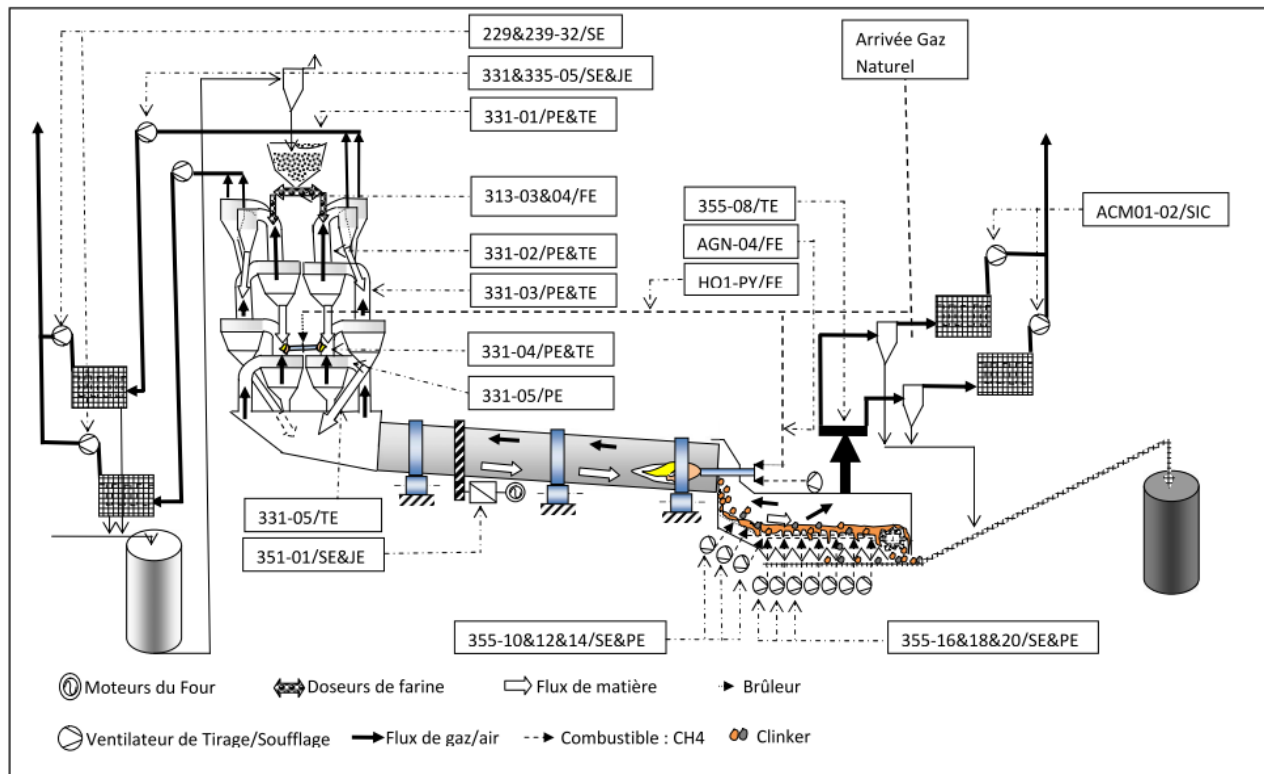


FIG. 2.5: Tour-Four-Refroidisseur

3 Problématique de la conduite d'un four rotatif de cimenterie

Le four sujet de cette étude est actuellement conduit d'une manière manuelle centralisée, à partir d'un centre de commandement (salle de contrôle) où toute l'information des capteurs installés sur le procédé est disponible et toutes les commandes vers la partie opératoire peuvent être envoyées, des opérateurs humains estiment à

chaque instant ces commandes en se basant sur les valeurs de mesures observées et leurs propres expériences.

Les signaux observés sont regardés séparément, un seuillage de type Shewhart est appliqué sur chaque signal seul pour en déduire s'il est en état normal (entre les deux seuils maximal et minimal). Une alarme est émise par le système de supervision en cas de dépassement de l'un des deux seuils pour le signal. Ce type de supervision est dit classique. Le nombre de signaux est très grand pour qu'il soit suivi par un seul opérateur, en réalité une centaine de signaux de mesures sont émis par l'équipement four seul, une cinquantaine sont des mesures reflétant directement l'état de son fonctionnement, dont chaque divergence peut avoir une conséquence néfaste sur la qualité du produit final du four et par la suite un coût économique négatif sur la trésorerie de l'entreprise. Les autres sont des mesures nécessaires pour sa propre sécurité tel que les températures des parties tournantes, de graissage, des huiles et autres.

Un développement de cette supervision est donc plus que nécessaire pour lui confier un aspect à la fois intelligent et ergonomique. Au lieu que chaque signal soit observé seul, il est préférable de résumer l'essentiel de ces signaux et de le fournir à l'opérateur sous une forme simple et efficace. Des techniques statistiques multi-variées ont cette capacité. L'opérateur ne sera plus contrarié par le grand nombre d'alarmes émises par chaque signal en cas de situation anormale, mais une seule alarme ou deux uniquement peuvent résumer la situation. Les opérateurs non encore qualifiés peuvent être énormément aidés par l'implémentation d'un aspect intelligent dans la supervision actuelle, qui peut comparer l'état du four normal appris auparavant avec chaque future situation du four et en déduire d'une manière automatique si celui-ci est normal ou non.

Annexe C

1 Identification des variables

1.1 Le tableau X : variables explicatives

X_1 : 1 – 229_32/SE : Vitesse en *tour/min* du ventilateur de tirage finale, tour *I*.

X_2 : 2 – 239_32/SE : Vitesse en *tour/min* du ventilateur de tirage finale, tour *II*.

X_3 : 3 – 313_03/FE : Débit de la matière première en *tonnes/heure*, entrant dans le four, tour *I*.

X_4 : 4 – 313_04/FE : Débit de la matière première en *tonnes/heure*, entrant dans le four, tour *II*.

X_5 : 5 – 331_01/PE : Dépression en mbar des gaz en sortie de la tour échangeur (sortie cyclone un), tour *I*.

X_6 : 6 – 331_01/TE : Température en C° des gaz en sortie de la tour échangeur (sortie cyclone un), tour *I*.

X_7 : 7 – 331_02/PE : Dépression en mbar des gaz en sortie du cyclone deux, tour *I*.

X_8 : 8 – 331_02/TE : Température en C° des gaz en sortie du cyclone deux, tour *I*.

X_9 : 9 – 331_03/PE : Dépression en mbar des gaz en sortie du cyclone trois, tour *I*.

X_{10} : 10 – 331_03/TE : Température en C° des gaz en sortie du cyclone trois, tour I.

X_{11} : 11 – 331_04_PE : Dépression en mbar des gaz en sortie du cyclone quatre, tour I.

X_{12} : 12 – 331_04/TE : Température en C° des gaz en sortie du cyclone quatre, tour I.

X_{13} : 13 – 331_05/JE(P) : Puissance en KW du moteur du ventilateur exhauster, tour I.

X_{14} : 14 – 331_05_PE : Dépression en mbar des gaz en entrée du cyclone quatre, tour I.

X_{15} : 15 – 331_05/SE : Vitesse en tour/min du ventilateur exhauster, tour I.

X_{16} : 16 – 331_05/TE : Température de la matière en entrée four, tour I.

X_{17} : 17 – 331_06/PE : Dépression en mbar, dans les gaz de sortie boîte à fumer, tour I.

X_{18} : 18 – 331_07/TE : Température des gaz en sortie boîte à fumer, tour I.

X_{19} : 19 – 335_05/JE(P) : Puissance en KW du moteur de la ventilateur exhauster, tour II.

X_{20} : 20 – 335_05/SE : Vitesse en tour/min du ventilateur exhauster, tour II.

X_{21} : 21 – 341_01/PE : Dépression en mbar des gaz en sortie de la tour échangeur, tour II.

X_{22} : 22 – 341_01/TE : Température en C° des gaz en sortie de la tour échangeur, tour II.

X_{23} : 23 – 341_02/PE : Dépression en mbar des gaz en sortie du cyclone deux, tour II.

X_{24} : 24 – 341_02/TE : Température en C° des gaz en sortie du cyclone deux, tour II.

X_{25} : 25 – 341_03/PE : Dépression en mbar des gaz en sortie du cyclone trois, tour II.

X_{26} : 26 – 341_03/TE : Température en C° des gaz en sortie du cyclone trois, tour II.

X_{27} : 27 – 341_04/PE : Dépression en mbar des gaz en sortie du cyclone quatre, tour II.

X_{28} : 28 – 341_04/TE : Température en C° des gaz en sortie du cyclone quatre, tour II.

X_{29} : 29 – 341_05/PE : Dépression en entrée du cyclone quatre, tour II.

X_{30} : 30 – 341_05/TE : Température de la matière en entrée four, tour II.

X_{31} : 31 – 314_06/PE : Dépression des gaz en sortie de la boîte à fumer, tour II.

X_{32} : 32 – 341_07 : Température des gaz en sortie de la boîte à fumer, tour II.

X_{33} : 34 – 351_03/SE : Vitesse de rotation du four en *tour/min*.

X_{34} : 35 – 355_08/TE : Température en C° d'air excédentaire du refroidisseur.

X_{35} : 36 – 355_10/PE : Pression en mbar de l'air sous la grille statique, (refoulement du ventilateur I).

X_{36} : 37 – 355_10/SE : Vitesse en *tour/min* du ventilateur de refroidissement, VI.

X_{37} : 39 – 355_12/PE : Pression en mbar de l'air sous la grille statique, (refoulement du ventilateur II).

X_{38} : 40 – 355_12/SE : Vitesse en *tour/min* du ventilateur de refroidissement, VII.

X_{39} : 41 – 355_14/PE : Pression en mbar de l'air sous la grille statique (refoulement du Ventilateur III).

X_{40} : 42 – 355_14/SE : Vitesse en *tour/min* du ventilateur de refroidissement, VIII.

X_{41} : 43 – 355_16/PE : Pression en mbar de l'air sous la chambre I de la grille dynamique, (refoulement du ventilateur IV).

X_{42} : 44 – 355_16/SE : Vitesse en *tour/min*, du ventilateur IV.

X_{43} : 45 – 355_18/PE : Pression en mbar de l'air sous la chambre II de la grille statique.

X_{44} : 46 – 355_18/SE : Vitesse en *tour/min* du ventilateur de refroidissement V.

X_{45} : 47 – 355_20/PE : Pression en mbar de l'air sous la chambre III, de la grille dynamique, (refoulemen du ventilateur de reforidissement VI).

X_{46} : 48 – 355_20/SE : Vitesse en *tour/min* du ventilateur de refroidissement VI.

X_{47} : 49 – 355_62/ST : Vitesse, en *coups/min* de la grille dynamique.

X_{48} : 50 – ACM01_02/SIC : Commande issue du régulateur de pression du capot de chauffe, pour les vitesses des ventilateurs de tirage du filtre refroidisseur, en *tour/min*.

X_{49} : 51 – AGN_04/FE : Débit de combustible en *Nm3/heure*, dans le brûleur four.

X_{50} : 52 – HO1_PY/FE : Débit de combustible en *m3/heure*, dans le brûleur de pré-calcination.

1.2 Le tableau Y : variables à expliquer

Y_1 : 33–351_01/JE : Somme des puissances en *KW*, des deux moteurs tournant le four.

Y_2 : 38 – 355_11/TE : Température en *C°* de l'air secondaire.

Avant de faire l'application de la régression PLS multiples nous avons commencé par faire une ACP.

2 Les sorties de l' ACP

2.1 Matrice de corrélation de l'ACP

	1	2	3	4	5	6
1	1					
2	0,004	1				
3	0,130	0,059	1			
4	0,002	-0,068	-0,184	1		
5	-0,001	0,025	-0,049	0,124	1	
6	-0,007	0,051	-0,121	-0,387	-0,063	1
7	-0,015	0,041	-0,016	0,085	0,645	0,004
8	0,027	-0,017	-0,136	-0,221	-0,023	0,334
9	0,029	-0,019	-0,089	0,089	0,527	-0,007
10	0,093	0,002	-0,051	-0,113	-0,053	0,536
11	-0,021	0,075	-0,006	0,052	0,448	-0,014
12	-0,068	0,043	-0,137	-0,413	-0,050	0,722
13	0,018	-0,006	0,037	0,023	0,025	0,072
14	-0,050	0,064	-0,015	0,030	0,342	-0,060
15	-0,020	0,051	0,019	-0,050	-0,157	0,192
16	0,005	-0,060	0,011	0,132	-0,025	-0,257
17	-0,098	0,070	0,012	0,016	0,113	-0,050
18	0,027	-0,082	0,066	0,155	0,042	-0,488
19	0,065	0,000	-0,075	-0,148	0,040	-0,052
20	-0,055	0,047	-0,056	0,052	0,143	-0,248
21	0,025	-0,132	0,027	0,060	0,049	-0,112
22	0,039	0,100	-0,078	-0,237	-0,079	0,796
23	0,051	-0,101	-0,019	-0,014	-0,066	-0,039
24	0,058	0,061	-0,079	-0,310	-0,077	0,801

25	0,013	-0,033	-0,025	-0,001	0,025	0,283
26	0,050	0,032	-0,125	-0,186	-0,073	0,535
27	-0,052	-0,110	-0,025	-0,057	0,048	0,113
28	-0,050	-0,026	0,169	0,195	0,063	-0,705
29	0,022	-0,064	-0,007	0,048	0,101	0,070
30	0,091	-0,088	0,121	0,295	0,099	-0,674
31	0,021	0,001	-0,049	-0,009	0,104	0,142
32	-0,002	-0,076	-0,054	0,144	-0,001	-0,166
34	-0,101	0,051	0,092	-0,079	-0,012	-0,009
35	-0,021	-0,060	0,109	0,304	0,092	-0,858
36	0,139	0,042	0,322	-0,133	-0,124	0,173
37	-0,130	0,070	-0,061	-0,089	-0,104	0,056
39	-0,054	-0,031	-0,122	0,057	-0,060	0,196
41	-0,062	-0,028	-0,160	-0,006	-0,135	0,352
42	0,057	0,088	0,054	0,022	0,102	-0,003
43	0,006	0,031	-0,085	-0,058	-0,007	0,461
45	-0,038	0,022	-0,109	-0,149	-0,063	0,726
47	-0,117	-0,056	0,013	0,118	0,116	-0,355
48	-0,013	-0,047	0,060	0,186	0,063	-0,670
49	0,036	-0,022	-0,062	-0,205	-0,113	0,641
50	-0,067	-0,047	0,039	0,085	0,068	-0,546
51	-0,039	-0,007	0,046	0,033	-0,088	-0,108
52	-0,005	-0,029	0,104	0,003	-0,044	-0,003
7	8	9	10	11	12	13

1						
0,012	1					
0,499	-0,109	1				
-0,003	0,716	-0,061	1			
0,396	-0,062	0,532	0,078	1		
-0,014	0,314	-0,086	0,276	-0,070	1	
-0,057	0,095	0,020	0,100	0,040	0,095	1
0,368	-0,056	0,428	0,023	0,844	-0,059	-0,022
0,005	0,107	-0,043	0,137	-0,065	0,223	0,109
-0,022	-0,071	0,013	-0,072	0,011	-0,299	-0,115
0,160	-0,040	0,214	0,049	0,473	-0,037	0,074
0,058	0,410	-0,014	0,091	-0,075	-0,536	0,012
0,050	0,048	-0,053	-0,011	0,129	-0,031	-0,118
0,083	-0,107	0,009	-0,110	0,021	-0,265	-0,091
-0,003	0,084	0,163	0,085	0,175	-0,045	0,087
-0,026	0,156	-0,010	0,485	0,088	0,645	0,049
-0,029	0,363	-0,116	0,225	-0,080	-0,156	-0,044
-0,028	0,200	-0,006	0,560	0,063	0,683	0,065
0,010	0,160	0,214	0,352	0,178	0,122	0,095
0,046	0,728	-0,039	0,822	-0,005	0,342	0,119
0,025	0,156	0,090	0,253	0,169	0,057	0,044
0,015	-0,261	-0,034	-0,602	-0,142	-0,611	-0,004
0,073	0,129	0,180	0,285	0,369	-0,001	0,053
0,074	0,123	0,023	-0,080	0,008	-0,745	-0,035
0,063	0,133	0,159	0,303	0,401	0,037	0,005
0,045	0,689	-0,045	0,416	-0,104	-0,285	0,077
-0,054	-0,018	0,005	0,017	-0,027	-0,018	0,070
0,039	-0,282	0,020	-0,620	-0,066	-0,723	-0,076

-0,093	-0,044	-0,044	0,246	-0,054	-0,125	0,095
-0,054	0,065	-0,002	-0,020	-0,045	0,169	0,010
-0,022	0,173	-0,055	0,071	-0,163	0,218	0,077
-0,094	0,174	-0,121	0,264	-0,107	0,354	0,086
-0,008	0,002	0,049	-0,017	0,031	0,026	0,010
-0,085	0,053	-0,045	0,215	-0,083	0,484	0,086
-0,097	0,268	-0,016	0,606	0,019	0,538	0,050
-0,031	-0,163	0,050	-0,253	-0,020	-0,241	-0,005
0,039	0,162	-0,013	-0,255	-0,136	-0,547	-0,021
-0,032	0,602	-0,047	0,758	-0,085	0,425	0,177
0,029	-0,013	-0,051	-0,430	-0,166	-0,411	-0,011
-0,063	-0,039	0,073	0,093	0,124	-0,340	-0,056
-0,069	0,044	0,007	-0,030	-0,038	0,033	0,029
14	15	16	17	18	19	20

1						
-0,060	1					
0,018	-0,170	1				
0,694	-0,068	0,069	1			
-0,034	-0,177	0,192	-0,030	1		
0,149	-0,128	0,104	0,116	0,149	1	
0,069	-0,133	0,060	0,079	0,215	-0,015	1
0,084	0,046	0,026	-0,026	0,175	0,090	-0,268
0,048	0,197	-0,267	0,014	-0,674	-0,177	-0,295
-0,122	-0,053	0,097	-0,141	0,527	0,255	0,078
0,017	0,215	-0,230	0,030	-0,693	-0,104	-0,344

0,046	0,064	-0,059	0,039	-0,102	-0,021	-0,124
-0,009	0,141	-0,131	0,022	0,081	-0,081	-0,183
0,103	-0,006	0,096	0,155	0,042	0,090	0,097
-0,060	-0,226	0,085	-0,102	0,590	0,118	0,281
0,299	0,012	,067	0,248	0,056	0,068	0,006
0,042	-0,229	0,288	0,036	0,870	0,174	0,211
0,351	-0,049	0,084	0,292	-0,004	0,090	-0,047
-0,080	-0,043	0,093	-0,073	0,879	0,053	0,062
0,013	0,054	-0,047	0,091	0,025	-0,215	0,259
-0,006	-0,236	0,240	-0,026	0,666	0,144	0,311
-0,103	0,002	-0,030	-0,015	-0,035	-0,012	-0,236
-0,027	0,035	-0,055	0,029	-0,019	-0,063	-0,022
-0,144	0,001	-0,088	-0,148	0,088	0,034	-0,134
-0,093	0,169	-0,116	-0,093	-0,164	0,012	-0,110
-0,010	-0,080	-0,024	-0,006	-0,046	-0,174	-0,018
-0,110	0,134	-0,105	-0,099	-0,453	-0,128	-0,196
-0,054	0,138	-0,235	-0,045	-0,467	-0,216	-0,254
-0,037	-0,145	0,071	-0,083	0,236	-0,010	0,154
-0,084	-0,168	0,216	-0,075	0,817	0,131	0,299
-0,113	0,150	-0,269	-0,090	0,015	-0,169	-0,155
-0,066	-0,208	0,115	-0,058	0,695	0,166	0,284
0,079	0,018	0,091	0,027	0,281	0,111	0,155
-0,053	0,047	-0,050	-0,038	0,025	-0,040	0,018
21	22	23	24	25	26	27

-0,101	1					
0,169	-0,246	1				
-0,047	0,889	-0,215	1			
0,430	0,282	0,205	0,289	1		
0,075	0,474	0,147	0,536	0,218	1	
0,160	0,090	0,208	0,095	0,388	0,164	1
0,022	-0,771	0,152	-0,845	-0,377	-0,521	-0,206
0,201	0,078	0,156	0,091	0,398	0,174	0,703
0,145	-0,732	0,422	-0,715	-0,186	-0,082	-0,041
0,148	0,136	0,142	0,151	0,275	0,213	0,447
0,164	-0,364	0,535	-0,381	0,057	0,434	0,087
-0,061	-0,019	-0,013	-0,010	-0,144	0,048	-0,037
0,055	-0,907	0,186	-0,961	-0,357	-0,575	-0,149
0,106	0,202	0,122	0,253	0,218	0,166	0,016
0,019	0,049	-0,062	0,021	0,053	-0,035	-0,052
-0,017	0,049	0,032	0,036	-0,007	0,071	-0,047
-0,012	0,309	-0,090	0,290	0,150	0,274	-0,029
-0,143	0,009	-0,154	0,029	-0,136	0,035	0,007
-0,088	0,471	-0,135	0,490	0,149	0,208	0,034
-0,024	0,763	-0,120	0,775	0,338	0,561	0,177
0,098	-0,336	0,072	-0,381	-0,105	-0,220	0,025
0,068	-0,796	0,298	-0,836	-0,208	-0,233	-0,024
0,057	0,555	0,203	0,550	0,389	0,729	0,203
0,051	-0,754	0,244	-0,812	-0,311	-0,367	-0,141
0,067	-0,147	0,283	-0,174	0,178	-0,076	0,097
-0,048	-0,030	-0,006	-0,049	-0,074	-0,042	0,009
28	29	30	31	32	34	35

1						
-0,221	1					
0,657	0,018	1				
-0,283	0,737	-0,020	1			
0,304	0,097	0,680	0,051	1		
0,033	-0,075	-0,004	-0,090	0,040	1	
0,900	-0,142	0,761	-0,209	0,318	0,009	1
-0,119	0,029	0,029	0,029	-0,012	-0,050	-0,210
-0,033	-0,062	-0,121	-0,082	-0,011	0,029	-0,043
0,019	-0,055	-0,094	-0,118	0,132	-0,283	-0,055
-0,278	-0,021	-0,325	-0,020	0,013	-0,268	-0,330
-0,034	-0,039	-0,005	0,009	-0,038	0,081	-0,022
-0,445	0,001	-0,491	0,056	-0,257	-0,062	-0,525
-0,748	0,199	-0,575	0,205	-0,141	0,001	-0,833
0,325	0,034	0,263	-0,008	0,091	0,057	0,370
0,703	-0,058	0,757	-0,128	0,610	0,007	0,808
-0,550	0,198	-0,231	0,211	0,346	0,058	-0,646
0,789	-0,167	0,609	-0,227	0,447	-0,034	0,812
0,128	0,147	0,252	0,107	0,180	-0,120	0,151
0,028	0,028	-0,006	-0,098	0,035	0,018	0,029
36	37	39	41	42	43	45

1	
-0,042	1

0,044	0,134	1				
0,078	0,038	0,511	1			
-0,056	0,104	-0,020	-0,075	1		
0,057	0,105	0,173	0,378	0,096	1	
0,215	0,049	0,114	0,300	0,065	0,589	1
-0,135	0,002	-0,044	-0,249	0,054	0,014	0,002
-0,205	-0,024	0,005	-0,240	-0,088	-0,430	-0,636
0,280	0,029	0,134	0,276	-0,036	0,280	0,648
-0,254	0,085	0,195	-0,089	-0,017	-0,376	-0,651
0,214	0,027	0,127	0,082	-0,013	-0,153	-0,068
-0,089	0,044	-0,034	-0,054	-0,101	0,070	-0,012
47	48	49	50	51	52	

1						
0,309	1					
-0,210	-0,278	1				
0,296	0,730	-0,404	1			
0,045	0,140	0,045	0,090	1		
0,082	0,091	0,023	0,047	-0,020	1	

2.2 Valeurs propres et vecteurs

	<i>Valeur propre</i>	<i>Variabilité(%)</i>	<i>%cumulé</i>
<i>F1</i>	10,571	22,492	22,491
<i>F2</i>	4,694	9,986	32,478
<i>F3</i>	4,042	8,600	41,078
<i>F4</i>	2,210	4,701	45,780
<i>F5</i>	1,994	4,242	50,021
<i>F6</i>	1,736	3,693	53,714
<i>F7</i>	1,646	3,502	57,216
<i>F8</i>	1,447	3,080	60,296
<i>F9</i>	1,414	3,009	63,305
<i>F10</i>	1,294	2,753	66,057
<i>F11</i>	1,227	2,611	68,669
<i>F12</i>	1,144	2,434	71,102
<i>F13</i>	1,046	2,227	73,329
<i>F14</i>	1,012	2,153	75,482
<i>F15</i>	0,938	1,997	77,479
<i>F16</i>	0,884	1,881	79,360
<i>F17</i>	0,868	1,847	81,207
<i>F18</i>	0,819	1,742	82,948
<i>F19</i>	0,735	1,564	84,513
<i>F20</i>	0,689	1,466	85,978
<i>F21</i>	0,649	1,381	87,359
<i>F22</i>	0,612	1,302	88,661
<i>F23</i>	0,559	1,190	89,851
<i>F24</i>	0,513	1,092	90,943
<i>F25</i>	0,497	1,059	92,002
<i>F26</i>	0,403	0,857	92,860

	<i>Valeur propre</i>	<i>Variabilité(%)</i>	<i>%cumulé</i>
<i>F27</i>	0,374	0,796	93,656
<i>F28</i>	0,356	0,756	94,412
<i>F29</i>	0,332	0,706	95,119
<i>F30</i>	0,293	0,624	95,743
<i>F31</i>	0,271	0,577	96,320
<i>F32</i>	0,264	0,563	96,883
<i>F33</i>	0,238	0,505	97,388
<i>F34</i>	0,170	0,362	97,751
<i>F35</i>	0,164	0,349	98,100
<i>F36</i>	0,159	0,338	98,438
<i>F37</i>	0,117	0,248	98,686
<i>F38</i>	0,109	0,232	98,918
<i>F39</i>	0,099	0,211	99,129
<i>F40</i>	0,090	0,191	99,320
<i>F41</i>	0,082	0,174	99,494
<i>F42</i>	0,077	0,164	99,659
<i>F43</i>	0,062	0,132	99,791
<i>F44</i>	0,048	0,103	99,893
<i>F45</i>	0,027	0,057	99,950
<i>F46</i>	0,017	0,037	99,987
<i>F47</i>	0,006	0,013	100,000

3 Les Tableaux des sorties PLS1

Les Tableaux des sorties PLS1 de la variable Y_1

Indice de qualité Q^2 :

<i>Composante</i>	Y_1	<i>Total</i>
<i>Comp1</i>	0,031	0,031
<i>Comp2</i>	0,001	0,001
<i>Comp3</i>	-0,027	-0,027
<i>Comp4</i>	-0,267	-0,267
<i>Comp5</i>	-0,195	-0,195
<i>Comp6</i>	-0,293	-0,293
<i>Comp7</i>	-0,336	-0,336
<i>Comp8</i>	-0,327	-0,327
<i>Comp9</i>	-0,323	-0,323
<i>Comp10</i>	-0,278	-0,278

Indice de qualité Q^2 cumulé :

<i>Composante</i>	Y_1	<i>Total</i>
<i>Comp1</i>	0,031	0,031
<i>Comp2</i>	0,032	0,032
<i>Comp3</i>	0,006	0,006
<i>Comp4</i>	-0,260	-0,260
<i>Comp5</i>	-0,506	-0,506
<i>Comp6</i>	-0,947	-0,947
<i>Comp7</i>	-1,601	-1,601
<i>Comp8</i>	-2,452	-2,452
<i>Comp9</i>	-3,565	-3,565
<i>Comp10</i>	-4,835	-4,835

R^2 et Redondance de (Y, t) :

<i>Variable</i>	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9	t_{10}
Y_1	0,251	0,063	0,039	0,027	0,010	0,012	0,010	0,004	0,003	0,002
<i>Redondance</i>	0,251	0,063	0,039	0,027	0,010	0,012	0,010	0,004	0,003	0,002

Redondance cumulée de (Y, t) :

<i>Variable</i>	$Cp1$	$Cp2$	$Cp3$	$Cp4$	$Cp5$	$Cp6$	$Cp7$	$Cp8$	$Cp9$	$Cp10$
Y_1	0,251	0,315	0,353	0,380	0,390	0,402	0,411	0,416	0,419	0,421
<i>Redondance</i>	0,251	0,315	0,353	0,380	0,390	0,402	0,411	0,416	0,419	0,421

R^2 et Redondance de $(Y, \tilde{u})$:

<i>Variable</i>	$\tilde{u}_1$	$\tilde{u}_2$	$\tilde{u}_3$	$\tilde{u}_4$	$\tilde{u}_5$	$\tilde{u}_6$	$\tilde{u}_7$	$\tilde{u}_8$	$\tilde{u}_9$	$\tilde{u}_{10}$
Y_1	1	1	1	1	1	1	1	1	1	1
<i>Redondance</i>	1	1	1	1	1	1	1	1	1	1

Redondance cumulée de $(Y, \tilde{u})$:

<i>Variable</i>	$Cp1$	$Cp2$	$Cp3$	$Cp4$	$Cp5$	$Cp6$	$Cp7$	$Cp8$	$Cp9$	$Cp10$
Y_1	1	2	3	4	5	6	7	8	9	10
<i>Redondance</i>	1	2	3	4	5	6	7	8	9	10

Paramètres du modèle :

<i>Variable</i>	Y_1	X_{25}	1,707
<i>Constante</i>	16872,359	X_{26}	-14,353
X_1	-1,550	X_{27}	1,786
X_2	-1,935	X_{28}	0,221
X_3	-21,132	X_{29}	6,444
X_4	6,425	X_{30}	-0,612
X_5	-0,872	X_{31}	5,618
X_6	-19,605	X_{32}	-0,940
X_7	-0,069	X_{33}	-277,904
X_8	5,200	X_{34}	-0,099
X_9	1,254	X_{35}	-4,029
X_{10}	7,456	X_{36}	0,580
X_{11}	1,126	X_{37}	-23,925
X_{12}	7,606	X_{38}	0,000
X_{13}	-0,144	X_{39}	-41,931
X_{14}	0,623	X_{40}	-3,570
X_{15}	-1,121	X_{41}	-4,965
X_{16}	-0,179	X_{42}	0,000
X_{17}	1,520	X_{43}	3,912
X_{18}	-3,316	X_{44}	0,000
X_{19}	0,255	X_{45}	-7,269
X_{20}	2,200	X_{46}	2,423
X_{21}	0,541	X_{47}	-9,543
X_{22}	4,882	X_{48}	10,597
X_{23}	14,453	X_{49}	-0,119
X_{24}	1,034	X_{50}	-2,545

Coefficients d'ajustement Y_1 :

<i>Observations</i>	254,000
<i>Sommedespoids</i>	254,000
<i>DDL</i>	243,000
R^2	0,421
<i>Ecart – type</i>	12,428
<i>MCE</i>	147,763
<i>RMCE</i>	12,156

Les Tableaux des sorties PLS1 de la variable Y_2

Indice de qualité Q^2 :

<i>Composante</i>	Y_2	<i>Total</i>
<i>Comp1</i>	0,076	0,076
<i>Comp2</i>	-0,571	-0,571
<i>Comp3</i>	-0,427	-0,427
<i>Comp4</i>	-0,393	-0,393
<i>Comp5</i>	-0,570	-0,570
<i>Comp6</i>	-0,595	-0,595
<i>Comp7</i>	-0,314	-0,314
<i>Comp8</i>	-0,652	-0,652
<i>Comp9</i>	-0,681	-0,681
<i>Comp10</i>	-0,510	-0,510

Indice de qualité Q^2 cumulé :

<i>Composante</i>	Y_2	<i>Total</i>
<i>Comp1</i>	0,076	0,076
<i>Comp2</i>	-0,451	-0,451
<i>Comp3</i>	-1,071	-1,071
<i>Comp4</i>	-1,884	-1,884
<i>Comp5</i>	-3,528	-3,528
<i>Comp6</i>	-6,220	-6,220
<i>Comp7</i>	-8,484	-8,484
<i>Comp8</i>	-14,668	-14,668
<i>Comp9</i>	-25,336	-25,336
<i>Comp10</i>	-38,757	-38,757

R^2 et Redondance de (Y, t) :

<i>Variable</i>	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9	t_{10}
Y_2	0,398	0,195	0,056	0,055	0,024	0,014	0,011	0,014	0,022	0,015
<i>Redondance</i>	0,398	0,195	0,056	0,055	0,024	0,014	0,011	0,014	0,022	0,015

Redondance cumulée de (Y, t) :

<i>Variable</i>	$Cp1$	$Cp2$	$Cp3$	$Cp4$	$Cp5$	$Cp6$	$Cp7$	$Cp8$	$Cp9$	$Cp10$
Y_2	0,398	0,593	0,649	0,704	0,728	0,742	0,753	0,767	0,789	0,804
<i>Redondance</i>	0,398	0,593	0,649	0,704	0,728	0,742	0,753	0,767	0,789	0,804

R^2 et Redondance de $(Y, \tilde{u})$:

<i>Variable</i>	$\tilde{u}_1$	$\tilde{u}_2$	$\tilde{u}_3$	$\tilde{u}_4$	$\tilde{u}_5$	$\tilde{u}_6$	$\tilde{u}_7$	$\tilde{u}_8$	$\tilde{u}_9$	$\tilde{u}_{10}$
Y_2	1	1	1	1	1	1	1	1	1	1
<i>Redondance</i>	1	1	1	1	1	1	1	1	1	1

Redondance cumulée de $(Y, \tilde{u})$:

<i>Variable</i>	$Cp1$	$Cp2$	$Cp3$	$Cp4$	$Cp5$	$Cp6$	$Cp7$	$Cp8$	$Cp9$	$Cp10$
Y_2	1	2	3	4	5	6	7	8	9	10
<i>Redondance</i>	1	2	3	4	5	6	7	8	9	10

Paramètres du modèle :

<i>Variable</i>	Y_2		
<i>Constante</i>	975,200		
X_1	0,130	X_{27}	-0,012
X_2	-0,022	X_{28}	-0,024
X_3	-1,419	X_{29}	0,105
X_4	-1,221	X_{30}	0,231
X_5	0,014	X_{31}	-0,232
X_6	-0,686	X_{32}	-0,201
X_7	-0,001	X_{33}	0,630
X_8	1,313	X_{34}	0,025
X_9	-0,017	X_{35}	-0,174
X_{10}	0,532	X_{36}	0,011
X_{11}	-0,049	X_{37}	-0,159
X_{12}	-0,016	X_{38}	0,000
X_{13}	0,001	X_{39}	0,036
X_{14}	0,036	X_{40}	0,073
X_{15}	-0,048	X_{41}	-0,165
X_{16}	0,014	X_{42}	0,000
X_{17}	-0,010	X_{43}	-1,032
X_{18}	0,667	X_{44}	0,000
X_{19}	0,000	X_{45}	0,597
X_{20}	0,015	X_{46}	0,186
X_{21}	0,038	X_{47}	-0,397
X_{22}	-0,378	X_{48}	-0,145
X_{23}	0,002	X_{49}	0,001
X_{24}	-0,312	X_{50}	0,482
X_{25}	-0,150		
X_{26}	-0,466		

Coefficients d'ajustement Y_2 :

<i>Observations</i>	254,000
<i>Sommedespoids</i>	254,000
<i>DDL</i>	243,000
<i>R²</i>	0,804
<i>Ecart – type</i>	0,417
<i>MCE</i>	0,167
<i>RMCE</i>	0,408

4 Les Tableaux des sorties PLS2

Paramètres du modèle :

<i>Variable</i>	Y_1	Y_2
<i>Constante</i>	16416,795	1262,375
X_1	2,182	0,092
X_2	1,088	-0,027
X_3	17,605	-0,235
X_4	7,394	-0,191
X_5	0,770	0,027
X_6	1,872	-0,864
X_7	0,162	0,006
X_8	3,622	0,091
X_9	0,363	0,003
X_{10}	0,556	-0,112
X_{11}	2,643	0,050
X_{12}	6,274	-0,111
X_{13}	-0,135	-0,002
X_{14}	2,397	0,066
X_{15}	-1,140	-0,040
X_{16}	-0,052	0,019
X_{17}	2,010	0,062
X_{18}	-1,748	0,303
X_{19}	0,248	0,007
X_{20}	1,376	0,076
X_{21}	0,752	0,016
X_{22}	0,185	-0,273
X_{23}	8,642	0,182
X_{24}	0,641	-0,170
X_{25}	1,669	-0,063
X_{26}	-4,613	-0,261
X_{27}	3,312	0,051

X_{28}	-0,965	0,047
X_{29}	5,452	0,058
X_{30}	-0,192	0,065
X_{31}	4,978	0,029
X_{32}	-0,581	0,004
X_{33}	-149,666	-2,479
X_{34}	-0,036	0,008
X_{35}	-6,829	-0,238
X_{36}	0,697	0,012
X_{37}	-22,273	-0,500
X_{38}	0,000	0,000
X_{39}	-30,134	-0,774
X_{40}	-3,341	-0,049
X_{41}	-4,784	-0,315
X_{42}	0,000	0,000
X_{43}	-2,799	0,491
X_{44}	0,000	0,000
X_{45}	-1,674	0,223
X_{46}	0,815	0,105
X_{47}	-4,274	-0,414
X_{48}	-6,937	0,685
X_{49}	-0,101	-0,001
X_{50}	0,473	0,025

Indice de qualité Q^2 :

<i>Composante</i>	Y_1	Y_2	<i>Total</i>
<i>Comp1</i>	-0,006	0,064	0,029
<i>Comp2</i>	0,083	-0,083	0,020
<i>Comp3</i>	-0,052	-0,573	-0,283

Indice de qualité Q^2 cumulé :

<i>Composante</i>	Y_1	Y_2	<i>Total</i>
<i>Comp1</i>	-0,006	0,064	0,029
<i>Comp2</i>	0,077	-0,013	0,049
<i>Comp3</i>	0,030	-0,593	-0,220

R^2 et Redondance de (Y, t) :

<i>Variable</i>	t_1	t_2
Y_1	0,003	0,257
Y_2	0,392	0,018
<i>Redondance</i>	0,198	0,137

Redondance cumulée de (Y, t) :

<i>Variable</i>	<i>Comp1</i>	<i>Comp2</i>
Y_1	0,003	0,260
Y_2	0,392	0,410
<i>Redondance</i>	0,198	0,335

 R^2 et Redondance de $(Y, \tilde{u})$:

<i>Variable</i>	$\tilde{u}_1$	$\tilde{u}_2$
Y_1	0,010	0,934
Y_2	0,992	0,071
<i>Redondance</i>	0,501	0,503

Redondance cumulée de $(Y, \tilde{u})$:

<i>Variable</i>	<i>Comp1</i>	<i>Comp2</i>
Y_1	0,010	0,944
Y_2	0,992	1,063
<i>Redondance</i>	0,501	1,004

Références

- [1] Abdi, H. (2010) . Partial Least Square Regression, PLS-Regression. *John Wiley & Sons, Inc.*
- [2] Belsley D.A., Kuh.E., Welsh.E. (1980). *Regression diagnostics : identifying influential data and sources of collinearity*, New York, Wiley.
- [3] Chavent, M. and Patouille, B. (2003). Calcul des coefficients de régression et du PRESS en régression PLS1. *La Revue de Modulad*, **Vol. 30**, pp. 1 – 9.
- [4] Ciarlet, P.G.(1989). *Introduction to numerical linear algebra and optimisation*, London, Cambridge University Press.
- [5] Cornell(1990).
- [6] De Jong, S. (1993), SIMPLS : An alternative approach to partial least squares regression, *Chemometrics and Intelligent Laboratory Systems*, **vol. 18**, pp. 251 – 263.
- [7] Foucart, T.(1984). *Analyse factorielle de tableaux multiples*, Paris, Masson.
- [8] Foucart, T.(1991). Transitivité du produit scalaire, *Rev. Statistique Appliquée*, XXXIX(3), pp. 57 – 68.
- [9] Foucart, T.(1992). Colinéarité dans une matrice de produit scalaire, *Rev. Statistique Appliquée*, XXXX (3), pp. 5 – 17.
- [10] Foucart, T. (2006) . Colinéarité et régression linéaire. *Math. & Sci. hum. Mathematics and Social Sciences*, n° 173, pp. 5 – 25.
- [11] Geladi, P., and Kowalski B. (1986). Partial least square regression : A tutorial. *Analytica Chimica Acta*, **35**, 1 – 17.
- [12] Helland I.S. (1990). PLS regression and statistical models. *Scandinavian ; Journal of Statistics*, **17**, 97 – 114.

- [13] Jolliffe, I.T., (1982). A note on the use of principal components in regression, *Appl.Statist.* 31, pp. 300 – 303.
- [14] Kendall, M.G., Stuart, A.(1961). The advanced theory of statistics, **vol 2**, *Inference and relationship*, Griffin, London.
- [15] Kettaneh-Wold (1992).
- [16] Mallows, C.L.(1973). Some comments on Cp, *Technometrics* 15, pp. 661 – 676.
- [17] Martens, H, and Naes, T. (1989). *Multivariate Calibration*. London :Wiley.
- [18] Menardoc.1993. Application de la régression PLS à la spectrométrie IRTF,Revue.
- [19] Naes, T., Helland, I.S., (1993). Relevant components in regression, *Scand. J. Statis.* 20, 1993, pp. 239 – 250.
- [20] Phatak, A., and de Jong, S. (1997). The geometry of partial least squares. *Journal of Chemometrics*, **11**, 311 – 338.
- [21] Stone, M. & Brooks, R. (1990), Continuum regression : Cross-validated sequentially constructed prediction embracing ordinary least squares, partial least squares, and principal components regression, *Journal of the Royal Statistical Society*, Series B, 52(2), pp. 237 – 269.
- [22] Tenenhaus, M.(1998). *La régression PLS, Théorie et Pratique*. Editions Technip.
- [23] Tenenhaus, M., Gauchi, J.P. and Ménardo. C. (1995).Régression PLS et applications. *Revue de statistique appliquée*, tome 43, n°1, p. 7 – 63.
- [24] Ter Braak, C.J.F., and de Jong, S. (1998). The objective function ofpartial least squares regression. *Journal of Chemometrics*, **12**,41 – 54.
- [25] Wold, H. (1966). Estimation of principal component and related models by iterative least squares, *Multivariate Analysis*, ed. P.R. Krishnaiah, Newyork : Academic Press, pp. 391 – 420.

- [26] Wold, S., Ruhe, A., Wold, H., Dunn, III W.J. (1984). The collinearity problem in linear regression : The partial least squares (PLS) approached to generalized inverses, *SIAM Sci .Stat. Comp.*, **Vol 5**, n° 3, pp. 735 – 743.
- [27] XLSTAT-PLS « Partial Least Squares regression - PLS», version 2014.
- [28] MATLAB, version 2010.