Université de Montréal

Comparaison de différentes méthodes de modélisation pour le traitement des eaux usées

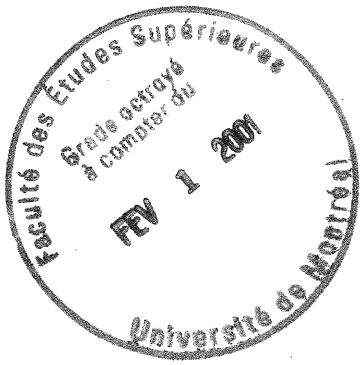
par

Janie Dufresne

Département de mathématiques et de statistique
Faculté des arts et des sciences

Mémoire présenté à la Faculté des études supérieures
en vue de l'obtention du grade de
Maître ès sciences (M.Sc.)
en statistique

Août 2000



Université de Montréal

Faculté des études supérieures

Ce mémoire intitulé

Comparaison de différentes méthodes de modélisation pour le traitement des eaux usées

présenté par

Janie Dufresne

a été évalué par un jury composé des personnes suivantes :

M. Robert Cléroux
(président-rapporteur)

M. Christian Léger
(directeur de recherche)

M. Martin Savoie
(co-directeur)

M. Roch Roy
(membre du jury)

Mémoire accepté le :
26 octobre 2000

SOMMAIRE

Le traitement des eaux usées sert à contrôler l'eau qui est rejetée dans l'effluent après avoir été utilisée par l'usine de fabrication. La prévision de certaines variables est importante puisque le traitement entraîne des coûts et des normes environnementales doivent être respectées. De nombreuses variables, souvent autocorrélées, sont mesurées dans le traitement chaque jour et elles sont également corrélées entre elles. De plus, le nombre de variables mesurées dans le traitement est important par rapport au nombre de données observées pour chaque variable. Afin d'essayer de solutionner efficacement ces difficultés, plusieurs méthodes de modélisation seront comparées : la régression linéaire multiple, la régression ridge, la régression avec composante principale, la régression avec racine latente, les moindres carrés partiels et les réseaux de neurones. Nous présenterons d'abord la théorie derrière chacune de ces méthodes, puis nous les appliquerons à un jeu de données provenant d'une usine de pâtes et papier.

REMERCIEMENTS

Je tiens à remercier ma famille pour toutes les petites attentions qui m'ont rendu la vie tellement plus facile et agréable, merci également pour vos encouragements et votre soutien. L'université est une étape qu'il faut souvent savoir traverser avec un brin d'humour. Celui-ci est souvent difficile à trouver en certaines périodes, la famille et les amis sont alors un cadeau inestimable. Je tiens donc à remercier en plus les meilleurs amis qu'on puisse avoir, mes "colocs" Isabelle, Michel et Marc-André ainsi que Marilène pour avoir rendu ma vie bien plus drôle ces cinq dernières années, pour avoir su supporter mes commentaires incessants, surtout devant la télévision, mais particulièrement pour le simple, mais très précieux, fait d'être toujours là et de me faire profiter de leur amitié.

Je remercie Christian Léger pour son temps et son aide dans la correction de ce mémoire ainsi que pour tous ses judicieux conseils. Je veux dire un merci particulier à Gina Bezeau sans qui ce projet n'aurait jamais eu lieu, merci de m'avoir donné ma chance. Je remercie également Martin Carignan pour avoir accepté de prendre la relève après le départ de Gina. Je souhaite la meilleure des chances aux deux futurs multimillionnaires de Bezeau, Carignan et associés. Je remercie Martin PLS Savoie pour son enthousiasme face à ce projet. Enfin, je tiens à exprimer ma reconnaissance à Abitibi-Consolidated et au Conseil de recherches en sciences naturelles et en génie pour leur soutien financier au long de ces deux dernières années.

Table des matières

Sommaire iii

Remerciements iv

Table des figures viii

Liste des tableaux xi

Introduction 1

Chapitre 1. Méthodes d'analyse 5

1.1. Régression linéaire multiple 7

1.1.1. Le modèle de régression linéaire 7

1.1.2. Sélection de variables 10

1.1.2.1. Addition par étape 10

1.1.2.2. Élimination par étape 11

1.1.2.3. Sélection pas-à-pas 12

1.2. Régression ridge 14

1.2.1. Le modèle de régression ridge 14

1.2.2. Validation croisée 17

1.3. Régression avec composante principale 19

1.3.1. Analyse en composante principale 19

1.3.2. Exemple 23

1.3.3. Le modèle de régression avec composante principale 26

1.4. Régression avec racine latente 29

1.5. Moindres carrés partiels 34

1.5.1. Algorithme 1 35

1.5.1.1. Cas univarié 35

1.5.1.2. Cas multivarié 41

1.5.2. Algorithme 2 42

1.6. Méthode non linéaire 46

1.6.1. Hypothèse de linéarité 46

1.6.2. Les réseaux de neurones 49

1.6.2.1. Construction du modèle 49

1.6.2.2. Descente du gradient 53

1.6.2.3. Méthode de Newton 57

1.6.2.4. Méthode du gradient conjugué 59

1.6.2.5. Interprétabilité des réseaux de neurones 61

1.6.2.6. Conclusion 62

Chapitre 2. Analyse préliminaire du jeu de données du traitement des eaux usées 64

2.1. Description du jeu de données 64

2.2. Le traitement des valeurs manquantes 65

2.2.1. Méthodes d'imputation 65

2.2.2. Les valeurs manquantes du jeu de données 68

2.2.3. Le traitement de la variable N-NO_2-NO_3 à l'affluent 74

Chapitre 3. Comparaison des méthodes 80

3.1. Les modèles et leur estimation 80

3.2. Description des statistiques utilisées 86

3.3. Comparaison des méthodes 89

3.3.1. Estimation/test sur les 146 variables 89

3.3.2. Validation croisée avec 10 blocs sur le jeu complet 95

3.3.3. Modèles ajustés à partir de sous-groupes de variables 99

3.3.4. Temps de calcul 104

3.4. Analyse des variables dépendantes 107

3.4.1. Analyse des variables IVB, DCO, DBO_5 et MES 107

3.4.1.1. Série chronologique 108

3.4.1.2. Meilleur modèle 111

3.4.1.3. Coefficients des modèles obtenus pour la DCO avec la méthode d'estimation/test sur les variables du groupe "Jour 1" 112

3.4.2. Analyse des autres variables 117

Conclusion 122

Annexe A. Description des variables 126

Annexe B. Variables dépendantes 130

Annexe C. Fonctions Splus 139

Bibliographie 149

Table des figures

1.2.1 Dilemme biais-variance 15

1.3.1 Le résultat de l'ACP donne de nouvelles variables maximisant la variation contenue dans les données. 21

1.6.1 Relation linéaire entre deux variables. 46

1.6.2 Suite de Fibonacci 47

1.6.3 Ajustement d'un modèle linéaire pour la suite de Fibonacci 49

1.6.4 La boîte noire: l'image la plus souvent associée aux réseaux de neurones. 50

1.6.5 Réseau de neurones à une couche cachée. 50

1.6.6 Réseau de neurones représentant la régression linéaire multiple. 54

1.6.7 Nombre de neurones et identification des poids pour un réseau avec une couche cachée. 55

1.6.8 Réseau de neurones univarié ayant une couche cachée contenant deux neurones et troix variables explicatives. 62

1.6.9 Effet de la variable x_1 pour le modèle de prévision $\hat{y} = tanh(-x_1 + x_2 - x_3) - tanh(-2x_1 + x_2 + x_3)$ lorsque $x_2 = x_3 = 0$. 62

1.6.10 Effet de la variable x_1 pour le modèle de prévision $\hat{y} = tanh(-x_1 + x_2 - x_3) - tanh(-2x_1 + x_2 + x_3)$ lorsque $x_2 = -2$ et $x_3 = 0$. 62

2.1.1 Schéma du traitement des eaux usées. 66

2.2.1 Distribution des valeurs manquantes dans le jeu de données original du traitement des eaux usées 69

2.2.2 Représentation de la relation linéaire justifiant l'élimination ou le traitement particulier de certaines variables 71

2.2.3 Distribution des valeurs manquantes dans le jeu de données modifié du traitement des eaux usées 73

2.2.4 Histogramme de la variable N-NO_2-NO_3 75

3.3.1 Somme des erreurs au carré pour différentes valeurs du paramètre de la RR 105

3.4.1 Représentation des variables dépendantes sur le jeu de validation ainsi que de leur prévision obtenue avec le meilleur modèle pour les variables IVB, DCO, DBO_5_lab et MES..... 119

3.4.2 Représentation des résidus sur le jeu de validation en fonction de leur prévision obtenue avec le meilleur modèle pour les variables IVB, DCO, DBO_5_lab et MES..... 120

3.4.3 Représentation de la DCO observée ainsi que les prévisions obtenues pour la RLM avec l'AÉ et pour LRR 121

B.0.1 Représentation des variables IVB et débit du 01/12/97 au 26/09/99.. 131

B.0.2 Représentation des variables DCO et DBO_5 du 01/12/97 au 26/09/99. 132

B.0.3 Représentation des variables MES et MES_lab du 01/12/97 au 26/09/99..... 133

B.0.4 Représentation des variables N-NH_3 et MESnd du 01/12/97 au 26/09/99. 134

B.0.5 Représentation des variables O-PO_4 et turbid du 01/12/97 au 26/09/99. 135

B.0.6 Représentation des variables pHcomp et pHmoy du 01/12/97 au 26/09/99..... 136

B.0.7 Représentation des variables Tmoy et cond du 01/12/97 au 26/09/99. 137

B.0.8 Représentation de la variable DCOsol du 19/10/98 au 26/09/99..... 138

Liste des tableaux

1.3.1 Matrice $X^{(6)}$, tirée de l'article de Webster, Gunst et Mason (1974) ... 25

1.3.2 Matrice des corrélations de $X^{(6)}$ ... 25

1.6.1 Suite de Fibonacci ... 47

1.6.2 Fonctions de lien les plus couramment utilisées pour construire les réseaux de neurones avec leur fonction d'activation correspondante et les contraintes qu'elles posent sur les sorties du réseau ... 53

2.1.1 Liste des variables dépendantes des modèles dont la description est fournie au tableau A.0.1 ... 66

2.2.1 Variables ayant un taux élevé de valeurs manquantes avec leur taux respectif ... 70

2.2.2 Liste des journées où la production de l'usine était nulle pendant la période du 01-12-97 au 26-09-99. Ces journées ont été excluses des analyses ... 72

2.2.3 Liste des variables imputées à l'aide de l'autocorrélation de délai 1 ... 72

2.2.4 Liste des variables imputées à l'aide de la régression linéaire simple .. 74

2.2.5 Pourcentage d'erreur obtenu par validation croisée, pour un seuil fixé, afin de prévoir la variable N-NO_2-NO_3 transformée à l'aide des autres variables dépendantes ... 77

2.2.6 Comparaison des modèles avec et sans la variable N-NO_2-NO_3 transformée grâce à la racine de la moyenne des erreurs au carré pour méthode d'estimation/test et la validation croisée en blocs avec la méthode PCR. 79

3.2.1 SSE_M pour la méthode d'estimation/test et la VC en 10 blocs sur les 629 observations. 87

3.2.2 SSE_T pour la méthode d'estimation/test et la VC en 10 blocs sur les 629 observations. 88

3.3.1 R^2_M obtenu avec la méthode estimation/test avec l'optimisation des paramètres par la VC avec une observation à la fois. 91

3.3.2 R^2_T obtenu avec la méthode estimation/test avec l'optimisation des paramètres par la VC avec une observation à la fois. 92

3.3.3 Rang moyen de la SSE des méthodes de modélisation pour la méthode d'estimation/test et VC en blocs sur toutes les observations. 94

3.3.4 Corrélation entre quelques variables dépendantes. 94

3.3.5 R^2_M obtenu avec la VC en 10 blocs sur le jeu complet. 96

3.3.6 R^2_T obtenu avec la VC en 10 blocs sur le jeu complet. 97

3.3.7 Racine de la SSE moyenne obtenue avec PLS univarié pour l'estimation des paramètres avec la méthode estimation/test, puis celle sur le jeu de validation pour le modèle ajusté avec les paramètres optimaux estimés ainsi que celle obtenue pour la VC en 10 blocs sur les 629 observations 99

3.3.8 R^2_M obtenu avec la méthode d'estimation/test à partir des variables sélectionnées par l'addition par étape sur les 300 premières observations et l'optimisation des paramètres par la VC avec une observation à la fois. 102

3.3.9 R_T^2 obtenu avec la méthode d'estimation/test à partir des variables sélectionnées par l'addition par étape sur les 300 premières observations et l'optimisation des paramètres par la VC avec une observation à la fois. 103

3.4.1 Liste des variables significatives lorsque la sélection par l'addition par étape est faite à partir du groupe "Jour 1" pour les variables IVB, DCO, DBO_5 et MES. 109

3.4.2 Racine de la SSE moyenne pour les variables IVB, DCO, DBO_5 et MES obtenue par la méthode d'estimation/test avec l'optimisation des paramètres par la VC avec une observation à la fois. 110

3.4.3 Description des méthodes ayant donné les meilleurs modèles pour les variables IVB, DCO, DBO_5 et MES. 112

3.4.4 Tableau des coefficients des modèles obtenus pour la variable DCO avec la méthode d'estimation/test avec l'optimisation des paramètres par la VC avec une observation à la fois sur toutes les variables du groupe "Jour 1". La moyenne et l'écart type de chacune des variables sont également donnés à titre de référence. 114

3.4.5 Description des méthodes ayant donné les meilleurs modèles 117

A.0.1 Tableau des variables mesurées au traitement des eaux usées, ainsi que leur abbréviation respective. 126

B.0.1 Statistiques descriptives pour les variables dépendantes 130

INTRODUCTION

Depuis 1995, certaines usines d'Abitibi-Consolidated ont mis en opération un traitement des eaux usées. Le gouvernement réglemente les rejets des usines dans les effluents et des normes environnementales sont à respecter. Les facteurs réglementés sont les matières en suspension (MES), la demande biochimique en oxygène (DBO) et la toxicité. Il faut donc contrôler le traitement afin de ne pas polluer dangereusement les effluents.

Avoir une bonne connaissance du traitement permet de n'agir que lorsque c'est nécessaire. Chaque action en vue de nettoyer davantage les eaux usées entraîne des coûts, par exemple, le coût de l'ajout de nutriments. Cependant, ne pas agir, et obtenir incidemment un dépassement des normes, entraîne également des coûts, comme une amende imposée par le gouvernement ou un arrêt temporaire de la production afin de rétablir la situation. Il est donc important d'agir lorsqu'il le faut, mais, en même temps, de n'agir que lorsqu'il le faut. La prévision des variables à la sortie du traitement devient donc un outil qui peut se révéler fort utile dans l'opération du traitement des eaux usées. De plus, les résultats des tests de la DBO ne sont obtenus qu'après plusieurs jours. Un modèle de prévision permettrait donc d'avoir un estimé de la DBO.

Le traitement de l'usine est du type réacteur biologique séquentiel (RBS). L'usine a un traitement composé de cinq RBS. Le jeu de données fourni par l'usine contient près de 175 variables mesurées à l'entrée du traitement, à chacun

des cinq RBS et à la sortie. Certaines variables ne sont pas calculées directement, mais obtenues à partir d'équations combinant d'autres variables. Les variables mesurées, en plus d'être autocorrélées, sont corrélées entre elles, parfois même fortement. Parmi les variables disponibles, une quinzaine d'entre elles sont des variables dépendantes alors que les autres peuvent être utilisées pour les expliquer. Ceci amène à la constatation suivante: étant donné le nombre de variables disponibles, un modèle statistique peut devenir rapidement très lourd et très complexe. L'idéal serait de parvenir à obtenir un modèle alliant simplicité et efficacité. Donc, un modèle qui, en plus de bien prévoir les variables de sortie, resterait simple.

Nous nous proposons d'utiliser six méthodes pour analyser le jeu de données.

- La régression linéaire multiple avec sélection de variables
- La régression ridge
- La régression avec composante principale
- La régression avec racine latente
- Les moindres carrés partiels
- Les réseaux de neurones

Pourquoi utiliser ces six méthodes en particulier? La régression linéaire multiple (RLM) est celle qui est la plus souvent utilisée. Elle est simple et, surtout, les résultats s'obtiennent très rapidement. Les autres méthodes ont l'avantage de tenir compte du problème de la collinéarité entre les variables, ce que ne fait pas la RLM, mais il y a des paramètres à optimiser. La régression ridge ressemble beaucoup à la RLM, mais tient compte de la collinéarité. Une façon simple d'identifier des collinéarités entre les variables est d'utiliser l'analyse en composante principale. Une méthode utilisant cette technique est donc simple et

peut possiblement donner de bons résultats: la méthode de régression avec composante principale (PCR) utilise directement les composantes trouvées par l'ACP pour faire la régression. La méthode de régression avec racine latente calcule un estimé du vecteur des coefficients en passant par l'analyse en composante principale tout comme PCR. PCR n'utilise que les variables explicatives pour définir les nouvelles composantes contrairement à LRR qui utilise en plus les variables dépendantes. Les moindres carrés partiels sont peu utilisés en statistique, cette méthode se retrouve surtout dans la littérature du domaine de la chimie ou de la chémométrie. Une analyse préliminaire du jeu de données avait déjà été effectuée avec cette méthode. Toutes ces méthodes utilisent des modèles linéaires. Les réseaux de neurones n'ont pas cette contrainte; ils sont plus flexibles en ne définissant pas une classe de modèles. Les réseaux de neurones sont souvent la meilleure méthode au niveau des prévisions, mais ils sont difficilement interprétables. Ainsi, nous tenterons de trouver parmi ces six méthodes celle qui offre le meilleur compromis entre efficacité, simplicité et interprétabilité.

Chacune des six méthodes sera utilisée afin de modéliser le traitement, puis elles seront comparées entre elles afin de déterminer laquelle donne de meilleurs résultats. Un bon modèle devrait pouvoir fournir une bonne prévision sur de nouvelles données, tout en étant le plus simple possible. D'après la littérature, mis à part les réseaux de neurones, PLS et RR devraient donner les meilleurs résultats au niveau de la qualité des prévisions. Parmi les trois autres méthodes, LRR devrait donner de meilleures estimations que PCR et toutes deux devraient être supérieures à la RLM à cause de l'effet de la collinéarité entre les variables.

Le premier chapitre introduit chacune des méthodes ainsi que leurs avantages et leurs inconvénients. En plus, différents outils utiles pour appliquer les différentes méthodes sont brièvement expliqués: la sélection de variables, l'analyse

en composante principale et la validation croisée. Le second chapitre présente l'analyse préliminaire qui a été effectuée sur le jeu de données provenant d'une usine de pâtes et papier d'Abitibi-Consolidated. Nous y décrivons le jeu de données initial, puis les variables qui ont été éliminées ainsi que l'imputation des valeurs manquantes. Finalement, le troisième chapitre présente la comparaison des différentes méthodes. Nous utilisons les variables du jeu de données ainsi que différents sous-groupes de celles-ci afin de voir l'efficacité relative de chaque méthode. Nous comparons également les modèles obtenus aux modèles où nous prévoyons y par sa moyenne estimée ou par un modèle AR d'ordre 2. Nous verrons la facilité d'ajustement de chaque méthode ainsi que le temps de calcul nécessaire pour estimer les paramètres et ajuster le modèle.

Chapitre 1

MÉTHODES D'ANALYSE

Le développement actuel des systèmes informatiques amène l'apparition de bases de données de plus en plus importantes. La simplification de la collecte et du stockage des données permettent de conserver de plus en plus d'information sur des procédés, des marchés, des clients, etc. Ceci nécessite des outils d'analyse toujours plus performants et offrant de plus en plus de flexibilité afin de pallier aux difficultés qui font surface. Les variables observées sont tantôt quantitatives, tantôt qualitatives, continues ou discrètes, mesurées à différentes fréquences, avec différentes unités, etc. Plusieurs outils statistiques ont déjà été développés jusqu'à ce jour afin de répondre à deux besoins au niveau de l'analyse: la classification et la prévision. Un tour d'horizon des différentes méthodes existantes et de leurs applications est présenté dans le livre de Lefébure et Venturi (1998).

La classification permet, par exemple, de créer des catégories de clients ou de déterminer si une action a plus de chance d'avoir un haut rendement. Le résultat est sous forme de classes discrètes: mauvais/bon/excellent client ou faible/haut rendement. Parmi les outils les plus fréquemment utilisés, il y a les analyses par grappes, les analyses d'associations et les arbres de décision. Ce type d'analyse ne sera pas discuté ici. Pour plus de détails, le livre de Everitt (1993) offre une introduction compréhensible sur les analyses par grappes, tandis que les arbres de régression sont abordés dans le livre de Breiman, Friedman, Olshen et Stone

(1984) et dans l'article de Bioch, van der Meer et Potharst (1997). Il existe plusieurs façons de faire des analyses d'associations, les plus courantes sont présentées dans l'article de Carter, Hamilton et Cercone (1997).

Le second type d'analyse, la prévision, permet, à partir d'une série d'observations, de trouver une estimation d'un facteur n'ayant pas encore été observé. Par exemple, nous pourrions prévoir quelles seront les ventes d'un produit le mois prochain à partir des ventes mensuelles des cinq dernières années ou les profits d'une entreprise à partir des données connues sur celles-ci, comme le nombre d'employés, leur salaire, les clients et leurs achats auprès de l'entreprise, les différents frais fixes, les dépenses en matériel, les investissements, etc. Plusieurs méthodes de prévision ont été développées, dont les plus connues sont l'analyse de séries chronologiques, qui est présentée en détail dans les livres d'Abraham et Ledolter (1983) et de Brockwell et Davis (1996), et la régression. Ces deux méthodes peuvent également être combinées. C'est ce dernier type d'analyse qui sera utilisé pour modéliser les données provenant du traitement des eaux usées. Les données sont des mesures prises quotidiennement et nous utiliserons de l'information ramassée sur plusieurs jours consécutifs afin de faire des prévisions. Le modèle lui-même sera ajusté avec des outils provenant de la régression. Dans les sections qui suivent, nous aborderons une à une les différentes méthodes qui seront ensuite utilisées pour modéliser les données: la régression linéaire multiple, la régression ridge, la régression avec composante principale, la régression avec racine latente, les moindres carrés partiels et, finalement, les réseaux de neurones.

1.1. RÉGRESSION LINÉAIRE MULTIPLE

1.1.1. Le modèle de régression linéaire

Bien souvent, il existe un lien entre des variables, par exemple, plus la première augmente et plus l'autre diminue ou alors que deux variables augmentent ou diminuent de façon proportionnelle. Ces relations peuvent être exprimées souvent comme une combinaison linéaire d'une ou plusieurs variable(s), combinaison qui explique le comportement d'une autre variable. La régression linéaire multiple (RLM) ajuste un modèle en formant des combinaisons linéaires des variables explicatives. Nous posons le modèle de la forme:

$$E[Y|X] = \beta_0 + \beta_1 X_1 + \ldots + \beta_P X_P. \tag{1.1.1}$$

Celui-ci est linéaire par rapport aux coefficients β_i. Ce modèle suppose que $(Y, X_1, \ldots, X_P)$ suit une distribution F multivariée avec moyenne $\underset{\sim}{\mu}$ et matrice de variances-covariances Σ. Il est possible d'inclure des variables transformées dans le modèle ou de tenir compte des interactions qui peuvent exister parmi les variables explicatives, autrement dit, il est possible de considérer des termes comme X_p^k, $\ln(X_p)$, $X_p X_q$, ...; le modèle reste linéaire en $\underset{\sim}{\beta}$. Au niveau de la notation, nous noterons les matrices et les variables aléatoires par des lettres majuscules alors que les observations ou les éléments calculés à partir de celles-ci seront représentés par des lettres minuscules.

Afin d'estimer ce modèle, nous avons N observations indépendantes de chacune des variables. Ces observations sont notées de la manière suivante: y_n est le n^{e} élément de $\underset{\sim}{y}$ et x_{np} est le n^{e} élément du vecteur $\underset{\sim}{x}_p$. De plus, $X_{N\times(P+1)} = [\underset{\sim}{1}, \underset{\sim}{x}_1, \ldots, \underset{\sim}{x}_P]$ est la matrice des variables explicatives observées et $\underset{\sim}{y}_{N\times 1}$ est le vecteur de la variable que nous voulons estimer par $\underset{\sim}{\hat{y}} = X\underset{\sim}{\hat{\beta}}$, où $\underset{\sim}{\hat{\beta}}$ est l'estimation du vecteur des coefficients. Ce dernier est déterminé en minimisant la somme des

erreurs au carré, i.e.

$$(\underset{\sim}{y} - X\underset{\sim}{\beta})'(\underset{\sim}{y} - X\underset{\sim}{\beta}).$$

Cette expression est minimisée par

$$\hat{\underset{\sim}{\beta}} = (X'X)^{-1}X'\underset{\sim}{y}, \tag{1.1.2}$$

qui devient l'estimateur du vecteur des coefficients. C'est l'estimateur de la RLM par les moindres carrés. Une démonstration de ce résultat est présentée, entre autres, dans Weisberg (1985) à l'annexe 2A.3. Par la suite, afin d'abréger la notation, nous ferons référence à cet estimateur comme étant celui de la RLM au lieu de dire "l'estimateur de la RLM par les moindres carrés".

Afin que cet estimateur soit défini de façon unique, il faut que la matrice des observations soit de plein rang de telle sorte que $X'X$ soit également de rang P et donc inversible. Si les variables ne sont pas indépendantes, c'est-à-dire si une variable explicative peut être exprimée de façon exacte par une combinaison linéaire des autres variables explicatives ($\underset{\sim}{x}_p = \alpha_1\underset{\sim}{x}_1 + \cdots + \alpha_{p-1}\underset{\sim}{x}_{p-1} + \alpha_{p+1}\underset{\sim}{x}_{p+1} + \ldots + \alpha_P\underset{\sim}{x}_P$, où au moins un des α_i n'est pas nul pour $i \neq p$) alors la matrice $X'X$ n'est pas inversible. La solution la plus simple à adopter, si un tel cas se présente, est d'éliminer les variables qui causent la dépendance. Si nous réussissons à enlever les variables qui sont des combinaisons linéaires de certaines des autres variables alors la difficulté disparaît. Le problème avec cette solution est qu'il est souvent difficile d'identifier de telles dépendances. De plus, une dépendance, même imparfaite, a une influence sur la qualité des estimations. Le variance des coefficients du vecteur $\hat{\underset{\sim}{\beta}}$ peut être très élevée lorsqu'il y a présence de collinéarité. L'estimateur obtenu par la RLM est sans biais et, parmi tous les estimateurs

linéaires ayant cette propriété, il est celui ayant la plus faible variance:

$$\begin{aligned} E[\hat{\underset{\sim}{\beta}}|X] &= \underset{\sim}{\beta} \\ Var[\hat{\underset{\sim}{\beta}}|X] &= \sigma^2(X'X)^{-1} = \sigma^2 \sum_{p=1}^{P} \lambda_p^{-1} \underset{\sim}{l}_p \underset{\sim}{l}'_p, \end{aligned} \tag{1.1.3}$$

où les λ_p sont les valeurs propres de la matrice $X'X$ et les $\underset{\sim}{l}_p$, les vecteurs propres normalisés associés à λ_p. Plus de détails sur les valeurs et vecteurs propres d'une matrice seront fournis à la section 1.3.1. La présence de collinéarité se traduit par de petites valeurs propres, ce qui a comme conséquence que la variance de $\hat{\underset{\sim}{\beta}}$ devient très grande

$$\lambda_p \text{ petite} \Rightarrow \lambda_p^{-1} \text{ grande} \Rightarrow \text{Var}[\hat{\underset{\sim}{\beta}}|\text{X}] \text{ grande.}$$

Ceci a pour conséquence de limiter l'utilisation du modèle résultant pour faire des prévisions. Celles-ci ne pourront être fiables que pour des observations situées dans la même région que les données ayant servi à construire le modèle.

Dans le jeu de données du traitement des eaux usées, dont nous parlerons plus en détails au chapitre 2, nous savons, d'après les informations fournies par les ingénieurs de l'usine, que les variables sont liées et donc qu'il y a de la collinéarité parmi les variables. Donc, l'estimateur trouvé par la RLM aura une grande variance.

Bien que cette méthode ait ses inconvénients lorsqu'il y a collinéarité, elle offre une approche simple et facile à réaliser et à comprendre. De plus, le modèle est aisément interprétable et s'obtient rapidement. Il permet de déterminer le sens de l'effet des variables sur les variables dépendantes ainsi que leur importance relative. La régression linéaire multiple par les moindres carrés reste donc une méthode intéressante à considérer.

1.1.2. Sélection de variables

Dans une base de données où il y a beaucoup de variables, il est fort possible, et souhaitable même dans une certaine mesure, que plusieurs d'entre elles contiennent peu d'information pouvant expliquer les variables dépendantes. Il est donc utile d'avoir une méthode permettant de sélectionner un sous-ensemble de variables contenant un maximum d'information. Plusieurs méthodes peuvent être trouvées dans la littérature comme celles basées sur les critères C_p de Mallow, MSE, R^2, R^2 ajusté, etc. Nous présentons ici trois méthodes séquentielles assez simples: l'addition par étape (AÉ), l'élimination par étape (ÉÉ) et la sélection pas-à-pas (SPAP); toutes trois sont basées sur le même principe. L'élimination par étape ne peut pas être utilisée directement lorsque la matrice des observations n'est pas de plein rang, mais nous la présentons tout de même ici puisque le principe est utilisé dans la procédure SPAP.

1.1.2.1. *Addition par étape*

Cette méthode de sélection ajoute les variables une à une au modèle jusqu'à ce qu'un critère d'arrêt soit atteint. Aussi longtemps que l'addition d'une nouvelle variable amène une amélioration notable au modèle, le processus de sélection continue. Il faut commencer par ajuster un modèle avec une seule variable. Ceci est effectué pour chacune des variables explicatives et celle qui optimise le critère choisi, par exemple X_j, est conservée pour les étapes suivantes. Nous obtenons ainsi une première estimation de Y:

$$\hat{Y}^{(1)} = \hat{a}_{10} + \hat{a}_{1j}X_j,$$

où les coefficients $\hat{a}_{10}$ et $\hat{a}_{1j}$ sont ceux obtenus en estimant le modèle $E[Y|X_j] = a_{10} + a_{1j}X_j$ avec la RLM comme nous l'avons vu à la section 1.1.1.

Plusieurs critères peuvent être utilisés, comme la valeur de la statistique F associée à la nouvelle variable, ou la valeur-p associée au test, ou encore l'augmentation du R^2 par rapport à l'étape précédente.

La procédure est reprise en incluant X_j dans le modèle et en ajoutant une seconde variable pour avoir une meilleure estimation de Y:

$$\hat{Y}^{(2)} = \hat{a}_{20} + \hat{a}_{2j}X_j + \hat{a}_{2i}X_i, \quad i \neq j.$$

Encore une fois, les variables qui n'ont pas encore été sélectionnées sont utilisées à tour de rôle, par exemple, dans cette seconde étape, toutes les variables sauf X_j seraient utilisées. La variable qui optimise le critère est retenue et elle est ajoutée au modèle de base pour les étapes subséquentes. La procédure se poursuit de la même façon. Un seuil minimal pour le critère est sélectionné et, lorsqu'aucune des variables non sélectionnées ne l'atteint, la procédure s'arrête.

1.1.2.2. *Élimination par étape*

L'élimination par étape est l'inverse de AÉ en ce sens qu'elle commence avec le modèle complet, celui contenant toutes les variables, pour ensuite les éliminer une à une jusqu'à ce que le seuil du critère choisi soit atteint. À chaque étape, la variable qui contribue le moins au modèle est enlevée, par exemple, celle qui a la plus grande valeur-p.

Comme mentionné plus tôt, cette méthode ne peut pas être utilisée lorsque la matrice X n'est pas de rang P. Il faut d'abord ajuster le modèle avec toutes les variables. Le vecteur des coefficients $\underset{\sim}{\beta}$ est estimé par (1.1.2). Or, dans le modèle complet, si $X'X$ n'est pas de plein rang à cause de la collinéarité entre les variables ou parce qu'il y a moins d'observations que de variables alors $X'X$ n'est pas inversible et nous ne pouvons pas estimer le modèle et évaluer le critère. Cependant, lorsqu'il est possible de l'utiliser, la méthode ÉÉ offre l'avantage, pour

certains critères, de n'avoir à ajuster qu'un seul modèle à chaque étape; donc si nous arrêtons après K étapes, K modèles auront été ajustés. Si nous éliminons les variables en nous basant sur la valeur-p ou sur la statistique F alors il suffit d'ajuster le modèle avec toutes les variables qui n'ont pas encore été éliminées afin de déterminer si une autre variable peut l'être et laquelle. L'ajustement du modèle nous donnera, par exemple, la valeur-p associée à chaque variable, puis nous éliminerons celle ayant la plus grande valeur-p, en autant que celle-ci soit supérieure au seuil d'arrêt fixé. Pour AÉ, il faut ajuster, à chaque étape, autant de modèles qu'il y a de variables restantes; donc, pour K étapes, il faut ajuster $\sum_{k=1}^{K}(P-k+1) = \frac{K}{2}(2P-K+1)$ modèles. Ainsi, si nous avons 100 variables et que le modèle final contient 5 variables, ÉÉ nécessitera 96 régressions (96 étapes) alors que, pour AÉ, nous aurons besoin d'effectuer 585 régressions (6 étapes).

1.1.2.3. *Sélection pas-à-pas*

Dans les deux méthodes précédentes, un retour en arrière n'est pas permis: une fois qu'une variable est choisie (éliminée) dans le modèle, il n'est pas possible de l'exclure (inclure) plus tard dans la procédure. Cependant, une interaction entre deux variables ou plus peut diminuer (augmenter) la contribution d'une autre au modèle de telle sorte que nous souhaitons avoir la possibilité de l'exclure (inclure). C'est ce que permet la méthode de sélection pas-à-pas. À tout moment, il est possible de reconsidérer l'importance de la contribution d'une variable au modèle. Pour ce faire, nous commençons la procédure comme avec AÉ, mais, à chaque fois qu'une nouvelle variable est ajoutée, nous vérifions avec la méthode ÉÉ pour voir si nous pourrions éliminer une des variables incluses aux étapes précédentes. Ceci permet de ce fait d'obtenir un modèle qui présente un compromis entre simplicité et efficacité au niveau de la prévision.

La sélection de variables est importante lorsque le nombre de variables est grand. Elle permet d'éliminer du modèle les composantes n'ayant que peu d'influence sur les variables dépendantes et donc de mettre en évidence les variables importantes. Bien qu'il soit possible souvent de construire un modèle sans faire de sélection, le temps supplémentaire à y consacrer permet d'obtenir des résultats plus simples qui gagnent en interprétabilité. Les méthodes séquentielles de sélection de variables ne garantissent pas que le modèle final sera le modèle optimal; d'ailleurs en utilisant deux méthodes de sélection différentes, nous obtenons souvent deux modèles différents. Pour trouver le modèle optimal selon un critère fixé, il faut utiliser la méthode exhaustive qui consiste à essayer tous les modèles possibles et à sélectionner le meilleur. Cette méthode de sélection de variables est très coûteuse en temps de calculs puisqu'il y a $2^P - 1$ modèles possibles. Les méthodes séquentielles sont donc une solution raisonnable lorsqu'il y a plusieurs variables et que nous désirons un modèle efficace, mais simple. Les méthodes AÉ, ÉÉ et SPAP sont donc de bons compromis entre efficacité et temps de calcul lorsque le nombre de variables est grand.

De façon générale, la sélection de variables séquentielle peut donner de bons résultats avec des variables corrélées puisqu'elle choisit les variables une à une et considère l'effet "individuel" de chacune des variables séparément avant de les inclure ou de les exclure dans le modèle. En particulier, ÉÉ élimine un peu de la collinéarité dans le modèle à chaque fois qu'une variable est éliminée. D'après Gunst et Mason (1977b), les variables qui sont corrélées ont tendance à avoir une petite valeur de la statistique t. Pour une variable $\underset{\sim}{X}_p$, celle-ci est donnée par

$$t_p = \frac{(1 - R_p^2)^{\frac{1}{2}} \hat{\beta}_p}{(\frac{1}{N-P-1} SSE)^{\frac{1}{2}}},$$

où R_p^2 est le coefficient de détermination du modèle

$$E[X_p|X^{(-p)}] = \alpha_0 + \alpha_1 X_1 + \ldots + \alpha_{p-1} X_{p-1} + \alpha_{p+1} X_{p+1} + \ldots + \alpha_P X_P,$$

où $X^{(-p)} = [X_1, \ldots, X_{p-1}, X_{p+1}, \ldots, X_P]$.

Plus la variable X_p est fortement corrélée avec les autres variables dépendantes et plus R_p^2 sera près de 1, ce qui aura pour conséquence que t_p aura une petite valeur. La sélection de variable aura tendance, bien souvent, à ne pas retenir ces variables fortement corrélées, mais elles peuvent tout de même avoir une valeur prédictive. Donc, les composantes ne sont pas enlevées en prenant pour base la magnitude de leur coefficient dans le modèle dans la mesure où la collinéarité a un impact sur la statistique F ou t du modèle.

1.2. RÉGRESSION RIDGE

1.2.1. Le modèle de régression ridge

La RLM, introduite à la section 1.1.1, permet d'obtenir l'estimateur de $\hat{\underset{\sim}{\beta}}$ ayant la plus faible variance parmi tous les estimateurs linéaires non biaisés. Cependant, rien ne garantit que cette variance soit raisonnablement petite. La variance du vecteur des coefficients est donnée par (1.1.3). Si les valeurs propres λ_p sont petites pour certains indices p, ou même pour une seule valeur de p, alors la variance de $\hat{\underset{\sim}{\beta}}$ sera très grande. Si nous enlevons la contrainte posée sur le biais de l'estimateur alors il est possible d'obtenir un gain au niveau de la variance. Ainsi, au lieu d'avoir un estimateur de $\underset{\sim}{\beta}$ non biaisé ayant la plus petite variance, un biais est permis et nous essayons d'obtenir un petit écart quadratique moyen (EQM), autrement dit, nous faisons un compromis entre biais et variance. Pour un estimateur $\hat{\theta}$, l'EQM est donné par:

$$EQM[\hat{\theta}] = Variance[\hat{\theta}] + (Biais[\hat{\theta}])^2.$$

Donc, si un petit biais permet d'obtenir un grand gain au niveau de la variance alors l'estimateur obtenu est plus "efficace" du point de vue de l'erreur quadratique moyenne comme illustré à la figure 1.2.1. Si nous avons deux estimateurs $\hat{\theta}_1$ et $\hat{\theta}_2$ pour estimer θ et que $\hat{\theta}_1$ est sans biais comparativement à $\hat{\theta}_2$ qui ne l'est pas, mais que $\hat{\theta}_2$ a une plus faible variance alors nous choisirons celui ayant le plus petit EQM, soit $\hat{\theta}_2$ dans la cas illustré.

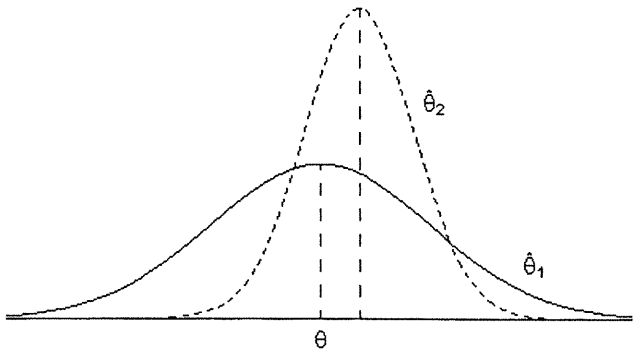


FIGURE 1.2.1. *L'estimateur $\hat{\theta}_1$ est sans biais, mais a une forte variance, comparativement à $\hat{\theta}_2$ qui a un petit biais, mais une variance qui est beaucoup plus faible. L'EQM de $\hat{\theta}_2$ est plus petite que celui de $\hat{\theta}_1$, il est donc plus "efficace" du point de vue de l'erreur quadratique moyenne.*

L'estimateur de $\underset{\sim}{\beta}$ obtenu par la RLM peut être réécrit de la façon suivante:

$$\hat{\underset{\sim}{\beta}}^{RLM} = \sum_{p=1}^{P} \lambda_p^{-1} \underset{\sim}{l}_p \underset{\sim}{l}'_p X' \underset{\sim}{y}.$$

(Si la matrice X est de plein rang, les valeurs propres sont toutes différentes de zéro.)

S'il y a une collinéarité parfaite entre des variables alors le problème se résout en enlevant une de ces variables. Si la collinéarité n'est pas parfaite alors nous ne pouvons pas enlever de variables sans perdre un peu d'information, mais certaines valeurs propres seront près de zéro ce qui entraîne le problème au niveau de la

variance. La régression ridge (RR) propose comme solution d'ajouter une petite constante, disons κ, aux valeurs propres de telle sorte que $(\lambda_p + k)^{-1}$ ne soit pas trop grand. Intuitivement, si λ_p est près de zéro, l'ajout de κ aidera à diminuer la variance tandis que, si ce n'est pas le cas, il y aura peu de changement, c'est-à-dire si λ_p n'est pas trop petit, mais que κ est petit, alors $\lambda_p^{-1} \approx (\lambda_p + \kappa)^{-1}$. Cette solution diminue donc la variance, mais un biais est tout de même introduit et il est proportionnel à la valeur de κ. Pour une petite valeur du paramètre κ, le biais est faible comparativement au gain apporté par la diminution de la variance. Lorsque κ est zéro, l'estimateur redevient celui de la RLM.

L'estimateur de $\underset{\sim}{\beta}$ donné par la régression ridge est

$$\hat{\underset{\sim}{\beta}}^{RR} = \sum_{p=1}^{P} (\lambda_p + \kappa)^{-1} \underset{\sim}{l_p} \underset{\sim}{l_p}' X' \underset{\sim}{y}.$$

Il peut aussi s'écrire sous forme matricielle

$$\hat{\underset{\sim}{\beta}}^{RR} = (X'X + \kappa I_P)^{-1} X' \underset{\sim}{y}$$

et, lorsque κ est fixe, sa variance est donnée par

$$Var[\hat{\underset{\sim}{\beta}}^{RR} | X] = \sigma^2 \sum_{p=1}^{P} \lambda_p (\lambda_p + \kappa)^{-2} \underset{\sim}{l_p} \underset{\sim}{l_p}'.$$

Le paramètre κ est déterminé à l'aide de la validation croisée, qui sera présentée à la section suivante. D'autres méthodes sont également proposées dans la littérature, dont la minimisation de la "trace ridge" (Gunst et Mason, 1980). Plusieurs autres méthodes sont expliquées dans Hoerl, Kennard et Baldwin (1975) et Hoerl et Kennard (1976).

La RR offre plusieurs des avantages de la RLM: nous obtenons un modèle simple et facilement interprétable. De plus, cette méthode tient compte de la collinéarité possible entre les différentes variables. Elle est cependant un peu

plus difficile à implanter que la RLM dans la mesure où il y a un paramètre, κ, à déterminer et qu'il faut de ce fait utiliser la validation croisée. Hoerl et Kennard (1970a) ont démontré que la somme des erreurs au carré obtenue à partir de la RR diminue d'abord lorsque κ augmente, puis elle atteint un minimum pour ensuite augmenter à mesure que κ augmente. Cette propriété facilite la détermination de la valeur du paramètre avec la validation croisée. Cependant, la meilleure technique pour choisir le paramètre κ n'est pas déterminée. Hoerl et Kennard (1970b) ont essayé de faire de la sélection de variable sur de vraies données et ont conclu que ce n'était pas recommandé avec la RR. Donc, RR ne réduit pas la dimension du problème. Ceci a également été observé par Kresta, MacGregor et Marlin (1991). Si le modèle final ne s'obtient pas aussi aisément et rapidement que la RLM, il gagne cependant en efficacité dans de nombreux cas.

1.2.2. Validation croisée

Dans un contexte de régression, il y a des méthodes où un paramètre doit être fixé afin de pouvoir estimer le modèle. Ce paramètre, disons k, doit être déterminé de façon à ce que le modèle offre la "meilleure" capacité de prévision possible. C'est ce que tente de faire la validation croisée (VC).

S'il y a suffisamment de données, le jeu de données peut être séparé en deux parties indépendantes, puis la méthode d'estimation/test est utilisée. *Pour chaque valeur de* k, la première partie sert à estimer le modèle et la seconde, à tester l'efficacité du dit modèle pour effectuer des prévisions. La performance de plusieurs modèles peut être comparée avec la somme des erreurs au carré (SSE) calculée à partir des prévisions faites pour les données contenues dans la seconde partie du jeu de données. Nous choisissons alors k associé au modèle ayant la meilleure capacité de généralisation, soit celui associé à la plus petite SSE. Si

le jeu de données est séparé en deux groupes, il reste moins d'observations pour estimer le modèle. Cette perte d'information peut faire en sorte que le modèle final dépendant du paramètre choisi soit moins "bon" pour prévoir de nouvelles données qu'il ne le serait si toutes les données étaient utilisées pour le construire. Ceci peut surtout se produire lorsque le nombre d'observations est faible. Nous souhaiterions avoir un maximum de données pour estimer le modèle, mais aussi pour le tester. Donc, si nous ne voulons pas séparer les données en deux groupes, nous pouvons utiliser la méthode qui suit. Le jeu de données est divisé en m sous-ensembles de taille semblable. Un à un, les sous-ensembles sont exclus et $\underset{\sim}{\beta}$ est estimé avec les $m-1$ autres sous-ensembles pour une valeur fixée du paramètre. Une prévision de chacune des observations contenues dans le sous-groupe exclu est ensuite effectuée à partir du modèle estimé. Après que tous les groupes aient été enlevés les uns après les autres, nous pouvons obtenir $\hat{\epsilon}^{(k)} = y - \hat{y}^{(k)}$ pour chaque observation et calculer la somme des erreurs au carré, $\text{SSE}(k) = \sum_1^N (\hat{\epsilon}_n^{(k)})^2$. Ici, $\hat{y}^{(k)}$ est la prévision de y obtenue à partir du modèle estimé avec le paramètre k. Cette procédure est répétée pour chaque valeur de k et la valeur choisie est celle associée à la plus petite valeur de $\text{SSE}(k)$

$$\hat{k} = \arg\min_k SSE(k). \tag{1.2.1}$$

Plus la valeur de m est grande plus le temps de calcul augmente. Dans le cas particulier où $m = N$, le nombre d'observations, chacune des observations est enlevée et estimée avec le modèle trouvé à l'aide des autres données. De façon générale, si nous cherchons la valeur optimale d'un paramètre, la VC est utilisée pour plusieurs valeurs de ce facteur et celle qui minimise la SSE est fixée pour estimer ensuite le modèle final. Si le paramètre prend des valeurs discrètes alors le critère est estimé pour toutes les valeurs possibles, alors que, quand le facteur est continu, il faut utiliser des valeurs réparties sur le domaine de définition du

facteur ou sur un intervalle dans lequel le facteur doit raisonnablement être situé. Ainsi, si le paramètre prend des valeurs sur l'intervalle [0, 1] alors nous pourrions estimer le critère pour $k = 0, 0{,}1, 0{,}2, \ldots, 0{,}9, 1$. Plus nous prenons de valeurs dans l'intervalle et plus la valeur optimale du paramètre sera précise, cependant le temps de calcul augmente également.

L'avantage de la validation croisée est que le modèle est testé à chaque itération avec des observations n'ayant pas servi à le construire, tout en conservant un maximum d'information. Elle permet donc de trouver un modèle maximisant autant que possible sa valeur prédictive. Le désavantage est l'augmentation du temps de calcul. Malgré cela, lorsque nous ne voulons pas avoir à séparer les données en un jeu de données pour construire le modèle et un autre pour le tester, la VC reste une méthode simple et efficace.

1.3. Régression avec composante principale

1.3.1. Analyse en composante principale

Lorsqu'il y a plusieurs variables, il est parfois utile de concentrer l'information. Nous voudrions pouvoir conserver moins de variables, mais sans perdre trop d'information. L'analyse en composante principale (ACP) permet de déterminer les combinaisons linéaires qui représentent un maximum de la variation contenue dans les données. Ainsi, s'il y a au départ une dizaine de variables très corrélées, par exemple, il est possible que l'ACP permette de trouver deux ou trois composantes qui expliquent 90% ou plus de la variation des données. L'élimination des autres composantes résulte en une petite perte d'information par rapport à la diminution du nombre de variables. De cette façon, nous n'éliminons pas de variables, mais nous les combinons pour former un nombre réduit de composantes.

De façon générale, supposons que nous avons un procédé pour lequel nous avons P variables aléatoires représentées par $\underset{\sim}{\mathcal{X}} = (X_1, \ldots, X_P)'$. La première composante $W_1 = a_{11}X_1 + a_{12}X_2 + \ldots + a_{1P}X_P = \underset{\sim}{a}_1'\underset{\sim}{\mathcal{X}}$ est trouvée en déterminant la combinaison des variables qui maximise la variance:

$$\underset{\sim}{a}_1 = \arg\max_{||\underset{\sim}{a}||=1} Var[\underset{\sim}{a}'\underset{\sim}{\mathcal{X}}].$$

La seconde composante est celle qui maximise la variance une fois la première composante considérée. Nous ajoutons comme contrainte l'orthogonalité entre les composantes:

$$\underset{\sim}{a}_2 = \arg\max_{||\underset{\sim}{a}||=1 \text{ et } \underset{\sim}{a}_1 \cdot \underset{\sim}{a}=0} Var[\underset{\sim}{a}'\underset{\sim}{\mathcal{X}}].$$

Nous continuons ainsi pour les autres composantes. La solution à ce système est donnée par les valeurs et vecteurs propres de la matrice de variances-covariances du vecteur $\underset{\sim}{\mathcal{X}}$. Par exemple, en deux dimensions, la figure 1.3.1 représente les composantes obtenues avec l'ACP. La première, W_1, est orientée dans la direction où la variance est maximale, tandis que la seconde, W_2, est perpendiculaire à la première.

Rappelons d'abord ce que sont les valeurs et vecteurs propres d'une matrice M. Les valeurs propres sont déterminées en premier en solutionnant l'équation suivante:

$$det[M - \lambda I] = 0. \tag{1.3.1}$$

La première valeur propre λ_1 est donnée par la plus grande solution de l'équation, la deuxième plus grande solution devient λ_2 et ainsi de suite. Notons que, lorsque M est définie positive, c'est-à-dire $\underset{\sim}{x}M\underset{\sim}{x} > 0, \forall \underset{\sim}{x} \neq 0$, les valeurs propres sont toutes positives. Puis, à chaque valeur propre différente de zéro, un vecteur

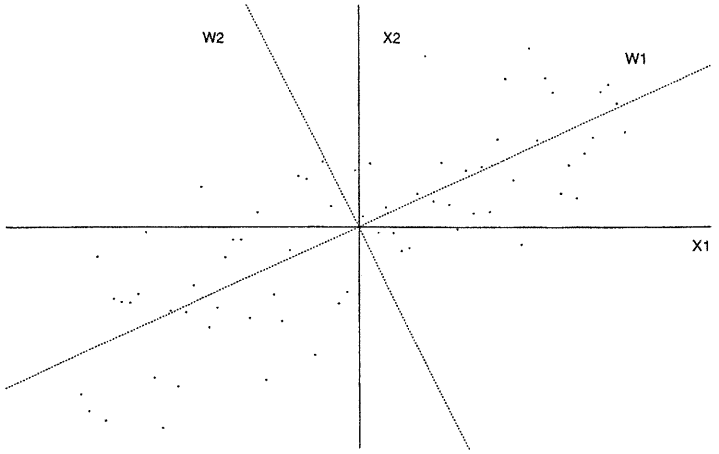


FIGURE 1.3.1. *Le résultat de l'ACP donne de nouvelles variables maximisant la variation contenue dans les données. Les variables X_1 et X_2 sont les variables originales alors que W_1 et W_2 sont les nouvelles composantes. La première composante est orientée dans le sens où la variation des observations est la plus grande tandis que la seconde est perpendiculaire à la première.*

propre, $\underset{\sim}{l}$, est associé tel que

$$M\underset{\sim}{l} = \lambda \underset{\sim}{l}. \tag{1.3.2}$$

Les vecteurs propres forment une base orthogonale. Les vecteurs et valeurs propres sont donc calculés de cette façon.

Il est possible de démontrer (voir Johnson et Wichern, 1992, p.358) que le vecteur $\underset{\sim}{a}_i$, défini au début de cette section, est donné par le i^{e} vecteur propre normalisé de la matrice de variances-covariances de $\underset{\sim}{\mathcal{X}}$. La i^{e} composante est

$$W_i = l_{i1}X_1 + l_{i2}X_2 + \ldots + l_{iP}X_P, \text{ avec } \underset{\sim}{l}_i = (l_{i1}, l_{i2}, \ldots, l_{iP})'$$

et la variance de chaque composante W_i est donnée par la valeur propre qui est associée à $\underset{\sim}{l}_i$:

$$Var[W_i|X] = \lambda_i.$$

Le pourcentage de variation expliquée par la composante W_i est donné par $\frac{\lambda_i}{\sum_{p=1}^{P} \lambda_p}$. Nous avons donc intérêt à éliminer les composantes associées à de petites valeurs propres. La matrice de variances-covariances est cependant inconnue. Elle doit donc être estimée à partir des données observées: la matrice de variances-covariances estimée est notée S et l'élément (i, j) est donné par $\frac{1}{N-1}(\underset{\sim}{x}_i - \bar{x}_i \underset{\sim}{1})'(\underset{\sim}{x}_j - \bar{x}_j \underset{\sim}{1})$, où $\bar{x}_i$ et $\bar{x}_j$ sont les moyennes observées des variables X_i et X_j respectivement. Conservons la même notation et notons maintenant les valeurs et vecteurs propres de la matrice de variances-covariances estimée λ_i et $\underset{\sim}{l}_i$ respectivement.

Les résultats de l'ACP ne sont pas les mêmes selon que nous effectuons l'analyse sur les données originales ou sur les données standardisées. Lorsque les différentes variables sont comparables (mêmes unités et même ordre de grandeur de la variance) alors nous utilisons la matrice S pour l'ACP. Si les unités des variables ne sont pas les mêmes alors nous obtenons des composantes W_i qui ne sont pas interprétables en ce sens que les nouvelles composantes sont la somme de quantités qui ne sont pas de dimensions comparables. D'un autre côté, si une variable X_j a une très grande variance par rapport à celle des autres alors la première composante principale sera essentiellement cette variable, i.e. $W_i = l_{1j} X_j$, et les autres ne seront pas considérées, puisque nous cherchons au départ une composante maximisant la variation contenue dans les données. Si un de ces cas se présente, l'ACP est faite plutôt sur la matrice de variances-covariances estimée avec les données standardisées. Ainsi le même "poids" est donné à toutes les variables puisqu'elles ont toutes la même variance une fois standardisées.

Il y a quelques problèmes avec l'ACP. Lorsqu'elle est effectuée sur les données originales, l'ACP n'est pas invariante par rapport aux changements d'échelle des variables. De plus, les composantes W_i sont souvent difficiles à interpréter autant pour l'ACP sur les données originales que sur celles standardisées. Les composantes représentent une somme de plusieurs variables, qui en elle-même ne veut rien dire généralement.

Lorsque l'analyse est faite sur les données originales alors l'ACP est effectuée sur la matrice des variances-covariances estimée de $\underset{\sim}{\mathcal{X}}$, tandis que si elle l'est sur les données standardisées, la matrice des corrélations, R, est utilisée. La matrice des données standardisées est appelée $Z^{(P)}$ et ses éléments sont représentés par z_{np}. Il est à noter que la standardisation des variables peut être effectuée de plus d'une façon. Si les données sont standardisées de telle sorte que $\sum_1^N z_{np}=0$ et $\sum_1^N z_{np}^2=1$, $p = 1, \ldots, P$, alors la matrice R est donnée par $Z^{(P)'}Z^{(P)}$ tandis que, si elles le sont telles que $\sum_1^N z_{np}=0$ et $\frac{1}{N}\sum_1^N z_{np}^2=1$, $p = 1, \ldots, P$, alors R est donnée par $\frac{1}{N}Z^{(P)'}Z^{(P)}$.

1.3.2. Exemple

Illustrons ce qui précède à l'aide d'un exemple. Prenons la matrice, notons-la $X^{(6)}$, tirée de l'article de Webster, Gunst et Mason (1974) dont les composantes sont données au tableau 1.3.1. Supposons que nous voulions faire une ACP basée sur la matrice des corrélations de $X^{(6)}$, notée R, donnée au tableau 1.3.2. Les valeurs propres sont les solutions du système donné par (1.3.1), où M est maintenant la matrice R, et elles sont λ_1 =2,429, λ_2 =1,546, λ_3 =0,922, λ_4 =0,794, λ_5 =0,308 et λ_6 =0,001.

Pour trouver les vecteurs propres, il faut appliquer la même procédure pour chacune des valeurs propres. Le vecteur propre associé à λ_1 est le vecteur normalisé, $\underset{\sim}{l}_1$ vérifiant $R\underset{\sim}{l}_1 = \lambda_1 \underset{\sim}{l}_1$. La solution est donnée par le vecteur

$$\underset{\sim}{l}_1 = (\text{-}0{,}391,\ \text{-}0{,}456,\ 0{,}483,\ 0{,}188,\ \text{-}0{,}498,\ 0{,}352)'.$$

Ce qui nous donne une première composante:

$$W_1 = \text{-}0{,}391X_1 - 0{,}456X_2 + 0{,}483X_3 + 0{,}188X_4 - 0{,}498X_5 + 0{,}352X_6.$$

Il faut procéder de la même manière pour trouver les autres vecteurs propres et les autres composantes.

Pour trouver les coordonnées des observations dans la nouvelle base, il suffit de faire le produit matriciel entre la matrice $X^{(6)}$ et les vecteurs propres: $W^{(P)} = X^{(P)}L$, où $L = (\underset{\sim}{l}_1, \underset{\sim}{l}_2, \underset{\sim}{l}_3, \underset{\sim}{l}_4, \underset{\sim}{l}_5, \underset{\sim}{l}_6)$. Le pourcentage de variation expliquée par chacune des composantes est donné par $\frac{\lambda_i}{\sum_{p=1}^{P} \lambda_p}$. Ainsi, les composantes expliquent respectivement 40,48%, 25,77%, 15,37%, 13,23%, 5,13%, et 0,02% de la variation contenue dans la matrice $X^{(6)}$. Le sixième vecteur propre conduit à une composante n'expliquant que très peu de variation, cette composante pourrait donc être éliminée sans qu'il y ait une grande perte d'information. En effet, nous pouvons vérifier que $X_2 \approx 10 - X_1 - X_3 - X_4$.

L'ACP permet donc de trouver de nouvelles variables. Celles-ci sont, en quelque sorte, "ordonnées" en ce sens que la première explique un maximum de la variation tandis que la seconde explique un maximum de la variation qui n'est pas déjà expliquée par la première, etc. Si nous conservons toutes les composantes, il n'y a aucune perte d'information. Nous pouvons perdre ou effacer les données, mais, si les composantes et les nouvelles coordonnées sont conservées, nous pouvons retrouver exactement les chiffres de départ. Donc, nous ne perdons de l'information qu'en effaçant des composantes, mais cette perte est contrôlée

X_1	X_2	X_3	X_4	X_5	X_6
8	1	1	1	0,541	-0,099
8	1	1	0	0,130	0,070
8	1	1	0	2,116	0,115
0	0	9	1	-2,397	0,252
0	0	9	1	-0,046	0,017
0	0	9	1	0,365	1,504
2	7	0	1	1,996	-0,865
2	7	0	1	0,228	-0,055
2	7	0	1	1,380	0,502
0	0	0	10	-0,798	-0,399
0	0	0	10	0,257	0,101
0	0	0	10	0,440	0,432

TABLEAU 1.3.1. *Matrice $X^{(6)}$, tirée de l'article de Webster, Gunst et Mason (1974)*

	X_1	X_2	X_3	X_4	X_5	X_6
X_1	1,00	0,05	-0,34	-0,50	0,42	-0,19
X_2	0,05	1,00	-0,43	-0,37	0,48	-0,32
X_3	-0,34	-0,43	1,00	-0,36	-0,51	0,49
X_4	-0,50	-0,37	-0,36	1,00	-0,21	-0,09
X_5	0,42	0,48	-0,51	-0,21	1,00	-0,12
X_6	-0,19	-0,32	0,49	-0,09	-0,12	1,00

TABLEAU 1.3.2. *Matrice des corrélations de $X^{(6)}$.*

dans la mesure où nous savons le pourcentage de variation qui est expliqué par

chacune d'elles. Donc, l'analyse en composante principale utilisée avec la régression permettrait de trouver un modèle contenant moins de variables lorsqu'il est exprimé en fonction des composantes et, donc, plus simple.

1.3.3. Le modèle de régression avec composante principale

Comme le décrit très bien le nom donné à cette méthode, la régression avec composante principale (PCR, de l'anglais "principal component in regression") utilise les nouvelles variables obtenues à l'aide de l'analyse en composante principale (ACP) sur les variables explicatives pour ajuster un modèle de la forme:

$$\underset{\sim}{\hat{y}} = W\underset{\sim}{\hat{\beta}}^* = \hat{\beta}_0^*\underset{\sim}{1} + \hat{\beta}_1^*\underset{\sim}{w}_1 + \hat{\beta}_2^*\underset{\sim}{w}_2 + \ldots + \hat{\beta}_K^*\underset{\sim}{w}_K,$$

où $\underset{\sim}{w}_i = a_{i1}\underset{\sim}{x}_1 + a_{i2}\underset{\sim}{x}_2 + \ldots + a_{iP}\underset{\sim}{x}_P$ et K est le nombre de composantes principales retenues. Il est à noter que $\underset{\sim}{\beta}^*$ n'est pas le même vecteurs de coefficients que celui utilisé pour la RLM et la RR. Les composantes obtenues à partir de l'ACP sont non corrélées; il n'y a donc pas de problème de collinéarité.

La méthode PCR utilise l'analyse en composante principale pour enlever le problème de la collinéarité. L'ACP est effectuée sur la matrice de variances-covariances estimée à partir de $X^{(P)}$ ou $Z^{(P)}$ qui est la matrice des données standardisées. C'est cette deuxième matrice que nous utiliserons. Nous obtenons ainsi $\underset{\sim}{w}_1, \underset{\sim}{w}_2, \ldots, \underset{\sim}{w}_P$, les nouvelles coordonnées des observations. Si la matrice de variances-covariances est de plein rang alors il y a P composantes, si elle est de rang $I \leq P$, il n'y en a que I. Comme mentionné à la section 1.3.1, le vecteur $\underset{\sim}{a}_i$ est le vecteur propre normalisé associé à la i^{e} plus grande valeur propre de la matrice $Z'^{(P)}Z^{(P)}$ et de petites valeurs propres identifient des corrélations entre les variables explicatives. Nous voudrions éliminer ces composantes puisqu'elles contiennent peu d'information.

Plusieurs méthodes peuvent être utilisées pour déterminer le nombre de composantes principales, K, à retenir. Un seuil peut être fixé et toutes les composantes dont la valeur propre associée est inférieure à cette limite sont éliminées. Il est également possible d'utiliser les méthodes de sélection de variables. Avec cette dernière méthode, le problème est toutefois compliqué par le fait que nous voulons éliminer les composantes contribuant à l'augmentation de la variance des estimations, ce qui n'est pas pris en compte par la sélection de variables. Nous pouvons aussi utiliser la validation croisée telle que définie dans la section 1.3.1 pour choisir de nombre "optimal" de composantes. Cependant, d'après Frank et Friedman (1993), cette méthode retiendra souvent plus de composantes qu'il n'est vraiment nécessaire.

Le vecteur des coefficients $\underset{\sim}{\beta}^*$ est estimé à l'aide des moindres carrés, mais avec les variables $W_1, W_2, \ldots, W_K$.

$$\hat{\underset{\sim}{\beta}}^{*PCR} = (W'W)^{-1}W'\underset{\sim}{y},$$

où W=$[\underset{\sim}{1}, \underset{\sim}{w}_1, \underset{\sim}{w}_2, \ldots, \underset{\sim}{w}_K]$. La variance de l'estimateur $\hat{\underset{\sim}{\beta}}^{*PCR}$ est donnée par

$$Var[\hat{\underset{\sim}{\beta}}^{*PCR}|X] = \sigma^2 \sum_{k=1}^{K} \lambda_k^{-1} \underset{\sim}{l}_k \underset{\sim}{l}'_k.$$

Si nous voulons ensuite représenter le modèle en fonction des variables de départ $(X_1, X_2, ..., X_P)$ alors il faut utiliser le fait que $W^{(P)} = X^{(P)}L$, où $L = (\underset{\sim}{l}_1, \underset{\sim}{l}_2, \ldots, \underset{\sim}{l}_K)$ et $\|\underset{\sim}{l}_k\| = 1$, $\forall k$. De cette façon, le modèle estimé peut être réécrit

$$\hat{\underset{\sim}{y}} = W\hat{\underset{\sim}{\beta}}^{*PCR} = [\underset{\sim}{1}, X^{(P)}L]\hat{\underset{\sim}{\beta}}^{*PCR} = X\hat{\underset{\sim}{\beta}}^{PCR},$$

où $\hat{\beta}_0^{*PCR} = \hat{\beta}_0^{PCR}$ et $L(\hat{\beta}_1^{*PCR}, \ldots, \hat{\beta}_K^{*PCR})' = (\hat{\beta}_1^{PCR}, \ldots, \hat{\beta}_P^{PCR})'$. Il est à noter que $\hat{\underset{\sim}{\beta}}^{*PCR}$ est un vecteur de longueur $K+1$ alors que $\hat{\underset{\sim}{\beta}}^{PCR}$ est de longueur $P+1$.

Le modèle obtenu avec PCR s'interprète plus difficilement lorsqu'il est représenté en fonction des nouvelles variables puisque celles-ci ne s'interprètent pas concrètement contrairement aux variables originales. Il est toujours possible de réécrire le modèle selon les variables initiales, mais alors, le modèle n'est plus simple, en ce sens qu'il contient généralement toutes les variables. Il est possible, ainsi, de déterminer en quel sens les variables indépendantes influencent les variables dépendantes ainsi que leur importance relative. Lorsque nous effaçons les k dernières composantes identifiant la collinéarité dans la matrice des variables explicatives, où k est fixé à l'avance, les propriétés de la distribution de l'estimateur de $\underset{\sim}{\beta}$ sont disponibles (voir Mansfield, 1975, et Manfield, Webster et Gunst, 1977). Tout comme la RR, il faut déterminer un paramètre, le nombre de composantes à retenir, et le modèle tient compte de la collinéarité lorsque nous éliminons les composantes selon la magnitude de la valeur propre qui leur est associée.

Il est possible de faire de la sélection de variables avec les composantes de l'ACP au lieu de choisir les K premières. Toutefois, en procédant de cette manière, les composantes sont sélectionnées en tenant compte de leur valeur prédictive. La collinéarité n'est pas considérée de telle sorte que les problèmes qu'elle entraîne ne sont pas éliminés.

La méthode PCR estime les variables *dépendantes* en se basant sur les composantes déterminées à partir des variables *explicatives* uniquement. Nous trouvons des composantes maximisant la variation contenue dans la matrice $X^{(P)}$ et nous supposons que cette variation va dans le même sens que celle de Y. Autrement dit, nous supposons que la composante maximisant la variation de $X^{(P)}$ maximise aussi celle de Y. Donc, l'hypothèse que les composantes éliminées ont peu de valeur prédictive doit être faite avec la méthode PCR lorsque l'élimination des

composantes est basée sur les valeurs propres. La méthode de régression présentée à la section suivante propose une alternative à PCR afin de ne pas avoir à faire cette supposition.

1.4. RÉGRESSION AVEC RACINE LATENTE

La méthode de régression avec racine latente (LRR, de l'anglais "latent root regression") est présentée dans la littérature par Gunst, Mason et Webster qui ont écrit plusieurs articles, ainsi que par Draper et Smith (1981). La section qui suit s'inspire principalement de Webster, Gunst et Mason (1974) et de Gunst et Mason (1980).

Avec LRR, nous voulons estimer $\underset{\sim}{\beta}$ dans le modèle:

$$E[Y|X] = \beta_0 + \underset{\sim}{\mathcal{X}}'\underset{\sim}{\gamma},$$

où $\underset{\sim}{\beta} = (\beta_0, \underset{\sim}{\gamma}')'$.

Introduisons d'abord quelques notations. Les n^e éléments de $\underset{\sim}{y}, \underset{\sim}{x}_1, \ldots, \underset{\sim}{x}_P$ sont notés respectivement $y_n, x_{n1}, \ldots, x_{nP}$. La moyenne observée de ces variables est $\bar{y}, \bar{x}_1, \ldots, \bar{x}_P$:

$$\begin{aligned} \bar{y} &= \frac{1}{N}\sum_{n=1}^{N} y_n \\ \bar{x}_p &= \frac{1}{N}\sum_{n=1}^{N} x_{np}, \quad p = 1, \ldots, P. \end{aligned}$$

Les variables observées standardisées seront notées $\underset{\sim}{y}^s$ et $\underset{\sim}{z}_p$ et leurs éléments, y_n^s et z_{np}. Les variables sont standardisées, rappelons-le, de telle sorte que $\sum_p z_{np} = 0$ et $\sum_p z_{np}^2 = 1$. La matrice contenant l'ensemble des variables explicatives est $Z^{(P)}$.

Nous trouvons un estimateur modifié par rapport à la RLM en utilisant l'ACP. Contrairement à PCR, LRR fait l'ACP en utilisant la matrice $A = [\underset{\sim}{y}^s, Z^{(P)}]$ au lieu de $Z^{(P)}$, où $\underset{\sim}{y}^s = \frac{1}{\eta}(\underset{\sim}{y} - \bar{y}\underset{\sim}{1})$ avec $\eta^2 = \sum_1^N (y_n - \bar{y})^2$. La procédure utilise ensuite les résultats obtenus pour déterminer les composantes significatives en tenant compte de leur valeur prédictive. Nous ajustons d'abord un modèle pour les variables standardisées

$$\hat{\underset{\sim}{y}}^s = \hat{\beta}_0^s \underset{\sim}{1} + Z^{(P)} \hat{\underset{\sim}{\gamma}}^s, \tag{1.4.1}$$

puis nous le transformons ensuite pour enlever la standardisation.

Nous calculons donc les valeurs propres λ_i et les vecteurs propres normalisés $\underset{\sim}{\delta}_i = (\delta_{0i}, \delta_{1i}, \delta_{2i}, \ldots, \delta_{Pi})'$ de $A'A$ afin d'obtenir les coordonnées des observations selon les nouvelles composantes:

$$\underset{\sim}{w}_i = \delta_{0i}\underset{\sim}{y}^s + \delta_{1i}\underset{\sim}{z}_1 + \ldots + \delta_{Pi}\underset{\sim}{z}_P, \quad i = 0, \ldots, \text{rang}(A).$$

Posons à partir de maintenant rang(A)=I. Pour une valeur propre λ_i, nous avons la relation suivante

$$\lambda_i = \underset{\sim}{\delta}_i' A' A \underset{\sim}{\delta}_i = (A\underset{\sim}{\delta}_i)'(A\underset{\sim}{\delta}_i),$$

qui découle de (1.3.2) et du fait que $\underset{\sim}{\delta}_i'\underset{\sim}{\delta}_i = 1$. Notons que $A\underset{\sim}{\delta}_i$ est donnée par

$$A\underset{\sim}{\delta}_i = \begin{pmatrix} y_1^s\delta_{0i} + \sum_{p=1}^P z_{1p}\delta_{pi} \\ y_2^s\delta_{0i} + \sum_{p=1}^P z_{2p}\delta_{pi} \\ \vdots \\ y_N^s\delta_{0i} + \sum_{p=1}^P z_{Np}\delta_{pi} \end{pmatrix} = \underset{\sim}{w}_i \quad i = 0, \ldots, I. \tag{1.4.2}$$

Donc, nous pouvons écrire de nouveau λ_i sous la forme d'une sommation en utilisant (1.4.2):

$$\lambda_i = \sum_{n=1}^{N}(y_n^s \delta_{0i} + \sum_{p=1}^{P} z_{np}\delta_{pi})^2 = \underset{\sim}{w}_i' \underset{\sim}{w}_i.$$

À partir de cette dernière équation, nous pouvons conclure que si λ_i est près de zéro alors $\underset{\sim}{w}_i$ l'est aussi. Cependant, si le coefficient δ_{0i} est différent de zéro alors la composante a une valeur prédictive:

$$\underset{\sim}{y}^s \approx \frac{1}{\delta_{0i}}(\delta_{1i}\underset{\sim}{z}_1 + \delta_{2i}\underset{\sim}{z}_2 + \ldots + \delta_{Pi}\underset{\sim}{z}_P).$$

Ainsi la composante peut être utile pour prévoir Y et elle n'est pas éliminée. Par contre, si $\lambda_i \approx 0$ et $\delta_{0i} \approx 0$ alors nous pouvons écrire, sans perte de généralité,

$$\underset{\sim}{z}_1 \approx \frac{1}{\delta_{1i}}(\delta_{2i}\underset{\sim}{z}_2 + \delta_{3i}\underset{\sim}{z}_3 + \ldots + \delta_{Pi}\underset{\sim}{z}_P)$$

et donc, $\underset{\sim}{z}_1$ est une combinaison linéaire des autres variables et nous pouvons éliminer la composante sans perdre trop d'information puisque celle-ci est peu utile pour expliquer Y: elle ne sert principalement qu'à expliquer des corrélations entre les variables explicatives.

Avec les composantes retenues, nous obtenons un estimateur de $\underset{\sim}{\gamma}^s$ pour le modèle (1.4.1). Pour justifier cette estimation de $\underset{\sim}{\gamma}^s$, Gunst, Webster et Mason (1974) utilisent le vecteur défini en (1.4.2) afin de déterminer $I+1$ estimations de $\underset{\sim}{y}^s$ en supposant d'abord que $\delta_{0i} \neq 0, \forall i$:

$$\underset{\sim}{\hat{y}}^s = -\delta_{0i}^{-1} Z^{(P)} \underset{\sim}{\delta}_i^0$$

ou, si nous enlevons la standardisation,

$$\underset{\sim}{\hat{y}}^{(i)} = \bar{y}\underset{\sim}{1} - \eta\delta_{0i}^{-1} Z^{(P)} \underset{\sim}{\delta}_i^0, \quad i = 0, \ldots, I.$$

Ces $I+1$ équations sont ensuite combinées pour fournir une unique estimation de Y

$$\underset{\sim}{\hat{y}} = \sum_{i=1}^{I} c_i \delta_{0i} \underset{\sim}{\hat{y}}^{(i)}.$$

Nous faisons une moyenne pondérée des $\underset{\sim}{\hat{y}}^{(i)}$ trouvés. Les poids sont arbitraires (δ_{0i} est fixe, mais les c_i sont quelconques), sauf que leur somme doit être égale à 1. Les coefficients c_i sont ceux minimisant la somme des erreurs au carré, $\sum_1^N (y_n - \hat{y}_n)^2 = \eta^2 \sum_0^I c_i^2 \lambda_i^2$, sous la contrainte $\sum_i c_i \delta_{0i} = 1$, c'est-à-dire

$$c_i = \frac{\delta_{0i} \lambda_i^{-1}}{\sum_{j=0}^{I} \delta_{0j}^2 \lambda_j^{-1}}.$$

Le même raisonnement est ensuite appliqué au cas où certaines composantes n'ont aucune valeur prédictive; ce qui conduit à l'estimateur $\underset{\sim}{\hat{\gamma}}^s$ suivant:

$$\underset{\sim}{\hat{\gamma}}^s = \sum_{i=0}^{I} f_i \underset{\sim}{\delta}_i^0,$$

où $\underset{\sim}{\delta}_i^0 = (\delta_{1i}, \delta_{2i}, \ldots, \delta_{Pi})'$

$$f_i = \begin{cases} 0 & \text{si } \lambda_i \approx 0 \text{ et } \delta_{0i} \approx 0 \\ \frac{-\eta \delta_{0i} \lambda_i^{-1}}{\sum_q \delta_{0q}^2 \lambda_q^{-1}} & \text{sinon,} \end{cases}$$

avec $\eta^2 = \sum_1^N (y_n - \bar{y})^2$.

La sommation est effectuée sur les q composantes sélectionnées. Une composante est éliminée si elle n'a qu'une faible valeur prédictive et qu'elle explique surtout la corrélation entre les variables; ce qui est le cas lorsque la valeur propre et le premier élément du vecteur propre associés à cette composante sont tous deux petits. Pour déterminer quelles composantes sont éliminées, un seuil doit être fixé pour les valeurs propres et pour le premier élément du vecteur propre.

Puis, toutes les composantes dont ces deux paramètres sont inférieurs (en valeur absolue) aux deux seuils sont éliminées. Pour fixer les seuils des paramètres, Webster, Gunst et Mason (1974) ont proposé de prendre $\lambda_i \leq 0{,}05$ et $|\delta_{0i}| \leq 0{,}10$. Les seuils critiques pour les paramètres peuvent être déterminés également par la VC.

Le modèle estimé est

$$\underset{\sim}{\hat{y}} = \hat{\beta}_0^s \underset{\sim}{1} + Z\underset{\sim}{\hat{\gamma}}^s,$$

où $\hat{\beta}_0^s = \bar{y}$.

Nous voulons maintenant exprimer le modèle en fonction des variables non standardisées, il faut donc modifier les coefficients obtenus.

$$\begin{aligned}
\hat{y}_n &= \hat{\beta}_0^s + \sum_{p=1}^{P} z_{np}\hat{\gamma}_p^s \\
&= \hat{\beta}_0^s + \sum_{p=1}^{P} \frac{x_{np} - \bar{x}_p}{\tau_p}\hat{\gamma}_p^s, \quad \text{où } \tau_p^2 = \sum_{n=1}^{N}(x_{np} - \bar{x}_p)^2 \\
&= (\hat{\beta}_0^s - \sum_{p=1}^{P} \frac{\bar{x}_p\hat{\gamma}_p^s}{\tau_p}) + \sum_{p=1}^{P} x_{np}\frac{\hat{\gamma}_p^s}{\tau_p}.
\end{aligned}$$

Ainsi, nous obtenons les estimations des coefficients

$$\begin{aligned}
\hat{\beta}_0 &= \hat{\beta}_0^s - \sum_{p=1}^{P} \frac{\bar{x}_p\hat{\gamma}_p^s}{\tau_p}, \\
\hat{\gamma}_p &= \frac{\hat{\gamma}_p^s}{\tau_p} \quad p = 1, \ldots, P,
\end{aligned}$$

et nous estimons finalement $\underset{\sim}{y}$ par $X\underset{\sim}{\hat{\beta}}^{LRR}$, où $\underset{\sim}{\hat{\beta}}^{LRR} = [\hat{\beta}_0, \underset{\sim}{\hat{\gamma}}']'$. Les propriétés de la distribution de cet estimateur sont inconnues: aucune expression pour l'espérance et la variance de $\hat{\beta}^{LRR} = (\hat{\beta}_0, \underset{\sim}{\hat{\gamma}}')'$ n'ont été développées jusqu'à présent.

LRR devrait donner de meilleurs résultats que la méthode précédente dans la mesure où la valeur prédictive des composantes est vérifiée à l'aide d'un paramètre supplémentaire avant de les éliminer, contrairement à PCR. De plus, LRR tient compte de la collinéarité. En se basant sur les valeurs propres obtenues, les méthodes PCR et LRR donnent une mesure indiquant si un estimateur biaisé devrait être utilisé plutôt que la RLM. Il y a deux paramètres à fixer, ce qui augmente cependant le temps de calcul et peut expliquer que LRR est une méthode peu utilisée jusqu'à ce jour. La meilleure manière de trouver les seuils critiques pour λ_j et δ_{0j} n'est pas connue, tout comme les propriétés de la distribution de l'estimateur.

1.5. Moindres carrés partiels

La méthode des moindres carrés partiels (PLS, de l'anglais "partial least squares" ou "projection to latent structure") reprend l'idée des méthodes PCR et LRR, en ce sens qu'elle trouve également des composantes et construit un modèle de prévision à partir de celles-ci. Cependant, les composantes sont trouvées différemment. Avec PLS, nous cherchons des composantes expliquant la variation en X, mais aussi celle en Y. Plusieurs algorithmes sont disponibles pour effectuer la régression, nous en présentons deux ici qui sont équivalents. Le premier, tiré de l'article de Garthwaite (1994), permet de comprendre les étapes de la procédure PLS. Les sections 1.5.1.1 et 1.5.1.2 s'inspirent principalement de cet article. L'auteur développe d'abord la procédure univariée, puis celle multivariée. Le second algorithme est celui trouvé le plus fréquemment dans la littérature et que nous retrouvons, entre autres, dans l'article de Höskuldsson (1988). Il est présenté ici dans le cas multivarié.

1.5.1. Algorithme 1

1.5.1.1. *Cas univarié*

Avec PLS, nous estimons $\underset{\sim}{y}$ à partir de variables non observables qui sont appelées "variables latentes". Ces variables sont considérées comme des facteurs cachés expliquant la corrélation entre les variables explicatives et dépendantes. Ces variables sont extraites des données observées. Chacune d'elles sera non corrélée avec toutes les autres variables latentes. La procédure PLS univariée ajuste le modèle suivant

$$Y = q_0 + q_1 T_1 + \ldots + q_K T_K + f = \mathcal{T}\underset{\sim}{q} + f. \tag{1.5.1}$$

Les T_i sont les variables latentes et f représente les erreurs du modèle. Les variables latentes sont estimées à partir des données observées, elles sont notées $\underset{\sim}{t_i}$ et sont des combinaisons linéraires des variables $X_1, \ldots, X_M$. Ainsi, les prévisions sont

$$\underset{\sim}{\hat{y}} = \hat{q}_0 \underset{\sim}{1} + \hat{q}_1 \underset{\sim}{t_1} + \ldots + \hat{q}_K \underset{\sim}{t_K}.$$

Supposons qu'il y ait M variables explicatives et N observations. Dans cette section, nous noterons M le nombre de variables explicatives afin qu'il n'y ait pas de confusion avec les éléments de la matrice P qui sera définie plus tard. Nous voudrions que chacune des composantes $\underset{\sim}{t_i}$ explique le mieux possible la variable Y de telle sorte que nous ayons à en conserver le moins possible. Le nombre de composantes à retenir, K, est déterminé, ici, par la validation croisée comme expliqué à la section 1.2.2, il existe cependant d'autres méthodes dans la littérature, par exemple, dans l'article de Lazraq et Cléroux (2000). La composante $\underset{\sim}{t_i}$ est calculée à la i^{e} étape et il y a au plus M composantes, soit le nombre de variables de départ. Les K composantes sont trouvées de manière itérative.

Chacune des variables est d'abord centrée par rapport à sa moyenne et nous obtenons

$$\begin{aligned} \underset{\sim}{v}_{1m} &= \underset{\sim}{x}_m - \bar{x}_m \underset{\sim}{1}, \quad m = 1, ..., M \\ \underset{\sim}{u}_1 &= \underset{\sim}{y} - \bar{y}\underset{\sim}{1}. \end{aligned} \tag{1.5.2}$$

Le vecteur $\underset{\sim}{v}_{1m}$ représente la variable V_m à l'étape 1 de la procédure itérative, de même que $\underset{\sim}{u}_1$ représente U_1 à la première étape. Les vecteurs $\underset{\sim}{v}_m$ et $\underset{\sim}{u}$ seront modifiés à chaque itération. De façon générale, l'indice i, associé aux différentes variables dans les algorithmes qui suivent, indique la i^e étape de la procédure.

Les variables explicatives sont toutes des prédicteurs possibles pour U_1. Nous pourrions donc ajuster un modèle de la forme suivante pour chacune de ces variables:

$$\underset{\sim}{u}_1 = \beta_m \underset{\sim}{v}_{1m} + \underset{\sim}{\epsilon}, \quad m = 1, ..., M.$$

Il n'y a pas de terme constant puisque $\underset{\sim}{u}$ et $\underset{\sim}{v}_{1m}$ sont centrées à l'origine. À partir de chaque vecteur $\underset{\sim}{v}_{1m}$, nous obtenons ainsi une estimation de $\underset{\sim}{u}_1$, notée $\hat{\underset{\sim}{u}}_1^{(m)}$. Sous forme matricielle, l'estimation de $\underset{\sim}{u}_1$ est la suivante:

$$\hat{\underset{\sim}{u}}_1^{(m)} = \underset{\sim}{v}_{1m}(\underset{\sim}{v}'_{1m}\underset{\sim}{v}_{1m})^{-1}\underset{\sim}{v}'_{1m}\underset{\sim}{u}_1.$$

Étant donné que chacune des variables prévoit U_1, la moyenne pondérée de chacune des prévisions obtenues $\sum_1^M w_{1m}\hat{\underset{\sim}{u}}_1^{(m)}$ devrait être un bon estimateur de U_1 également. Les poids w_{im} peuvent être choisis de plus d'une façon. Premièrement, nous pouvons donner le même poids à chaque $\hat{\underset{\sim}{u}}_i^{(m)}$. La prévision devient donc une moyenne simple: $\frac{1}{M}\sum_1^M \hat{\underset{\sim}{u}}_i^{(m)}$. Nous pouvons également considérer des poids proportionnels à la variance de $\underset{\sim}{v}_{im}$. De cette façon, les composantes peuvent être exprimées comme étant des combinaisons linéaires des $\underset{\sim}{v}_{im}$ avec des coefficients

proportionnels à la covariance entre $\underset{\sim}{v}_{im}$ et $\underset{\sim}{u}_i$, comme nous le verrons plus loin. Avec cette seconde méthode, les poids deviennent:

$$w_{im} = \underset{\sim}{v}'_{im}\underset{\sim}{v}_{im}. \tag{1.5.3}$$

La différence entre des poids constants et des poids proportionnels à la variance est que les premiers donnent des estimations invariantes par rapport aux changements d'échelles contrairement aux seconds qui ne sont invariants que par rapport aux transformations orthogonales de la matrice X. Cependant, la première méthode est peu utilisée puisque des simulations ont montré que la seconde donnait de meilleurs résultats généralement. C'est donc cette seconde méthode qui sera utilisée.

La moyenne pondérée obtenue sera la première composante du modèle, i.e. $\underset{\sim}{t}_1$:

$$\underset{\sim}{t}_1 = \sum_{m=1}^{M} w_{1m}\hat{\underset{\sim}{u}}_1^{(m)}.$$

La première composante est donc une combinaison linéaire des prévisions trouvées à l'aide de chacune des variables.

Pour trouver la seconde composante, nous utilisons la même procédure que celle utilisée pour trouver $\underset{\sim}{t}_1$, mais en changeant $\underset{\sim}{u}_1$ et les $\underset{\sim}{v}_{1m}$, car nous voulons enlever l'influence de $\underset{\sim}{t}_1$. Le modèle (1.5.1) peut être réécrit de la façon suivante:

$$\begin{aligned} \underset{\sim}{y} &= q_0\underset{\sim}{1} + q_1\underset{\sim}{t}_1 + q_2\underset{\sim}{t}_2 + ... + q_K\underset{\sim}{t}_K + \underset{\sim}{f} \\ &= q_0\underset{\sim}{1} + q_1\underset{\sim}{t}_1 + \underset{\sim}{f}^*. \end{aligned}$$

La variabilité de la variable dépendante qui n'est pas déjà expliquée par $\underset{\sim}{t}_1$ est celle contenue dans $\underset{\sim}{f}^*$. Donc, pour trouver la seconde composante, le vecteur $\underset{\sim}{y}$ est remplacé par $\hat{\underset{\sim}{f}}^*$, les résidus de la régression de $\underset{\sim}{y}$ sur $\underset{\sim}{t}_1$. De la même façon,

nous ne voulons pas considérer la variabilité des variables explicatives contenues dans $\underset{\sim}{t}_1$. Chacun des $\underset{\sim}{x}_i$ est donc remplacé par les résidus de la régression de $\underset{\sim}{x}_i$ sur $\underset{\sim}{t}_1$. Ce qui précède est équivalent à définir $\underset{\sim}{u}_2$ et $\underset{\sim}{v}_{2m}$ comme suit:

$$\begin{aligned}\underset{\sim}{u}_2 &= \underset{\sim}{u}_1 - (\underset{\sim}{t}_1'\underset{\sim}{u}_1)/(\underset{\sim}{t}_1'\underset{\sim}{t}_1)\ \underset{\sim}{t}_1 \\ \underset{\sim}{v}_{2m} &= \underset{\sim}{v}_{1m} - p_{1m}\underset{\sim}{t}_1, \quad m = 1, ..., M, \text{ où } p_{1m} = \underset{\sim}{t}_1'\underset{\sim}{v}_{1m}/\underset{\sim}{t}_1'\underset{\sim}{t}_1.\end{aligned}$$

De façon générale, lorsque nous avons $\underset{\sim}{u}_i$ et $\underset{\sim}{v}_{im}$, la composante $\underset{\sim}{t}_i$ est trouvée de la façon suivante:

$$\begin{aligned}\underset{\sim}{t}_i &= \sum_{m=1}^{M} w_{im}\hat{\underset{\sim}{u}}_i^{(m)} \\ &= \sum_{m=1}^{M} w_{im}b_{im}\underset{\sim}{v}_{im} \qquad (1.5.4) \\ &= \sum_{m=1}^{M} (\underset{\sim}{v}_{im}'\underset{\sim}{u}_i)\underset{\sim}{v}_{im}, \qquad (1.5.5)\end{aligned}$$

où $b_{im} = (\underset{\sim}{v}_{im}'\underset{\sim}{u}_i)/(\underset{\sim}{v}_{im}'\underset{\sim}{v}_{im})$ et w_{im} est donné par (1.5.3). La composante $\underset{\sim}{t}_i$ est donc une combinaison linéaire des $\underset{\sim}{v}_{im}$ comme le montre l'équation (1.5.5). Les coefficients associés aux $\underset{\sim}{v}_{im}$, $\underset{\sim}{v}_{im}'\underset{\sim}{u}_i$, sont proportionnels à la covariance entre $\underset{\sim}{v}_{im}$ et $\underset{\sim}{u}_i$. Donc, plus la covariance sera forte entre $\underset{\sim}{v}_{im}$ et $\underset{\sim}{u}_i$ et plus le coefficient associé à $\underset{\sim}{v}_{im}$ sera élevé.

Puis, $\underset{\sim}{u}_{i+1}$ et $\underset{\sim}{v}_{(i+1)m}$ sont déterminés tels que:

$$\begin{aligned}\underset{\sim}{u}_{i+1} &= \underset{\sim}{u}_i - (\underset{\sim}{t}_i'\underset{\sim}{u}_i)/(\underset{\sim}{t}_i'\underset{\sim}{t}_i)\ \underset{\sim}{t}_i \\ \underset{\sim}{v}_{(i+1)m} &= \underset{\sim}{v}_{im} - p_{im}\underset{\sim}{t}_i, \quad \text{où } p_{im} = \underset{\sim}{t}_i'\underset{\sim}{v}_{im}/\underset{\sim}{t}_i'\underset{\sim}{t}_i,\ m = 1, ..., M. \qquad (1.5.6)\end{aligned}$$

La procédure est répétée jusqu'à ce qu'un critère d'arrêt soit atteint (par exemple, si la norme du vecteur des résidus $\underset{\sim}{f}$ est plus petite qu'un seuil fixé) ou que

chacun des $\underset{\sim}{v}_{(i+1)m}$ soit le vecteur nul. Si nous utilisons la validation croisée, nous choisirons le nombre de composantes qui minimise la somme des erreurs au carré sur un jeu de données de validation. Le modèle est ensuite estimé à l'aide de la RLM en faisant une régression de $\underset{\sim}{y}$ en fonction des composantes $\underset{\sim}{t}_i$ et nous estimons $\underset{\sim}{q}$ dans (1.5.1) par

$$\underset{\sim}{\hat{q}} = (T'T)^{-1}T'\underset{\sim}{y}.$$

De cette procédure, nous conservons chacun des $\underset{\sim}{t}_i$, w_{im}, p_{im} et b_{im}, ainsi que la moyenne de chacune des variables $\underset{\sim}{x}_m$ afin de pouvoir effectuer des prévisions pour de nouvelles observations.

Pour faire des prévisions à partir de nouvelles observations, il faut d'abord trouver les nouvelles composantes $\underset{\sim}{t}_i^*$ associées aux nouvelles observations. Elles sont déterminées à l'aide des équations (1.5.2), (1.5.4) et (1.5.6). Puis, l'observation est estimée par $T^*\underset{\sim}{\hat{q}}$. Notons la matrice des nouvelles observations X^* qui est de taille $n^* \times M, n^* \geq 1$. La première composante $\underset{\sim}{t}_1^*$ est donnée par

$$\underset{\sim}{t}_1^* = \sum_{m=1}^{M} w_{1m}b_{1m}\underset{\sim}{v}_{1m}^*, \quad \text{où } \underset{\sim}{v}_{1m}^* = \underset{\sim}{x}_m^* - \bar{x}_m\underset{\sim}{1}.$$

Les éléments $\bar{x}_m$, w_{1m} et b_{1m}, $m = 1, \ldots, M$, ont été déterminés précédemment au cours de l'ajustement du modèle. L'indice * indique que l'élément est déterminé à partir des nouvelles observations, tandis que l'absence de cet indice indique que l'élément a été déterminé en ajustant le modèle. Pour les autres composantes, nous utilisons d'abord l'équation (1.5.6) afin de trouver $v_{(i+1)m}^*$

$$v_{(i+1)m}^* = v_{im}^* - p_{im}\underset{\sim}{t}_i^*, \quad m = 1, \ldots, M.$$

Ainsi, pour obtenir v^*_{2m}, nous utilisons v^*_{1m} et $\underset{\sim}{t}^*_1$ trouvés à l'étape précédente pour obtenir

$$\underset{\sim}{v}^*_{2m} = \underset{\sim}{v}^*_{1m} - p_{1m}\underset{\sim}{t}^*_1, \quad m = 1, \ldots, M.$$

La composante $\underset{\sim}{t}^*_{i+1}$ est ensuite calculée à partir de (1.5.4) et $v^*_{(i+1)m}$

$$\underset{\sim}{t}^*_{i+1} = \sum_{m=1}^{M} w_{(i+1)m} b_{(i+1)m} \underset{\sim}{v}^*_{(i+1)m}.$$

Si nous mettons ensemble toutes les étapes précédentes, nous pouvons trouver directement les composantes en fonction de X^*. Notons d'abord

$$\underset{\sim}{x}^*_0 = \underset{\sim}{1} \text{ et } d_{im} = w_{im} b_{im}$$

afin de simplifier la notation. La première composante peut être réécrite

$$\underset{\sim}{t}^*_1 = \sum_{j=0}^{M} a_{1j} \underset{\sim}{x}^*_j,$$

où

$$a_{1j} = \begin{cases} -\sum_{m=1}^{M} d_{1m}\bar{x}_m & \text{si } j = 0 \\ d_{1m} & \text{si } j = 1, \ldots, M \end{cases}$$

$$d_{i0} = -\sum_{m=1}^{M} d_{im}\bar{x}_m \quad i = 1, \ldots, M.$$

Les autres composantes sont trouvées de façon itérative

$$\underset{\sim}{t}^*_i = \sum_{j=0}^{M} \left(d_{ij} - \sum_{m=1}^{M} \sum_{k=1}^{i-1} d_{im} p_{km} a_{kj} \right) \underset{\sim}{x}^*_j = \sum_{j=0}^{M} a_{ij} \underset{\sim}{x}^*_j, \quad i = 2, \ldots, K.$$

La prévision est finalement trouvée

$$\hat{Y}^* = T^* \underset{\sim}{\hat{q}},$$

où $T^* = [\underset{\sim}{1}, \underset{\sim}{t}^*_1, \ldots, \underset{\sim}{t}^*_K]$. Il est facile, en se servant des coefficients a_{ij}, d'exprimer les prévisions en fonction des variables initiales.

1.5.1.2. *Cas multivarié*

Supposons, cette fois-ci qu'il y a L variables dépendantes. Pour chaque variable dépendante, un modèle de la forme suivante est ajusté:

$$\begin{aligned} \hat{\underset{\sim}{y}}_l &= \hat{q}_{l0}\underset{\sim}{1} + \hat{q}_{l1}\underset{\sim}{t}_1 + \ldots + \hat{q}_{lK}\underset{\sim}{t}_K = T\hat{\underset{\sim}{q}}_l, \quad l = 1, ..., L \text{ et } L \geq 2 \\ \hat{Y} &= T\hat{Q}, \quad \text{où } \hat{Q} = [\hat{\underset{\sim}{q}}_1, \ldots, \hat{\underset{\sim}{q}}_L]. \end{aligned} \tag{1.5.7}$$

La matrice $\hat{Y}$ est de taille $N \times L$ alors que $\hat{Q}$ et T sont respectivement de taille $(K+1) \times L$ et $N \times (K+1)$.

Les procédures PLS univariée et multivariée ne diffèrent que dans la façon dont $\underset{\sim}{u}_i$ est défini. Pour appliquer la procédure univariée, il faut que $\underset{\sim}{u}_i$ soit un vecteur. Nous ne pouvons donc pas utiliser la procédure univariée sans modification puisque, ici, Y n'est pas un vecteur, mais une matrice.

Nous définissons $\underset{\sim}{r}_{1l} = \underset{\sim}{y}_l - \bar{y}_l\underset{\sim}{1}$, $l = 1, ..., L$ et, comme dans le cas univarié, $\underset{\sim}{v}_{1m} = \underset{\sim}{x}_m - \bar{x}_m\underset{\sim}{1}$, $m = 1, ..., M$. Nous obtenons $R_1 = (\underset{\sim}{r}_{11}, ..., \underset{\sim}{r}_{1L})$ et $V_1 = (\underset{\sim}{v}_{11}, ..., \underset{\sim}{v}_{1M})$. Nous définissons ensuite $\underset{\sim}{u}_1 = R_1\underset{\sim}{c}_1$, où $\underset{\sim}{c}_1$ est le vecteur propre associé à la plus grande valeur propre de la matrice $R_1'V_1V_1'R_1$. Le choix de cette matrice est justifié dans Höskuldsson (1988). Nous appliquons ensuite la procédure décrite dans le cas univarié avec $\underset{\sim}{u}_1$ et V_1 pour obtenir $\underset{\sim}{t}_1$.

Pour le cas général, lorsque nous avons $\underset{\sim}{t}_i$, nous effectuons les étapes suivantes pour trouver $\underset{\sim}{u}_{i+1}$ et V_{i+1}:

Étape 1. $\underset{\sim}{r}_{(i+1)l} = \underset{\sim}{r}_{il} - (\underset{\sim}{t}_i'\underset{\sim}{r}_{il})/(\underset{\sim}{t}_i'\underset{\sim}{t}_i)\underset{\sim}{t}_i$, l=1, ... , L

Étape 2. Nous trouvons $\underset{\sim}{c}_{i+1}$, le vecteur propre associé à la plus grande valeur propre de $R_{i+1}'V_{i+1}V_{i+1}'R_{i+1}$

Étape 3. $\underset{\sim}{u}_{i+1} = R_{i+1}\underset{\sim}{c}_{i+1}$

Étape 4. Les colonnes de la matrice V_{i+1} sont définies par l'équation (1.5.6).

Nous itérons ensuite avec la procédure définie plus haut pour trouver la composante suivante. La matrice Q est ensuite trouvée en ajustant le modèle (1.5.7) par les moindres carrés. Les prévisions sont effectuées de la même façon que pour la méthode univariée.

La procédure multivariée de PLS sera utilisée de préférence lorsque les variables dépendantes sont corrélées, même si nous pourrions appliquer la procédure univariée pour chacune d'entre elles. Si deux variables sont non corrélées alors une composante $\underset{\sim}{t}_i$ expliquant bien la première variable n'aura que peu d'efficacité pour prévoir la seconde. Une composante ne sera éliminée que si elle n'a aucune valeur prédictive pour toutes les variables dépendantes. La procédure multivariée ne sera donc pas efficace dans une telle situation puisqu'elle retiendra beaucoup de composantes. Dans le cas où les variables dépendantes sont faiblement corrélées, il vaut mieux appliquer la procédure univariée.

1.5.2. Algorithme 2

Cet algorithme est présenté sous la forme simplifiée, mais le raisonnement est le même que pour l'algorithme de Garthwaite (1994). Dans son article, Garthwaite démontre que les deux procédures conduisent aux mêmes résultats. Le modèle ajusté utilise des variables reliant X et Y qui sont appellées "variables latentes":

$$\begin{aligned} Z^{(M)} &= TP + E, \quad \text{où } T = [\underset{\sim}{t}_1, \ldots, \underset{\sim}{t}_K] \\ Y^s &= TQ + F. \end{aligned}$$

Les $\underset{\sim}{t}_i$, qui sont les colonnes de la matrice T de taille $N \times K$, sont les nouvelles composantes qui représentent la variation en X ainsi que celle en Y. Les matrices $Z^{(M)}$ et Y^s sont les matrices des variables standardisées explicatives et

dépendantes respectivement. Nous conservons la moyenne de chacune des variables $\bar{x}_m$ et $\bar{y}_l$ ainsi que $\tau_m^2 = \sum_{n=1}^{N}(x_{nm} - \bar{x}_m)^2$ et $\eta_l^2 = \sum_1^N(y_{nl} - \bar{y}_l)^2$ pour $m = 1, \ldots, M$ et $l = 1, \ldots, L$. Rappelons que M est le nombre de variables explicatives et L, le nombre de variables dépendantes. Les erreurs de chacun des modèles sont représentées par E et F. Les matrices P et Q sont les "loadings" du modèle et sont respectivement de taille $K \times M$ et $K \times L$.

L'algorithme présenté dans l'article de Höskuldsson (1988) est le suivant:

1) Poser $i = 1$, $E_1 = Z^{(M)}$ et $F_1 = Y^s$

À la i^{e} itération:

2) Poser $\underset{\sim}{u}_i = \underset{\sim}{f}_1$, où $\underset{\sim}{f}_1$ est la première colonne de la matrice F_i

3) Jusqu'à convergence de $\underset{\sim}{t}_i$:

$$\begin{aligned}
\underset{\sim}{w}_i &\leftarrow E_i'\underset{\sim}{u}_i / \underset{\sim}{u}_i'\underset{\sim}{u}_i \\
\underset{\sim}{w}_i &\leftarrow \underset{\sim}{w}_i / \|\underset{\sim}{w}_i\| \\
\underset{\sim}{t}_i &\leftarrow E_i\underset{\sim}{w}_i \\
\underset{\sim}{c}_i &\leftarrow F_i'\underset{\sim}{t}_i / \underset{\sim}{t}_i'\underset{\sim}{t}_i \\
\underset{\sim}{c}_i &\leftarrow \underset{\sim}{c}_i / \|\underset{\sim}{c}_i\| \\
\underset{\sim}{u}_i &\leftarrow F_i\underset{\sim}{c}_i
\end{aligned}$$

4) Loadings du modèle

$$\begin{aligned}
X\text{-loadings:} \quad \underset{\sim}{p}_i &= E_i'\underset{\sim}{t}_i / \underset{\sim}{t}_i'\underset{\sim}{t}_i \\
Y\text{-loadings:} \quad \underset{\sim}{q}_i &= b_i\underset{\sim}{c}_i, \text{ où } b_i = \underset{\sim}{u}_i'\underset{\sim}{t}_i / \underset{\sim}{t}_i'\underset{\sim}{t}_i
\end{aligned}$$

5) Matrices des résidus

$$E_{i+1} = E_i - \underset{\sim}{t}_i\underset{\sim}{p}_i'$$

$$F_{i+1} = F_i - b_i \underset{\sim}{t}_i \underset{\sim}{c}_i'$$

La procédure (étapes 2 à 5) recommence avec les nouvelles matrices E_{i+1} et F_{i+1}, définies à l'étape 5, jusqu'à ce qu'un critère d'arrêt soit atteint, par exemple si la norme de chacune des colonnes de E_{i+1} est plus petite qu'un seuil fixé, ou que $E_{i+1} = 0_{N \times M}$. Encore une fois, le nombre de composantes à retenir peut être déterminé par la validation croisée. À chaque itération, il faut conserver $\underset{\sim}{p}_i$, $\underset{\sim}{q}_i$ et $\underset{\sim}{t}_i$. À la fin, les matrices P, Q et T sont formées, leur i^{e} colonne représentent, respectivement, $\underset{\sim}{p}_i$, $\underset{\sim}{q}_i$ et $\underset{\sim}{t}_i$ obtenus à la i^{e} itération de l'algorithme.

Pour effectuer des prévisions sur de nouvelles observations X^*, il faut d'abord standardiser les données avec les paramètres des observations ayant servi à estimer le modèle, i.e. $\bar{x}_m$ et τ_m^2. Il faut ensuite trouver T^*. Les composantes $\underset{\sim}{p}_i$ sont orthogonales entre elles, de telle sorte que $P'P = I$. Donc, nous pouvons trouver T^* pour une nouvelle observation de la façon suivante:

$$T^* = Z^{(M)*} P.$$

Ensuite, les prévisions sont obtenues par

$$\hat{Y}^{s*} = T^* Q.$$

Nous pouvons également exprimer directement $\hat{Y}^{s*}$ en fonction de $Z^{(M)*}$

$$\hat{Y}^{s*} = Z^{(M)*} PQ.$$

Il ne reste ensuite qu'à enlever l'effet de la standardisation avec les paramètres $\bar{y}_l$ et η_l^2 pour obtenir $\hat{Y}^*$.

PLS se rapproche de PCR et LRR. Il utilise des composantes tout en tenant compte de la variabilité des variables dépendantes et indépendantes. Cependant, tout comme LRR, les propriétés des estimateurs des moindres carrés partiels sont

inconnues. PLS offre une bonne capacité de prévision qui, d'après la littérature, est comparable à celle de la régression ridge. PLS devrait donner de bons résultats, d'après Garthwaite (1994), lorsqu'il y a beaucoup de variables et que ces dernières ont une grande variance. D'après Frank et Friedman (1993), PLS retiendra moins de composantes que PCR en sélectionnant celles-ci avec la VC. PLS n'utilise pas que l'information contenue dans la matrice X, mais aussi celle de la matrice Y ce qui conduit à un meilleur ajustement. Avec PLS, un modèle avec k composantes sera moins biaisé qu'un modèle avec PCR avec le même nombre de composantes, par contre, il sera plus variable.

Pour identifier l'efficacité des modèles présentés dans les sections précédentes, plusieurs études ont été effectuées à partir de données réelles et simulées. Pour n'en mentionner que quelques-unes, l'article de Frank et Friedman (1993) contient une comparaison des méthodes RLM, PCR, PLS et RR avec différents niveaux de collinéarité; Webster, Gunst et Mason (1974) comparent LRR et RLM avec des données simulées à partir d'un modèle connu; Garthwaite (1994) étudie PLS avec des poids égaux et avec des poids donnés par (1.5.3) ainsi que RLM et PCR avec des données simulées pour différents nombre de variables explicatives. Dans tous les articles, il y a toujours plus d'observations que de variables explicatives étant donné qu'il y a des comparaisons avec la RLM qui nécessite cette condition. De façon générale, la méthode qui donne les "meilleurs" résultats est la RR suivie de près par PLS. La méthode la moins efficace est la RLM. Pour PCR et LRR, les auteurs concluent d'après leurs études que LRR est meilleur que PCR. Donc, d'après les différents articles, lorsqu'il y a de la collinéarité parmi les données, les méthodes les plus fiables devraient être RR et PLS.

1.6. MÉTHODE NON LINÉAIRE

1.6.1. Hypothèse de linéarité

Les méthodes utilisées nécessitent toutes, à l'exception des réseaux de neurones, l'hypothèse de la linéarité de la relation entre les variables explicatives et dépendantes. L'hypothèse de cette relation linéaire est importante, car elle met une grande contrainte sur le modèle et les résultats obtenus peuvent être très mauvais si elle est fausse. Il peut exister une relation parfaite entre deux variables, mais si celle-ci n'est pas linéaire, les résultats obtenus par un modèle qui, lui, est linéaire ne seront pas bons. Dans le cas d'un modèle où il n'y a qu'une seule variable explicative une relation linéaire est représentée comme à la figure 1.6.1:

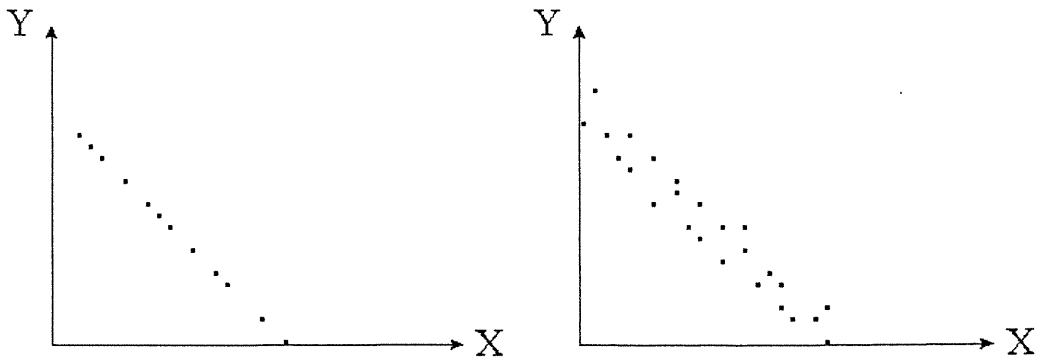


FIGURE 1.6.1. *Relation linéaire entre deux variables. À gauche: relation linéaire parfaite. À droite: relation linéaire bruitée.*

Par exemple, nous pourrions essayer de modéliser la croissance d'une population en fonction du temps. Prenons les premiers éléments de la suite de Fibonacci, énumérés au tableau 1.6.1, pour illustrer cette situation.

Temps	Population
0	0
1	1
2	1
3	2
4	3
5	5
6	8
7	13
8	21
9	34

TABLEAU 1.6.1. *Suite de Fibonacci*

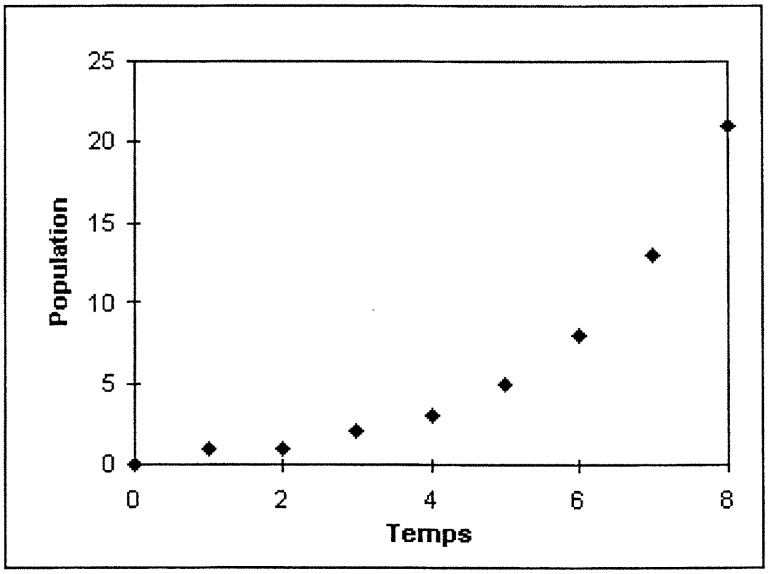


FIGURE 1.6.2. *Suite de Fibonacci*

Ici, le modèle est clairement non linéaire (voir la figure 1.6.2), alors l'utilisation d'une fonction de la forme

$$Taille = a_0 + a_1 Temps$$

ne donnera pas une bonne estimation de la population. Donc, si un modèle linéaire ne donne pas de bons résultats, cela n'implique pas qu'il n'y a pas de relation entre les variables, mais plutôt que le modèle ne trouve pas de relation parmi le type de relations qu'il a considéré. Pourquoi alors utiliser des modèles mettant de telles contraintes alors que l'hypothèse de linéarité est bien souvent difficile à vérifier? Parce que ces méthodes linéaires sont plus faciles à appliquer et à comprendre. Nous utilisons habituellement les méthodes non linéaires lorsque nous voulons un modèle et qu'il a déjà été vérifié que les méthodes linéaires ne donnent pas de bons résultats.

Si nous reprenons l'exemple de la suite de Fibonnaci, nous pourrions considérer que la relation est localement linéaire jusqu'au temps 7 et ajuster un modèle avec ces données, comme représenté à la figure 1.6.3. La droite semble faire une assez bonne approximation des données. Localement, les prévision seront bonnes. Si nous voulons prévoir la population au temps 8 alors la prévision obtenue est loin d'être précise. Nous constatons que plus nous nous éloignons du voisinage [0, 7], plus la prévision est sous-estimée. Un modèle avec une fonction non linéaire serait peut-être plus approprié.

Ce petit exemple illustre bien l'importance de l'hypothèse de linéarité. Un des principaux désavantages des modèles non linéaires est qu'il faut déterminer la forme du modèle que nous estimons. Ainsi, pour la suite de Fibonacci, nous pourrions décider d'ajuster un modèle quadratique (qui est linéaire par rapport aux coefficients du modèle) ou alors un modèle exponentiel. Ce choix est très important, car il a un grand impact sur le modèle estimé. De plus, les méthodes non linéaires ne permettent pas d'estimer les paramètres directement. Il faut optimiser des fonctions non linéaires et cette opération ne s'effectue que par des approximations. Nous rencontrons immanquablement le problème des maximums

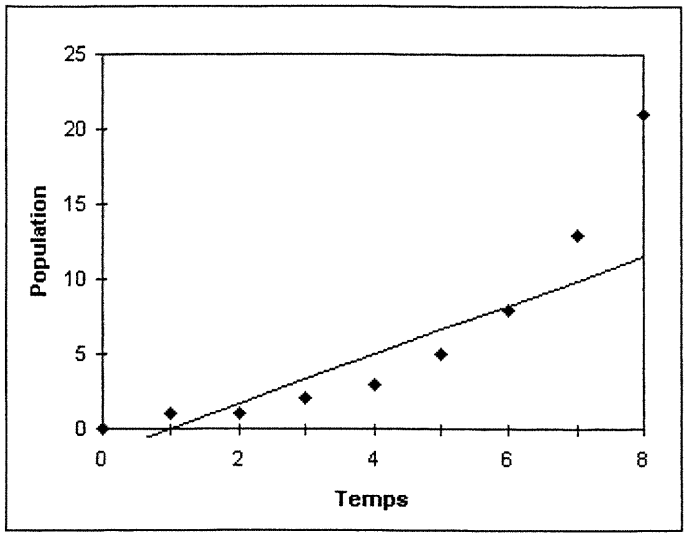


FIGURE 1.6.3. *Ajustement d'un modèle linéaire pour la suite de Fibonacci*

et des minimums locaux. Il faut donc choisir un algorithme et en vérifier les résultats afin de s'assurer que la solution trouvée est bien celle qui est optimale. Donc, il y a de nombreux problèmes, ce qui a comme conséquence que les modèles linéaires sont plus couramment utilisés.

1.6.2. Les réseaux de neurones

1.6.2.1. *Construction du modèle*

Les méthodes abordées précédemment utilisent toutes des modèles linéaires pour expliquer les variables dépendantes. Les réseaux de neurones n'ont pas cette restriction: ils peuvent être composés d'éléments linéaires et non linéaires. Leur plus grande flexibilité permet souvent une meilleure efficacité au niveau des prévisions. Nous pouvons également entraîner des réseaux de neurones aussi bien pour la prévision que pour la classification. Cependant, les réseaux de neurones (RN) ont pour désavantage de fournir des modèles généralement difficiles à interpréter. De plus, étant donné que la théorie derrière cette méthode n'est pas très connue,

les RN sont souvent associés à des "boîtes noires" dans lesquelles se produisent quelques tours de passe-passe. Nous présentons dans cette section ce qui se passe à l'intérieur de ces "boîtes noires".

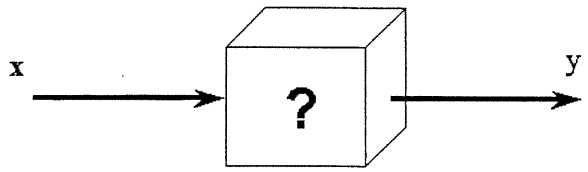


FIGURE 1.6.4. *La boîte noire: l'image la plus souvent associée aux réseaux de neurones.*

L'analogie entre les réseaux de neurones et le cerveau est souvent faite. Par exemple, nos différentes perceptions du monde extérieur envoient un signal qui est traité par les neurones et se propage d'un neurone à un autre à travers le cerveau et il en résulte une réaction. Dans les RN, les observations traversent le réseau en se propageant à travers les unités cachées (neurones) et la sortie est la prévision que nous souhaitons obtenir. Un RN peut être représenté avec un schéma comme la figure 1.6.5. La présente section s'inspire des livres de Bishop (1995) et de Haykin (1999) ainsi que d'un cours dont les notes sont disponibles à l'adresse électronique "http://www.iro.umontreal.ca/~bengioy/ift6266/".

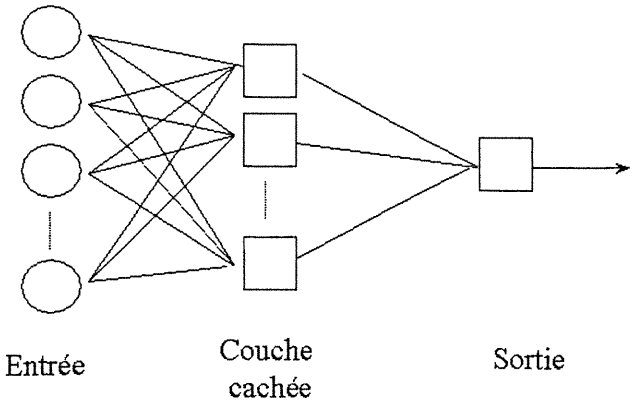


FIGURE 1.6.5. *Réseau de neurones à une couche cachée.*

La première couche représente les entrées et la dernière, les sorties. Les couches intermédiaires sont appelées "couches cachées" et elles contiennent des "unités cachées" ou des neurones. Chaque unité cachée traite l'information à sa manière et envoie le résultat à la couche suivante qui traite à son tour l'ensemble de l'information qu'elle reçoit. Un réseau de neurones avec une seule couche cachée est considéré comme un approximateur universel en ce sens que, s'il y a une infinité de données, alors, avec un nombre assez grand d'unités cachées, il est possible de faire une excellente approximation de n'importe quelle fonction qu'elle soit linéaire ou non (Haykin, 1999, p.208). Cependant, en pratique, l'obtention d'une bonne approximation nécessite parfois un modèle plus complexe. Souvent, un réseau de neurones avec deux couches cachées est suffisamment complexe pour donner une bonne approximation des variables dépendantes.

En pratique, nous pouvons fixer le nombre de couches cachées avant d'ajuster le modèle, par exemple deux couches cachées, puis déterminer par VC le nombre de neurones optimal à inclure dans chacune d'elle. Nous pouvons aussi considérer le nombre de couches comme un autre paramètre à déterminer par VC, mais il ne faut pas oublier que le temps de calcul augmente considérablement à chaque fois que nous ajoutons un paramètre supplémentaire à optimiser.

Supposons $\underset{\sim}{x} = (x_1, x_2, ..., x_P)$, le vecteur des observations à l'entrée et $\underset{\sim}{y} = (y_1, y_2, ..., y_L)$, les variables dépendantes observées. Chaque unité cachée transforme les données qu'elle reçoit:

$$H_i = w_{i0} + \sum_{p=1}^{P} w_{ip} x_p.$$

Puis, elle envoie aux unités de la couche suivante une transformation non linéaire de H_i, la fonction utilisée est appelée "fonction d'activation" et elle peut être différente d'une couche à une autre. L'inverse de la fonction d'activation est appelée

"fonction de lien". Les principales fonctions de lien et leur fonction d'activation correspondante sont fournies au tableau 1.6.2. La fonction d'activation la plus souvent utilisée est la tangente hyperbolique:

$$tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}.$$

Avec cette fonction d'activation, si les valeurs de x ne sont pas dans l'intervalle (-1, 1), $tanh(x)$ vaut toujours 1. Ceci peut entrainer des difficultés au niveau de l'estimation du modèle, c'est pourquoi il vaut mieux transformer les données d'entrée de telle sorte qu'elles soient dans cet intervalle. Une façon de faire ceci est de standardiser les observations de chacune des variables, mais en divisant par trois écarts type au lieu d'un seul. Une fois transformées, nous utilisons les nouvelles observations pour estimer le modèle, celles-ci auront une moyenne de zéro et un variance d'une unité. Dans la suite de cette section, nous noterons la p^e entrée du réseau par x_p, mais elle représentera désormais les observations de la variable X_p transformées.

Donc, s'il y a une seule couche cachée, les unités de la couche de sortie reçoivent comme entrées (tanh(H_1), tanh(H_2), ..., tanh(H_I)), où I est le nombre d'unités cachées. Elles traitent l'information de la même façon que la couche précédente. Les prévisions à la sortie de réseau sont

$$\hat{y}_l = f_2(w^*_{l0} + \sum_{i=1}^{I} w^*_{li} f_1(w_{i0} + \sum_{p=1}^{P} w_{ip} x_p)), \tag{1.6.1}$$

où les w^*_{li} sont les poids entre la couche cachée et la couche de sortie et f_1 et f_2 sont les fonctions d'activation du réseau, f_1 est la fonction liée à la couche cachée et f_2 est celle liée à la couche de sortie du réseau. La fonction d'activation à la couche de sortie dépend bien souvent des contraintes que nous voulons mettre sur

Fonction de lien	Fonction d'activation	Contrainte
identité $f(z) = z$	identité $f(x) = x$	aucune ($x \in \mathbb{R}$)
exponentielle $f(z) = e^z$	logarithme $f(x) = ln(x)$	$x \in (0, \infty)$
sigmoïde $f(z) = \frac{1}{1+e^z}$	logistique $f(x) = ln(\frac{x}{1-x})$	$x \in (0, 1)$
tangente hyperbolique inverse $f(z) = arctanh(z)$	tangente hyperbolique $f(x) = tanh(x)$	aucune ($x \in \mathbb{R}$)

TABLEAU 1.6.2. *Fonctions de lien les plus couramment utilisées pour construire les réseaux de neurones avec leur fonction d'activation correspondante et les contraintes qu'elles posent sur les sorties du réseau.*

les variables de sortie. Afin d'ajuster le modèle, il faut estimer chacun des poids du réseau.

Les RN contiennent plusieurs modèles connus. Ainsi, la RLM est un RN sans couche cachée et avec la fonction d'activation identité tel que représenté à la figure 1.6.6. Le modèle

$$\hat{y} = w_0 + \sum_{p=1}^{P} w_p x_p$$

est le même modèle que celui donné par la RLM avec une seule variable dépendante, où les β_p sont remplacés par les poids w_p.

1.6.2.2. *Descente du gradient*

Pour un nombre de couches cachées et un nombre de neurones par couche fixes, les inconnues dans un réseau de neurones sont les poids que nous retrouvons dans chacune des unités du réseau. Afin d'optimiser l'efficacité du modèle, par

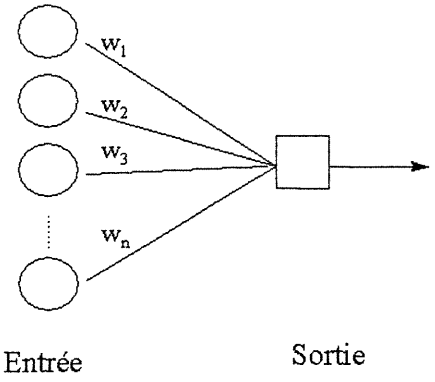


FIGURE 1.6.6. *Réseau de neurones représentant la régression linéaire multiple.*

rapport à une fonction de coût, par exemple la somme des erreurs au carré, l'algorithme de propagation à reculons (de l'anglais "back propagation algorithm") a été développé. Cet algorithme utilise la méthode de descente du gradient qui consiste à modifier récursivement la solution dans la direction où la descente est la plus abrupte:

$$\underset{\sim}{w}^{[k+1]} \leftarrow \underset{\sim}{w}^{[k]} + \varepsilon^{[k]} \left.\frac{\partial C}{\partial w}\right|_{\underset{\sim}{w}^{[k]}}, \quad k = 0, 1, 2, \ldots . \tag{1.6.2}$$

La fonction C (fonction de perte ou de coût) représente le coût associé aux erreurs effectuées. Plus l'erreur entre la prévision et la "vraie" réponse est grande, plus la perte augmente. Par exemple, la fonction de perte peut être définie par la fonction suivante:

$$C(\underset{\sim}{x}, \underset{\sim}{y}) = \frac{1}{2} \sum_{l=1}^{L} \sum_{n=1}^{N} (y_l^{(n)} - \hat{y}_l^{(n)})^2.$$

où l'indice (n) indique que l'élément est associé à la n^{e} observation. À une constante près, cette fonction représente la somme des erreurs au carré, c'est-à-dire la même fonction qui est minimisée pour ajuster les modèles linéaires.

Le paramètre ε contrôle le pas de gradient. Il est souhaitable que ε diminue à mesure que le nombre d'étapes augmente. Ainsi, plus nous nous approchons de la solution optimale, plus nous avançons à petits pas. Ceci peut éviter de "passer tout droit" ou de faire des étapes inutiles à cause du fait que nous tournons autour de la solution sans l'atteindre, car le pas de gradient est trop grand. Dans certains cas, le pas de gradient peut être une matrice au lieu d'un scalaire comme nous le verrons avec la méthode de Newton.

L'algorithme de propagation à reculons commence par calculer les gradients à la fin du réseau et utilise ensuite les résultats trouvés pour trouver les autres gradients en se dirigeant vers le début. Développons l'algorithme dans le cas où il n'y a qu'une seule couche cachée, comme à la figure 1.6.7.

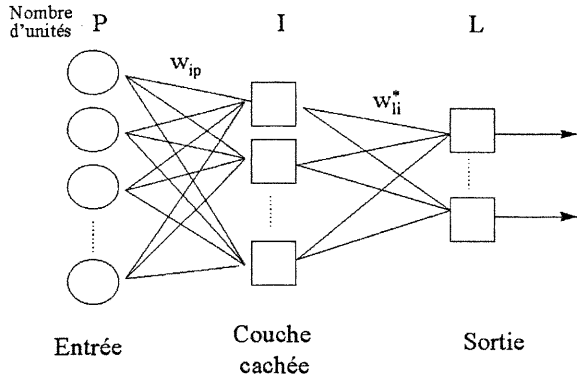


FIGURE 1.6.7. *Nombre de neurones et identification des poids pour un réseau avec une couche cachée.*

Supposons que la fonction d'activation utilisée est f_1 pour la couche d'entrée et f_2 pour la couche de sortie.

$$\text{Couche cachée} \quad : \quad H_i^{(n)} = w_{i0} + \sum_{p=1}^{P} w_{ip} x_p^{(n)} \text{ et } z_i^{(n)} = f_1(H_i^{(n)})$$

$$\text{Couche de sortie} \quad : \quad Z_l^{(n)} = w_{l0}^* + \sum_{i=1}^{I} w_{li}^* z_i^{(n)} \text{ et } \hat{y}_l^{(n)} = f_2(Z_l^{(n)}),$$

où l'indice (n) indique que l'élément dépend de la n^{e} observation. L'information qui traverse les couches cachées est la même pour toutes les variables dépendantes, ce n'est qu'au dernier niveau que les entrées sont combinées différemment pour chacune des variables de sortie. La fonction de coût est définie de la façon suivante:

$$C(\underset{\sim}{w}_i, \underset{\sim}{w}_l^*) = \frac{1}{2}\sum_n \sum_l (y_l^{(n)} - f_2(Z_l^{(n)}))^2, \tag{1.6.3}$$

où $\underset{\sim}{w}_i=(w_{i0}, w_{i1}, \ldots, w_{iP})$ et $\underset{\sim}{w}_l^*=(w_{l0}, w_{l1}, \ldots, w_{lI})$.

Pour calculer le gradient, il faut dériver la fonction de coût par rapport à chacun des paramètres. Les fonctions f_1 et f_2 étant connues, f_1' et f_2' le sont également. Les dérivées de la fonction de coût par rapport aux paramètres sont trouvées en commençant par la dernière couche:

$$\begin{aligned}
\frac{\partial C}{\partial w_{li}^*} &= -\sum_n (y_l^{(n)} - f_2(Z_l^{(n)}))f_2'(Z_l^{(n)})\frac{\partial Z_l^{(n)}}{\partial w_{li}^*} \\
&= -\sum_n \Lambda_l^{(n)} z_i^{(n)}, \quad \text{où } \Lambda_l^{(n)} = (y_l^{(n)} - f_2(Z_l^{(n)}))f_2'(Z_l^{(n)}) \\
\frac{\partial C}{\partial w_{l0}^*} &= -\sum_n (y_l^{(n)} - f(Z_l^{(n)}))f_2'(Z_l^{(n)})\frac{\partial Z_l^{(n)}}{\partial w_{l0}^*} \\
&= -\sum_n \Lambda_l^{(n)}.
\end{aligned}$$

Lorsque celles-ci sont calculées, elles sont conservées en mémoire afin d'accélérer le calcul des dérivées restantes. Nous calculons ensuite les dérivées de la couche précédant celle déjà calculée.

$$\begin{aligned}
\frac{\partial C}{\partial w_{ip}} &= -\sum_n \sum_l (y_l^{(n)} - f_2(Z_l^{(n)}))f_2'(Z_l^{(n)})\frac{\partial Z_l^{(n)}}{\partial z_i^{(n)}}\frac{\partial z_i^{(n)}}{\partial H_i^{(n)}}\frac{\partial H_i^{(n)}}{\partial w_{ip}} \\
&= -\sum_n \sum_l \Lambda_l^{(n)} w_{li}^* f_1'(H_i^{(n)}) x_p^{(n)}
\end{aligned}$$

$$
\begin{aligned}
&= -\sum_n \sum_l \Lambda_{li}^{*(n)} x_p^{(n)}, \quad \text{où } \Lambda_{li}^{*(n)} = \Lambda_l^{(n)} w_{li}^* f_1'(H_i^{(n)}) \\
\frac{\partial C}{\partial w_{i0}} &= -\sum_n \sum_l (y_l^{(n)} - f_2(Z_l^{(n)})) f_2'(Z_l^{(n)}) \frac{\partial Z_l^{(n)}}{\partial z_i^{(n)}} \frac{\partial z_i^{(n)}}{\partial H_i^{(n)}} \frac{\partial H_i^{(n)}}{\partial w_{i0}} \\
&= -\sum_n \sum_l \Lambda_l^{(n)} w_{li}^* f_1'(H_i^{(n)}) \\
&= -\sum_n \sum_l \Lambda_l^{*(n)}.
\end{aligned}
$$

Nous obtenons ainsi les gradients nécessaires pour calculer les estimations des paramètres. Si nous avons plus d'une couche cachée, le calcul des gradients pour les deux dernières couches (la dernière couche cachée et celle de sortie) serait le même que celui défini précédemment. Donc, il suffirait de faire les calculs pour les autres couches de la même façon.

1.6.2.3. *Méthode de Newton*

Il est important de choisir judicieusement le pas de gradient. Lorsque ε est petit, la convergence est lente, mais "lisse" alors que, lorsque ε est grand, l'algorithme se dirige vers l'optimum en zigzaguant, mais la convergence est rapide. Il faut faire attention à ce que ε ne soit pas trop grand, car la procédure peut alors diverger.

Plusieurs méthodes existent afin de déterminer le pas de gradient. Nous pouvons simplement fixer ε et conserver la même valeur tout au long de l'optimisation. Cependant, il serait intéressant de converger rapidement, i.e. un pas de gradient élevé serait approprié, mais en zigzaguant vers l'optimum, il y a des risques de ne jamais l'atteindre lorsque nous nous en approchons. Donc, si nous avions un grand pas de gradient au début et que nous le réduisions à chaque étape, la procédure serait plus performante que pour un ε fixe. La méthode de

Newton propose de prendre la matrice Hessienne comme pas de gradient

$$\underset{\sim}{w}^{[k+1]} \leftarrow \underset{\sim}{w}^{[k]} - \left(H^{-1} \frac{\partial C}{\partial \underset{\sim}{w}} \right) \Bigg|_{\underset{\sim}{w}^{[k]}} \quad , \text{ où } H_{ij} = \frac{\partial^2 C}{w_i w_j} \text{ et } k = 0, 1, 2, \ldots .$$

Avec cette méthode, chaque itération conduit les poids dans la direction d'un minimum, qui peut cependant être local.

Pour commencer l'algorithme, il faut prendre de petites valeurs initiales aléatoires. Les poids initiaux sont choisis de telle sorte que, pour la première couche du réseau, la fonction d'activation ne soit pas dans la zone de saturation, c'est-à-dire la région où elle vaut toujours la même valeur. Si cette contrainte n'est pas satisfaite, il peut y avoir des problèmes au niveau de l'algorithme d'optimisation. Pour la fonction *tanh*, la zone de saturation est à l'extérieur de l'intervalle (-1, 1) où la fonction vaut toujours 1. Dans la mesure où les observations sont transformées pour avoir une moyenne de zéro et une variance d'une unité, Bishop (1995) suggère de prendre des poids aléatoires petits avec la contrainte que leur somme soit égale à 1. Nous pouvons aussi prendre des poids uniformes avec une moyenne de zéro et une variance telle que la fonction d'activation de la première couche de soit pas dans la zone de saturation, cette méthode est suggérée par Hawkin (1999). Pour l'ajustement des RN avec nos données, les poids initiaux sont tirés d'une distribution uniforme sur l'intervalle [-1/$\sqrt{P}$, 1/$\sqrt{P}$], où P est le nombre de variables explicatives du modèle estimé, sauf pour les poids associés à un terme constant qui sont tirés d'une distribution uniforme sur [-0,1 , 0,1]. Pour minimiser le problème des optimums locaux, la procédure peut être effectuée plus d'une fois, avec des poids initiaux différents.

1.6.2.4. *Méthode du gradient conjugué*

D'après Haykin (1999), la matrice hessienne, utilisée comme pas de gradient dans l'algorithme de la descente de gradient, a un grand impact sur la procédure à cause du fait qu'elle a souvent de très petites et de très grandes valeurs propres. De plus, afin de pouvoir utiliser l'algorithme de Newton, celle-ci doit également être inversible. L'algorithme a tendance à zigzaguer vers un minimum de telle sorte que la convergence est lente, particulièrement dans le cas où le nombre de variables est grand. L'algorithme du gradient conjugué utilise le gradient, mais la procédure converge vers un minimum plus rapidement. De façon générale, cette procédure fait une recherche linéaire dans une direction en particulier afin de trouver le pas de gradient.

L'algorithme du gradient conjugué utilise une modification de l'équation (1.6.2). À chaque itération, les poids sont donnés par

$$\underset{\sim}{w}^{[k+1]} \leftarrow \underset{\sim}{w}^{[k]} + \varepsilon^{[k]}\underset{\sim}{d}^{[k]}, \quad k = 0, 1, 2, \ldots .$$

Dans cette équation, $\varepsilon^{[k]}$ représente le pas de gradient à l'étape k et $\underset{\sim}{d}^{[k]}$ est la direction vers laquelle les poids sont modifiés et celle-ci dépend du gradient dont le calcul a été expliqué à la section 1.6.2.2. Les poids initiaux sont choisis comme pour la méthode de Newton vue à la section 1.6.2.3.

La direction est d'abord initialisée à $\underset{\sim}{d}^{[0]} = -\underset{\sim}{g}^{[0]}$, où $\underset{\sim}{g}^{[0]}$ est le gradient de la fonction de coût évalué avec les poids initiaux aléatoires $\underset{\sim}{w}^{[0]}$. Le pas de gradient est déterminé comme étant celui qui minimise la fonction de coût dans la direction donnée par $\underset{\sim}{d}^{[0]}$. De façon générale, à l'étape k,

$$\varepsilon^{[k]} = \arg\min_{\varepsilon} C(\underset{\sim}{w}^{[k]} + \varepsilon\underset{\sim}{d}^{[k]}),$$

où C représente la fonction de coût et $\underset{\sim}{w}^{[k]}$ et $\underset{\sim}{d}^{[k]}$ sont fixes. Cette expression ne peut être calculée numériquement et nous devons donc en faire l'approximation à l'aide d'un algorithme d'optimisation. Lorsque nous avons obtenu le pas de gradient, nous pouvons calculer les poids $\underset{\sim}{w}^{[k+1]}$. Après avoir calculé chaque nouvelle série de poids, nous déterminons si la procédure s'arrête ou non en vérifiant si $\underset{\sim}{g}^{[k]}$ satisfait un critère fixé. Par exemple, la procédure peut être arrêtée lorsque la norme de $\underset{\sim}{g}^{[k]}$ représente un petit pourcentage fixé de la norme de $\underset{\sim}{g}^{[0]}$. Ainsi, lorsque la norme du gradient est très petite, les poids ne sont presque plus modifiés d'une étape à l'autre et la procédure s'arrête.

Pour effectuer l'étape suivante, nous devons calculer le gradient

$$\underset{\sim}{g}^{[k+1]} = \left.\frac{\partial C}{\partial \underset{\sim}{w}}\right|_{\underset{\sim}{w}^{[k+1]}}$$

et la nouvelle direction est une combinaison linéaire de la direction précédente et du gradient

$$\underset{\sim}{d}^{[k+1]} = \beta^{[k+1]}\underset{\sim}{d}^{[k]} - \underset{\sim}{g}^{[k+1]},$$

où $\beta^{[k+1]}$ est un facteur d'échelle défini par

$$\beta^{[k+1]} = max\left\{\frac{\underset{\sim}{g}'^{[k+1]}(\underset{\sim}{g}^{[k+1]} - \underset{\sim}{g}^{[k]})}{\underset{\sim}{g}'^{[k]}\underset{\sim}{g}^{[k]}}, 0\right\}.$$

Le chapitre 4 du livre de Fletcher (1987) fournit une justification pour cet estimateur nommé "l'estimateur de Polak-Ribière". La motivation part d'un contexte de minimisation d'une fonction de coût quadratique avec l'algorithme du gradient conjugué.

La méthode du gradient conjugué ne nécessite pas le calcul de la matrice hessienne, ce qui économise du temps et épargne les difficultés liées au calcul de

son inverse. Cette méthode est recommandée lorsque le nombre de variables est grand.

1.6.2.5. *Interprétabilité des réseaux de neurones*

L'équation (1.6.1) exprime les prévisions en fonction des variables dépendantes lorsque le réseau de neurones ne contient qu'une seule couche cachée. Avec l'ajout de couches cachées supplémentaires, l'expression des prévisions en fonction des variables dépendantes s'alourdit et se complexifie encore davantage. Les variables explicatives sont traitées sur plusieurs niveaux, ce qui rend difficile, voir même impossible, de déterminer l'effet d'une variable sur les prévisions. Cet impact n'est pas nécessairement monotone étant donné la fonction d'activation appliquée à la sortie de chaque neurone. Prenons l'exemple d'un RN à une couche cachée contenant deux neurones comme illustré à la figure 1.6.8. La fonction *tanh* est monotone croissante donc cette transformation ne change pas l'effet de H_1 et H_2 sur les prévisions. Dans H_1, la variable x_i peut avoir un effet positif ($w_{i1} > 0$) sur $\hat{y}$, alors que l'effet peut être négatif dans H_2 ($w_{i2} < 0$). L'effet total sera tantôt positif tantôt négatif. Si le modèle ajusté est

$$\hat{y} = tanh(-x_1 + x_2 - x_3) - tanh(-2x_1 + x_2 + x_3),$$

alors, pour $x_2 = x_3 = 0$, l'effet de x_1 sur $\hat{y}$ n'est pas linéaire comme nous pouvons le constater à la figure 1.6.9. Bref, l'effet n'étant ni linéaire ni monotone, il est difficile de l'interpréter. De plus, l'effet dépend également des valeurs prises par les autres variables dépendantes. La figure 1.6.10 nous montre que, lorsque x_2 vaut -2 au lieu de 0, l'effet de la variable x_1 est différent par rapport au cas précédent.

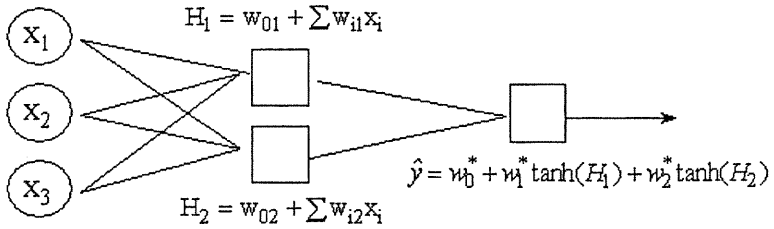


FIGURE 1.6.8. *Réseau de neurones univarié ayant une couche cachée contenant deux neurones et trois variables explicatives.*

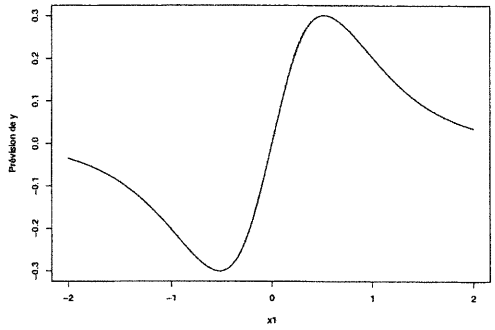


FIGURE 1.6.9. *Effet de la variable x_1 pour le modèle de prévision $\hat{y} = tanh(-x_1 + x_2 - x_3) - tanh(-2x_1 + x_2 + x_3)$ lorsque $x_2 = x_3 = 0$.*

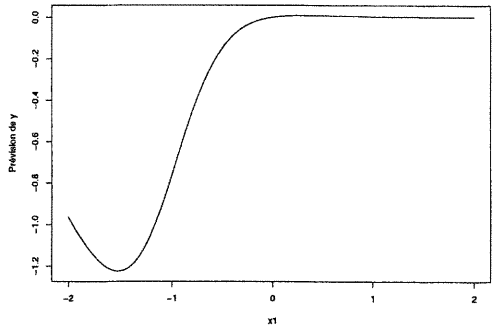


FIGURE 1.6.10. *Effet de la variable x_1 pour le modèle de prévision $\hat{y} = tanh(-x_1 + x_2 - x_3) - tanh(-2x_1 + x_2 + x_3)$ lorsque $x_2 = -2$ et $x_3 = 0$.*

1.6.2.6. *Conclusion*

Du moment que l'interprétabilité du modèle n'est pas une condition *sine qua non*, les RN sont une solution efficace pour effectuer des prévisions. Il est possible

de ne poser aucune contrainte sur le modèle de départ, ce qui permet une grande flexibilité. Bien sûr, l'obtention du modèle final nécessite du temps et des efforts, mais les résultats peuvent en valoir la peine. Les RN ne peuvent se targuer d'avoir l'efficacité du cerveau, cependant, à mesure que la théorie se développe, cette technique offre des possibilités de plus en plus intéressantes.

Chapitre 2

ANALYSE PRÉLIMINAIRE DU JEU DE DONNÉES DU TRAITEMENT DES EAUX USÉES

Dans ce second chapitre, nous verrons l'application des différentes méthodes introduites au chapitre 1 aux données du traitement des eaux usées de l'usine d'Abitibi-Consolidated. Dans un premier temps, nous présentons les différentes variables du jeu de données. Nous décrivons ensuite le traitement des valeurs manquantes qui a été fait avant les analyses, soit les variables et les observations qui ont été éliminées ainsi que la description des méthodes d'imputation utilisées.

2.1. DESCRIPTION DU JEU DE DONNÉES

Le jeu de données qui sera étudié contient des mesures prises quotidiennement sur un traitement des eaux usées à la sortie d'une usine de pâtes et papier d'Abitibi-Consolidated. L'eau est utilisée par l'usine dans le procédé de fabrication du papier. À la sortie de l'usine, avant d'être rejetée dans la rivière, l'eau polluée doit être traitée. Des normes environnementales sont à respecter, il est donc important d'avoir un bon contrôle sur les différentes variables du traitement. Le jeu de données fourni par les ingénieurs du traitement contenait au départ 175 variables mesurées à différents endroits dans le traitement: à l'affluent (i.e. juste

avant le traitement), à chacun des cinq réacteurs biologiques séquentiels (RBS) et à l'effluent (i.e. à la sortie du traitement, avant le rejet des eaux traitées dans la rivière). En plus de la valeur des variables dans chacun des cinq bassins, la moyenne de chaque variable par rapport aux bassins était également fournie. La figure 2.1.1 représente le traitement des eaux usées de l'usine. Pour les données mesurées à chacun des cinq bassins, la moyenne des observations a été utilisée puisqu'il n'y avait aucune raison de croire que les mesures prises dans un bassin en particulier pourraient mieux représenter les variables dépendantes que celles prises dans un autre. De plus, certaines variables n'étaient que des combinaisons linéaires d'autres variables de telle sorte qu'elles ont été éliminées également. Les variables restantes, ainsi que l'abbréviation par laquelle elles seront ensuite identifiées, sont fournies au tableau A.0.1 de l'annexe A. Il reste donc, à ce point, 90 variables dans le jeu de données. Les variables 1 à 3 sont mesurées avant le traitement, celles numérotées de 4 à 33 le sont à l'affluent, les variables 34 à 47 sont des moyennes pour les cinq bassins, tandis que les variables numérotées de 48 à 70 sont mesurées à l'effluent. Les variables restantes (71 à 90) sont mesurées un peu partout dans le traitement, mais sont, en majorité, des fonctions non linéaires des 70 premières variables, par exemple la DCO en T/j est, à une constante près, le produit de la DCO en mg/L et du débit à l'effluent. Aves toutes ces variables, nous sommes intéressés à construire un modèle pouvant prévoir les variables listées au tableau 2.1.1; leur description est donnée au tableau A.0.1.

2.2. Le traitement des valeurs manquantes

2.2.1. Méthodes d'imputation

Les méthodes d'analyse nécessitent souvent que le jeu de données à l'étude soit complet, c'est-à-dire que chaque observation soit observée pour chacune des

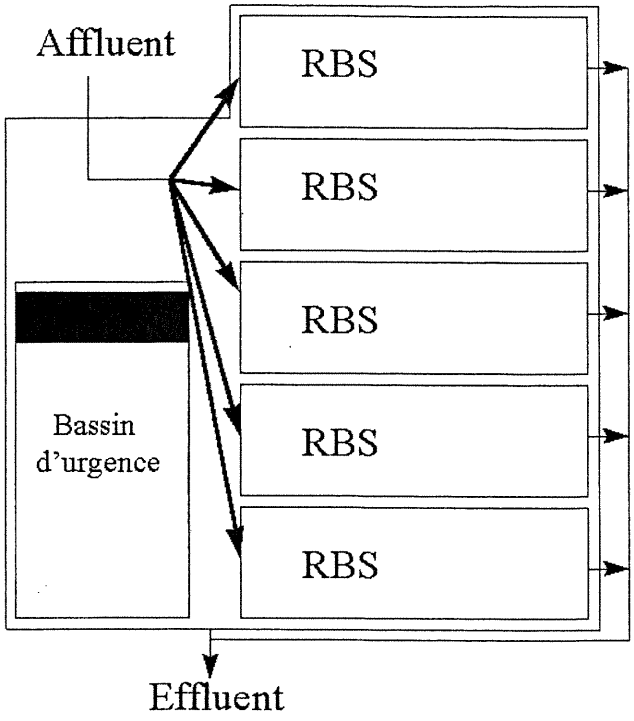


FIGURE 2.1.1. *Schéma du traitement des eaux usées.*

1. IVB_m	2. débit_e	3. DCO_e
4. DBO_lab_e	5. MES_e	6. MES_lab_e
7. N-NH_3_e	8. MESnd_e	9. O-PO4_e
10. turbid_e	11. pHcomp_e	12. pHmoy_e
13. Tmoy_e	14. cond_e	15. DCOsol_e

TABLEAU 2.1.1. *Liste des variables dépendantes des modèles dont la description est fournie au tableau A.0.1.*

variables. Bien souvent en pratique, différentes raisons ont pour conséquence que certaines données sont manquantes. Il est possible, par exemple, lors de la collecte, que certaines données ne puissent être observées à cause d'un bris d'équipement ou d'une erreur de manipulation ou, lors de la vérification, que certaines

valeurs soient considérées comme aberrantes et aient ensuite été effacées. Différentes techniques existent pour contourner ou remplacer les valeurs manquantes et effectuer l'analyse. Il ne faut cependant pas oublier d'en tenir compte lors de l'analyse des résultats, car, selon les cas, le biais et/ou la variance des estimateurs augmente(nt). Ceci est dû au fait que le remplacement des valeurs manquantes est effectué de façon à minimiser un facteur, par exemple minimiser les résidus d'un modèle, ce qui a un impact sur les estimations.

Lorsqu'il y a peu de valeurs manquantes, la méthode la plus simple à adopter consiste à éliminer tout simplement les observations contenant des valeurs manquantes. Ceci permet d'obtenir un jeu de données complet et, lorsque les observations sont indépendantes, les résultats ne sont pas faussés. Lorsqu'il y a trop de valeurs manquantes ou que les observations ne sont pas indépendantes, il faut avoir recours à d'autres méthodes.

Dans le cas où les données sont indépendantes et équidistribuées et que les observations manquantes se produisent de façon aléatoire, plusieurs méthodes d'imputation ont été proposées. Bello (1995) et Little (1992) présentent différentes méthodes permettant d'imputer des valeurs manquantes, tout comme dans l'article de Donner (1982) qui donne, en plus, des expressions pour le biais et la variance des estimateurs obtenus dans le cas d'un modèle particulier. La série d'articles d'Afifi et Elashoff (1966, 1967, 1969a, 1969b) présentent différentes méthodes, en plus de donner leurs propriétés asymptotiques et d'étudier le biais et l'efficacité pour de petits échantillons. Little et Rubin (1987) proposent différentes façons d'estimer les paramètres comme la moyenne multivariée ou la matrice de variances-covariances soit en imputant les valeurs manquantes soit en effectuant les calculs de façon à contourner celles-ci. Finalement, Efron (1996) présente l'utilisation du bootstrap afin d'effectuer des analyses et de produire

des estimations de la variance des estimateurs. Cependant, le problème demeure lorsque les données ne sont pas indépendantes et équidistribuées, comme c'est le cas avec le présent jeu de données.

2.2.2. Les valeurs manquantes du jeu de données

Tout d'abord, les données du traitement des eaux usées sont ramassées tous les jours et elles sont donc autocorrélées dans bien des cas. De plus, les variables sont corrélées entre elles. Les méthodes suggérées dans les articles mentionnés précédemment ne peuvent être utilisées efficacement. Ensuite, les valeurs manquantes ne sont pas aléatoires. Si nous regardons la disposition des valeurs manquantes représentée à la figure 2.2.1, nous remarquons qu'il y a plusieurs variables qui présentent un fort taux de valeurs manquantes pouvant aller jusqu'à 90%, certaines variables ne sont pas disponibles sur de longues périodes. Sur l'axe horizontal, les variables sont représentées par leur numéro d'indentification fourni au tableau A.0.1. L'axe vertical représente le temps. Chaque point du graphique est une valeur manquante. Les variables ayant un taux de valeurs manquantes supérieur à 10% sont données au tableau 2.2.1 avec leur taux respectif. Nous observons également un taux de défection très élevé durant les journées où l'usine était fermée.

Pour les variables TAO et TSUO (TAOfin_m, TSUOfin_m, TAOendo_m et TSUOendo_m), les données n'étaient que rarement mesurées par les employés de l'usine. Ces variables, d'après eux, ne sont pas essentielles pour assurer le bon fonctionnement du traitement ; elles ont donc été éliminées. Le débit de l'usine (débit_us), les MES aux émissaires 3, 4 et 5 (MESno345), le désencré (désencré), la pâte de meules (PMM), la pâte thermomécanique (PTM) et la pâte chimique

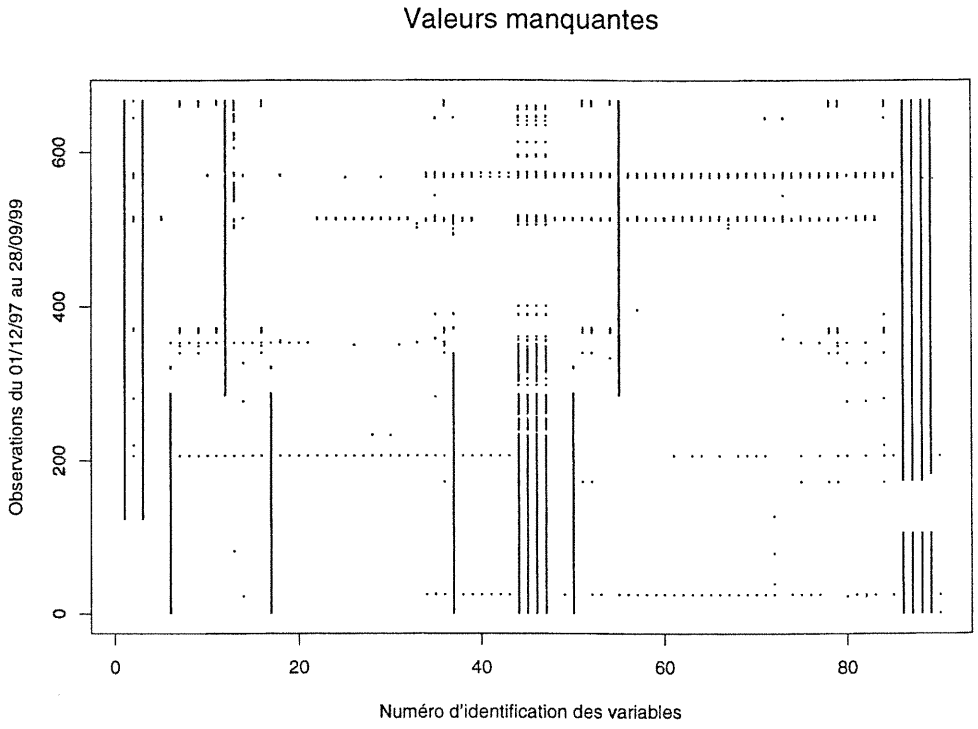


FIGURE 2.2.1. *Distribution des valeurs manquantes dans le jeu de données original du traitement des eaux usées en fonction des numéros d'identification des variables donnés au tableau A.0.1.*

à haut rendement (PCHR) ont également été éliminés puisque leur taux de valeurs manquantes était beaucoup trop élevé. Pour les variables DCO soluble à l'affluent (DCOsol_a), MVS à l'affluent (MVS_a) et MVS à l'effluent (MVS_e), nous pouvons constater à la figure 2.2.2 qu'elles sont fortement corrélées avec les variables DCO à l'affluent (corrélation de 0,98), MES à l'affluent (corrélation de 0,99) et MES à l'effluent (corrélation de 1,00) respectivement, donc, l'information étant déjà contenue dans ces variables, les premières ont été éliminées. La DCO soluble en T/j (DCOsol_T_a) a également dû être enlevée puisqu'elle était calculée à partir de la DCO soluble à l'affluent. Quant aux journées où la production

Variable	Taux de défection (%)	Variable	Taux de défection (%)
débit_us	82	MESno345	82
DCOsol_a	44	MVS_a	58
N-NH_3_a	10	DCOsol_T_a	44
N-NO_2-NO_3_m	55	TAOendo_m	54
TAOfin_m	51	TSUOendo_m	54
TSUOfin_m	51	DCOsol_e	45
MVS_e	58	désencré	90
PMM	90	PTM	90
PCHR	88		

TABLEAU 2.2.1. *Variables ayant un taux élevé de valeurs manquantes avec leur taux respectif.*

de l'usine était nulle, elles ne correspondent pas à ce que nous sommes intéressés à modéliser. Le modèle représente le traitement des eaux usées et ce qui s'y passe lorsque l'usine est arrêtée est de moindre intérêt ici. De plus, les données observées durant ces journées se comportent possiblement différemment des jours d'opération normale, alors mettre toutes ces données ensemble pour former un unique modèle ne donnerait peut-être pas d'aussi bons résultats. S'il devient intéressant de trouver un modèle pour les journées où l'usine est fermée alors il vaudrait mieux utiliser des données correspondant à de tels jours et bâtir un modèle différent de celui qui sera développé ici. Pour ce projet, les observations correspondant à ces journées ont été éliminées. La liste des journées qui n'ont pas été considérées est donnée au tableau 2.2.2, nous remarquons qu'il s'agit surtout de jours fériés ou de périodes de vacances. Une fois ces données et ces variables

disparues, le pourcentage de valeurs manquantes a considérablement diminué globalement, passant de 13% à 1%. Il reste ici 76 variables et 629 observations dans le jeu de données.

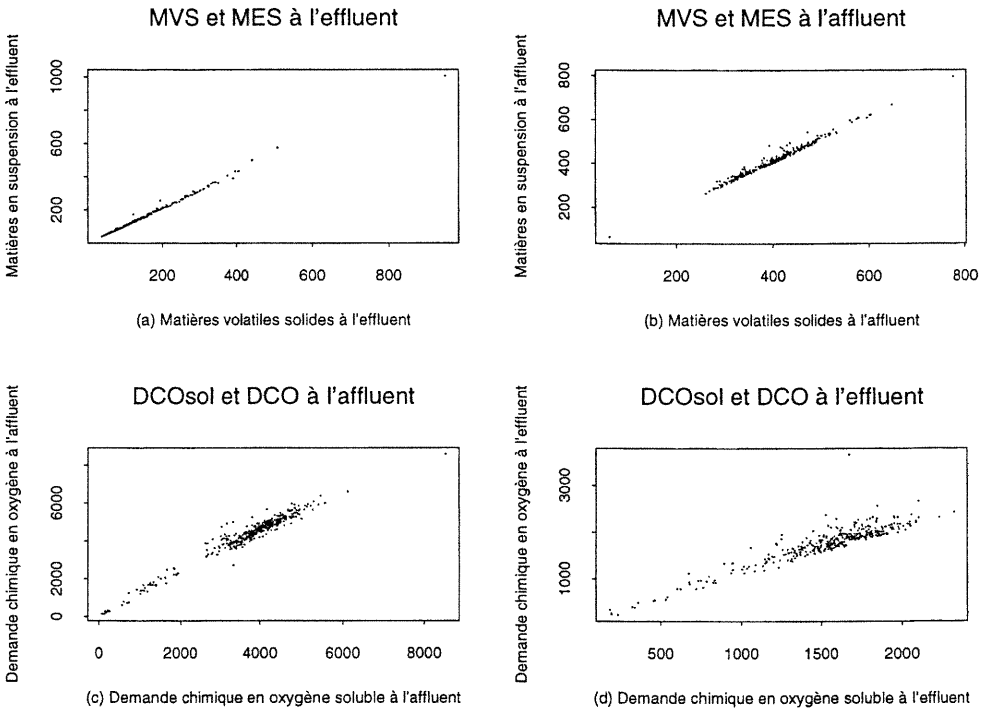


FIGURE 2.2.2. *(a) Relation entre les MVS et les MES à l'effluent, (b) Relation entre les MVS et les MES à l'affluent, (c) Relation entre la DCO soluble et la DCO à l'affluent, (d) Relation en la DCO soluble et la DCO à l'effluent.*

La distribution des données manquantes restantes est illustrée à la figure 2.2.3. Mises à part les variables N-NH_3 à l'affluent, N-NO_2-NO_3 et DCO soluble à l'effluent, il reste peu de valeurs manquantes. La forte corrélation entre les variables et l'autocorrélation ont été utilisées pour imputer les données manquantes. Si une variable est fortement autocorrélée, la donnée manquante peut être imputée à l'aide de l'observation de la journée précédente. Ce procédé a été utilisé

1997	1998	1999
23 au 25 décembre	22 au 24 juin	23 avril au 1er mai
	5 au 7 septembre	18 au 26 juin
	16 au 18 novembre	4 au 6 septembre
	23 au 25 décembre	

TABLEAU 2.2.2. *Liste des journées où la production de l'usine était nulle pendant la période du 01-12-97 au 26-09-99. Ces journées ont été exclues des analyses.*

Variable imputée	Autocorrélation de délai 1	Nombre de données imputées
DBO5_l_a	0,69	18
MESnd_a	0,53	3
MESbex	0,86	11
DBO5_l_e	0,82	17
vol_air	0,45	5

TABLEAU 2.2.3. *Liste des variables imputées à l'aide de l'autocorrélation de délai 1*

pour traiter les valeurs manquantes des variables données au tableau 2.2.3. Si une variable est fortement corrélée avec une autre variable alors une régression linéaire simple peut être utilisée pour prévoir les valeurs manquantes. Ainsi, les variables données au tableau 2.2.4 étaient fortement corrélées avec d'autres variables disponibles et la régression a été utilisée. Dans tous les cas, nous avons calculé l'autocorrélation et la corrélation avec les autres variables dépendantes et nous avons retenu la solution donnant le résultat le plus élevé.

Rendu à ce point, il reste le problème des variables N-NH_3 à l'affluent et N-NO_2-NO_3, qui sont des variables explicatives, et DCO soluble à l'effluent, qui est une variable dépendante. L'étude de la variable N-NH_3 à l'affluent montre que

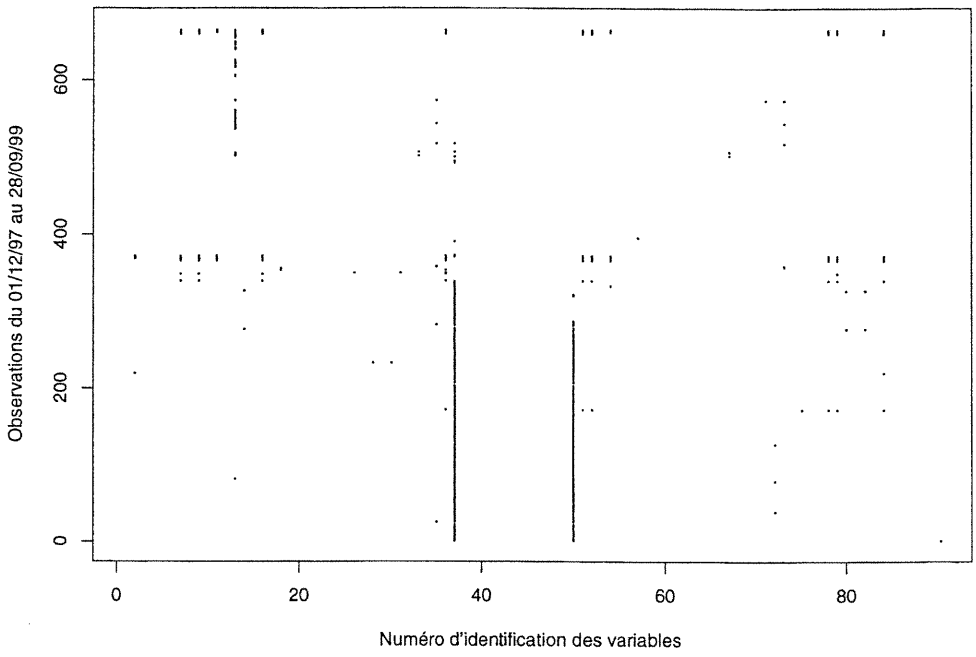


FIGURE 2.2.3. *Distribution des valeurs manquantes dans le jeu de données modifié du traitement des eaux usées en fonction des numéros d'identification des variables donnés au tableau A.0.1.*

sa corrélation avec chacune des autres variables est inférieure à 0,15. De plus, en faisant une régression pour prévoir chacune des variables dépendantes, nous constatons que N-NH_3 à l'affluent n'est significative dans aucun des cas. Donc, cette variables contenant 60 valeurs manquantes a été éliminée.

La variable DCO soluble à l'effluent est une variable dépendante et le but du modèle est donc de prévoir cette variable. Cependant, il faut que cette variable soit disponible pour construire le modèle et elle a 45% de valeurs manquantes. L'imputation de ces données a peu de sens dans ce cas étant donné que la DCO soluble n'était pas mesurée avant le 14 septembre 1998. En imputant 45% des

Variable imputée	Variable la plus fortement corrélée	Corrélation	Nombre de données imputées
DCO_a	DCOsol_e	0,79	1
MES_l_a	MES_a	0,88	13
Tmin_a	Tmoy_a	0,83	1
Tmoy_a	Tmin_a	0,83	1
Condmoya	Condcoma	0,83	1
Condmina	Condmoya	0,64	1
Condmoya	Condmoya	0,83	2
MES_l_e	MES_e	0,95	15
Tmin_e	Tmoy_e	0,72	2
Cond_e	Condmoye	0,91	2
bouessec	debitbex	0,72	9
F_Mba_m	DBO5_a	0,90	19

TABLEAU 2.2.4. *Liste des variables imputées à l'aide de la régression linéaire simple*

données afin de trouver un modèle permettant de prévoir la variable, il est fort probable que le résultat ne serait pas très bon pour généraliser, puisque l'imputation aurait introduit trop de variabilité dans le modèle. La variable DCO soluble à l'effluent sera donc traitée séparément et seule la période où elle est disponible sera utilisée. Cette variable n'était pas mesurée durant les premiers mois, mais, à partir du 14 septembre 1998, les données sont toutes disponibles.

2.2.3. Le traitement de la variable N-NO_2-NO_3 à l'affluent

Pour la variable N-NO_2-NO_3, une régression linéaire de cette variable en fonction de toutes les autres (avec seulement les données disponibles) donne un R^2

de 0,67, ce qui est faible si nous voulons imputer efficacement 350 données. C'est d'ailleurs le principal problème avec cette variable : le taux de valeurs manquantes est de 55%. Malheureusement, contrairement à N-NH_3 à l'affluent, N-NO_2-NO_3 est significative pour prévoir certaines des variables dépendantes d'après la RLM sur le modèle complet; l'effacer pourrait éliminer de l'information importante. La variable N-NO_2-NO_3 est distribuée selon la distribution donnée à la figure 2.2.4. Dans la mesure où la régression ne donnait pas un excellent résultat pour prévoir cette variable, nous l'avons transformée en variable dichotomique prenant les valeurs 0 ou 1 en mettant une valeur de 1 lorsque y vaut plus qu'un seuil fixé. En procédant ainsi, nous espérons trouver de meilleures prévisions pour cette variable. Donc, après avoir fait cette transformation, les valeurs manquantes sont estimées grâce à la régression logistique.

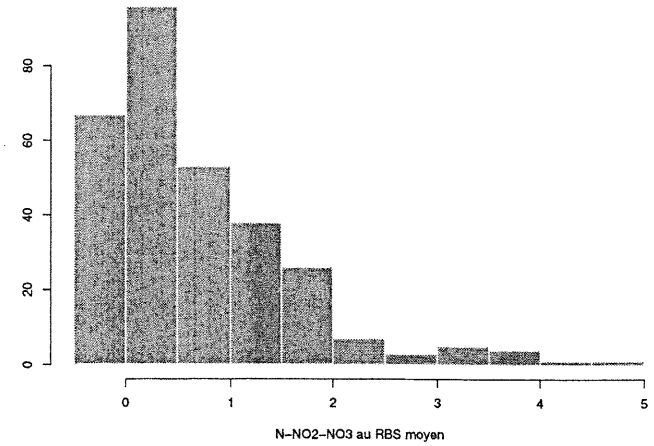


FIGURE 2.2.4. *Histogramme de la variable N-NO_2-NO_3*

Faisons un bref arrêt sur la régression logistique. Supposons que $p(Y)$ est la probabilité que la variable Y soit égale à un succès, i.e. $P(Y = 1)$, dans le cas où Y est une variable dichotomique prenant ses valeurs dans l'ensemble $\{0, 1\}$. Les

variables explicatives sont des variables quantitatives ou qualitatives. Le modèle de régression logistique est

$$logit(p(\underset{\sim}{x})) = log\left(\frac{p(\underset{\sim}{x})}{1-p(\underset{\sim}{x})}\right) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p.$$

Notons que $logit(\underset{\sim}{x})$ prend ses valeurs dans l'ensemble des réels alors que $p(\underset{\sim}{x})$ appartient à l'intervalle [0, 1]. Le modèle est ajusté en minimisant le même critère que pour la RLM, soit la somme des erreurs au carré. Cependant, contrairement à la RLM, la solution optimale ne peut pas être déterminée algébriquement; il faut utiliser des méthodes numériques d'approximation. Lorsque le modèle est trouvé, nous faisons la prévision de la valeur d'une nouvelle observation y^* en estimant, à l'aide du modèle, sa probabilité de valoir 1. Si cette probabilité est supérieure à 0,5 alors $\hat{y}^* = 1$ sinon $\hat{y}^* = 0$.

Pour transformer N-NO_2-NO_3 en variable dichotomique afin d'ajuster le modèle, il faut fixer le seuil séparant les valeurs 0 et 1. Ensuite, si $p(\underset{\sim}{x})$ est supérieur à 0,5, nous donnons la valeur 1 à $\hat{y}^*$, sinon, la valeur 0. Par exemple, si nous fixons le seuil à 0,25 alors

$$y^* = \begin{cases} 0 & \text{si } y<0{,}25 \\ 1 & \text{sinon.} \end{cases}$$

En transformant ainsi, 56,6 % des valeurs ont la valeur 1 et les autres sont égales à 0. Ensuite, nous estimons $p(\underset{\sim}{x})$ avec le modèle logistique et $\hat{y}^*$ est estimée par

$$\hat{y}^* = \begin{cases} 0 & \text{si la probabilité estimée est inférieure à 0,50} \\ 1 & \text{sinon.} \end{cases}$$

Afin de trouver un seuil le plus optimal possible, nous avons, pour plusieurs valeurs de seuil, effectué la régression logistique de la variable transformée en

Seuil	% d'erreur	% de 1
0	-	77,4
0,25	24,0	56,6
0,50	22,9	45,5
0,75	23,6	35,8
1,00	25,0	28,8
1,25	-	20,5

TABLEAU 2.2.5. *Pourcentage d'erreur obtenu par validation croisée, pour un seuil fixé, afin de prévoir la variable N-NO_2-NO_3 transformée à l'aide des autres variables dépendantes.*

fonction des autres variables dépendantes en utilisant les observations où la variable n'était pas manquante. La validation croisée a été utilisée pour trouver une prévision de chaque observation, puis nous avons calculé le pourcentage d'erreur pour chaque seuil fixé. Les résultats pour les différentes valeurs de seuil utilisées sont donnés au tableau 2.2.5.

Lorsque le seuil est fixé à 0 ou à 1,25, le pourcentage de valeurs transformées égales respectivement à 1 et à 0 devient très grand (supérieur à 75%). Si le seuil est fixé par exemple à 0, un modèle donnant toujours 1 comme prévision est alors très performant en apparence, mais très mauvais lorsque la "vraie" valeur de y vaut 0. Nous n'avons donc pas calculé le pourcentage d'erreur pour de tels seuils. Parmi les seuils étudiés, les meilleurs résultats sont obtenus lorsque le seuil est fixé à 0,50 ou 0,75. Dans le mesure où le pourcentage de 1 et de 0 est plus balancé lorsque le seuil est 0,50, nous avons choisi ce dernier. Nous avons donc transformé la variable N-NO_2-NO_3 et les valeurs manquantes ont été estimées à l'aide de la régression logistique.

Si nous ajustons un modèle pour estimer les variables dépendantes avec la RLM en utilisant les variables explicatives avec la variable transformée, N-NO_2-NO_3 est significative pour prévoir le débit, la DCO, la DBO_5 pour les valeurs du laboratoire, les MES, les MES non dissoutes, N-NH_3, O-PO_4, la turbidité et la température moyenne. En faisant la même chose avec N-NO_2-NO_3 non transformée et en utilisant seulement les observations sans valeurs manquantes alors la variable est significative pour prévoir la DCO, la DBO_5 pour les valeurs du laboratoire, les MES, N-NH_3, O-PO_4 et la turbidité. Donc, la variable, transformée avec le seuil 0,50, est significative pour toutes les variables où N-NO_2-NO_3 l'est.

Afin de vérifier si la variable transformée vaut la peine d'être incluse dans le jeu de données utilisé pour trouver le modèle final, nous avons fait une petite analyse avec la méthode PCR. Nous avons ajusté un modèle pour expliquer quelques-unes des variables pour lesquelles N-NO_2-NO_3 était significative, ainsi que pour la variable IVB, afin de voir si l'ajout de N-NO_2-NO_3 transformée change beaucoup la somme des erreurs au carré résultant de l'ajustement du modèle. Ceci a été effectué pour le modèle ajusté avec la validation croisée en blocs, puis pour celui ajusté avec méthode d'estimation/test. Ces méthodes seront décrites plus en détails à la section 3.1. Les résultats, pour chaque type d'analyse, sont donnés au tableau 2.2.6. Étant donné le peu de différence, nous avons tout simplement éliminé la variable du jeu de données: celle-ci n'entraîne pas une grande augmentation de la qualité du modèle, par contre, elle apporte une complication pour trouver de nouvelles prévisions avec le modèle dans la mesure où la variable devait toujours être transformée avant de pouvoir déterminer de nouvelles prévisions.

Estimation/test

Variable	Avec N-NO_2-NO_3 transformée	Sans N-NO_2-NO_3 transformée
IVB	16,05	7,44
DCO	252,81	249,35
MES	61,42	61,43
N-NH_3	2,59	2,60

Validation croisée en blocs

Variable	Avec N-NO_2-NO_3 transformée	Sans N-NO_2-NO_3 transformée
IVB	6,63	6,66
DCO	166,10	165,69
MES	66,85	66,87
N-NH_3	2,50	2,50

TABLEAU 2.2.6. *Comparaison des modèles avec et sans la variable N-NO_2-NO_3 transformée grâce à la racine de la moyenne des erreurs au carré pour méthode d'estimation/test et la validation croisée en blocs avec la méthode PCR.*

Le reste du jeu de données est maintenant "complet", c'est-à-dire qu'il n'y a plus aucune donnée manquante non imputée. Chacune des méthodes peut être utilisée pour analyser ce jeu de données modifié.

Chapitre 3

COMPARAISON DES MÉTHODES

Le chapitre 1 introduit plusieurs méthodes de modélisation qui ont été utilisées avec différents sous-groupes de variables pour estimer les variables dépendantes. Ces méthodes sont, rappelons-le, la régression linéaire multiple (RLM), la régression ridge (RR), la régression avec composante principale (PCR), la régression avec racine latente (LRR), les moindres carrés partiels (PLS) et les réseaux de neurones (RN). Cette section présente les résultats obtenus pour l'ajustement de différents modèles. Nous comparons également les méthodes de modélisation entre elles.

3.1. LES MODÈLES ET LEUR ESTIMATION

Les méthodes RR, PCR, LRR, PLS et RN nécessitent l'optimisation d'un ou de plusieurs paramètres. Nous avons utilisé trois façons pour estimer les modèles avec chacune des méthodes. Premièrement, le jeu de données a été séparé en deux parties et nous avons utilisé la méthode d'estimation/test pour estimer le modèle. La première partie du jeu de données a servi à estimer le modèle. Les paramètres à optimiser l'ont été par validation croisée avec une observation enlevée à la fois sur cette première partie du jeu de données. Nous rappelons que la VC a été décrite à la section 1.2.2. Lorsque les paramètres ont été optimisés, le modèle résultant a été ajusté sur les données de la première partie et il a été utilisé pour

prévoir les données de la deuxième partie et, ainsi, tester de façon indépendante l'efficacité du modèle en fonction de la somme des erreurs au carré. La première partie contenait 300 observations et le reste, soit 329 observations, formait la seconde.

Deuxièmement, la première procédure a été reprise, mais la VC sur la première partie a été effectuée en séparant celle-ci en dix blocs de 30 observations au lieu de prendre 300 blocs d'une observation. Parfois, les paramètres optimaux étaient identiques à ceux trouvés à l'aide de la première méthode, ce qui a mené à des modèles estimés identiques. Dans les autres cas, le modèle résultant était différent.

Avec la méthode d'estimation/test, les 300 premières observations sont utilisées pour optimiser les paramètres tel que décrit précédemment, sauf pour la RLM où il n'y a pas de paramètres à optimiser. À l'exception de la RLM, le modèle est estimé à partir des 300 premières observations avec les paramètres optimaux trouvés. La somme des erreurs au carré (SSE) est ensuite calculée avec les observations contenues dans le jeu de validation, soit les 329 observations restantes. L'ajustement du modèle avec la RLM est expliqué plus loin. Pour une méthode de modélisation choisie, la SSE peut donc être exprimée de la façon suivante :

$$SSE_{\text{méth}} = \sum_{i=301}^{629} (y_i - \hat{y}_i)^2, \tag{3.1.1}$$

où $\hat{y}_i$ est l'estimation de y_i, qui appartient au jeu de validation, avec le modèle construit à partir des 300 premières observations.

Troisièmement, la VC en blocs a été utilisée sur tout le jeu de données. Le jeu de données a été séparé en dix blocs d'environ 63 observations chacun. Chacun des blocs est enlevé un à un et nous estimons un modèle avec toutes les autres

observations pour une valeur fixée des paramètres. Le modèle est ensuite utilisé pour prévoir les données du bloc n'ayant pas servi à construire le modèle. La procédure est répétée pour les dix blocs. Lorsque tous les blocs ont été enlevés, nous avons une prévision de toutes les observations et nous calculons la SSE. Nous changeons ensuite la valeur des paramètres et nous recommençons les étapes qui viennent d'être décrites. Nous conservons les valeurs des paramètres ayant produit la plus petite SSE. Le modèle final, ajusté avec les paramètres trouvés avec la VC en blocs, est estimé avec les 629 observations du jeu de données. Il faut noter que, pour cette méthode d'estimation des paramètres, l'évaluation du modèle n'est pas indépendante de son estimation dans la mesure où les mêmes observations sont utilisées pour les deux. La SSE servant à comparer les méthodes est celle calculée lors de l'estimation des paramètres

$$SSE_{\text{méth}} = \sum_{j=1}^{10} \sum_{i=1}^{n_j} (y_{i,j} - \hat{y}_{i,j})^2, \tag{3.1.2}$$

où $\hat{y}_{i,j}$ est l'estimation de $y_{i,j}$, qui est la i^{e} observation du j^{e} bloc, à partir du modèle construit sans le j^{e} bloc et n_j est le nombre d'observations dans le j^{e} bloc ($\sum_j n_j = 629$). Cette SSE est la somme des erreurs au carré *minimale* obtenue par VC parmi toutes les SSE obtenues pour chacune des valeurs testées des paramètres.

La RLM ne nécessite pas l'optimisation de paramètres. Pour effectuer la sélection de variables, il faut fixer un seuil et une statistique pour l'élimination ou l'ajout des variables. Nous avons utilisé la statistique de Fisher et fixé sa valeur limite à 4. Afin de comparer la RLM aux autres méthodes, nous devons calculer des mesures d'efficacité comparables à celles des autres méthodes. Pour la méthode d'estimation/test avec VC en 10 ou 300 blocs, le choix des variables et l'ajustement du modèle par RLM sont effectués à partir des 300 premières

observations. Le modèle est ensuite testé sur le jeu de validation. Pour la VC en 10 blocs sur les 629 observations, les prévisions pour les observations d'un bloc proviennent du modèle ajusté avec la RLM à partir des observations des neuf autres blocs. Puis, nous calculons la SSE. Il est à noter que les variables choisies par la RLM pour prévoir les observations d'un bloc peuvent changer d'un bloc à l'autre.

Les deux méthodes d'estimation/test offrent une mesure d'efficacité où la validation est indépendante de l'ajustement. La seconde méthode d'estimation/test est rapide à calculer puisque l'optimisation des paramètres est rapide. La troisième méthode, la VC en 10 blocs sur les 629 observations, n'offre pas une mesure indépendante et pourrait donc être trop optimiste. Mais, dans la mesure où elle ajuste un modèle contenant près de 150 coefficients avec environ 560 observations, soit les observations de neuf des dix blocs, comparativement à la méthode d'estimation/test qui en utilise seulement 300, cette méthode devrait donner de meilleurs modèles. Nous verrons plus tard quelle mesure d'efficacité est la plus fiable.

La DCO soluble à l'effluent n'était pas mesurée sur une longue période comme mentionné à la section 2.2.2. Nous n'avons que 316 observations disponibles pour cette variable. Nous avons estimé les modèles de la même façon que pour les autres variables. Pour la méthode d'estimation/test, le jeu d'estimation contenait 150 observations et celui de validation, 166 observations. Étant donné que nous avons moins de données pour la DCO soluble, la méthode PLS multivariée a été utilisée sur les autres variables seulement.

Nous venons d'expliquer de quelle manière les modèles seront estimés pour un ensemble de variables donné. Voyons maintenant quels sous-groupes de variables seront utilisés pour prévoir les variables dépendantes. Dans la mesure où toutes

les variables sont récoltées durant la journée, les variables explicatives d'une journée ne peuvent être utilisées pour faire des prévisions des variables dépendantes du même jour. Il faut utiliser les données de la journée précédente pour prévoir les variables d'une journée en particulier. De plus, étant donné la forte autocorrélation de certaines variables, nous avons utilisé également l'avant-dernière journée comme variable explicative. Dans la mesure où nous utilisons chacune des variables sur deux journées, le nombre de variables explicatives double. Nous avons en tout 146 variables explicatives. Notons "Jour 1" le groupe des variables de la journée précédente et "Jour 2", celui des variables de l'avant-dernière journée. Chacun de ces deux groupes contient 73 variables. Nous avons d'abord estimé des modèles avec toutes ces 146 variables en optimisant les paramètres avec la méthode estimation/test avec validation croisée (VC) avec 10 et 300 blocs sur le jeu d'estimation puis avec la validation croisée avec 10 blocs sur le jeu complet de 629 observations. Une question peut ensuite être posée, à savoir si le groupe de variables "Jour 2" apporte une amélioration au modèle. Nous avons donc utilisé, dans un deuxième temps, le groupe "Jour 1" seulement pour estimer les variables dépendantes avec la VC en 10 blocs sur les 629 observations, soit la méthode d'estimation la plus rapide.

Les modèles précédents contiennent beaucoup de variables explicatives. Mise à part la RLM utilisée avec la sélection de variables, les modèles résultants sont très longs. Étant donné le nombre de variables, il est plus que probable que plusieurs d'entre elles n'aient que peu de valeur prédictive pour l'estimation de certaines variables dépendantes. Nous avons donc estimé des modèles avec les méthodes univariées en utilisant les variables sélectionnées à l'aide de l'addition par étapes (AÉ) appliquée sur les 300 premières observations et les 146 variables explicatives, en bref, les variables ayant été choisies avec la RLM avec l'AÉ pour la

méthode estimation/test. Le sous-groupe de variables significatives étant différent pour chacune des variables dépendantes, l'utilisation de méthodes de sélection de variables multivariées conduira à des modèles contenant beaucoup de variables explicatives. Pour cette raison, nous ne les utiliserons pas ici. Une méthode de sélection de variables multivariée est présentée dans l'article de Lazraq et Cléroux (1988). Nous avons utilisé les sous-groupes formés pour estimer des modèles avec la VC en 10 blocs sur les 629 observations et la méthode estimation/test avec VC pour une observation à la fois sur le jeu d'estimation.

Les précédents sous-groupes contiennent souvent des variables du groupe "Jour 2" sans contenir la variable correspondante du groupe "Jour 1", ce qui n'est guère logique au niveau de l'interprétation. Il y a de bonnes chances que l'information de la journée précédente soit pertinente si celle d'avant-hier est importante. Lorsque de tels cas se présentaient, nous avons ajouté la variable équivalente du groupe "Jour 1" afin de créer de nouveaux sous-groupes qui ont servi à estimer des modèles avec la VC en 10 blocs sur les 629 observations et la méthode estimation/test avec VC pour 1 observation à la fois sur le jeu d'estimation.

Nous avons ensuite refait ceci en appliquant la sélection de variables AÉ de la RLM sur les 300 premières observations et uniquement sur les variables du groupe "Jour 1". Les nouveaux sous-groupes sont formés des variables significatives.

Nous comparerons tous ces modèles entre eux pour voir l'efficacité relative des différentes méthodes de modélisation. Nous vérifierons l'amélioration de ces modèles par rapport au modèle très simple qui est d'estimer une variable Y par sa moyenne estimée $\bar{y}$. Nous regarderons également l'amélioration de l'estimation par rapport au modèle estimant Y à l'aide des valeurs observées pour Y seulement

durant les deux journées précédentes, ce qui équivaut à estimer une série chronologique à l'aide d'un modèle autorégressif d'ordre 2, AR(2). Ceci permet de voir la contribution des 144 autres variables aux modèles. Cette dernière comparaison offre des résultats assez surprenants.

3.2. Description des statistiques utilisées

Afin de comparer l'efficacité des différentes méthodes pour estimer les variables dépendantes, nous calculerons différentes statistiques. Nous allons comparer les modèles par rapport au modèle $\hat{y} = \bar{y}$ et par rapport au modèle estimant y à partir des observations de y sur les deux journées précédentes :

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 y_{i-1} + \hat{\beta}_2 y_{i-2}, \tag{3.2.1}$$

où i représente le temps et les $\hat{\beta}_i$ sont obtenus par les moindres carrés . Pour faire cette comparaison, nous calculerons deux statistiques que nous noterons R^2_M, pour la comparaison avec l'estimation par la moyenne, et R^2_T, pour la comparaison avec l'estimation par les deux derniers jours. La première statistique, R^2_M, s'inspire du coefficient de détermination de la RLM représentant le pourcentage de variabilité expliqué par le modèle. La deuxième statistique nous permettra de voir la contribution au modèle des variables explicatives autres que celles correspondant à la variable dépendante.

Lorsque nous estimons les variables par leur moyenne, la SSE, notée SSE_M, est obtenue de la façon suivante :

– pour l'estimation/test

$$SSE_M = \sum_{i=301}^{629} (y_i - \bar{y}^{(esti)})^2,$$

SSE_M		
Variables	Estimation/test	VC en 10 blocs
IVB	28,6300	36,7770
débit	5265,537	5453,551
DCO	414,9244	416,9921
DBO_5	22,0556	21,7214
MES	70,0225	84,3229
MES_lab	70,0760	82,6493
$N\text{-}NH_3$	2,4351	2,6027
MES_nd	32,5066	31,1567
$O\text{-}PO_4$	1,3844	1,0847
turbid	31,7037	37,3934
pHcomp	0,1495	0,1555
pHmoy	0,1498	0,1385
Tmoy	2,4988	2,1607
cond	333,1760	285,2336
DCOsol	242,5793	310,1344

TABLEAU 3.2.1. *SSE_M pour la méthode d'estimation/test et la VC en 10 blocs sur les 629 observations.*

où les y_i sont les observations du jeu de validation et $\bar{y}^{(esti)}$ est la moyenne de y sur le jeu d'estimation, soit les 300 premières observations.

- pour la VC en 10 blocs sur les 629 observations

$$SSE_M = \sum_{i=1}^{629} (y_i - \bar{y})^2,$$

où $\bar{y}$ est la moyenne des 629 observations.

Les résultats obtenus pour les deux méthodes sont donnés au tableau 3.2.1.

Pour le calcul de la SSE du modèle en fonction des deux derniers jours, notée SSE_T, nous utilisons les équations (3.1.1) et (3.1.2). Dans le cas présent, la

SSE_T		
Variables	Estimation/test	VC en 10 blocs
IVB	5,4634	5,3602
débit	3230,546	3082,521
DCO	190,6411	199,7216
DBO_5	12,6879	12,6090
MES	55,1085	68,5351
MES_lab	51,8866	64,4790
$N\text{-}NH_3$	2,0951	2,4890
MES_nd	22,1575	21,6889
$O\text{-}PO_4$	0,6291	0,6417
turbid	23,1332	30,8259
pHcomp	0,1294	0,1250
pHmoy	0,1564	0,1078
Tmoy	1,4440	1,1981
cond	102,1590	94,5319
DCOsol	151,0998	160,8408

TABLEAU 3.2.2. *SSE_T pour la méthode d'estimation/test et la VC en 10 blocs sur les 629 observations.*

prévision de y_i est celle obtenue en ajustant le modèle donné par (3.2.1). Les résultats obtenus pour la méthode d'estimation/test et la VC en 10 blocs sur toutes les observations sont donnés au tableau 3.2.2. Nous calculons finalement R^2_M et R^2_T

$$R^2_M = \frac{SSE_M - SSE_{\text{méth}}}{SSE_M}$$

et

$$R^2_T = \frac{SSE_T - SSE_{\text{méth}}}{SSE_T}.$$

Il est à noter que ces deux statistiques sont inférieures ou égales à 1, mais elles peuvent être négatives si la SSE d'une méthode est supérieure à SSE_M ou SSE_T. En régression, le coefficient de détermination est toujours compris entre 0 et 1, cependant, ici, les prévisions utilisées dans les calculs des SSE sont indépendantes des observations servant à construire le modèle utilisé pour calculer ces prévisions.

3.3. Comparaison des méthodes

Nous allons maintenant comparer les différentes méthodes de modélisation en nous appuyant sur la SSE ainsi que sur les différentes statistiques calculées en se basant sur cette dernière. Nous verrons les résultats pour le différents sous-groupes de variables utilisés.

3.3.1. Estimation/test sur les 146 variables

Les tableaux 3.3.1 et 3.3.2 donnent les R^2_M et les R^2_T pour la méthode d'estimation/test sur les 146 variables explicatives, soit les variables des groupes "Jour 1" et "Jour 2". Mises à part les variables N-NH_3, turbidité et pH composé, il y a toujours au moins un modèle qui apporte une amélioration sur le modèle $\hat{y} = \bar{y}$. Pour les variables IVB et conductivité, les R^2_M sont très élevés atteignant plus de 0,90 pour certains modèles. Ces statistiques sont également élevées, entre 0,60 et 0,80, pour le débit, la DCO, O-PO_4, la température moyenne et la DCO soluble et elles sont plus faibles pour les variables restantes. Donc, les variables explicatives contribuent à l'amélioration des prévisions. Cependant, parmi les variables explicatives, deux d'entre elles correspondent aux mesures de la variable dépendante prises durant les deux dernières journées. Nous pouvons nous demander si ce ne sont pas principalement ces deux variables qui expliquent la variable dépendante.

Si nous regardons maintenant le deuxième tableau contenant les R_T^2 des modèles, nous constatons que la majorité des résultats sont négatifs, ce qui indique que les modèles trouvés sont moins bons au niveau de la SSE qu'un simple modèle AR(2). Ainsi, la variable IVB, qui avait un très fort R_M^2, n'a que des R_T^2 négatifs. C'est également le cas pour d'autres variables qui avait de bons R_M^2. Il y a toutefois quelques variables, comme le débit, la température moyenne, la conductivité et la DCO soluble qui, en plus d'avoir des R_M^2 positifs ont également des R_T^2 non nuls. Ces résultats peuvent s'expliquer de plus d'une façon. Les variables explicatives, autres que celles correspondant à la variable dépendante mesurée durant les deux jours précédents, peuvent ne pas avoir d'impact sur la variable dépendante ou alors l'information qu'elles contiennent est déjà incluse dans les deux variables incluses dans le modèle AR(2). Ensuite, il peut y avoir tant de bruit dans les 146 variables que les méthodes de modélisation ne parviennent pas à trouver les signaux contenus dans le jeu de données.

La méthode PCR se compare à la RLM quant à la SSE et à PLS univarié quant à l'erreur moyenne. L'erreur moyenne est définie ici comme étant $\frac{1}{N}\sum_i(y_i - \hat{y}_i)$, où les $\hat{y}_i$ sont les prévisions obtenues par la VC. Comme nous venons de le dire indirectement, la RLM a une SSE plus faible que PLS, mais une erreur moyenne généralement plus élevée. Bref, dans l'ensemble, aucune de ces deux dernières méthodes ne surpasse vraiment l'autre. Nous avons comparé les méthodes en calculant les rangs moyens. Pour ce faire, pour chaque variable, nous trions en ordre décroissant les SSE des différentes méthodes et nous leur assignons un rang. Ensuite, pour chaque méthode, nous faisons la moyenne des rangs. Les rangs moyens sont donnés au tableau 3.3.3 pour la méthode d'estimation/test et pour la VC en 10 blocs. Au niveau des rangs moyens de la SSE donnés au tableau 3.3.3, la RLM arrive en premier, suivi de PCR, puis de PLS univarié un

TABLEAU 3.3.1. R^2_M *obtenu avec la méthode estimation/test avec l'optimisation des paramètres par la VC avec une observation à la fois.*

	IVB_m	débit_e	DCO_e	DBO5_lab_e	MES_e
RLM (AÉ)	0,9419	0,6360	0,7789	0,4921	-0,2693
RLM (SPAP)	0,9419	0,6080	0,7782	0,4921	-0,1136
RR	0,9382	0,6468	0,6136	0,4485	0,2640
PCR	0,9324	0,6815	0,6389	0,3897	0,3135
LRR	0,7845	0,2677	0,6065	-0,3568	-2,1865
PLS (uni)	0,9489	0,6929	0,6398	0,4798	0,1785
PLS (multi)	0,2634	0,5783	0,6873	-0,0560	0,0807

	MES_lab_e	$N\text{-}NH_3$	MES_nd_e	$O\text{-}PO_4$_e	Turbid_e
RLM (AÉ)	-0,1119	-1,9527	0,3615	0,5939	-2,2596
RLM (SPAP)	-0,1119	-2,0244	0,3615	0,6528	-2,1955
RR	0,2651	-0,2684	0,3155	-1,1362	-2,3238
PCR	0,3146	-0,1364	0,4097	-0,0756	-2,5461
LRR	-0,1768	-1,0163	-0,1115	0,5669	-3,6379
PLS (uni)	0,2570	-0,0304	0,3604	-0,2730	-2,6280
PLS (multi)	0,1517	-0,6134	0,1502	-0,5762	-0,0546

	pHcomp_e	pHmoy_e	Tmoy_e	Cond_e	DCOsol_e
RLM (AÉ)	-0,0958	-0,0605	0,7111	0,9271	0,6846
RLM (SPAP)	-0,0958	-0,0605	0,7164	0,9322	0,6846
RR	-1,8617	-1,1547	0,3152	0,8856	0,6547
PCR	-2,2337	-0,0014	0,7035	0,8852	0,5949
LRR	-0,3517	0,0720	0,6987	0,4904	-1,7731
PLS (uni)	-1,2946	-0,1164	0,6833	0,8935	0,6547
PLS (multi)	-0,1025	0,1344	0,1446	0,7697	-

TABLEAU 3.3.2. *R_T^2 obtenu avec la méthode estimation/test avec l'optimisation des paramètres par la VC avec une observation à la fois.*

	IVB_m	débit_e	DCO_e	DBO5_lab_e	MES_e
RLM (AÉ)	-0,5957	0,0331	-0,0474	-0,5347	-1,0492
RLM (SPAP)	-0,5957	0,0331	-0,0508	-0,5347	-0,7117
RR	-0,6979	0,0617	-0,8302	-0,6666	-0,2965
PCR	-0,8565	0,1540	-0,7107	-0,8443	-0,2427
LRR	-4,9181	-0,9154	-0,8639	-3,1000	-4,6510
PLS (uni)	-0,4024	0,1843	-0,7063	-0,5720	-0,3263
PLS (multi)	-19,2283	-0,1202	-0,4811	-2,1910	-0,4842

	MES_lab_e	N-NH_3	MES_nd_e	O-PO_4_e	Turbid_e
RLM (AÉ)	-1,0281	-2,9889	-0,3742	-0,9667	-5,1222
RLM (SPAP)	-1,0281	-3,0858	-0,3742	-0,6817	-5,0018
RR	-0,3405	-0,7136	-0,4733	-9,3459	-5,2429
PCR	-0,2503	-0,5352	-0,2705	-4,2094	-5,6605
LRR	-1,1465	-1,7239	-1,3923	-1,0975	-7,7110
PLS (uni)	-0,3552	-0,3920	-0,3767	-5,1651	-5,8142
PLS (multi)	-0,5474	-1,1796	-0,8291	-6,6338	-0,9808

	pHcomp_e	pHmoy_e	Tmoy_e	Cond_e	DCOsol_e
RLM (AÉ)	-0,4618	0,0271	0,1351	0,2241	0,1870
RLM (SPAP)	-0,4618	0,0271	0,1509	0,2792	0,1870
RR	-2,8176	-0,9768	-1,0506	-0,2163	0,1101
PCR	-3,3139	0,0813	0,1123	-0,2210	-0,0442
LRR	-0,8033	0,1487	0,0977	-4,4206	-6,1472
PLS (uni)	-2,0611	-0,0242	0,0518	-0,1331	0,0604
PLS (multi)	-0,4708	0,2059	-1,5614	-1,4497	-

peu plus loin. Avec ce jeu de données, le modèle complet ne peut pas être ajusté avec la RLM à cause des problèmes de singularité dans le calcul de l'estimateur des coefficients du modèle donné à l'équation (1.1.2). En effet, la matrice $X'X$ n'est pas inversible lorsque X représente toutes les variables explicatives. Du côté de la sélection de variables avec la RLM, l'AÉ et la SPAP donnent des résultats semblables.

De façon générale, pour la méthode d'estimation/test sur toutes les variables explicatives avec l'optimisation des paramètres par la VC avec une observation à la fois, la RR est parmi les méthodes donnant les moins bons résultats lorsque les variables ont un petit écart type, i.e. pour N-NH_3, O-PO_4, le pH moyen, le pH composé et la température moyenne à l'effluent. Si nous regardons les R^2_M données au tableau 3.3.1, nous constatons que les statistiques pour la RR prennent de petites valeurs comparativement à celles obtenues avec les autres méthodes. Le tableau 3.3.3 contient le rang moyen de la SSE de chaque méthode de modélisation pour la méthode d'estimation/test: la RR n'est pas parmi les meilleures méthodes si nous regardons les résultats de la première colonne du tableau. Il faut faire attention avant de comparer cette première colonne avec les résultats des deux suivantes puisque celle-ci ne contient que sept méthodes alors que les deux autres en contiennent huit.

Les variables dépendantes n'étaient pas toutes fortement corrélées les unes avec les autres, comme le montre le tableau 3.3.4 contenant les corrélations entre quelques-unes des variables dépendantes, de telle sorte que nous nous attendions à ce que PLS multivarié ne donne pas de très bons résultats ; ce qui est confirmé par la pratique. Que ce soit au niveau de la SSE ou de l'erreur moyenne, PLS multivarié donne de moins bons résultats. La méthode LRR donne également de piètres résultats, pires que PLS multivarié dans bien des cas.

Méthode	Estimation/test		VC en 10 blocs
	1 obs. à la fois	VC en 10 blocs	sur toutes les obs.
RLM (AÉ)	3,0	3,3	3,5
RLM (SPAP)	3,1	3,3	3,9
RR	4,0	4,4	4,5
PCR	3,3	4,3	3,7
LRR	5,7	5,5	5,1
PLS (uni)	3,8	3,3	3,1
PLS (multi)	4,9	5,4	6,1
RN	-	6,2	5,8

TABLEAU 3.3.3. *Rang moyen de la SSE des méthodes de modélisation pour la méthode d'estimation/test et VC en blocs sur toutes les observations.*

	$O\text{-}PO_4$_e	Turbid_e	pHcomp_e	pHmoy_e	Tmoy_e	cond_e
$O\text{-}PO_4$_e	1,000	0,269	-0,110	0,337	-0,124	-0,353
Turbid_e		1,000	0,046	0,183	-0,306	0,002
pHcomp_e			1,000	0,153	0,024	0,143
pHmoy_e				1,000	0,100	-0,170
Tmoy_e					1,000	0,338
cond_e						1,000

TABLEAU 3.3.4. *Corrélation entre quelques variables dépendantes*

Avec l'optimisation des paramètres par la VC en 10 blocs sur le jeu d'estimation, nous obtenons des résultats semblables aux précédents au niveau de la SSE, parfois même identiques, car les paramètres optimaux estimés sont quelques fois les mêmes dans les deux cas. Nous avons ajusté un modèle avec les RN avec deux

couches cachées de cette façon, mais les résultats ne sont pas bons. Pour l'estimation/test avec VC pour une observation à la fois, cette méthode de modélisation n'avait pas été utilisée, car le temps de calcul était trop grand. Souvent, les RN donnent les pires résultats, comme nous pouvons le constater au tableau 3.3.3 où les RN ont un rang moyen de 6,2 (sur 8 méthodes). LRR et PLS multivarié sont également des méthodes qui ne donnent pas de bons résultats comme nous pouvons le constater d'après les rangs moyens de la SSE du tableau 3.3.3. Avec la VC en 10 blocs sur le jeu d'estimation, PLS univarié et la RLM donnent d'aussi bons résultats l'un que l'autre tandis que PCR fait moins bien et devient équivalent à la RR au niveau des rangs moyens de la SSE.

3.3.2. Validation croisée avec 10 blocs sur le jeu complet

Si nous regardons maintenant les résultats des tableaux 3.3.5 et 3.3.6 obtenus avec la validation croisée, nous arrivons aux mêmes conclusions en ce qui concerne PLS multivarié et LRR : les résultats ne sont pas très bons comparativement aux autres méthodes. De plus, pour la validation croisée, nous avons ajusté un modèle à l'aide des réseaux de neurones avec deux couches cachées. Les résultats obtenus sont cependant les pires. La RR ne donne pas non plus de très bons résultats.

Pour PCR, nous notons encore une fois qu'il est moins bon que PLS univarié ou que la RLM qui donnent une SSE et une erreur moyenne en général plus faibles. Entre PLS et RLM, nous remarquons que PLS donne généralement une SSE et une erreur moyenne qui sont meilleures que dans le cas de la RLM. Au niveau des rangs moyens donnés à la première colonne du tableau 3.3.3, la meilleure méthode est PLS univarié, suivi de la RLM avec l'AÉ, puis PCR. Ici, l'AÉ et la SPAP ne sont pas équivalentes au niveau des rangs, mais en regardant les tableaux 3.3.5 et 3.3.6, nous constatons que les résultats sont très semblables, la

TABLEAU 3.3.5. R^2_M *obtenu avec la VC en 10 blocs sur le jeu complet.*

	IVB_m	débit_e	DCO_e	DBO5_lab_e	MES_e
RLM (AÉ)	0,9787	0,6958	0,8264	0,5973	0,3377
RLM (SPAP)	0,9787	0,7051	0,8264	0,5965	0,3411
RR	0,9726	0,7379	0,8399	0,6360	0,3644
PCR	0,9672	0,7445	0,8421	0,6320	0,3712
LRR	0,9688	0,6918	0,8179	0,5828	0,3657
PLS (uni)	0,9768	0,7457	0,8373	0,6450	0,3620
PLS (multi)	0,6040	0,7203	0,7175	0,0083	0,2758
RN	0,9431	0,6524	0,7979	0,4913	0,2253

	MES_lab_e	$N\text{-}NH_3$	MES_nd_e	$O\text{-}PO_4$_e	Turbid_e
RLM (AÉ)	0,4330	0,0407	0,4969	0,6250	0,3165
RLM (SPAP)	0,4029	0,0691	0,4991	0,6242	0,2896
RR	0,4282	-0,1456	0,4903	0,1788	0,3171
PCR	0,4223	0,0755	0,4940	0,2523	0,3521
LRR	0,4006	0,0850	0,4609	0,6086	0,3412
PLS (uni)	0,4285	0,0868	0,4901	0,2755	0,3148
PLS (multi)	0,3126	-0,0746	0,2004	0,2591	0,1556
RN	0,4153	-0,1618	0,3456	0,5232	0,3717

	pHcomp_e	pHmoy_e	Tmoy_e	Cond_e	DCOsol_e
RLM (AÉ)	0,3679	0,4864	0,7479	0,9247	0,8733
RLM (SPAP)	0,3700	0,4821	0,7505	0,9248	0,8734
RR	-0,4021	-0,5148	0,4929	0,9247	0,8733
PCR	0,0879	0,1749	0,7346	0,9215	0,8679
LRR	0,3279	0,4508	0,7513	0,9059	0,8416
PLS (uni)	0,0857	0,2068	0,7515	0,9258	0,8750
PLS (multi)	-0,0145	0,1237	0,2082	0,8148	-
RN	0,1732	0,3422	0,6961	0,8818	0,5985

TABLEAU 3.3.6. *R^2_T obtenu avec la VC en 10 blocs sur le jeu complet.*

	IVB_m	débit_e	DCO_e	DBO5_lab_e	MES_e
RLM (AÉ)	-0,0012	0,0478	0,2432	-0,1952	-0,0026
RLM (SPAP)	-0,0012	0,0771	0,2531	-0,1974	0,0026
RR	-0,2900	0,1795	0,2061	-0,0804	0,0378
PCR	-0,5438	0,2003	0,3117	-0,0921	0,0481
LRR	-0,4684	0,0352	0,3019	-0,2380	0,0397
PLS (uni)	-0,0936	0,2041	0,2910	-0,0534	0,0342
PLS (multi)	-17,6427	0,1247	-0,2317	-1,9430	-0,0962
RN	-1,6792	-0,0881	0,1190	-0,5096	-0,1727

	MES_lab_e	N-NH_3	MES_nd_e	O-PO_4_e	Turbid_e
RLM (AÉ)	0,0684	-0,0489	-0,0383	-0,0715	-0,005
RLM (SPAP)	0,0189	-0,0179	-0,0336	-0,0737	-0,0453
RR	0,0605	-0,2526	-0,0518	-1,3464	-0,0048
PCR	0,0509	-0,0109	-0,0442	-1,1366	0,0466
LRR	0,0151	-0,0005	-0,1125	-0,1183	0,0306
PLS (uni)	0,0611	0,0015	-0,0523	-1,0703	-0,0082
PLS (multi)	-0,1294	-0,1750	-0,6501	-1,1171	-0,2425
RN	0,0393	-0,2704	-0,3505	-0,3623	0,0754

	pHcomp_e	pHmoy_e	Tmoy_e	Cond_e	DCOsol_e
RLM (AÉ)	0,0212	0,1527	0,1802	0,3146	0,1685
RLM (SPAP)	0,0244	0,1456	0,1886	0,3156	0,1793
RR	-1,1712	-1,4988	-0,6495	0,3148	0,2466
PCR	-0,4125	-0,3611	0,1368	0,2856	0,1825
LRR	-0,0408	0,0940	0,1911	0,1436	-0,0117
PLS (uni)	-0,4158	-0,3085	0,1918	0,3240	0,2581
PLS (multi)	-0,5709	-0,4456	-1,5754	-0,6865	-
RN	-0,2803	-0,0852	0,0115	-0,0761	-0,4928

différence s'explique par le fait que les méthodes de modélisation donnent souvent des résultats n'ayant qu'un petit écart les uns par rapport aux autres.

Nous obtenons des modèles différents avec la VC en 10 blocs et avec la méthode d'estimation/test, mais quelles statistiques sont préférables? Les racines des SSE moyennes obtenues avec la VC en 10 blocs sont souvent inférieures à celles de la méthode d'estimation/test. Ceci nous donne des R^2_M et des R^2_T plus élevées, par exemple, pour la DCO, le meilleur R^2_T avec la méthode d'estimation/test sur les 146 variables explicatives est de -0,04 alors que, pour la VC en 10 blocs, il est de 0,31. Donc, devons-nous conclure que les variables explicatives contribuent à la prévision des observations ou non? Comme nous l'avons mentionné à la section 3.1, la SSE de la VC en 10 blocs est celle calculée lors de l'optimisation des paramètres, c'est-à-dire que c'est la plus petite SSE trouvée. Nous pouvons donc nous demander si les résultats obtenus avec cette méthode sont biaisés. Il est également possible que la VC en 10 blocs donne de meilleurs résultats dans la mesure où il y a deux fois plus observations qui sont utilisées pour estimer les paramètres et construire le modèle que pour la méthode d'estimation/test.

Afin de répondre à ces questions, nous pouvons comparer la racine de la SSE moyenne de la VC en 10 blocs à celle trouvée avec la méthode d'estimation/test lors de l'optimisation des paramètres, toute deux sont présentées au tableau 3.3.7. La première colonne de chiffres de ce tableau contient la racine de la SSE moyenne trouvée en faisant la VC pour une observation à la fois sur les 300 premières observations avec la valeur du paramètre optimal estimé alors que la seconde contient la racine de la SSE moyenne évaluée sur le jeu de validation. La dernière colonne représente la racine de la SSE moyenne calculée avec la VC en 10 blocs sur toutes les observations avec la valeur du paramètre optimal estimé. Dans tous les cas, le paramètre optimal est celui donné par (1.2.1) sur le jeu de données servant

Variables	Estimation/test		VC en 10 blocs
	Estimation	Validation	
IVB	5,5675	6,3677	5,6054
débit	2664,9808	2917,7760	2749,9600
DCO	166,5878	249,0289	168,1735
DBO_5	12,4449	23,4258	12,9413
MES	73,2268	63,4659	67,3526

TABLEAU 3.3.7. *Racine de la SSE moyenne obtenue avec PLS univarié pour l'estimation des paramètres avec la méthode estimation/test, puis celle sur le jeu de validation pour le modèle ajusté avec les paramètres optimaux estimés ainsi que celle obtenue pour la VC en 10 blocs sur les 629 observations*

à construire le modèle. Nous constatons, comme le montre le tableau 3.3.7, que la racine de la SSE moyenne de la VC en 10 blocs et celle trouvée avec la méthode d'estimation/test lors de l'optimisation des paramètres sont semblables. La VC en 10 blocs sur tout le jeu de données donne donc une évaluation optimiste du modèle estimé. Les statistiques trouvées grâce à la méthode d'estimation/test seront donc plus fiables que celles de la VC en 10 blocs, il vaut donc mieux se fier aux résultats des tableaux 3.3.1 et 3.3.2.

3.3.3. Modèles ajustés à partir de sous-groupes de variables

Les modèles obtenus par VC en 10 blocs à partir des 73 variables du groupe "Jour 1" sont semblables à ceux obtenus avec toutes les variables explicatives, tantôt un peu mieux tantôt un peu moins bons. Dans tous les cas, la différence n'est jamais très importante. Pour la méthode d'estimation avec VC pour une observation à la fois, les modèles n'ont été ajustés que pour les variables IVB, DCO, DBO_5 et MES, qui seront étudiées plus en détails à la section 3.4.1. Bien

que les différences soient un peu plus importantes, encore une fois, les résultats sont parfois meilleurs parfois pires. Donc, pour certaines variables dépendantes, les variables du groupe "Jour 2" n'ont que peu d'importance, par exemple, pour la DCO, l'ensemble des modèles sont un peu mieux lorsqu'il n'y a que des variables du groupe "Jour 1".

Parmi les variables explicatives, plusieurs d'entre elles n'ont que peu de valeur prédictive pour les variables dépendantes. Ainsi, la RLM avec sélection de variables ne donne aucun modèle ayant plus de vingt variables significatives. Rappelons que le modèle complet contient 146 variables explicatives. Donc, avec les autres méthodes, étant donné qu'au départ toutes les variables explicatives sont incluses dans le modèle, il y a beaucoup de bruit que le modèle doit réussir à distinguer de l'information pertinente pour effectuer les prévisions. Nous pourrions donc nous demander quel serait le résultat d'un modèle ajusté avec chacune des méthodes, mais ne contenant, au départ, que les variables sélectionnées par l'addition par étape effectuée avec la RLM.

La sélection de variables fournit un sous-groupe influençant les variables dépendantes. Il faut cependant se rappeler qu'une méthode de sélection différente peut donner un sous-groupe différent. La sélection permet toutefois d'appliquer les méthodes et d'obtenir un modèle plus court. Étant donné que nous enlevons plusieurs variables n'ayant pas une influence significative sur les variables dépendantes une fois considérées les variables du sous-groupe retenu, nous éliminons une grande partie du bruit contenu dans le jeu de données. Une fois ce bruit ôté, nous obtenons de meilleurs modèles dans bien des cas comme nous pouvons le constater en comparant les tableaux 3.3.1 et 3.3.8 par exemple qui donnent les R^2_M pour la méthode d'estimation/test : les R^2_M sont plus élevés lorsqu'il y a d'abord une sélection de variables. Les RN, ajustés avec la VC en 10 blocs sur les

629 observations, ne donnent pas de meilleurs résultats que la RLM. Toutes les méthodes linéaires donnent, en général, des résultats meilleurs que la RLM. Plus le sous-groupe de variables est petit et moins la corrélation entre les variables explicatives est importante alors plus les résultats de toutes les méthodes sont semblables à ceux de la RLM. Pour la méthode d'estimation/test, PCR, LRR et PLS univarié donnent la même chose que la RLM alors que la RR fait mieux que ce soit au niveau de la SSE ou de l'erreur moyenne.

En ajoutant les variables du groupe "Jour 1" lorsque les variables correspondantes du "Jour 2" sont incluses seulement, nous ajoutons du bruit au modèle, ces variables n'étant pas considérées assez significatives pour être incluses dans le sous-groupe retenu. Les modèles sont donc passablement équivalents bien qu'ils gagnent au niveau de l'interprétabilité, car, pour toutes les variables significatives du groupe "Jour 2", nous retrouvons la variable équivalente du groupe "Jour 1" dans le modèle. Les modèles sont parfois un peu mieux dans quelques cas, ce qui illustre bien le fait que les variables éliminées peuvent contribuer à l'amélioration du modèle. Nous n'avons pas ajouté les tableaux de résultats pour ce sous-groupe de variables, ainsi que pour les suivants, étant donné qu'il y a peu de différence avec les tableaux 3.3.8 et 3.3.9.

Nous avons constaté précédemment que l'ajout du groupe "Jour 2" n'amenait que peu d'information aux modèles en comparant les résultats obtenus avec la VC en 10 blocs sur toutes les observations et celle sur le groupe "Jour 1" seulement. Si nous comparons les modèles ajustés à partir des variables significatives auxquelles nous ajoutons les variables du groupe "Jour 1" lorsque les variables correspondantes du groupe "Jour 2" sont incluses et ceux ajustés à partir des variables significatives dans le groupe "Jour 1" uniquement alors nous constatons que les modèles sont assez équivalents. Seule la variable débit a un modèle qui

TABLEAU 3.3.8. R^2_M *obtenu avec la méthode d'estimation/test à partir des variables sélectionnées par l'addition par étape sur les 300 premières observations et l'optimisation des paramètres par la VC avec une observation à la fois.*

	IVB_m	débit_e	DCO_e	DBO5_lab_e	MES_e
RLM	0,9419	0,6360	0,7789	0,4921	-0,2693
RR	0,9573	0,6719	0,7929	0,5640	-0,3678
PCR	0,9419	0,6360	0,7789	0,4921	-0,2693
LRR	0,9420	0,6449	0,7789	0,4495	-0,2693
PLS (uni)	0,9419	0,6360	0,7789	0,4978	-0,2693

	MES_lab_e	N-NH_3	MES_nd_e	O-PO_4_e	Turbid_e
RLM	-0,1119	-1,9527	0,3615	0,5939	-2,2596
RR	-0,1119	-1,7038	0,3945	0,0000	0,2295
PCR	-0,1143	-1,9527	0,3615	0,5929	0,2174
LRR	-0,1119	-1,9527	0,3615	0,6113	0,2174
PLS (uni)	-0,1119	-1,9527	0,3615	0,5924	0,2174

	pHcomp_e	pHmoy_e	Tmoy_e	Cond_e	DCOsol_e
RLM	-0,0958	-0,0605	0,7111	0,9271	0,6846
RR	0,0151	0,0591	0,7095	0,9296	0,7038
PCR	-0,0958	-0,0605	0,7111	0,9271	0,6846
LRR	-0,0958	-0,0655	0,7111	0,9289	0,6844
PLS (uni)	-0,0958	-0,0605	0,7111	0,9271	0,6846

s'améliore si nous incluons des variables du groupe "Jour 2". Si nous comparons les modèles AR(2) et AR(1), nous constatons que le débit du groupe "Jour 2" est très important pour prévoir cette variable. Inversement, les MES pour les valeurs du laboratoire ont un meilleur modèle lorsqu'il n'y a que des variables du groupe "Jour 1".

TABLEAU 3.3.9. *R_T^2 obtenu avec la méthode d'estimation/test à partir des variables sélectionnées par l'addition par étape sur les 300 premières observations et l'optimisation des paramètres par la VC avec une observation à la fois.*

	IVB_m	débit_e	DCO_e	DBO5_lab_e	MES_e
RLM	-0,5957	0,0331	-0,0474	-0,5347	-1,0492
RR	-0,1737	0,1282	0,0189	-0,3173	-1,2082
PCR	-0,5957	0,0331	-0,0474	-0,5347	-1,0492
LRR	-0,5928	0,0567	-0,0474	-0,6636	-1,0492
PLS (uni)	-0,5957	0,0331	-0,0474	-0,5176	-1,0492

	MES_lab_e	$N\text{-}NH_3$	MES_nd_e	$O\text{-}PO_4$_e	Turbid_e
RLM	-1,0281	-2,9889	-0,3742	-0,9667	-5,1222
RR	-1,0281	-2,6527	-0,3033	-3,8430	-0,4472
PCR	-1,0324	-2,9889	-0,3742	-0,9714	-0,4699
LRR	-1,0281	-2,9889	-0,3742	-0,8826	-0,4699
PLS (uni)	-1,0281	-2,9889	-0,3742	-0,9742	-0,4699

	pHcomp_e	pHmoy_e	Tmoy_e	Cond_e	DCOsol_e
RLM	-0,4618	0,0271	0,1351	0,2241	0,1870
RR	-0,3139	0,1368	0,1303	0,2514	0,2366
PCR	-0,4618	0,0271	0,1351	0,2241	0,1870
LRR	-0,4618	0,0225	0,1351	0,2434	0,1867
PLS (uni)	-0,4618	0,0271	0,1351	0,2241	0,1870

Dans l'ensemble, la RLM avec sélection de variables, PCR et PLS univarié donnent les meilleurs résultats lorsque le nombre de variables explicatives est grand. Dans la littérature, PLS est reconnue comme donnant de bons résultats lorsqu'il y a beaucoup de variables explicatives et que la variance de celles-ci est élevée. Les modèles gagnent toutefois en efficacité lorsqu'une sélection de

variables est d'abord faite. Lorsque le nombre de variables (significatives) est plus petit, la RR devient souvent la meilleure méthode. Les autres donnent souvent des résultats très semblables à ceux de la RLM. Les RN et LRR donnent les moins bons résultats. Malgré tout, pour plusieurs variables dépendantes, les modèles donnant les meilleurs résultats sont les modèles AR d'ordre 1 ou 2. Il est possible, sur une même journée, que les variables explicatives servent à décrire le comportement des variables dépendantes, mais sur deux jours ou plus, l'information n'est pas toujours utile. Donc, bien souvent, il est inutile de s'embêter à utiliser des méthodes compliquées puisque le modèle le plus simple est le meilleur.

3.3.4. Temps de calcul

Tous les modèles ont été ajustés avec des fonctions Splus dont le code est donné à l'annexe C. Au niveau de l'efficacité pour le temps de calcul, aucune méthode ne se compare bien sûr avec la RLM puisque cette méthode n'a aucun paramètre à optimiser. La RR s'ajuste facilement puisque la SSE est quadratique comme le montre la figure 3.3.1 qui représente la SSE obtenue en fonction de différentes valeurs du paramètre k pour la DCO avec la méthode d'estimation/test sur les 146 variables explicatives. Tous les résultats ont cette allure, à savoir une fonction quadratique lisse. Il est donc facile de cibler la région où le paramètre estimé optimal de la régression se situe. Il suffit d'essayer une série de valeur très espacées d'abord, puis de cibler la région où se situe le paramètre optimal et ce recommencer en prenant une autre série de valeur du paramètre k plus rapprochées et plus proches de la valeur optimale. Pour PLS et PCR, il est relativement facile de trouver le paramètre optimal dans la mesure où il suffit d'essayer toutes les possibilités. Cette méthode demande cependant un peu plus

de temps que la RR. Les pires méthodes au niveau du temps et de la difficulté de détermination des paramètres sont LRR et les réseaux de neurones.

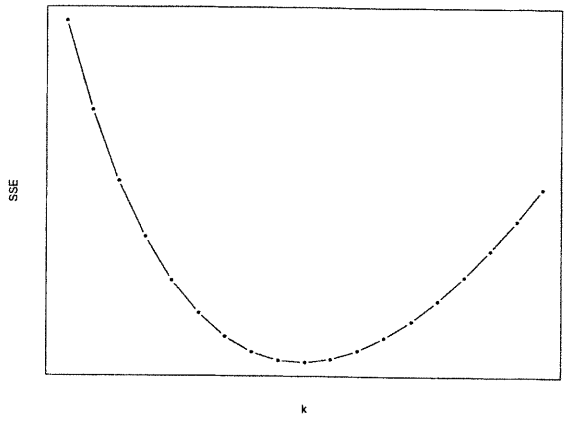


FIGURE 3.3.1. *Somme des erreurs au carré pour différentes valeurs du paramètre de la RR pour la DCO avec la méthode d'estimation/test sur les 146 variables explicatives. Toutes les courbes ont l'allure d'une fonction quadratique lisse ce qui facilite la recherche du paramètre k optimal.*

La première méthode, LRR, nécessite l'estimation de la SSE pour plusieurs valeurs des deux paramètres à ajuster. Le résultat n'est ni quadratique ni monotone en fonction de l'un des deux paramètres. Il est donc très difficile de déterminer les paramètres optimaux selon les données. La seule manière consisterait à regarder les valeurs propres (pour le premier paramètre) et la première composante des vecteurs propres (pour le second paramètre) et à essayer de fixer les seuils entre chacune des valeurs observées pour les deux paramètres et à essayer ensuite tous les couples possibles. Donc, pour M variables explicatives et une variable dépendante qui composent la matrice servant à calculer les valeurs et vecteurs propres, il y a M seuils potentiels pour chacun des deux paramètres, donc il y a M^2 possibilités. Lorsque le nombre de variables est grand, il est pratiquement impensable d'essayer toutes ces possibilités. Par exemple, pour 150

variables, comme c'était le cas avec le jeu de données étudié, il y a 22500 combinaisons possibles. Nous n'avons pas essayé toutes ces possibilités évidemment. Pour le premier paramètre, nous avons essayé une séquence de douze valeurs tandis que, pour le second, la séquence utilisée contenait sept valeurs. Nous avons choisi ces valeurs en regardant l'ordre de grandeur des valeurs propres et de la première composante des vecteurs propres pour quelques-unes des variables dépendantes. Le fait que nous avons utilisé une petite séquence pour chacun des paramètres pourrait, bien sûr, expliquer pourquoi cette méthode ne donne pas de très bons résultats. D'un autre côté, plus le nombre de combinaisons testées est grand, plus le temps de calcul augmente.

Pour les réseaux de neurones, il faut déterminer le nombre de neurones par couche cachée. Cependant, il faut aussi tester pour une couche cachée, deux couches cachées, ..., et nous décidons d'arrêter après un certain nombre de couches. Tout comme pour LRR, si nous essayons, par exemple pour deux couches cachées, toutes les combinaisons du nombre de neurones contenu dans chacune des couches, il y a plusieurs possibilités. De plus, étant donné que l'ajustement du modèle se fait par des algorithmes d'optimisation non linéaire, chacun des ajustements d'un modèle nécessite plus de temps de calcul que les méthodes linéaires. Finalement, l'ajustement du modèle dépend des paramètres aléatoires fixés au départ. Donc, un modèle avec les mêmes paramètres peut être ajusté plus d'une fois sans jamais fournir le même résultat. Afin d'essayer de trouver un modèle le plus efficace au niveau des prévisions, il faut ajuster le même modèle plus d'une fois et retenir le meilleur. Donc, en résumé, il faut déterminer le nombre de couches, le nombre de neurones dans chacune des couches et refaire l'ajustement du modèle plus d'une fois à chaque essai. Déjà, le temps de calcul est considérable. Nous pouvons aussi essayer de changer les fonctions d'activations et

recommencer. Bref, il est possible de s'amuser très longtemps avant d'arriver au "meilleur" résultat avec cette méthode. Ici, nous avons ajusté des RN avec deux couches cachées et essayer de mettre de 5 à 45 neurones par couche cachée, en comptant par multiples de 5. Nous avons toujours utilisé la même fonction d'activation, soit la fonction Tanh() pour les couches cachées et la fonction identité pour la couche de sortie.

3.4. Analyse des variables dépendantes

3.4.1. Analyse des variables IVB, DCO, DBO_5 et MES

Nous avons analysé plus en détails les variables IVB, DCO, DBO_5 et MES. Ces quatre variables ont été choisies par les ingénieurs de l'usine en raison de leur importance dans l'opération du traitement des eaux usées de l'usine. L'augmentation de la variable IVB est une conséquence d'une mauvaise opération dans le traitement, c'est un signe qu'il y a un problème. D'après les ingénieurs de l'usine, lorsque cette variable est à un niveau correct alors le traitement va bien et vice-versa. Les variables DBO_5 et MES sont également importantes puisqu'elles doivent respecter des normes environnementales. La première variable est traitée via la DCO avec laquelle elle est fortement corrélée. La variable DCO est utilisée comme référence, car le test pour obtenir le niveau de DBO nécessite 5 jours. À cause de la durée des tests, s'il y a un problème avec le niveau de DBO celui-ci ne sera identifié que plusieurs jours plus tard. Afin d'opérer le traitement pour diminuer la DBO et faire en sorte qu'elle respecte les normes environnementales, le niveau de DCO est donc utilisé à la place de la DBO. Ceci explique pourquoi nous regardons également la DCO.

Les modèles trouvés à partir des sous-groupes de variables sélectionnées par l'addition par étapes étaient assez semblables malgré les différences dans les sous-groupes. Le tableau 3.4.1 contient la liste des variables significatives lorsque la sélection est faite à partir du groupe "Jour 1". Les variables sont identifiées par les abbréviations décrites à l'annexe A.

Le tableau 3.4.2 donne la racine de la SSE moyenne pour tous les modèles trouvés avec la méthode d'estimation/test avec l'optimisation des paramètres par la VC avec une observation à la fois. Les meilleurs résultats sont en caractères gras dans le tableau. Si nous comparons les meilleurs modèles obtenus à partir des 146 variables, nous constatons que, mises à part les MES, les résultats sont les moins bons. Les racines de la SSE moyenne des modèles obtenus à partir des sous-groupes de variables sont passablement équivalents entre elles, sauf pour la variable IVB ajustée avec les variables sélectionnées à partir des 146 variables explicatives avec l'ajout des variables du groupe "Jour 1" lorsqu'il n'a que la variable correspondante du groupe "Jour 2" où le résultat est moins bon. Mises à part les MES, tous les modèles apportent une grande amélioration par rapport au modèle $\hat{y} = \bar{y}$ ($R_M^2 > 0$), mais les modèles AR sont assez équivalents aux autres modèles, souvent mêmes meilleurs comme dans le cas de la DCO, de la DBO_5 et des MES.

3.4.1.1. *Série chronologique*

Nous avons estimé une série chronologique avec les 150 premières observations de chacune de ces quatre variables. Afin d'obtenir des séries stationnaires, c'est-à-dire ne présentant pas de tendances croissantes ou décroissantes durant la période étudiée, nous avons dû prendre la différence entre les observations afin de modéliser la série $w_t = y_t - y_{t-1}$, où y_t est la variable que nous voulons modéliser

IVB

Tmin_a	IVB_m	pH_ba_m	O-PO_4_m
débit_e	DCODBO_e	MES_lab_e	enlevDBO

DCO

DCO_a	turbid_a	pHmoycorr_a	Tmoy_a
DCO_e	DCODBO_e	O-PO_4_e	Tmin_e
MES_T_e	enlevMES		

DBO_5

condcomp_a	DCO_e	DBO_lab_e	débit_bex
vol_air	enlevDBO		

MES

débit_d	F/M_m	VB30_m	O-PO_4_m
MES_lab_e	Tmax_e	bouessec	DBO_T_e

TABLEAU 3.4.1. *Liste des variables significatives lorsque la sélection par l'addition par étape est faite à partir du groupe "Jour 1" pour les variables IVB, DCO, DBO_5 et MES.*

mesurée au temps t. Nous avons d'abord vérifié si les séries différenciées contenaient seulement du bruit afin de voir si l'ajustement d'un modèle plus complexe que $y_t = y_{t-1}$ était nécessaire. Plusieurs tests existent pour vérifier si une série chronologique contient un signal ou si elle ne représente que du bruit. Nous avons utilisé le test de Ljung et Box qui se base sur la statistique

$$Q_{LB} = n(n+2)\sum_{h=1}^{H}\frac{\rho^2(h)}{n-h},$$

où n est le nombre d'observations dans la série et

$$\rho(k) = \frac{cov(w_t, w_{t-k})}{\sqrt{var(w_t)var(w_{t-k})}}$$

Variables	Modèle 146 variables		Modèle 73 variables	
	Méthodes	SSEmoy$^{\frac{1}{2}}$	Méthodes	SSEmoy$^{\frac{1}{2}}$
IVB	RR	7,1189	PLS (uni)	6,3486
DCO	RLM (AÉ)	195,1090	**RLM (AÉ)**	**180,4313**
DBO_5	RLM (AÉ)	15,7180	RLM (AÉ)	14,8116
MES	PCR	61,4335	PCR	59,1049

Variables	AÉ sur 146 variables		AÉ + "Jour 1" si juste "Jour 2"	
	Méthodes	SSEmoy$^{\frac{1}{2}}$	Méthodes	SSEmoy$^{\frac{1}{2}}$
IVB	RR	5,9188	Toutes	7,0382
DCO	RR	188,8350	PLS (uni)	184,4637
DBO_5	RR	14,5626	RR	14,3578
MES	Toutes sauf RR	78,8888	PLS (uni)	78,6683

Variables	AÉ sur "Jour 1"		Modèle $\hat{y} = \bar{y}$	AR(1)	AR(2)
	Méthodes	SSEmoy$^{\frac{1}{2}}$	SSEmoy$^{\frac{1}{2}}$	SSEmoy$^{\frac{1}{2}}$	SSEmoy$^{\frac{1}{2}}$
IVB	RR	5,9772	28,6300	5,4634	**5,4607**
DCO	Toutes sauf RR	180,4313	414,9244	190,6411	190,4390
DBO_5	RR	14,7315	22,0556	12,6879	**12,6755**
MES	Toutes sauf RR	74,8598	70,0225	**55,1085**	56,2141

TABLEAU 3.4.2. *Racine de la SSE moyenne pour les variables IVB, DCO, DBO_5 et MES obtenue par la méthode d'estimation/test avec l'optimisation des paramètres par la VC avec une observation à la fois.*

est l'autocorrélation de délai k, soit la corrélation entre la série au temps t et celle au temps $t - k$.

Si la série n'est que du bruit alors cette statistique est asymptotiquement une $\mathcal{X}^2$ avec H degrés de liberté. La statistique est calculée pour une petite suite de délais, par exemple $H = 6, 12, 18, 24$. Le test est rejeté pour de grandes valeurs de la statistique Q_{LB}.

Pour les séries représentant les variables IVB, DBO_5 et MES, le test conclut que les séries ne sont que de bruit et qu'elles peuvent être modélisées par $\hat{w}_t = 0$, i.e. $\hat{y}_t = y_{t-1}$. En faisant une sélection de variables avec l'addition par étape à partir des 150 premières observations et des 146 variables explicatives, nous constatons que, pour ces trois variables, la variable correspondante du groupe "Jour 2" n'est pas incluse parmi les variables significatives. La variable DCO n'est pas considérée comme étant du bruit uniquement. Nous avons ajusté un modèle à la série différenciée et le modèle obtenu le plus simple est un AR(9). Ceci nous laisse croire que les observations de plusieurs jours précédents peuvent être utiles pour prévoir cette série. Si nous regardons les variables sélectionnées comme pour les trois autres, nous notons cependant que la DCO du groupe "Jour 2" n'est pas significative. Il y aurait donc lieu d'approfondir un peu plus l'analyse de cette série. Nous avons toutefois arrêté ici notre analyse des variables avec les séries chronologiques.

3.4.1.2. *Meilleur modèle*

Pour la DCO, le modèle ajusté avec les variables sélectionnées à partir de l'addition par étape sur les variables du groupe "Jour 1" seulement donne le meilleur résultat. Le fait d'utiliser une des autres méthodes avec les variables sélectionnées par l'AÉ sur le groupe "Jour 1" n'améliore pas les résultats. Pour la DBO_5 et IVB, le meilleur modèle est celui n'utilisant que la DBO_5 et IVB respectivement de la journée précédente (AR(1)) alors que, pour les MES, c'est celui utilisant les

Variable	Méthode	Racine de la SSE moyenne	R^2_M	R^2_T
IVB	AR(1)	5,4607	0,9636	0,0010
DCO	RLM avec AÉ sur "Jour 1"	180,4313	0,8109	0,1042
DBO_5	AR(1)	12,6755	0,6697	0,0019
MES	AR(2)	55,1085	0,3806	0

TABLEAU 3.4.3. *Description des méthodes ayant donné les meilleurs modèles pour les variables IVB, DCO, DBO_5 et MES.*

MES des deux journées précédentes (AR(2)). Les SSE minimales obtenues pour chacune des quatre variables précédentes sont fournies au tableau 3.4.3. Chacune des quatre variable dépendantes ainsi que leur prévision sur le jeu de validation avec le meilleur modèle sont illustrées à la figure 3.4.1 pour le jeu de validation. Les résidus sont illustrés à la figure 3.4.2. Nous constatons que l'ajustement des variables IVB et DBO_5 est très bon alors que celui de la DCO et des MES l'est un peu moins. En particulier, le modèle pour les MES a tendance à surestimer les observations. Mise à part la DCO, la variabilité des résidus augmente lorsque les prévisions sont plus grandes. Lorsque ces variables sont élevées, le modèle est moins efficace pour les prévoir.

3.4.1.3. *Coefficients des modèles obtenus pour la DCO avec la méthode d'estimation/test sur les variables du groupe "Jour 1"*

Regardons d'un peu plus près les différences entre les modèles ajustés à partir de toutes les observations obtenus pour la variable DCO. Le tableau 3.4.4 présente les coefficients des modèles obtenus à partir des variables du groupe "Jour 1". Les coefficients associés à chaque variable sont donnés pour les cinq méthodes de modélisation univariées. La racine de la SSE moyenne obtenues avec

la méthode d'estimation/test avec l'optimisation des paramètres par la VC avec une observation à la fois est également donnée à la fin du tableau. Il ne faut pas oublier qu'un coefficient identique pour deux variables différentes ne veut pas dire la même chose étant donné que les variables ont des moyennes et des écarts type différents, c'est pourquoi ces derniers sont donnés en référence dans les deux dernières colonnes du tableau 3.4.4.

La RLM avec l'AÉ donne le meilleur résultat par rapport à la SSE suivi de près par LRR qui, pour une fois, donne de bons résultats. Les autres méthodes font un peu moins bien. Nous remarquons au tableau 3.4.4 que les coefficients associés à chaque variable explicative sont souvent assez différents d'une méthode à une autre. Par exemple, pour la variable pH moyen corrigé à l'affluent, la variable n'est pas considérée comme étant significative pour la RLM avec l'AÉ, son coefficient est petit pour la RR, PCR et PLS alors que, pour LRR, le coefficient est important. Nous observons ceci pour plus d'une variable. D'un autre côté, il y a des coefficients qui n'ont pas le même signe d'un modèle à un autre, comme pour la DBO en kg/T mesurée à l'effluent. Bref, pour un modèle la variable a un effet positif sur la DCO alors que pour un autre, elle a un effet négatif. Ce fait plutôt contradictoire s'explique probablement par la présence de collinéarité parmi les variables, collinéarité qui est traitée différemment d'une méthode à une autre. Prenons un exemple simple pour illustrer ce fait. Si nous observons $\underset{\sim}{x}_1 = (2,4,6,8,10)'$ et $\underset{\sim}{x}_2 = (1,2,3,4,5)'$ alors $0{,}5\underset{\sim}{x}_1+0{,}5\underset{\sim}{x}_2=1{,}0\underset{\sim}{x}_1-0{,}5\underset{\sim}{x}_2$. Les signes des coefficients ne sont pas les mêmes, mais l'effet sur les prévisions le sera. Ceci peut également expliquer les différences entre les coefficients obtenus d'un modèle à un autre.

TABLEAU 3.4.4. *Tableau des coefficients des modèles obtenus pour la variable DCO avec la méthode d'estimation/test avec l'optimisation des paramètres par la VC avec une observation à la fois sur toutes les variables du groupe "Jour 1". La moyenne et l'écart type de chacune des variables sont également donnés à titre de référence.*

coefficients	RLM-AÉ	RR	PCR	LRR	PLS (uni)	Moyenne	Écart type
constante	142,826	0,001	-10,157	216,573	-5088,842		
papier	-	0,055	0,088	-0,028	0,047	1019,645	108,294
débit_a	-	-0,002	0,001	-0,001	0,092	40891,846	5352,500
débit_d	-	0,001	-0,002	-0,004	-0,100	4235,200	2077,865
DBO_5_lab_a	-	0,082	0,069	0,005	0,048	1181,392	231,176
DCO_a	0,119	0,080	0,116	0,094	0,003	4417,862	853,952
DCODBO_a	-	0,006	0,352	-35,704	0,000	3,767	0,462
MES_a	-	-0,477	-0,450	-0,323	-0,463	399,561	85,646
MES_lab_a	-	0,337	0,287	0,410	0,316	368,827	87,092
MESnd_a	-	-0,004	-0,128	0,092	-0,065	169,471	42,020
DCO_T_a	-	0,784	-0,069	1,319	0,486	180,786	38,763
DBO_T_a	-	-0,098	0,040	-1,314	0,004	48,371	10,622
$O\text{-}PO_4$_a	-	-0,078	-2,680	-7,816	-0,035	1,023	0,486
turbid_a	-	0,132	0,185	0,142	0,133	474,002	101,912
pHcomp_a	-71,321	-0,087	-1,083	-58,654	-0,040	5,289	0,481
pHmoy_a	-	-0,030	-1,932	9,537	-0,013	5,588	0,481
pHmin_a	-	-0,068	-3,543	-11,385	-0,027	4,313	0,510
pHmax_a	16,486	0,284	5,324	14,914	0,112	7,303	1,166
pHmoycorr_a	**-**	**-0,074**	**-2,966**	**-72,996**	**-0,037**	**8,467**	**0,359**
Tmin_a	-	-0,024	-0,666	-1,267	0,008	36,467	4,463
Tmax_a	3,918	0,712	-0,275	3,030	0,350	57,023	6,855
Tmoy_a	-	0,198	-0,411	5,694	0,089	43,295	4,173
Tmoycorr_a	-	-0,074	1,687	-14,998	-0,042	31,661	1,543
condmoy_a	0,149	0,086	0,069	0,131	0,066	1257,471	255,266
condmin_a	-	0,018	0,024	0,022	0,002	677,887	227,105
condmax_a	-	0,007	0,007	0,004	-0,155	7324,064	2936,149
condcomp_a	-	0,210	0,212	0,045	0,190	1126,431	231,420

coefficients	RLM-AÉ	RR	PCR	LRR	PLS (uni)	Moyenne	Écart type
MESLM_ba_m	-	-0,040	-0,031	-0,018	-0,061	5119,316	533,766
MES_bex_m	-	-0,004	-0,005	0,008	0,013	10403,617	2046,874
F/M_m	-	-0,002	-0,064	-434,200	-0,001	0,201	0,043
VB30_m	0,263	0,349	0,314	0,063	0,356	441,460	172,895
IVB_m	-	0,088	0,282	0,537	0,012	87,515	36,849
pH_ba_m	-	0,021	0,238	206,060	0,007	6,700	0,119
Temp_ba_m	-	0,117	1,364	1,936	0,044	32,765	1,934
$N\text{-}NH_3$_m	-12,942	-0,775	-2,652	-13,790	-0,349	3,842	2,097
$O\text{-}PO_4$_m	-	-0,224	-6,301	-4,869	-0,099	1,841	1,097
débit_e	-	0,001	-0,002	-0,010	0,160	36649,280	5468,216
DCO_e	0,545	0,698	0,627	0,323	0,622	1824,889	419,417
DBO_lab_e	1,842	0,957	1,133	2,126	1,444	53,296	21,762
DCODBO_e	-	-0,900	-1,481	-0,516	-1,146	37,634	11,504
MES_e	-	-0,498	-0,089	-0,468	-0,514	139,378	84,406
MES_lab_e	-	-0,261	-0,229	-0,155	-0,318	137,436	82,749
$N\text{-}NH_3$_e	-	-0,519	-3,597	2,338	-0,253	3,254	2,614
MESnd_e	-	0,638	0,208	0,850	0,562	71,606	31,071
$O\text{-}PO_4$_e	-18,233	-0,243	-6,224	-11,376	-0,107	1,912	1,086
turbid_e	-	-0,059	-0,696	0,622	-0,180	61,054	37,384
pHcomp_e	-	-0,002	0,261	-63,443	-0,004	7,138	0,155
pHmoy_e	-	-0,020	0,001	-157,636	-0,011	6,594	0,139
pHmin_e	-	-0,003	-0,556	51,992	-0,005	6,461	0,123
pHmax_e	-	-0,021	-0,500	-30,104	-0,012	**6,742**	**0,170**
Tmin_e	10,706	0,725	0,629	10,20	9 0,334	30,448	3,389
Tmax_e	-10,974	0,037	1,569	-12,672	0,008	35,131	2,118
Tmoy_e	-	0,136	1,668	12,147	0,057	33,842	2,176
cond_e	-	-0,236	-0,194	-0,020	-0,267	1526,674	286,358
condmoy_e	-	0,076	0,104	0,068	0,039	1518,946	309,792
condmin_e	-	-0,022	-0,024	-0,044	-0,059	1003,518	482,499
condmax_e	-	-0,053	-0,069	-0,050	-0,064	1699,785	345,212
débit_bex	-	-0,012	-0,011	0,028	-0,057	3893,049	991,424

coefficients	RLM-AÉ	RR	PCR	LRR	PLS (uni)	Moyenne	Écart type
vol_air	-	0,000	0,000	0,000	-0,001	1474075	182692
bouessec	-	0,213	-0,068	-3,356	0,082	39,866	9,804
DCO_T_e	-	-0,173	0,350	5,939	-0,030	66,636	15,958
enlevDCO	-	0,759	0,013	0,240	0,368	62,696	7,863
MES_T_a	-	-0,148	-0,414	-2,684	-0,059	16,301	3,756
MES_T_e	-11,016	-0,095	-0,278	-14,606	-0,055	4,979	2,956
enlevMES	-	-0,031	-0,145	-1,248	-0,027	67,558	21,028
DBO_T_e	-	0,033	1,166	27,17	3 0,052	1,931	0,747
enlevDBO	-	0,144	-0,235	13,671	-0,072	95,793	2,298
MESnd_T_a	-	-0,024	-0,082	0,871	-0,009	6,947	1,970
MESnd_T_e	-	-0,022	0,137	-1,127	-0,001	2,573	1,067
enlevMESnd	-	-0,012	-0,014	-0,508	-0,113	61,540	17,125
MESnd/MES_e	-	-0,657	0,131	-1,331	-0,808	57,765	22,177
DBO_kg_e	**-**	**0,041**	**1,101**	**-33,790**	**0,053**	**1,915**	**0,790**
MESLM_T_m	-	0,097	0,200	-1,310	0,049	50,186	5,390
MES5jrs_a	-	0,657	0,873	0,513	0,604	80,981	16,386

Racine de la SSE moyenne	180,4313	251,1123	208,5122	186,8947	234,6131		

La RLM ajustée avec l'AÉ donne le modèle le plus simple avec seulement quatorze coefficients non nuls. La RR, PCR et PLS donnent des modèles où toutes les variables ont de petits coefficients, mis à part le terme constant. Pour ces modèles, il y a des variables qui ont des coefficients pratiquement nuls et n'ont que peu d'impact sur les prévisions si nous considérons l'ordre de grandeur des variables. LRR a un terme constant élevé et les coefficients de certaines variables sont importants. La SSE des modèles obtenus par la RLM avec l'AÉ et par LRR est semblable, cependant, les modèles obtenus sont très différents. Toutefois, les prévisions sont comparables dans les deux cas, comme nous pouvons le constater

Variable	Méthode	Racine de la SSE moyenne	R^2_M	R^2_T
débit	RR, 146 variables	2914,8130	0,6468	0,0617
MES_lab	AR(2)	51,8866	0,4518	0
N-NH_3	AR(2)	2,0951	0,2598	0
MES_nd	AR(1)	22,0110	0,5415	0,0132
O-PO_4	AR(1)	0,6270	0,7949	0,0067
turbid	AR(2)	23,1332	0,4676	0
pHcomp	AR(2)	0,1294	0,2504	0
pHmoy	PLSmulti 146 variables	0,1394	0,1344	0,2059
Tmoy	RLM (SPAP) 146 variables	1,3307	0,7164	0,1509
cond	RLM (SPAP) 146 variables	86,7329	0,9322	0,2792
DCOsol	RR, variables sélectionnées parmi "Jour1"	128,4051	0,7198	0,2778

TABLEAU 3.4.5. *Description des méthodes ayant donné les meilleurs modèles*

à la figure 3.4.3 qui représente la DCO observée ainsi que les prévisions de la RLM avec l'AÉ et LRR. En particulier, même lorsque les prévisions s'éloignent des valeurs observées, les deux modèles continuent d'avoir des prévisions semblables.

3.4.2. Analyse des autres variables

Les racines des SSE moyennes minimales obtenues pour chacune des autres variables sont fournies au tableau 3.4.5. Encore une fois, les meilleurs modèles sont souvent les modèles AR.

La représentation de chacune des variables dépendantes est donnée aux figures B.0.1 à B.0.8 de l'annexe B. Sur ces graphiques, nous remarquons que la variable débit a subi un fort changement autour du 20 mai 1998: sa moyenne

passe de 44000 m^3/j à 34000m^3/j, tandis que son écart type augmente de 2600 m^3/j à 3275 m^3/j. Cette baisse du niveau a affecté certains modèles, ce qui peut expliquer pourquoi les résultats pour cette variable sont différents selon les sous-groupes utilisés. Si nous comparons la racine de la SSE moyenne donnée tableau 3.4.5 à l'écart type donné plus haut, nous remarquons qu'elles ont le même ordre de grandeur. Le modèle détecte le palier du 20 mai, mais n'arrive pas à bien expliquer la variation du débit, seules les tendances à long terme sont identifiées.

Les variables N-NH_3 et pH moyen ont également des résultats qui sont comparables à leur écart type, les modèles ne parviennent pas à prévoir efficacement ces variables. Pour le pH moyen, il faut en plus noter que le "meilleur" modèle est supérieur de beaucoup à tous les autres modèles ajustés pour cette variable. Le modèle donnant les deuxièmes meilleurs résultats n'a qu'un R^2_M et un R^2_T de 0,0591 et 0,1368 respectivement.

Parmi les autres variables, la DCO soluble et la conductivité ont des racines de la SSE moyenne représentant le tiers de leur écart type alors que, pour les autres, elles représentent entre 60 % et 80 % de leur écart type. Dans tous les cas, les meilleurs modèles apportent une amélioration par rapport au modèle $\hat{y} = \bar{y}$.

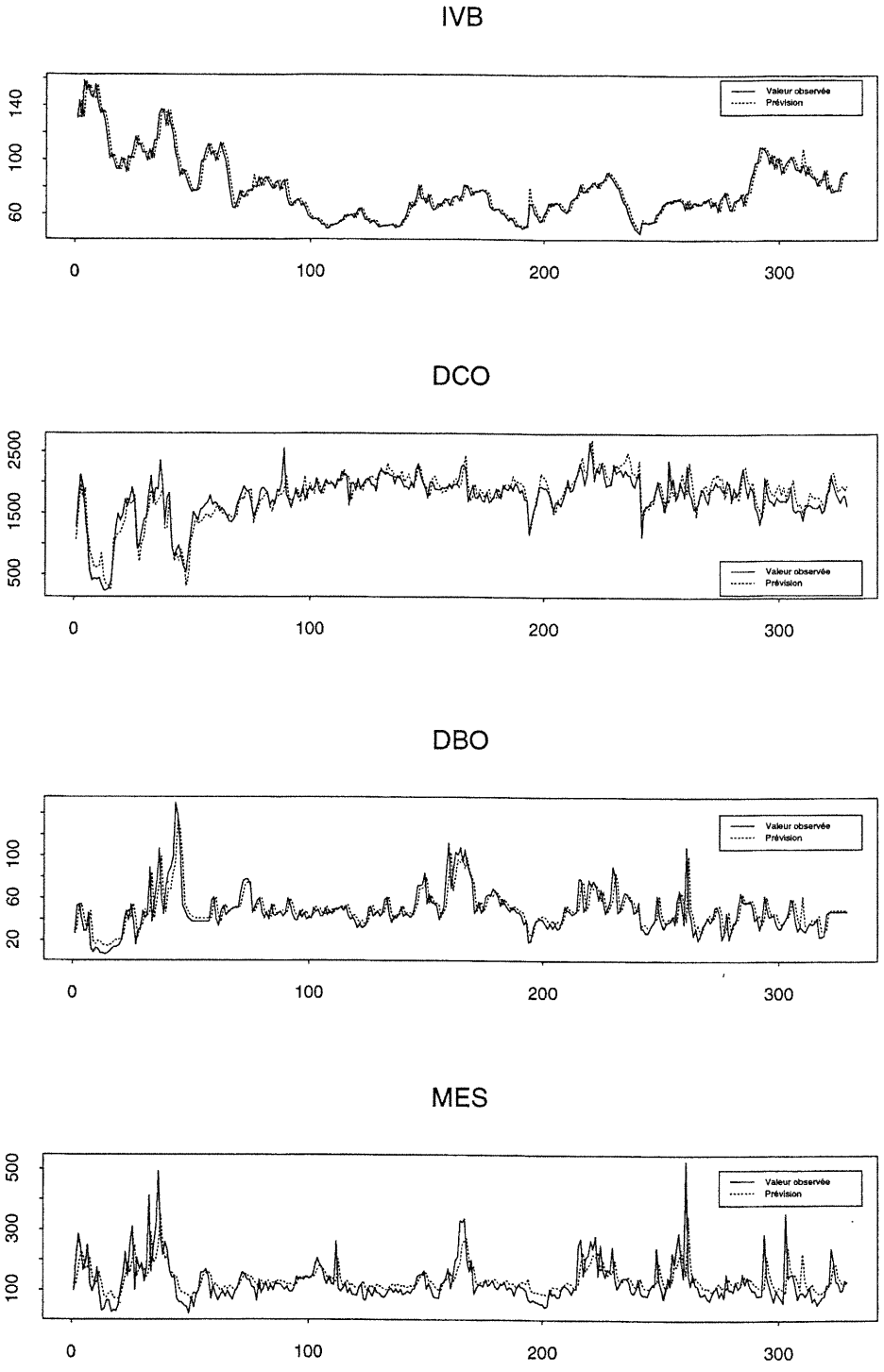


FIGURE 3.4.1. *Représentation des variables dépendantes sur le jeu de validation ainsi que de leur prévision obtenue avec le meilleur modèle pour les variables IVB, DCO, DBO_5_lab et MES.*

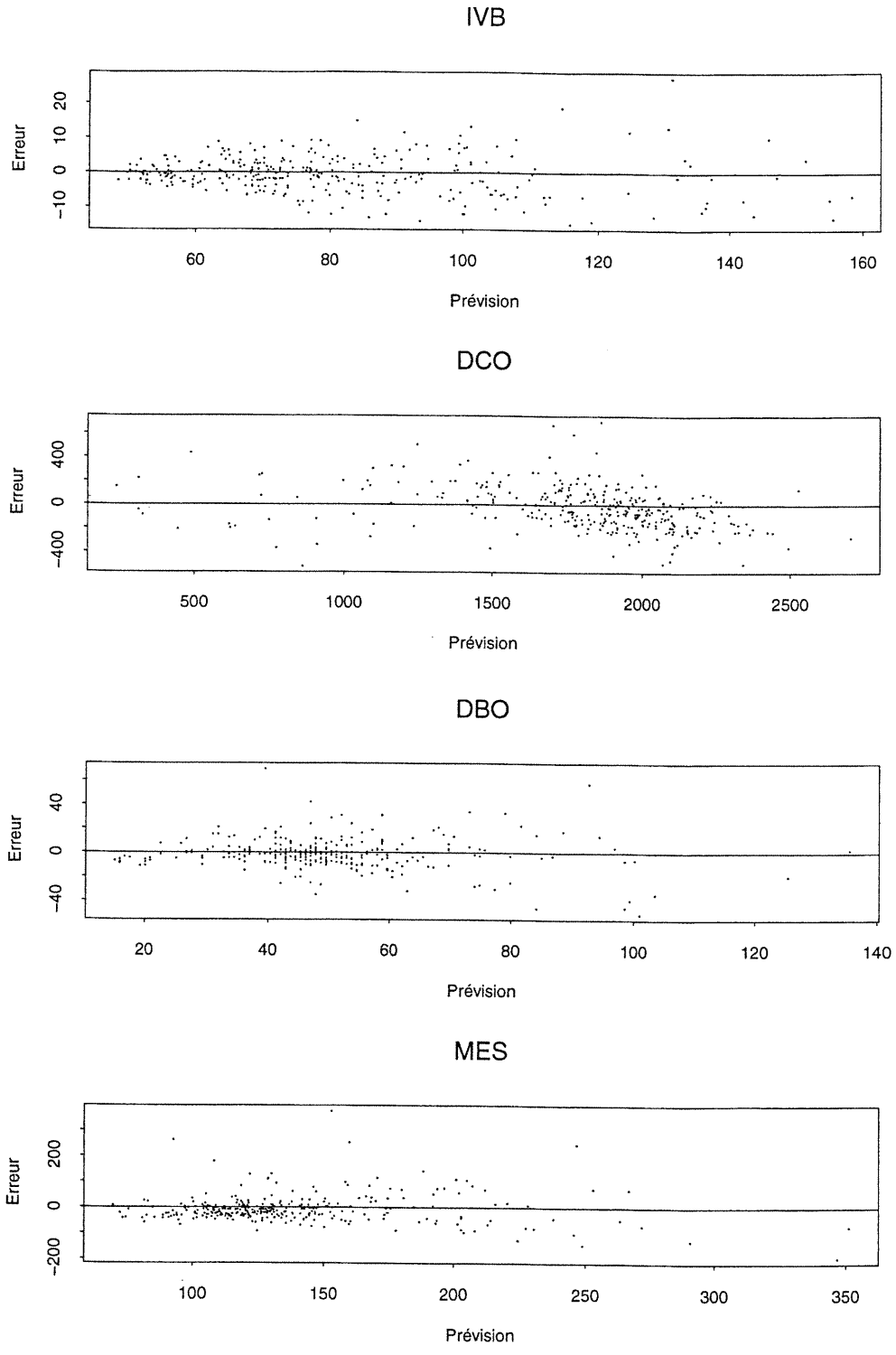


FIGURE 3.4.2. *Représentation des résidus sur le jeu de validation en fonction de leur prévision obtenue avec le meilleur modèle pour les variables IVB, DCO, DBO_5_lab et MES.*

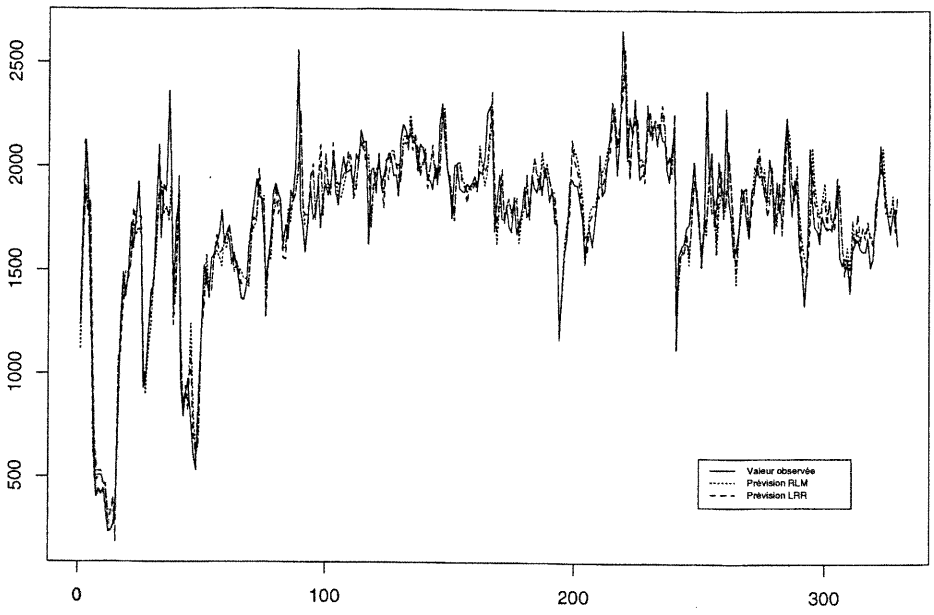


FIGURE 3.4.3. *Représentation de la DCO observée ainsi que les prévisions obtenues pour la RLM avec l'AÉ et pour LRR*

CONCLUSION

Le but de ce mémoire était de comparer différentes méthodes de modélisation, soit la RLM, la RR, PCR, LRR, PLS et les RN. Le jeu de données étudié se caractérisait par un taux de valeurs manquantes important et un grand nombre de variables. La quantité importante de variables observées fait en sorte que le jeu de données contient beaucoup de bruit et de collinéarité entre les variables et qu'il devient difficile d'identifier les signaux expliquant les variables dépendantes.

L'analyse du jeu de données avec les différentes méthodes de modélisation met en avant la RLM avec sélection de variables, PCR et PLS univarié lorsque toutes les variables explicatives sont utilisées. La première méthode conduit souvent à des modèles contenant peu de variables alors que les autres méthodes estiment des coefficients pour toutes les variables explicatives.

Nous avons estimé des modèles avec la méthode d'estimation/test et avec la VC, la question est ensuite de savoir sur quels résultats nous basons-nous ensuite. Les statistiques calculées à partir des résutats de la méthode d'estimation/test sont préférées à celles obtenues grâce à la VC en 10 blocs, car ces dernières ont tendance à être trop optimistes. Cependant, le modèle final est calculé à partir de toutes les observations.

Dans l'analyse du jeu de données de l'usine, nous constatons qu'il est avantageux de faire une sélection de variables. Mis à part le débit, les variables du groupe "Jour 2" n'apportent que peu de nouvelle information aux modèles.

Lorsque nous ne conservons que les variables significatives de la sélection de variables, la majeure partie de la corrélation disparaît souvent de telle sorte que celle-ci n'est plus vraiment un problème et alors toutes les méthodes univariées donnent approximativement les mêmes résultats. En effet, si deux variables sont très fortement corrélées, la sélection de variables ne choisira qu'une seule de ces variables la majorité du temps. La RR permet parfois d'obtenir un modèle un peu mieux qu'avec les autres méthodes.

Dans l'ensemble, au niveau du temps de calcul, les méthodes les plus rapides sont le RLM suivie de la RR. Ensuite, il y a PCR puis PLS qui sont surtout avantagées par rapport à LRR et RN par le fait qu'elles n'ont qu'un seul paramètre à optimiser et que celui-ci peut prendre un nombre fini de valeurs. La méthode LRR a deux paramètres à optimiser qui sont théoriquement continus. Les RN ont un très grand nombre de paramètres à estimer: le nombre de couches cachées, le nombre de neurones par couche cachée et le type de fonction d'activation. De plus, les RN sont ajustés avec des algorithmes d'optisation dont le résultat dépend des valeurs initiales prises par les poids du modèle. Bref, ce dernier modèle prend beaucoup plus de temps de calcul que toutes les autres méthodes.

Au niveau de l'interprétabilité, la RLM avec sélection de variables fournit un modèle très court par rapport à ceux obtenus à partir des autres méthodes. Cependant, il faut se rappeler qu'une méthode de sélection différente peut donner un sous-groupe de variables significatives différent et donc modifier l'interprétation que nous faisons des résultats. Les autres méthodes linéaires donnent un coefficient à chaque variable. Il est donc possible d'interpréter ces modèles de la même façon que pour la RLM, mais les modèles sont longs et contiennent toutes les variables. En fait, seuls les RN posent de sérieuses difficultés quant à l'interprétation des résultats étant donné que l'effet de chaque variable n'est

pas linéaire. La force des RN réside dans leur pouvoir de prévision d'après la littérature, bien qu'ils n'aient cependant pas été très bons avec le jeu de données étudié.

Bref, pour des modèles rapidement obtenus et relativement efficaces, la RLM avec sélection de variables est un candidat à ne pas écarter. Il y a avantage à sélectionner des sous-groupes de variables significatives avant d'appliquer les méthodes sur beaucoup de variables contenant trop de bruit pour que les méthodes de modélisation soient efficaces. Il n'y a que peu de différence entre les modèles ajustés avec la RLM avec l'AÉ ou avec la SPAP. Souvent, les modèles estimés sont les mêmes. D'après la littérature, lorsqu'il y a beaucoup de variables explicatives et que leur variance est grande, PLS donne de bons résultats. Effectivement, avec 146 variables, l'efficacité de PLS se compare à celle de la RLM avec sélection, bien qu'il existe de meilleurs modèles plus courts. PLS multivarié ne donne pas de très bons résultats avec le jeu de données étudié. Certaines variables sont faiblement corrélées alors il n'est pas surprenant que ce soit le cas. Les procédures univariées sont plus appropriées dans un tel cas et, en effet, elles donnent de meilleurs résultats.

Nous avons constaté que l'utilisation d'un modèle permet, dans presque tous les cas, d'obtenir une amélioration par rapport au modèle où nous estimons les variables dépendantes par leur moyenne estimée. Cependant, très souvent, un simple modèle AR d'ordre 1 ou 2 est celui donnant les meilleurs résultats: les variables explicatives, autres que celles correspondant à la variable dépendante, n'apportent pas d'information supplémentaire significative.

Nous avons utilisé les observations des journées précédentes pour prévoir les variables dépendantes, car, pour une même journée, celles-ci ne sont pas mesurées après les variables explicatives. Donc, pour une même journée, l'information pour

prévoir la suivante est déjà contenue dans la variable correspondant à la variable dépendante mesurée durant les jours précédents. Si nous voudrions identifier les variables qui influencent les variables dépendantes sur une même journée alors il faudrait construire de nouveaux modèles en utilisant toutes les variables mesurées durant la même journée. Ce modèle n'aurait cependant aucune utilité au niveau des prévisions.

Pour les données de l'usine, nous trouvons de bonnes prévisions pour les variables IVB, DBO_5, la DCO soluble et la conductivité. Les variables DCO, MES, MES_lab, MESnd, O-PO_4, turbidité, pH moyen et température moyenne sont aussi assez bien prévues par les modèles. Pour le débit, N-NH_3 et le pH composé, les modèles trouvés n'arrivent qu'à identifier les tendances qui se maintiennent sur de longues périodes. Les oscillations à court terme sont peu ou pas détectées par les modèles.

Il est clair qu'il y a de nombreuses autres méthodes de modélisation, entre autres des méthodes de modélisation non linéaires que nous n'avons que très peu abordées ici. De plus, dans la littérature, nous retrouvons plusieurs variations pour les méthodes qui ont été présentées, particulièrement pour les RN. Certains articles discutent également de façons de faire de la sélection de variables avec les méthodes autres que la RLM. Bref, il existe encore plusieurs autres voies à explorer.

Annexe A

DESCRIPTION DES VARIABLES

Tout d'abord, voici quelques abbréviations utiles:

- MES: Matières en suspension
- DCO: Demande chimique en oxygène
- DBO: Demande biochimique en oxygène
- MVS: Matières volatiles en suspension

TABLEAU A.0.1. *Tableau des variables mesurées au traitement des eaux usées, ainsi que leur abbréviation respective.*

ID	Variable	Abbréviation
1	Débit mesuré à l'usine en m^3/j	débit_us
2	Quantité de papier produite par l'usine en T/j	papier
3	MES mesurées aux émissaires 3, 4 et 5 en mg/L	MESno345
4	Débit à l'affluent en m^3/j	débit_a
5	Débit à la déviation en m^3/j	débit_d
6	DCO soluble à l'affluent en mg/L	DCOsol_a
7	DBO à l'affluent, valeur du laboratoire, en mg/L	DBO_5_lab_a
8	DCO à l'affluent en mg/L	DCO_a
9	Rapport DCO/DBO à l'affluent	DCODBO_a
10	MES à l'affluent en mg/L	MES_a
11	MES à l'affluent, valeur du laboratoire, en mg/L	MES_lab_a
12	MVS à l'affluent en mg/L	MVS_a
13	N-NH_3 à l'affluent en mg/L	N-NH_3_a

ID	Variable	Abbréviation
14	MES non dissoutes à l'affluent en mg/L	MESnd_a
15	DCO à l'affluent en T/j	DCO_T_a
16	DBO à l'affluent en T/j	DBO_T_a
17	DCO soluble à l'affluent en T/j	DCOsol_T_a
18	MES à l'affluent en T/j	MES_T_a
19	O-PO_4 à l'affluent en mg/L de P	O-PO_4_a
20	Turbidité à l'affluent en UTN	turdib_a
21	pH composé à l'affluent	pHcomp_a
22	pH moyen à l'affluent	pHmoy_a
23	pH minimal à l'affluent	pHmin_a
24	pH maximal à l'affluent	pHmax_a
25	pH moyen corrigé à l'affluent	pHmoycorr_a
26	Température minimale à l'affluent en degrés Celsius	Tmin_a
27	Température maximale à l'affluent en degrés Celsius	Tmax_a
28	Température moyenne à l'affluent en degrés Celsius	Tmoy_a
29	Température moyenne corrigée à l'affluent en degrés Celsius	Tmoycorr_a
30	Conductivité moyenne à l'affluent en μS/cm	condmoy_a
31	Conductivité minimale à l'affluent en μS/cm	condmin_a
32	Conductivité maximale à l'affluent en μS/cm	condmax_a
33	Conductivité composée à l'affluent en μS/cm	condcomp_a
34	MES en mg/L, liqueur mixte, boues activées, RBS moyen	MESLM_ba_m
35	MES en mg/L, boues extraites, RBS moyen	MES_bex_m
36	Rapport des nutriments/masse par jour, boues extraites, RBSmoyen	F/M_m
37	N-NO_2-NO_3 en mgN/L, RBS moyen	N-NO_2-NO_3_m
38	Volume des boues en ml, RBS moyen	VB30_m
39	Indice de volume des boues en ml/g, RBS moyen	IVB_m
40	pH, boues extraites	pH_ba_m
41	Température en degrés Celsius, boues extraites	Temp_ba_m
42	N-NH_3 en mg/L, RBS moyen	N-NH_3_m
43	O-PO_4 en mg/L de P, RBS moyen	O-PO_4_m
44	TAO endogène en mg/L/h , RBS moyen	TAOendo_m

ID	Variable	Abbréviation
45	TAO fin en mg/L/h, RBS moyen	TAOfin_m
46	TSUO endogène en mg/L/h/g, RBS moyen	TSUOendo_m
47	TSUO fin en mg/L/h/g, RBS moyen	TSUOfin_m
48	Débit à l'effluent en mg/L	débit_e
49	DCO à l'effluent en mg/L	DCO_e
50	DCO soluble à l'effluent en mg/L	DCOsol_e
51	DBO à l'effluent, valeur du laboratoire, en mg/L	DBO_lab_e
52	Rapport DCODBO à l'effluent	DCODBO_e
53	MES à l'effluent en mg/L	MES_e
54	MES à l'effluent, valeur du laboratoire, en mg/L	MES_lab_e
55	MVS à l'effluent en mg/L	MVS_e
56	N-NH_3 à l'effluent en mg/L	N-NH_3_e
57	MES non dissoutes à l'effluent en mg/L	MESnd_e
58	O-PO_4 à l'effluent en mg/L de P	O-PO4_e
59	Turbidité à l'effluent en UTN	turdib_e
60	pH composé à l'effluent	pHcomp_e
61	pH moyen à l'effluent	pHmoy_e
62	pH minimal à l'effluent	pHmin_e
63	pH maximal à l'effluent	pHmax_e
64	Température minimale à l'effluent en degrés Celsius	Tmin_e
65	Température maximale à l'effluent en degrés Celsius	Tmax_e
66	Température moyenne à l'effluent en degrés Celsius	Tmoy_e
67	Conductivité à l'effluent en μS/cm	cond_e
68	Conductivité moyenne à l'effluent en μS/cm	condmoy_e
69	Conductivité minimale à l'effluent en μS/cm	condmin_e
70	Conductivité maximale à l'effluent en μS/cm	condmax_e
71	Débit, boues extraites en m^3/ j	débit_bex
72	Volume d'air total en Sm^3/ j	vol_air
73	Boues secondaires en T/j	bouessec
74	DCO à l'effluent en T/j	DCO_t_e
75	Pourcentage d'enlèvement de la DCO	enlevDCO

ID	Variable	Abbréviation
76	MES à l'effluent en T/j	MES_T_e
77	Pourcentage d'enlèvement des MES	enlevMES
78	DBO à l'effluent en T/j	DBO_T_e
79	Pourcentage d'enlèvement de la DBO	enlevDBO
80	MES non dissoutes à l'affluent en T/j	MESnd_T_a
81	MES non dissoutes à l'effluent en T/j	MESnd_T_e
82	Pourcentage d'enlèvement des MES non dissoutes	enlevMESnd
83	Rapport des MES non dissoutes par rapport aux MES à l'effluent	MESnd/MES_e
84	DBO à l'effluent en kg/T	DBO_kg_e
85	MES, liqueur mixte en T/j, RBS moyen	MESLM_T_m
86	Quantité de désencré produite	désencré
87	Quantité de pâte de meules produite	PMM
88	Quantité de pâte thermomécanique produite	PTM
89	Quantité de pâte chimique à haut rendement produite	PCHR
90	Total des MES en T/j sur les 5 derniers jours à l'affluent	MES5jrs_a

Annexe B

VARIABLES DÉPENDANTES

Cette section contient des statistiques descriptives pour les variables dépendantes ainsi que la représentation des chacune des variables dépendantes.

Variable	Minimum	Moyenne	Maximum	Écart type	Unités
IVB_m	38,6	87,4	202,8	36,8	mg/L
debit_e	18433	366010	56036	5458	m^3/j
DCO_e	235	1831	3244	417,3	mg/L
DBO5_lab_e	7	53	150	22	mg/L
MES_e	23	139	1005	84	mg/L
MES_lab_e	23	137	930	83	mg/L
N-NH3_e	0,00	3,24	37,00	2,60	mg/L
MES_nd_e	10	711	170	31	mg/L
O-PO_4_e	0,02	1,92	7,00	1,09	mg/L de P
Turbid_e	11,7	61,0	505,0	37,4	UTN
pHcomp_e	6,71	7,14	7,70	0,16	
pHmoy_e	5,50	6,59	6,97	0,14	
tmoy_e	27,0	33,9	40,3	2,2	°C
cond_e	581	1531	2346	285	μS/cm
DCOsol_e	185	1555	2324	338	mg/L

TABLEAU B.0.1. *Statistiques descriptives pour les variables dépendantes*

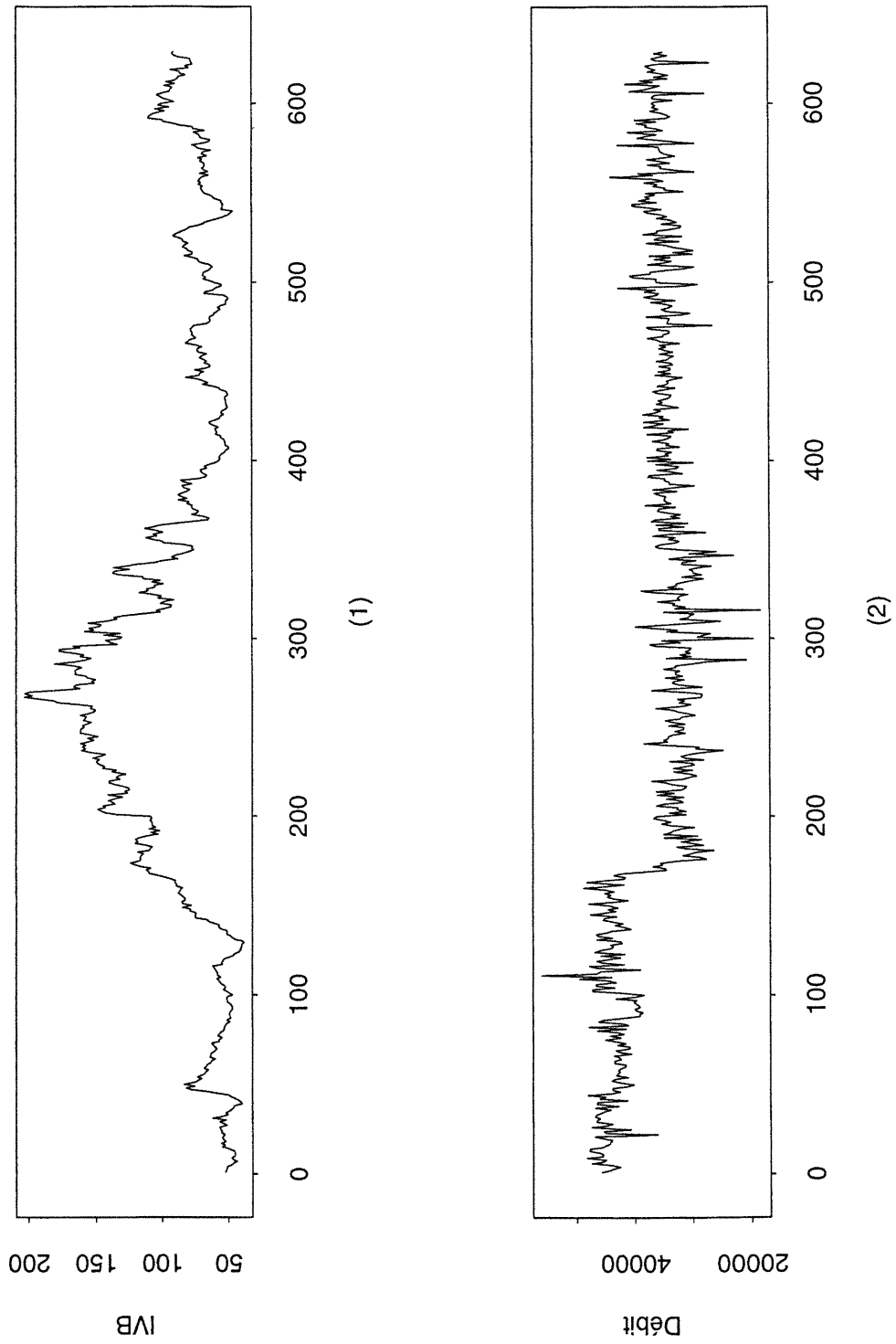


FIGURE B.0.1. *Représentation des variables IVB (1) et débit (2) du 01/12/97 au 26/09/99.*

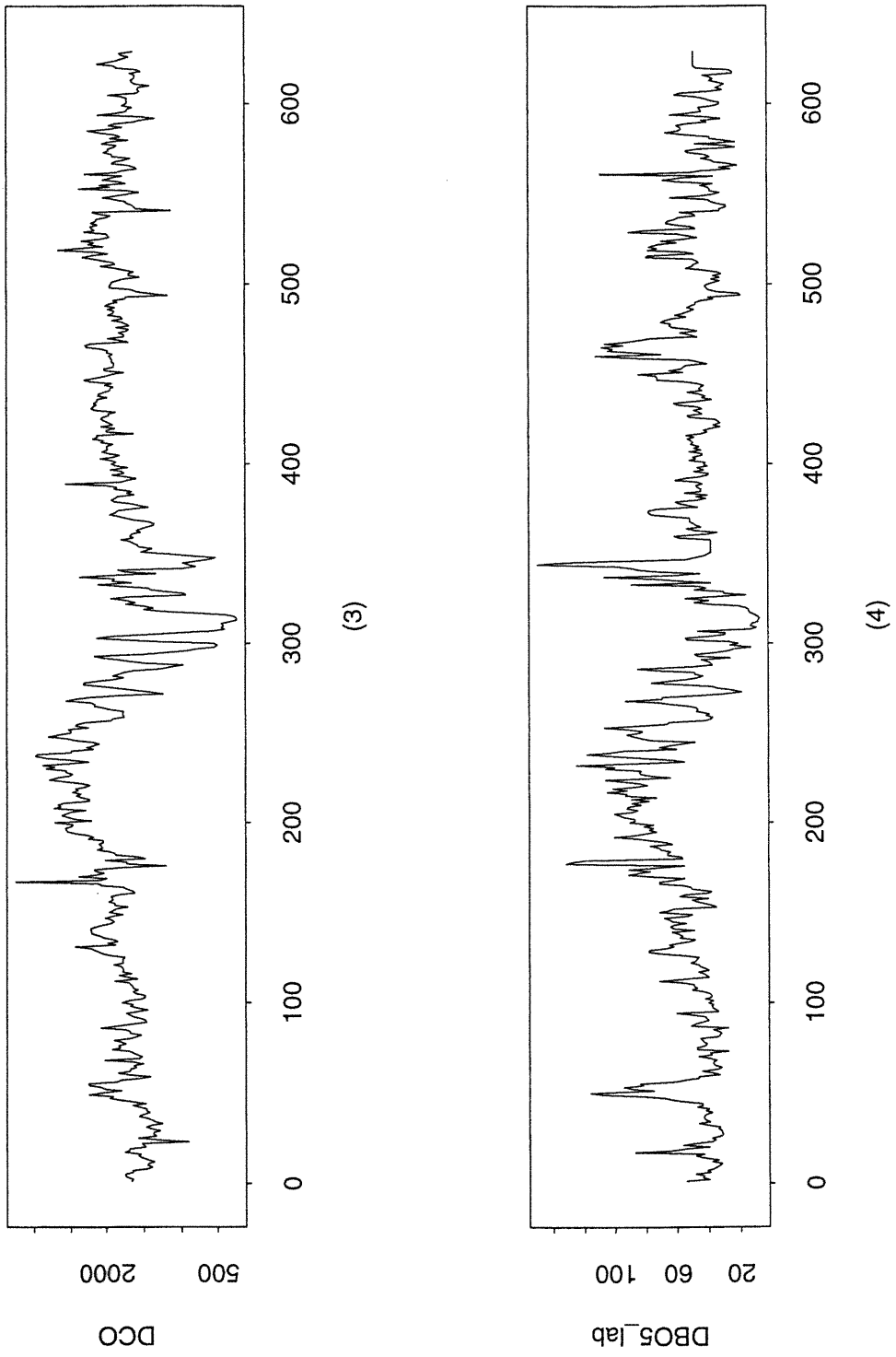


FIGURE B.0.2. *Représentation des variables DCO (3) et DBO_5 (4) du 01/12/97 au 26/09/99.*

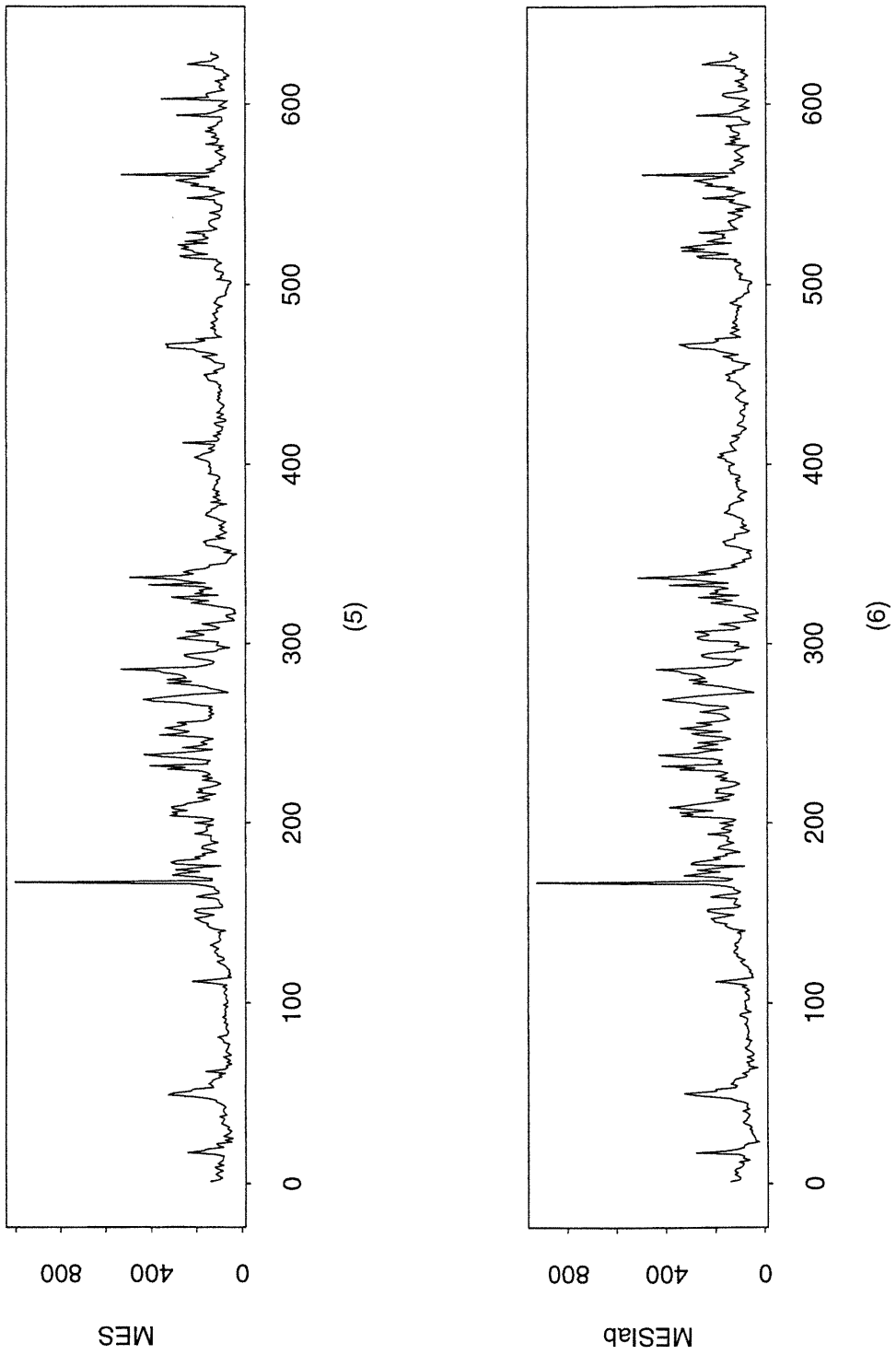


FIGURE B.0.3. *Représentation des variables MES (5) et MES_lab (6) du 01/12/97 au 26/09/99.*

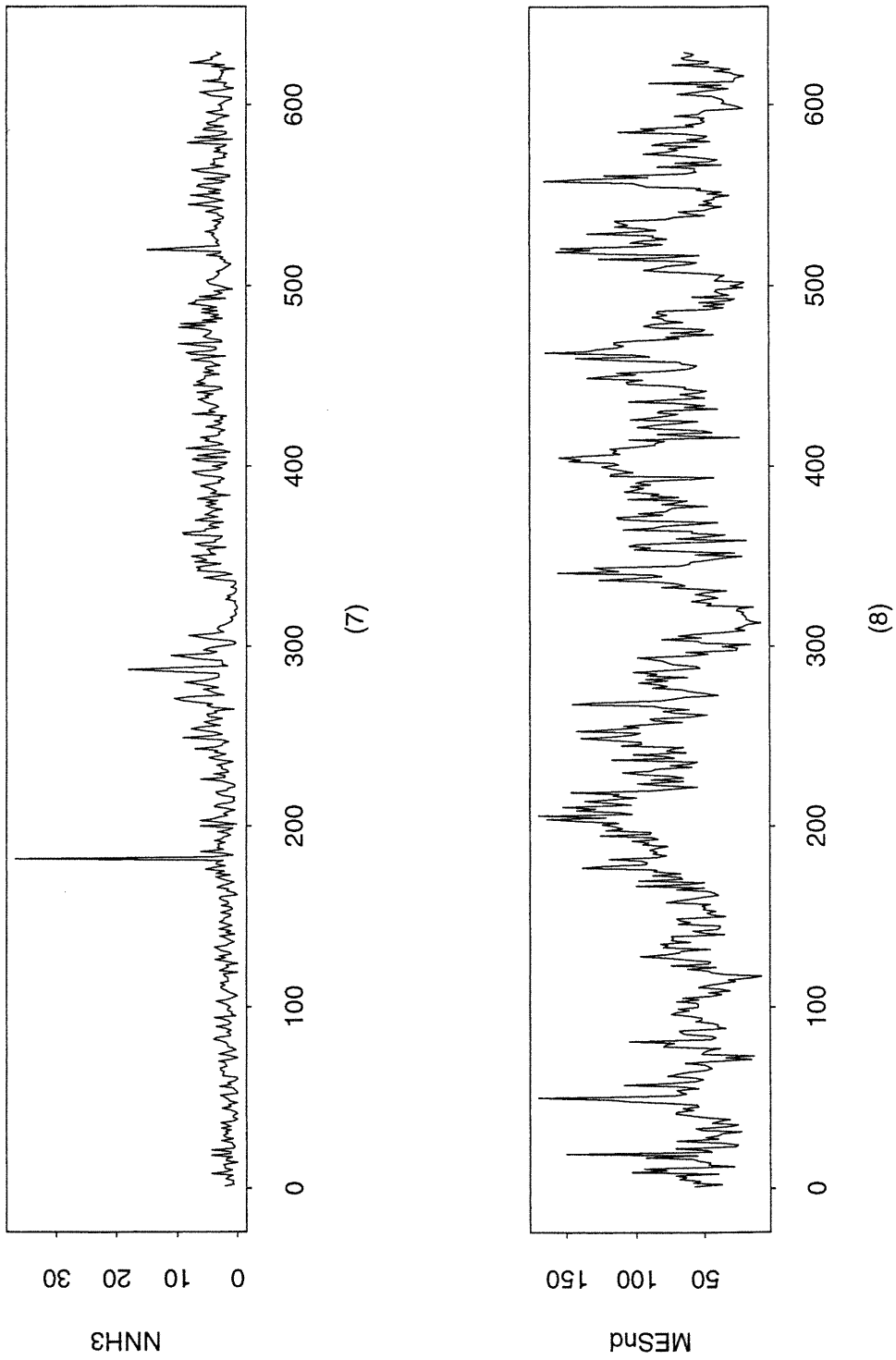


FIGURE B.0.4. *Représentation des variables N-NH_3 (7) et MESnd (8) du 01/12/97 au 26/09/99.*

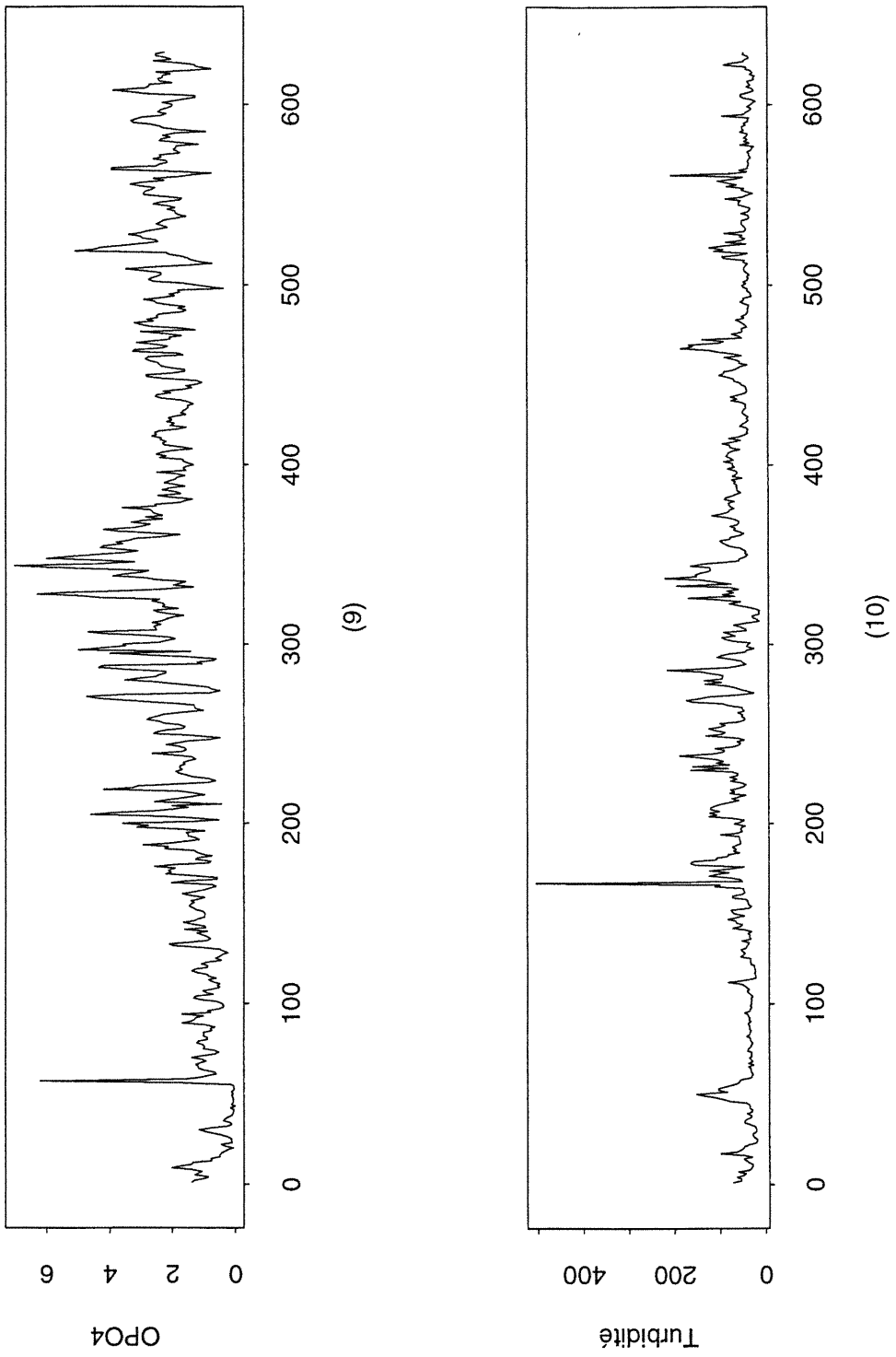


FIGURE B.0.5. *Représentation des variables O-PO_4 (9) et turbid (10) du 01/12/97 au 26/09/99.*

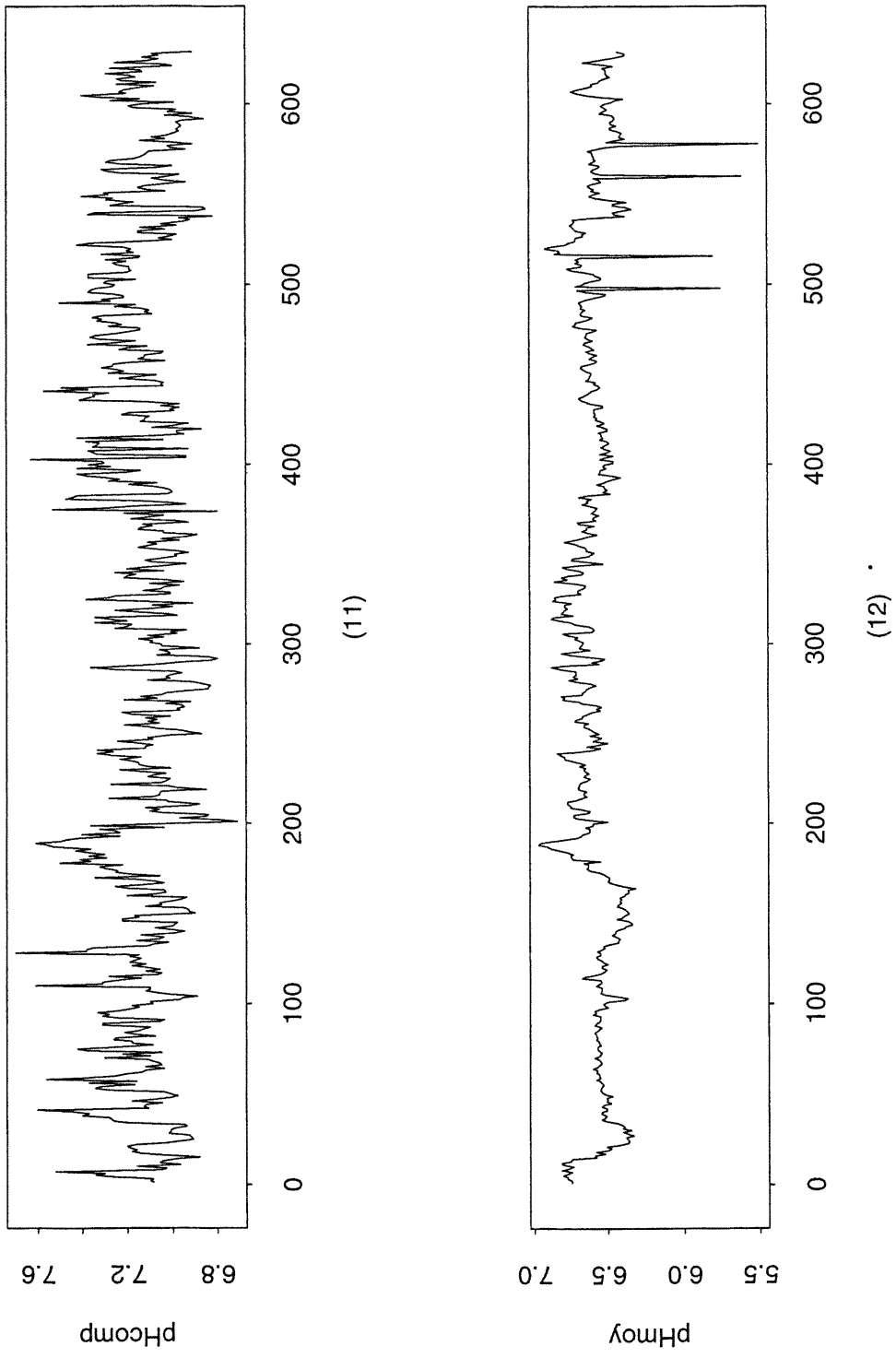


FIGURE B.0.6. *Représentation des variables pHcomp (11) et pHmoy (12) du 01/12/97 au 26/09/99.*

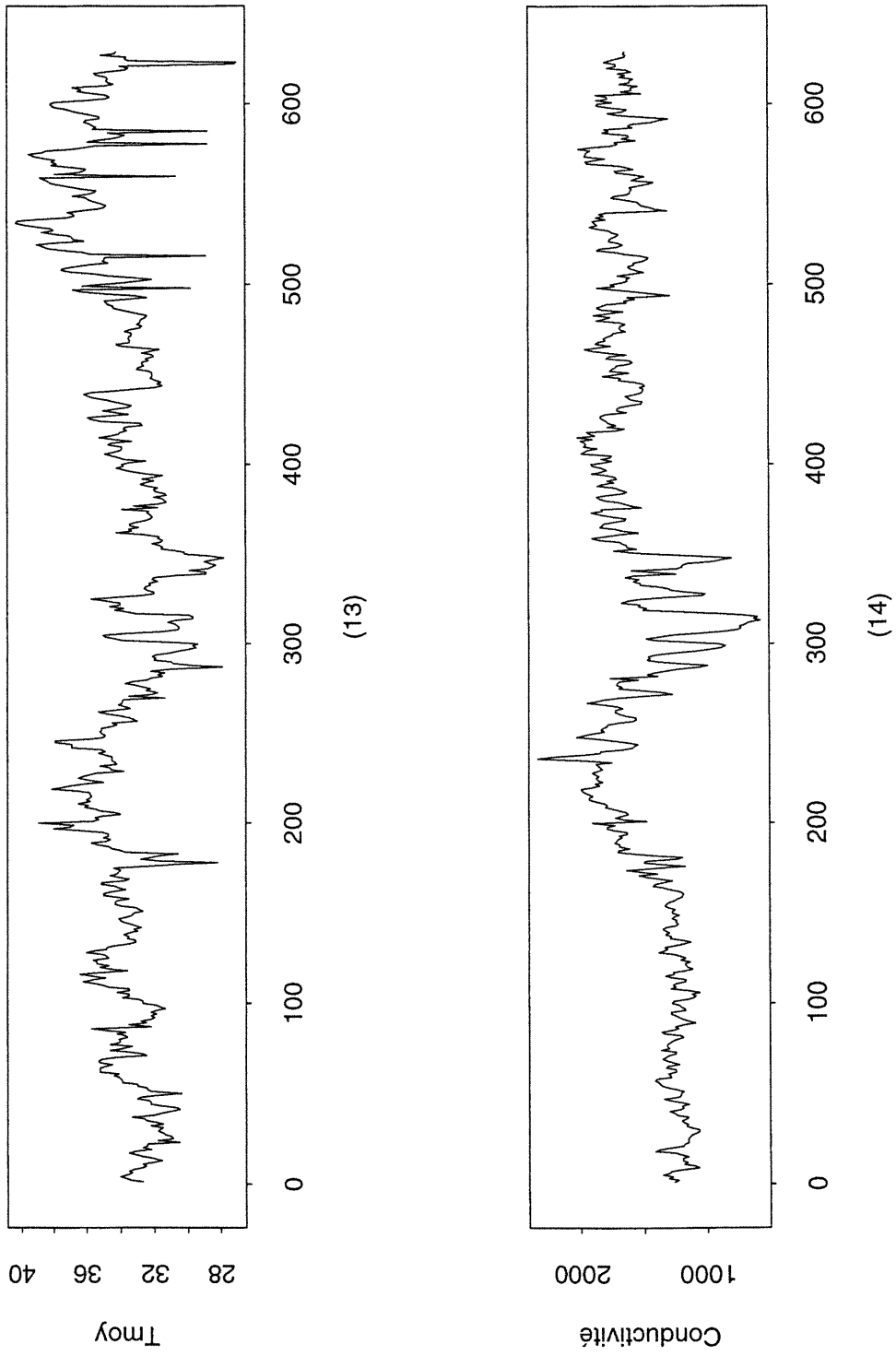


FIGURE B.0.7. *Représentation des variables Tmoy (13) et cond (14) du 01/12/97 au 26/09/99.*

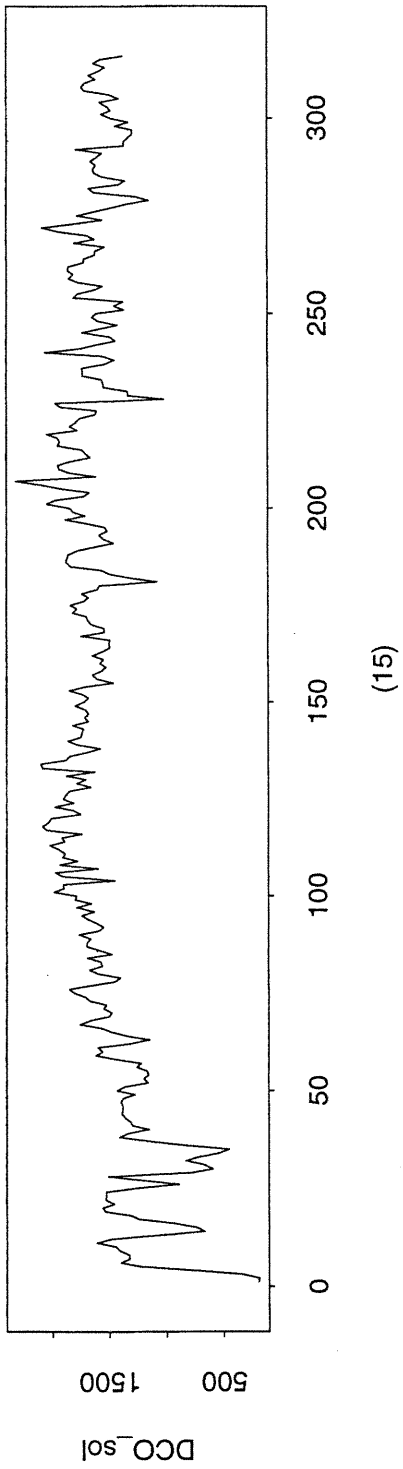


FIGURE B.0.8. *Représentation de la variable DCOsol (15) du 19/10/98 au 26/09/99.*

Annexe C

FONCTIONS SPLUS

Cette annexe contient les principales fonctions Splus (version 3,4) utilisées pour obtenir les résultats.

Liste des fonctions:

1. MatStan()
2. VecStan()
3. RLMselection()
4. RR()
5. PCR()
6. LRR()
7. PLSuni()
8. PLSmulti()
9. previsionPLSuni()
10. previsionPLSmulti()
11. RN2ccStan

```
# ====================================================================
# 1. Standardisation d'un vecteur
# ====================================================================

VectStan_function(Y)
        {
        (Y-mean(Y))/var(Y)^(1/2)
        }

# ====================================================================
# 2. Standardisation d'une matrice
# ====================================================================

MatStan_function(X)
        {
        nbcol_dim(X)[[2]]
        nbrow_dim(X)[[1]]
        Xmoy_matrix(apply(X, 2 ,mean), ncol=nbcol, nrow=nbrow, byrow=T)
        Xvar_matrix(apply(X, 2 ,var), ncol=nbcol, nrow=nbrow, byrow=T)
        (X-Xmoy)/Xvar^(1/2)
        }
```

```
# ====================================================================
# 3. Régression linéaire multiple avec sélection de variables
# ====================================================================

RLMselection_function(Y,X,methode, fcritique)
        {
# Le paramètre "methode" doit prendre une des valeurs suivantes:
# "forward", "backward", "efroymson" ou "exhaustive".

        # Sélection de variables avec la méthode choisie.
        selection_stepwise(X,Y, method=methode,plot=F, f.crit=fcritique)

        # Détermination d'une nouvelle matrice contenant seulement
        # les variables sélectionnées.
        test_selection$which[dim(selection$which)[1],]
        Xnew_matrix(NA,ncol=sum(test),nrow=dim(X)[1])
        colonne_1
        for (i in 1:dim(X)[2])
                if (test[i]==1)
                        {
                        Xnew[,colonne]_X[,i]
                        colonne_colonne+1
                        }

        # Estimation des coefficients du modéle réduit.
        Bhat_lsfit(Xnew,Y)$coef

        # Sorties de la fonction.
        list(Bhat=Bhat, VariablesChoisies=test)
        }

# ====================================================================
# 4. Régression ridge
# ====================================================================

RR_function(Y,X,k)
        {
        #Estimation du modèle
        X1_cbind(1,X)
        ID_diag(dim(X1)[[2]])
        Bhat_solve(t(X1)%*%X1+k*ID) %*% t(X1)%*%Y

        #Sorties de la fonction.
        Bhat
         }
```

```
# ======================================================================
# 5. Régression avec composante principale (PCR)
# ======================================================================

PCR_function(Y,X,P)
        {
        # Standardisation de la matrice des variables explivatives
        Xstan <- MatStan(X)
        mat <- t(Xstan) %*% Xstan
        # NOTE:  mat = (n-1)*cor(Xstan)
        # Le résultat ne change pas si on utilise mat=cor(A)

        # Valeurs et vecteurs propres
        vp <- eigen(mat)
        n <- dim(X)[1]          # nombre de données
        valeurspropres_vp$values
        vecteurspropres_vp$vectors

        # Combinaison des observations selon les nouvelles composantes
        nouvellesvariables_X %*% vecteurspropres[,1:P]

        # Ajustement du modèle
        resultat_lm(Y~nouvellesvariables)$coef

        # Sortie de la fonction
        list(Bhat=resultat,vectp=vecteurspropres[,1 :P])
        }

# ======================================================================
# 6. Régression avec racines latentes (LRR)
# ======================================================================

LRR_function(Y, X,limlambda, limdelta)
        {
        # Standardisation des variables
        Ystan_VectStan(Y)
        Xstan_MatStan(X)

        A_cbind(Ystan,Xstan)
        mat_t(A) %*% A
        # NOTE:  mat = (n-1)*cor(A)
        # Le résultat ne change pas si on utilise mat=cor(A)

        # Valeurs et vecteurs propres
        vp_eigen(mat)
```

```
        # nombre de vecteurs latents = (nb de variables X) + variable Y
        p_dim(mat)[2]

        # Détermination des facteurs importants et élimination des facteurs
        # non significatifs   *** Limites fixées ***
        comb_rbind(vp$values, vp$vectors)
        test_test1_test2_rep(NA, length(comb[1,]))
        test1_abs(comb[1,])<limlambda
        test2_abs(comb[2,])<limdelta
        test_test1*test2

        # Calcul du vecteur des coefficients
        nhu_sqrt(sum((Y-mean(Y))^2))
        num_ -nhu*comb[2,] / comb[1,]*(1-test)
        denom_sum(comb[2,]^2 / comb[1,]*(1-test))
        f_num/denom
        fmat_t(matrix(rep(f,p-1), ncol=p-1, nrow=p))
        delta_comb[-(1:2),]
        Bhat_apply(fmat * delta, 1, sum)
        B0_mean(Y)

        # Prévision de la variable Y
        yhat_ B0+ Xstan %*% (Bhat/sqrt(n-1))
        # REMARQUE:
        # Pourquoi faut-il diviser par (n-1)?  Dans Gunst et Mason,
        # ils stansdardisent avec sum(xi)=0 et sum(xi^2)=1, mais les
        # fonctions MatStan() et VectStan() standardisent avec
        # sum(xi)=0 et var(xi^2)=1.

        #Sorties de la fonction
        list(Bhat =  Bhat, B0 = B0,comb=comb[1 :2,])
        }

# ================================================================
# 7. Moindres carrés partiels (PLS) univarié
# ================================================================

PLSuni_function(Y,X,P)
        {
        # Centrer les variables
        U_Y-mean(Y)
        nbcol_dim(X)[[2]]
        nbrow_dim(X)[[1]]
        Xmoy_matrix(apply(X, 2, mean), ncol = nbcol, nrow =nbrow, byrow=T)
        V_X-Xmoy
```

```
Tcomp_matrix(NA,ncol=P,nrow=nbrow)
Pmat_matrix(NA,ncol=P, nrow=nbcol)
W_matrix(NA,ncol=P,nrow=nbcol)
b_matrix(NA, ncol= P, nrow=nbcol )

# Estimation des composantes
for (k in 1:P)
        {
        # b est une estimation du vecteur des coefficients pour chacun des
        # modèles Y=X[,i]
        ecarttype_diag(var(cbind(U,V)))^0.5
        correlation_cor(V,U)
        b[,k]_correlation*ecarttype[1]/ecarttype[-1]

        # Estimation de U pour chacune des variables
        Uhat_ V* matrix(b[,k],nrow=nbrow,ncol=nbcol,byrow=T)

        # Détermination des poids
        w_diag(t(V)%*%V)
        w_w/sum(w)
        W[,k]_w
        w_matrix(w,nrow=nbrow,ncol=nbcol,byrow=T)

        # Détermination de la kè composante
        Tcomp[,k]_(w*Uhat)%*%matrix(1,ncol=1, nrow=nbcol)

        # Résidus à utiliser pour la boucle suivante
        Tk_ matrix(Tcomp[,k],ncol=nbcol, nrow=nbrow)
        cte_matrix(t(Tcomp[,k])%*%V/sum(Tcomp[,k]^2), ncol=nbcol,
     nrow=nbrow,byrow=T)
        Pmat[,k]_cte[1,]
        V_V-cte*Tk
        U_U-((t(Tcomp[,k])%*%U)/( sum(Tcomp[,k]^2) ))*Tcomp[,k]
        }

# Ajustement du modèle
resultat_lsfit(Tcomp,Y)

# Calcul de la somme des erreurs au carré
sse_sum(resultat$residuals^2)

# Sorties de la fonction
list(sse=sse,Bhat=resultat$coef,T=Tcomp, A=W*b, Pmat=Pmat, Xmoy=Xmoy[1,])
 }
```

```
# ==========================================================================
# 8. Moindres carrés partiels (PLS) multivarié
# ==========================================================================

PLSmulti_function(Y,X,P)
        {
        nbcol_dim(X)[[2]]
        nbrow_dim(X)[[1]]

        # Centrer les variables
        Xmoy_matrix(apply(X, 2, mean), ncol = nbcol, nrow =nbrow, byrow=T)
        V_X-Xmoy
        Ymoy_matrix(apply(Y, 2, mean), ncol = dim(Y)[[2]], nrow=nbrow, byrow=T)
        R_Y-Ymoy

        Pmat_matrix(NA, ncol=P, nrow=nbcol)
        Tcomp_matrix(NA, ncol=P, nrow=nbrow)
        W_matrix(NA,ncol=P, nrow=nbcol)
        b_matrix(NA, ncol= P, nrow=nbcol)

        # Estimation des composantes
        for (k in 1:P)
                {
                c1_eigen(t(R) %*%V%*%t(V)%*%R)$vectors[,1]
                U_R%*%c1

                # On construit Tcomp comme dans le cas univarié avec U et V.
                # b est une estimation du vecteur des coefficients pour chacun des
                # modèles Y=X[,i]+erreur.
                for (i in 1 :nbcol) b[i,k]_lsfit(V[,i],U)$coef[2]

                # Estimation de U pour chacune des variables
                Uhat_ V* matrix(b[,k],nrow=nbrow,ncol=nbcol,byrow=T)

                # Détermination des poids
                w_apply(V*V,2,sum)
                w_w/sum(w)
                W[,k]_w
                w_matrix(w,nrow=nbrow,ncol=nbcol,byrow=T)

                # Détermination de la kè composante
                Tcomp[,k]_apply(w*Uhat,1,sum)

                # Résidus à utiliser pour la boucle suivante :
                Tk_matrix(Tcomp[,k],ncol=dim(Y)[[2]],nrow=nbrow)
                cte_matrix(t(Tcomp[,k])%*%R /(t(Tcomp[,k])%*%Tcomp[,k]) [1,1],
```

```
        ncol=dim(Y)[[2]], nrow=nbrow,byrow=T)
           R_R-cte*Tk

           Tk_ matrix(Tcomp[,k],ncol=nbcol, nrow=nbrow)
           cte_matrix(t(Tcomp[,k])%*%V / (t(Tcomp[,k])%*%Tcomp[,k]) [1,1],
        ncol=nbcol, nrow=nbrow,byrow=T)
           Pmat[,k]_cte[1,]
           V_V-cte*Tk
           }

     # Ajustement du modèle
     beta_lsfit(Tcomp,Y)$coef

     # Estimation des variables dépendantes
     Yhat_cbind(1,Tcomp) %*% beta

     # Calcul de la somme des erreurs au carré
     sse_sum((Y-Yhat)^2)

     # Sorties de la fonction
     list(sse=sse,Bhat=beta,T=Tcomp,A=W*b,Pmat=Pmat,Xmoy=Xmoy[1,]
     }

# ================================================================
# 9. Prévision univariée pour PLS
# ================================================================

previsionPLSuni_function(Xnouveau, Bhat,A, Pmat, Xmoy, P)
     {
     Vetoile_matrix(NA,ncol=length(Xnouveau),nrow=1)
     Tetoile_matrix(NA,ncol=P,nrow=1 )
     Vetoile2_Vetoile

     # Étape 1
     # Calcul des coordonnées de la première composante
     Vetoile_Xnouveau-Xmoy
     Tetoile[,1]_sum(A[,1]*Vetoile)

     # Étapes suivantes
     # Calcul des coordonnées des autres composantes
     if (P>1) for (i in 2:P)
           {
           for (j in 1:length(Vetoile)) Vetoile2[j]_Vetoile[j]-Pmat[j,i-1]*
        Tetoile[i-1]
           Tetoile[i]_sum(A[,i]*Vetoile2)
           Vetoile_Vetoile2
```

```
        }

    # Prévision et sortie de la fonction
    c(1,Tetoile)%*%Bhat
    }


# ======================================================================
# 10. Prévision multivariée pour PLS
# ======================================================================

previsionPLSmulti_function(Xnouveau,Bhat, A, Pmat, Xmoy, P)
    {
    Vetoile_matrix(NA,ncol=dim(Xnouveau)[2],nrow=dim(Xnouveau)[1])
    Tetoile_matrix(NA,ncol=P,nrow=dim(Xnouveau)[1] )
    Vetoile2_Vetoile

    # Étape 1
    # Calcul des coordonnées de la première composante
    Vetoile_Xnouveau-matrix(Xmoy, ncol =dim(Xnouveau)[2] , nrow=dim(Xnouveau)[1],
  byrow=T)
    Tetoile[,1]_apply(matrix(A[,1], ncol=dim(Xnouveau)[2], nrow=dim(Xnouveau)[1],
  byrow=T)*Vetoile,1,sum)

    # Étapes suivantes
    # Calcul des coordonnées des autres composantes
    if (P>1) for (i in 2:P)
            {
            for (j in 1:(dim(Vetoile)[2])) Vetoile2[,j]_Vetoile[,j]-
        Pmat[j,i-1]* Tetoile[,i-1]
            Tetoile[,i]_apply(matrix(A[,i], ncol=dim(Xnouveau)[2],
        nrow=dim(Xnouveau)[1], byrow=T)*Vetoile2,1,sum)
            Vetoile_Vetoile2
            }

    # Prévision et sortie de la fonction.
    cbind(1,Tetoile)%*%Bhat
    }
```

```
# ------------------------------------------------------------------
# NOTE:  La fonction suivante est disponible avec la version 5 de
# Splus.  Il faut entrer les commandes suivantes à chaque fois que
# Splus5 est ouvert et que la fonction doit être utilisée.
        attach("/home/bengioy/ad/adhome5/.Data/")
        dyn.open("/home/bengioy/ad/adhome5/SOLARIS/ad_splus.so")
# ------------------------------------------------------------------

# ==================================================================
# 11. Réseau de neurones avec deux couches cachées
# ==================================================================

RN2ccStan_function(Y,X,nb.hidden1, nb.hidden2, type1, type2)
        {
        # nb.hidden1: nombre de neurones de la 1ère couche cachée
        # nb.hidden2: nombre de neurones de la 2è couche cachée
        # type1: fonction d'activation de la 1ère couche cachée
        # type2: fonction d'activation de la 2è couche cachée

        Xmoy_apply(X,2,mean)
        Xvar_apply(X,2,var)
        X_MatStan(X)
        Ymoy_mean(Y)
        Yvar_var(Y)
        Y_cbind(VectStan(Y))/3

        # Initialisation du générateur aléatoire
        seed()

        # Définition du réseau de neurones.
        nb.variables_dim(X)[2]
        nb.outputs_dim(Y)[2]
        nb.par.couche_c(nb.variables,nb.hidden1, nb.hidden2, nb.outputs)
        type.de.couche_c(type1, type2, layer.linear)
        machine_new.Mlp(nb.par.couche,type.de.couche)

        # Type d'optimisation (Descente de gradient conjugué)
        opt_setup.optimizer(inner.opt.type = CG)

        # Détermination des paramètres du modèle en utilisation le type
        # d'optimisation défini précédemment.
        params_mse.train(X,Y,opt,machine)

        # Prévisions obtenues à partir du modèle
        YhatStan_mse.test(X,nb.outputs,params,machine)
        Yhat_YhatStan*3*Yvar^0.5+Ymoy
```

```
   # Calcul de la somme des erreurs au carré
   SSE_sum((Y-Yhat)^2)

   # Sorties de la fonction
   list(Ymoy=Ymoy, Yvar=Yvar, Xmoy=Xmoy, Xvar=Xvar, SSE=SSE, parametres=params,
machine=machine, nb.outputs=nb.outputs)
   }
```

BIBLIOGRAPHIE

[1] Abraham, Bovas et Ledolter, Johannes (1983), *Statistical Methods for Forecasting*, Wiley, New-York, 445 p.

[2] Afifi, A. A. et Elashoff, R. M. (1966), *Missing Observations in Multivariate Statistics I, Review of the litterature*, Journal of the American Statistical Association, **61**, 595-604.

[3] Afifi, A. A. et Elashoff, R. M. (1967), *Missing Observations in Multivariate Statistics II, Point Estimation in Simple Linear Regression*, Journal of the American Statistical Association, **62**, 10-29.

[4] Afifi, A. A. et Elashoff, R. M. (1967), *Missing Observations in Multivariate Statistics III, Large Sample Analysis of Simple Linear Regression*, Journal of the American Statistical Association, **64**, 337-358.

[5] Afifi, A. A. et Elashoff, R. M. (1967), *Missing Observations in Multivariate Statistics IV, A Note on Simple Linear Regression*, Journal of the American Statistical Association, **64**, 359-365.

[6] Bello, A. L. (1995), *Imputation Techniques in Regression Analysis: Looking Closely at Their Implementation*, Computational Statistics and Data Analysis, **20**, 45-57.

[7] Belsley, David A., Kuh, Edwin et Welsch, Roy E. (1980), *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*, Wiley, New-York, 292 p.

[8] Bioch, Jan C., van der Meer, Onno et Potharst, Rob (1997), *Bivariate Decision Tree*, Principles of Data Mining and Knowledge Discovery, First European Symposium, PKDD'97 Trondheim, Norway, June 24-27 1997, Proceedings, Springer, Berlin, 396 p.

[9] Bishop, Christopher M. (1995), *Neural Networks for Pattern Recognition*, Oxford University Press, Oxford, 482 p.

[10] Breiman, Friedman, Olshen et Stone (1984), *Classification and Regression Trees*, Wadsworth, Pacific Grove, 358 p.

[11] Brockwell, Peter J. et Davis, Richard A.(1996), *Introduction to Time Series and Forecasting*, Springer, New-York, 420 p.

[12] Buck, S. F. (1960), *A Method of Estimation of Missing Values in Multivariate Data Suitable for Use With an Electronic Computer*, JRSS(B), **22**, 302-306.

[13] Carter, Colin L., Hamilton, Howard J. et Cercone, Nick (1997), *Share Based Measures for Itemsets*, Principles of Data Mining and Knowledge Discovery, First European Symposium, PKDD'97 Trondheim, Norway, June 24-27 1997, Proceedings, Springer, Berlin, 396 p.

[14] Donner, Allan (1982), *The Relative Effectiveness of Procedures Commonly Used in Multiple Regression Analysis for Dealing With Missing Values*, The American Statistician, **36**(4), 378-381.

[15] Draper, N. R. et Smith, H. (1981), *Applied Regression Analysis*, Wiley, New-York, 709 p.

[16] Efron, Bradley (1996), *Missing Data, Imputation, and the Bootstrap*, Journal of the American Statistical Association, 89, 463-479.

[17] Everitt, Brian S. (1993), *Cluster Analysis*, 3^{e} édition, Halsted Press, London, 170 p.

[18] Fletcher, Roger (1987), *Practical Methods of Optimization*, 2^{e} édition, Wiley, New-York, 436 p.

[19] Frank, Ildiko E., Friedman, Jerome H. (1993), *A Statistical View of Some Chemometrics Regression Tools*, Technometrics, **35**(2), 109-135.

[20] Funahashi, Ken-Ichi (1989), *On the Approximate Realization of Continuous Mappings by Neural Networks*, Neural Networks, **2**, 183-192.

[21] Garthwaite, Paul H. (1994), *An Interpretation of Partial Least Squares*, Journal of the American Statistical Association, **89**, 122-127.

[22] Geladi, Paul et Kowalski, Bruce R. (1986), *Partial Least Squares Regression: a Tutorial*, Analytica Chimica Acta, **185**, 1-17.

[23] Geladi, Paul et Kowalski, Bruce R. (1986), *An Example of 2-Block Predictive Partial Least-Squares Regression With Simulated Data*, Analytica Chimica Acta, **185**, 19-32.

[24] Gunst, Richard F. et Mason, Robert L. (1977a), *Biased Estimation if Regression: An Evaluation Using Mean Square Error*, Journal of the American Statistical Association, **72**, 616-628.

[25] Gunst, Richard F. et Mason, Robert L. (1977b), *Advantages of Examining Multicollinearities in Regression Analysis*, Biometrics, **33**, 249-260.

[26] Gunst, Richard F. et Mason, Robert L. (1980), *Regression Analysis and its Application, a Data-Oriented Approach*, Marcel-Dekker, New-York, 402 p.

[27] Gunst, Richard F., Webster, J. T. et Mason, R. L. (1976), *A Comparison of Least Squares and Latent Root Regression Estimators*, Technometrics, **18**, 75-83.

[28] Hawkins, Douglas M. (1973), *On the Investigation of Alternative Regressions by Principal Component Analysis*, Applied Statistics, **22**, 275-286.

[29] Hawkin, Simon (1999), *Neural Networks, a Comprehensive Foundation*, 2^{e} édition, Prentice-Hall, New-Jersey, 842 p.

[30] Helland, I. S. (1990), *Partial Least Squares Regression and Statistical Methods*, Scandinavian Journal of Statistics, **17**, 97-114.

[31] Hill, Carter R., Fomby, Thomas B. et Johnson, S. R. (1977), *Component Selection Norms for Principal Components Regression*, Communications in Statistics, Theory and Methods, **A6**, 309-334.

[32] Hoerl, Arthur E. et Kennard, Robert W. (1970a), *Ridge Regression: Biased Estimation for Nonorthogonal Problems*, Technometrics, **12**, 55-68.

[33] Hoerl, Arthur E. et Kennard, Robert W. (1970b), *Ridge Regression: Ridge Regression: Application to Nonorthogonal Problems*, Technometrics, **12**, 69-82.

[34] Hoerl, Arthur E. et Kennard, Robert W. (1976), *Ridge Regression, Iterative Estimation of the Biasing Parameter*, Communications in statistics, theory and methods, **A5**(1), 77-88.

[35] Hoerl, Arthur E. et Kennard, Robert W., Baldwin, Kent F. (1975), *Ridge Regression: Some Simulations*, Communications in statistics, **4**(2), 105-123.

[36] Höskuldsson, Agnar (1988), *PLS Regression Methods*, Journal of Chemometrics, **2**, 211-228.

[37] Jolliffe, I. T. (1986), *Principal Component Analysis*, Springer-Verlag, New-York, 271 p.

[38] Johnson, Richard A. et Wichern, Dean W. (1992), *Applied Multivariate Statistical Analysis*, 3^{e} édition, Prentice-Hall, New-Jersey, 642 p.

[39] Kresta, James V., MacGregor, John F. et Marlin, Thomas E. (1991), *Multivariate Statistical Monitoring of Process Operating Performance*, The Canadian Journal of Chemical Engineering, **69**, 35-47.

[40] Lazraq, Aziz et Cléroux, Robert (1988), *Un Algorithme pas à pas de Sélection de Variables en Régression Linéraire Multivariée*, Statistique et Analyse des Données, **13**, 39-58.

[41] Lazraq, Aziz et Cléroux, Robert (2000), *The PLS Multivariate Regression Model: Testing the Significance of Successive PLS Components*, À paraître dans Journal of Chemometrics.

[42] Lefébure, René et Venturi, Gilles (1998), *Le Data Mining*, Eyrolles, Paris, 330 p.

[43] Little, Roderick J. A (1992), *Regression with Missing X's: A Review*, Journal of the American Statistical Association, **87**, 1227-1237.

[44] Little, Roderick J. A. et Rubin, Donald B. (1987), *Statistical Analysis with Missing Data*, Wiley, New-York, 278 p.

[45] Mansfield, E. R., Webster, J. T. et Gunst, R. F., (1977), *An Analytic Variable Selection Technique for Principal Component Regression*, Applied Statistics, **26**, 34-40.

[46] Masmoudi, Radwan A. (1999), *Rapid Prediction of Effluent Biochemical Oxygen Demand for Improved Environmental Control*, Tappi Journal, **82**(10), 111-119.

[47] Mason, Robert L., Gunst, R. F. et Webster, J. T. (1975), *Regression Analysis and Problems of Multicollinearity*, Communications in Statistics, **4**, 277-292

[48] Massy, William F. (1965), *Principal Components Regression in Exploratory Statistical Research*, Journal of the American Statistical Association, **60**, 234-256.

[49] Montgomery, Douglas C. et Peck, Elisabeth A. (1992), *Introduction to Linear Regression Analysis*, 2^{e} édition, Wiley, New-York, 527 p.

[50] Maüller, B. et Reinhardt, J. (1990), *Neural Networks. An Introduction*, Springer-Verlag, Berlin, 266 p.

[51] Neter, John, Kutner, Michael H., Nachtsheim, Christopher J. et Wasserman, William (1996), *Applied Linear Statistical Models*, 4^{e} édition, Irwin,Chicago, 1408 p.

[52] Randall D. Tobias, *An Introduction to Partial Least Squares Regression*, SAS Institute, N.C. , 8 p.

[53] Rawlings, John O. (1988), *Applied Regression Analysis: a Research Tool*, Wadsworth, Belmont, 553 p.

[54] Sampson, Allan R., (1974), *A Tale of two Regressions*, JASA, **69**, 682-689.

[55] Sjöström, Michael, Wold, Svante, Lindberg, Walter, Persson, Jan-Åke et Martens, Harald (1983), *A Multivariate Calibration Problem in Analytical Chemistry Solved by Partial Least-Squares Models in Latent Variables*, Analytica Chimica Acta, **150**, 61-70.

[56] Webster, J. T., Gunst, R. F. et Mason, R. L. (1974), *Latent Root Regression*, Technometrics, **16**, 513-522.

[57] Weisberg, Stanford (1985), *Applied Linear Regression*, Wiley, New-York, 324 p.

[58] Wold, H. (1985), *PLS regression*, Encyclopedia of Statistical Sciences, **6**, 581-591.