Record Number: 21470
Author, Monographic: St-Hilaire, A.//Lachance, M.//Bobée, B.
Author Role:
Title, Monographic: Avis sur l'application de différentes méthodes statistiques pour la prédiction de bris de conduites d'eau de la ville de Montréal
Translated Title:
Reprint Status:
Edition:
Author, Subsidiary:
Author Role:
Place of Publication: Québec
Publisher Name: INRS-Eau
Date of Publication: 2001
Original Publication Date: Mars 2001
Volume Identification:
Extent of Work: v, 50
Packaging Method: pages incluant un annexe
Series Editor:
Series Editor Role:
Series Title: INRS-Eau, rapport de recherche
Series Volume ID: 585
Location/URL:
ISBN: 2-89146-452-4
Notes: Rapport annuel 2000-2001
Abstract:
Call Number: R000585
Keywords: rapport/ ok/ dl

Avis sur l'application de différentes méthodes statistiques pour la prédiction de bris de conduites d'eau de la ville de Montréal

Rapport de recherche R-585

Avis sur l'application de différentes méthodes statistiques pour la prédiction de bris de conduites d'eau de la ville de Montréal.

Par :

André St-Hilaire

Marius Lachance

Bernard Bobée

Chaire en hydrologie statistique

INRS-Eau

2800 rue Einstein, C.P. 7500, Sainte-Foy (Québec) G1V 4C7

Rapport présenté à :

BPR Groupe-conseil

Rapport de recherche R-585

Mars 2001

TABLE DES MATIÈRES

Liste des tableaux iii

Liste des figures iv

Équipe de Projet v

1.0 Introduction 1

2.0 Analyse critique des données 2

3.0 Régression multiple 3

3.1 Sélection d'un échantillon aléatoire 3

3.2 Méthode de la régression multiple 4

3.3 Résultats 7

3.3.1 Régression dans le cas où le type de sol est l'argile 7

3.3.2 Régression pour les autres types de sol 9

4.0 Analyse multivariée 17

4.1 Méthodologie 17

4.2 Résultats de l'analyse des correspondances 19

4.2.1 Tableau analysé 19

4.2.2 Résultats de l'AFC 20

4.3 Résultats de la classification ascendante hérarchique 24

4.4 Résultats de la classification par les Nuées dynamiques 26

4.4.1 Comparaison des classifications obtenues par analyse multivariée avec celles obtenues par BPR 26

4.4.2 Caractérisation des classes de tronçons obtenues par analyse multivariée 29

5.0 Analyse de variance 31

5.1 Introduction 31

5.2 Méthodologie 32

5.2.1 Approche générale 32

5.2.2 Test *a posteriori* 33

5.3 Exemple d'application de l'analyse de variance : critère des diamètres de conduites 35

TABLE DES MATIÈRES

6.0 Utilisation de un ou deux paramètres de calage pour les courbes exponentielles...38

7.0 Calcul des intervalles de confiance 40

8.0 Conclusions et recommandations 42

9.0 Références 45

ANNEXE A 46

LISTE DES TABLEAUX

Tableau 1 Coefficients de corrélation (r) entre les variables de la régression pour les tronçons ayant l'argile comme type de sol. La valeur critique du coefficient de corrélation (r_c) est au bas du tableau. 7

Tableau 2 Analyse des résultats de la régression multiple dans le cas où le type de sol est l'argile. Les coefficients, de même que leur écart-type (s) sont donnés avec les résultats de l'analyse de variance. 8

Tableau 3 Coefficients de corrélation (r) entre les variables de la régression. La valeur critique du coefficient de corrélation (r_c) est au bas du tableau 10

Tableau 4 Analyse des résultats des régressions multiples par type de sol (sable, till, tourbe et roche). 11

Tableau 5 Comparaison des regroupements par classification avec ceux obtenus par BPR 28

Tableau 6. Caractérisation des huit classes de tronçons obtenues par analyse multivariée 29

Tableau 7 Résumé des moyennes, écart-types et variances par classe de diamètre 35

Tableau 8 Analyse de variance des taux de bris moyens par classe de diamètre 35

Tableau 9 Valeurs de p pour les tests de Scheffé entre chaque classe de diamètre 36

Tableau 10 Résumé des moyennes, écart-types et variances par classe de diamètre avec regroupement. 37

Tableau 11 Analyse de variance des taux de bris moyens par classe de diamètre avec regroupement 37

LISTE DES FIGURES

Figure 1 Exemple de la distribution de Fisher avec zones de rejet de l'hypothèse nulle. 5

Figure 2 Relation entre les valeurs observées et les valeurs prédites, selon la régression multiple présentée au tableau 2. 9

Figure 3 Relation entre les valeurs observées et les valeurs prédites, selon la régression multiple présentée au tableau 4 pour le sable. 13

Figure 4 Relation entre les valeurs observées et les valeurs prédites, selon la régression multiple présentée au tableau 4 pour le till. 14

Figure 5 Relation entre les valeurs observées et les valeurs prédites, selon la régression multiple présentée au tableau 4 pour le roc. 15

Figure 6 Relation entre les valeurs observées et les valeurs prédites, selon la régression multiple présentée au tableau 4 pour la tourbe. 16

Figure 7 Tracé de la variance expliquée par les axes factoriels 20

Figure 8 Représentation des points colonnes dans le plan des axes 1 et 2 22

Figure 9 Classification ascendante hiérarchique des tronçons 25

Figure 10 Comparaison des histogrammes des deux classifications 27

Figure 11 Représentation des centres de gravité des classes dans le plan des axes 1 et 2 31

ÉQUIPE DE PROJET

BPR:

Luc Aubertin, ing	Chargé de projet
Denis Curodeau, M.B.A., ing.	Ingénieur municipal

INRS-Eau, Chaire en hydrologie statistique :

Bernard Bobée, Ph.D.	Professeur et titulaire de la Chaire
Marius Lachance, Ph.D.	Professeur
André St-Hilaire, Ph.D.	Associé de recherche

Référence :

St-Hilaire, A., M. Lachance, et B. Bobée (2001). Avis sur l'application de différentes méthodes statistiques pour la prédiction de bris des conduites d'eau de la ville de Montréal. Rapport de recherche R-585, préparé par la Chaire en hydrologie statistique à l'INRS-Eau. 48 p.

1.0 INTRODUCTION

BPR Groupe-conseil a été mandaté par la ville de Montréal pour mettre en place un système permettant de sélectionner les conduites qui devront être remplacées dans le réseau d'aqueduc de la ville.

Dans le cadre de ce mandat, BPR a demandé à la Chaire en hydrologie statistique (INRS-EAU) de proposer les approches les plus appropriées pour prédire les réparations des tronçons du réseau. Les étapes du mandat sont divisées de telle sorte qu'un avis a été donné sur plusieurs questions ayant trait à l'approche statistique proposée pour l'analyse des données de la ville de Montréal.

Les questions portaient sur les thèmes suivants :

1. La pertinence de l'utilisation de l'approche régressive multiple afin de prédire le taux de bris à partir de variables indépendantes.
2. La possibilité d'utiliser l'approche multivariée pour déterminer des sous-groupes de tronçons ayant le même comportement pour la prévision des taux de bris.
3. L'élaboration d'une méthodologie de regroupement basée sur l'analyse de variance.
4. La pertinence de caler des courbes exponentielles de prédiction de taux de bris à l'aide d'un ou deux paramètres.
5. La détermination d'intervalles de confiance sur les courbes exponentielles de prédiction.
6. La possibilité de transférer l'information contenue dans les courbes de prédiction pour un groupe à des tronçons jugés problématiques

Ce document collige les avis émis sur chacune des questions soulevées. En premier lieu, une brève analyse critique des données est présentée, mettant ainsi en lumière certains problèmes associés à l'information fournie. Par la suite, chaque grand thème est traité à l'intérieur d'une section. Les conclusions et recommandations qui se dégagent de cette information sont aussi produites dans la dernière section du rapport.

2.0 ANALYSE CRITIQUE DES DONNÉES

Les données de bris de conduites de différentes régions de la ville de Montréal ont été colligées par tronçon, pour une population totale de 18174 tronçons. Pour chaque tronçon, on a défini la longueur, le diamètre, la région d'appartenance, le type de sol, la pression, l'année d'installation et le nombre de bris pour chaque année.

Les bris sur les tronçons de l'ensemble des régions de la ville de Montréal sont connus seulement pour quatre années (1993-1996), sauf pour la région d'Ahuntsic où les bris sont enregistrés depuis 1973. On dispose donc de plus de vingt ans de données pour cette région. Étant donné qu'un des objectifs de ce projet consiste à établir des courbes permettant de prédire le taux de bris en fonction du temps, le fait que les 20 premières années de données soient limitées à la seule région d'Ahuntsic cause une difficulté. Il faut s'assurer que les regroupements qui seront faits pour l'ensemble du territoire soient aussi adéquats pour la région d'Ahuntsic. Cette vérification devra en outre être faite lorsque les groupes seront déterminés par analyse de variance (voir section 3).

La longueur des tronçons varie entre 4 m et 2,7 km, mais ces extrêmes représentent un très faible pourcentage (<1%) de la population. Treize pourcent de la population consistent en des tronçons dont la longueur est inférieure à 50 m, 42% des tronçons ont des longueurs comprises entre 50 m et 100 m, 24% ont une longueur entre 100 m et 150 m et 21% ont des longueurs supérieures à 150 m. Pour des fins de comparaison, les taux de bris ont été normalisés en divisant par la longueur du tronçon (bris/km). Cette normalisation sur des longueurs non uniformes peut introduire un biais dans la base de données. Par exemple, un bris sur un tronçon de 10 m correspond au même taux de bris que 10 bris sur une conduite de 100 m. Pourtant, ce même taux décrit deux réalités physiques différentes. Idéalement, l'information provenant des plus petits tronçons aurait dû, si possible, être agrégée avec celle des tronçons voisins pour produire des taux de bris sur des longueurs de conduites similaires.

3.0 RÉGRESSION MULTIPLE

Il a fallu procéder à des essais préliminaires afin de juger de l'opportunité d'utiliser la régression multiple comme outil de prédiction du taux de bris par unité de longueur.

Les essais effectués sont résumés dans le texte qui suit. L'approche utilisée y est décrite, et une interprétation des résultats est donnée, avec les conclusions que l'on peut en tirer.

Comme il était prévu que ces calculs préliminaires ne seraient pas faits sur l'ensemble des données, un échantillon représentatif de la base de données a été extrait. Il a fallu ensuite procéder à différents tests sur l'adéquation de l'approche de régression.

3.1 SÉLECTION D'UN ÉCHANTILLON ALÉATOIRE

Les données qui composent l'échantillon ayant servi aux tests ont été sélectionnées aléatoirement parmi la population de 18 174 tronçons. Au départ, 400 tronçons ont ainsi été choisis, avec représentation de toutes les régions, et toutes les catégories de variables explicatives (année d'installation, pression, diamètre, type de sol). Le critère utilisé pour vérifier la représentativité de ce sous-échantillon de tronçons consiste à vérifier si la proportion de bris pour chacune des quatre années (1993-1996) est représentative. De façon générale, nous cherchons à tester l'hypothèse nulle (H_0) suivante :

$$H_0 : p_i = \pi_i$$

où : $p_i = \frac{x_i}{n_i}$ = proportion de l'échantillon i ;

x_i = nombre de bris total dans l'échantillon i ;
n_i = taille de l'échantillon;

$\pi_i = \frac{X_i}{N_i}$ = proportion de bris dans la population pour l'année i;

X_i= nombre de bris dans la population pour l'année i;
N_i= taille de la population.

Afin de tester cette hypothèse, il faut calculer la statistique du test z' telle que :

$$Z' = \frac{(p_i - \pi_i)}{\sqrt{\left[\frac{N_i + N_i\pi_i}{n_i + N_i}\right]\left[1 - \frac{N_i + N_i\pi_i}{n_i + N_i}\right]\left[\frac{1}{n_i} + \frac{1}{N_i}\right]}}$$

L'hypothèse nulle est rejetée si $|z'| > z_{0.025}$.

Où : $z_{0.025}$ est la valeur de z tirée des tables de la loi normale centrée réduite pour une probabilité au dépassement de 0,025. Le niveau de signification α est fixé à 5%.

Il a fallu ensuite ajouter 301 tronçons supplémentaires aux 400 tronçons initialement sélectionnés pour arriver à un échantillon représentatif pour chacune des quatre années selon ce critère. L'échantillon représente 4% de la population.

3.2 MÉTHODE DE LA RÉGRESSION MULTIPLE

La régression linéaire multiple cherche à établir une relation linéaire entre une variable que l'on veut prédire (dans notre cas, le nombre de bris total sur un tronçon donné) et les variables indépendantes explicatives. Afin de trouver la fonction qui représente le mieux cette relation, on utilise la méthode des moindres carrés, qui consiste à minimiser les carrés des déviations entre les valeurs observées et les valeurs prédites.

La régression testée dans chaque cas prenait la forme suivante :

$$E(Y) = \beta_0 + \sum_{i=1}^{4} \beta_i X_i + \varepsilon$$

où : β_0= constante;
β_i= coefficients de régression ;
Y= variable dépendant (nombre de bris);
E(Y) = Estimation de la variable dépendante;
X_i= variables prédictives;
ε = résidus.

Les variables prédictives sélectionnées sont les variables quantitatives fournies par BPR. Elles incluent la longueur du tronçon, l'année de construction, le diamètre et la pression. Le type de sol est une variable non numérique et ne peut être utilisée comme variable indépendante prédictive pour la régression. Comme cette variable est jugée importante dans l'analyse des bris de conduite, les régressions ont été testées pour chaque type de sol.

L'adéquation de la régression a été vérifiée par une analyse de variance dans laquelle on vérifie si la valeur des coefficients de régression calculés β_i est significativement différente de zéro. On teste l'hypothèse H_0 :

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$$

l'hypothèse alternative est :

H_1 : Un ou plusieurs des β_k (k=1,...4) non nul(s).

Cette hypothèse est vérifiée pour l'ensemble des paramètres par le calcul de F, qui est le rapport entre les écarts moyens expliqués par la régression et les écarts moyens résiduels.

$$F = \frac{\left(\sum (y-\bar{y})^2 - \sum (y-\hat{y})^2\right)/k}{\sum (y-\hat{y})^2/(n-k-1)}$$

où : $\bar{y}$ = moyenne de la variable dépendante ;
n = taille de l'échantillon ;
k = nombre de variables indépendantes.

Cette valeur de F suit une distribution de Fisher avec k et (n-k-1) degrés de libertés. Pour un niveau de signification donné (α) et des degrés de liberté connu, si la valeur de F calculée dépasse la valeur critique établie par les tables de la loi de Fisher (α = 0.05, par exemple), il faut rejeter l'hypothèse nulle (figure 1). On rejette donc H_0 si :

$$F > F(1-\alpha, \nu_1, \nu_2)$$

où :
ν1 = k degrés de liberté
ν2 = n-k-1 degrés de liberté

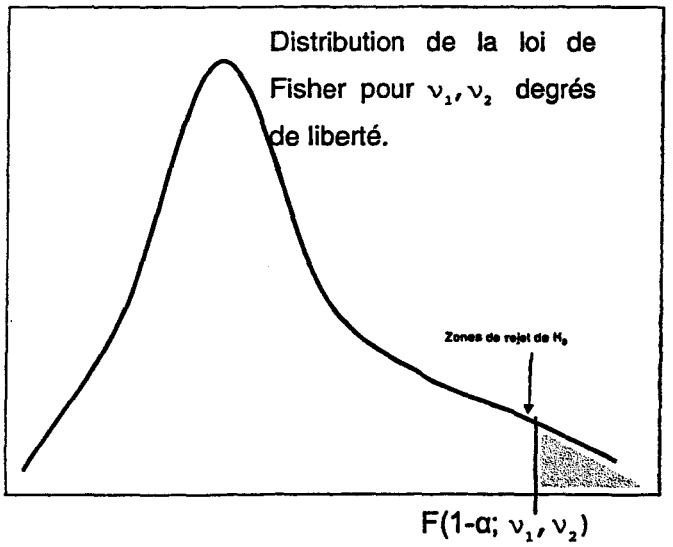


Figure 1 Exemple de la distribution de Fisher avec zones de rejet de l'hypothèse nulle.

On associe à cette valeur de F une valeur de p (« p value ») qui représente la probabilité au dépassement de la valeur calculée F selon la distribution de Fisher.

Lorsque l'analyse de variance permet de conclure à une régression dont les coefficients sont significativement différents de zéro, on peut alors comparer le poids relatif de chaque variable dépendante à l'aide des coefficients B'_i, qui représentent la valeur des coefficients lorsque les variables sont centrées-normalisées (moyenne nulle et écart-type de 1). Si le B'_i associé à la variable X_1 est le double de celui associé à la variable X_2, alors X_1 a un poids double pour expliquer la variance de Y.

De plus, chaque coefficient B_i est testé indépendamment à l'aide de l'hypothèse nulle suivante :

$$H_0 : \beta_i = 0$$

Cette hypothèse est testée par le calcul de la valeur de t :

$$t = \frac{b_i}{S_{b_i}}$$

où : b_i= estimation du paramètre β_i;
S_{bi}= écart type de l'estimé du paramètre B_i de l'équation de régression.

L'hypothèse nulle est acceptée si :

$$-t_{\alpha/2;n-k-1} < t < t_{\alpha/2;n-k-1}$$

Encore une fois, on associe à la valeur de t calculée une valeur de p correspondant à la probabilité au non dépassement de la valeur calculée selon la loi de Student.

La régression multiple présuppose que les variables indépendantes ne sont pas corrélées entre elles. Les coefficients de corrélation (r) ont donc été calculés afin de vérifier cette hypothèse. Le coefficient r calculé pour un échantillon de points est une estimation de la véritable valeur du coefficient de corrélation (ρ) associé à la population. On peut tester l'hypothèse suivante :

$$H_0 : \rho = 0$$

en calculant la valeur de t : $t = \frac{r}{\sqrt{\frac{(1-r^2)}{N-k-1}}}$

l'hypothèse nulle est acceptée si :

$$-t_{\alpha/2,n-k-1} < t < t_{\alpha/2,n-k-1}$$

pour des valeurs de t qui suivent une distribution de Student avec n-k-1 degrés de liberté.

On peut calculer la valeur de r correspondant au seuil de rejet et d'acceptation de H_0 à partir des valeurs de t compilées dans les tables statistiques pour $\alpha = 0.05$ et n-k-1 degrés de liberté. Les coefficients de corrélation dépassant cette valeur critique sont significatifs.

3.3 RÉSULTATS

Les coefficients de régression multiple, de même que les tableaux de coefficients de corrélations et les analyses de variances ont été calculés pour les tronçons correspondant à un type de sol donné, pour les données provenant de toutes les régions (1993-1996), y compris Ahuntsic.

3.3.1 Régression dans le cas où le type de sol est l'argile

Le tableau 1 montre les coefficients de corrélation (r) entre les variables indépendantes et la variable dépendante. Les coefficients sont faibles dans tous les cas ($|r| < 0.14$). Ces faibles valeurs sont une indication que l'hypothèse d'indépendance entre les variables indépendantes de la régression est acceptable. Par contre, la variable dépendante (nombre de bris totaux par tronçon) n'est pas fortement corrélée avec aucune des variables indépendantes. Aucun des coefficients de corrélation ne dépasse (en valeur absolue) le coefficient r_c, ce qui indique qu'il n'y a pas de relation linéaire significative entre les variables prédictives et la variable prédite.

Tableau 1. Coefficients de corrélation (r) entre les variables de la régression pour les tronçons ayant l'argile comme type de sol. La valeur critique du coefficient de corrélation (r_c) est au bas du tableau.

	Longueur	Diamètre	Année	Pression	Bris
Longueur	1.000	0.053	-0.009	0.081	0.089
Diamètre	0.053	1.000	-0.069	0.128	-0.070
Année	-0.009	-0.069	1.000	-0.131	-0.006
Pression	0.081	0.128	-0.131	1.000	-0.005
Bris totaux	0.089	-0.070	-0.006	-0.005	1.000

$r_c = 0.119$

Les coefficients de régression (B), les coefficients de régression centrés-normalisés (B') et les valeurs de t et p (« p value ») associées à chaque coefficient sont résumés dans le tableau 2.

Tableau 2. Analyse des résultats de la régression multiple dans le cas où le type de sol est l'argile. Les coefficients, de même que leur écart-type (s) sont donnés avec les résultats de l'analyse de variance.

	B'	$S_{B'}$	B	S_B	t(188)	p
			0.641	3.205	0.200	0.842
Longueur	0.093	0.073	0.226	0.176	1.282	0.201
Diamètre	-0.076	0.073	-0.014	0.014	-1.033	0.303
Année	-0.011	0.073	0.000	0.001	-0.150	0.881
Pression	-0.004	0.074	-0.001	0.023	-0.056	0.956

$R^2 = 0.014$ F(4,188)=.65364 p<.62500

Le tableau 2 donne aussi le résumé de l'analyse de variance. Le coefficient de détermination demeure faible (R^2= 0.014) et l'analyse de variance démontre que l'hypothèse nulle est confirmée ($F = 0.65$, $p < 0.63$).

On peut donc conclure que, pris individuellement (t et valeur de p associée) ou conjointement (F et valeur de p associée), les statistiques démontrent que les coefficients de régression ne sont pas significativement différents de 0.

La figure 2 présente la droite de régression entre les valeurs observées et les valeurs prédites par la régression multiple. Une régression qui représenterait parfaitement les valeurs observées (R^2 = 1) produirait une droite, sur laquelle seraient alignés tous les points (valeur prédite = valeur observée). La dispersion des points quantifiée par la faible valeur de R^2 (0.014) confirme visuellement l'interprétation quantitative du tableau 2. La grande majorité des points se retrouvent à l'extérieur de l'intervalle de confiance de 95%. La régression multiple ne permet pas de prédire significativement le taux de bris pour ce type de sol.

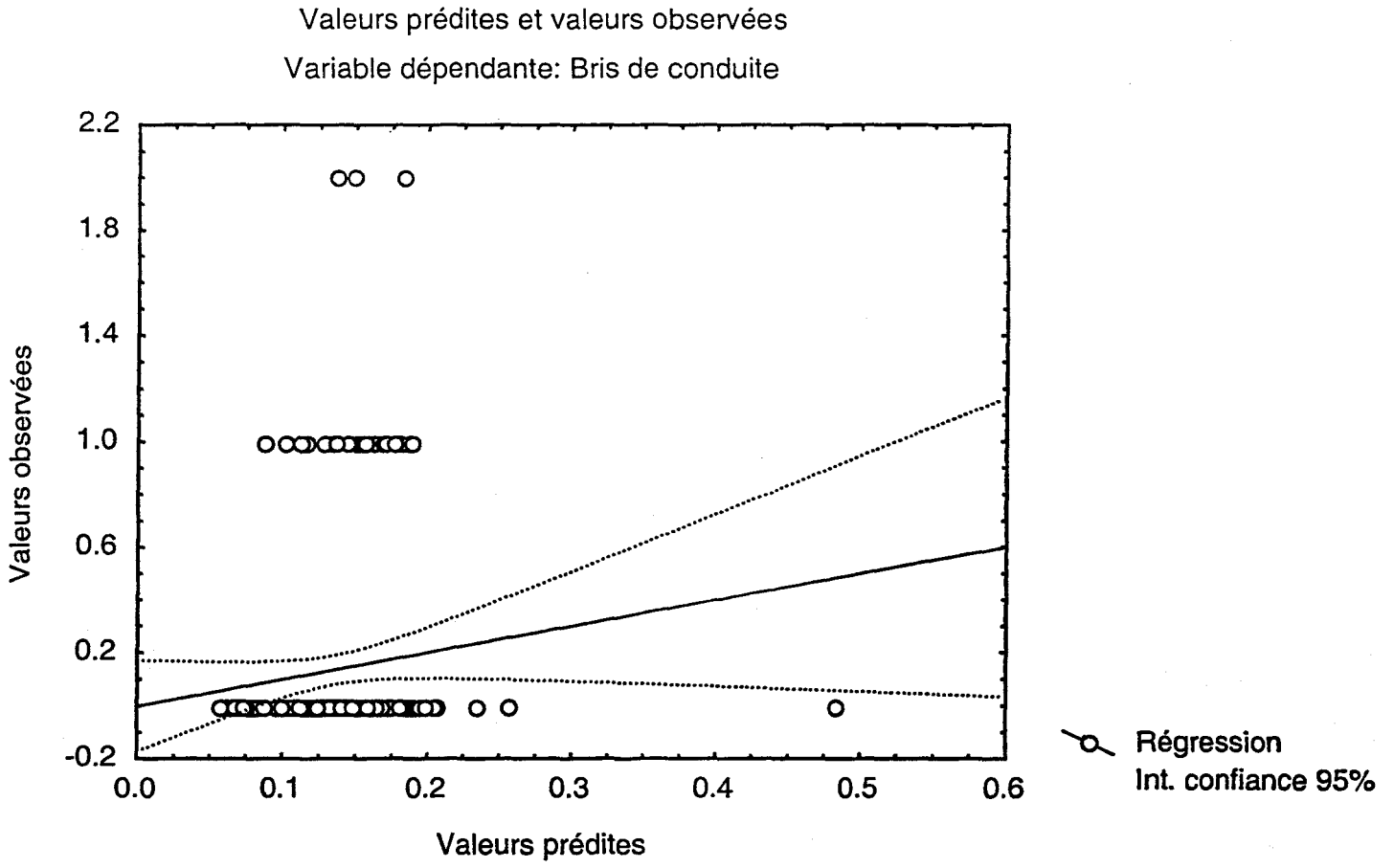


Figure 2 Relation entre les valeurs observées et les valeurs prédites, selon la régression multiple présentée au tableau 2.

3.3.2 Régression pour les autres types de sol

L'analyse présentée dans la section 3.3.1 a été reprise pour les autres types de sol. Les coefficients de corrélations présentés au tableau 3 demeurent relativement faibles entre les variables indépendantes. Les valeurs de r entre le nombre de bris et les variables indépendantes sont aussi généralement faibles ($|r| < 0.2$), sauf pour quelques exceptions. Ainsi, le nombre de bris est positivement corrélé avec la longueur du tronçon pour le roc ($r = 0.43$) de même que pour le till ($r = 0.26$) . Le nombre total de bris est négativement corrélé avec l'année d'installation pour les tronçons installés dans la tourbe.

Tableau 3. Coefficients de corrélation (r) entre les variables de la régression. La valeur critique du coefficient de corrélation (r_c) est au bas du tableau.

Sable	Longueur	Diamètre	Année	Pression	Bris totaux
Longueur	1.000	-0.078	0.005	-0.042	0.056
Diamètre	-0.078	1.000	0.030	0.073	0.126
Année	0.005	0.030	1.000	0.167	-0.080
Pression	-0.042	0.073	0.167	1.000	0.112
Bris totaux	0.056	0.126	-0.080	0.112	1.000
$r_c = 0.116$					
Till	Longueur	Diamètre	Année	Pression	Bris totaux
Longueur	1.000	0.071	0.208	0.018	0.255
Diamètre	0.071	1.000	0.054	0.102	0.048
Année	0.208	0.054	1.000	-0.189	-0.018
Pression	0.018	0.102	-0.189	1.000	0.033
Bris totaux	0.255	0.048	-0.018	0.033	1.000
$r_c = 0.116$					
Roc	Longueur	Diamètre	Année	Pression	Bris totaux
Longueur	1.000	-0.170	-0.058	0.104	0.431
Diamètre	-0.170	1.000	0.137	-0.111	-0.010
Année	-0.058	0.137	1.000	-0.026	-0.045
Pression	0.104	-0.111	-0.026	1.000	0.257
Bris totaux	0.431	-0.010	-0.045	0.257	1.000
$r_c = 0.169$					
Tourbe	Longueur	Diamètre	Année	Pression	Bris totaux
Longueur	1.000	-0.005	-0.303	-0.050	0.190
Diamètre	-0.005	1.000	0.100	0.200	-0.400
Année	-0.303	0.100	1.000	-0.267	-0.329
Pression	-0.050	0.200	-0.267	1.000	0.000
Bris totaux	0.190	-0.400	-0.329	0.000	1.000
$r_c = 0.521$					

Les résultats des régressions pour les types de sol autres que l'argile (sable, till, roc et tourbe) sont résumés au tableau 4. Les coefficients de régressions et les valeurs des statistiques t et F, de même que les valeurs de p associées y sont colligées. Les analyses de variance faites pour l'ensemble des coefficients montrent que seuls deux types de sols ont au moins un coefficient de régression dont la valeur est significativement différente de zéro. Il s'agit des conduites retrouvées dans le till ($F = 3.4$, $p < 0.01$) et dans le roc ($F = 3.4$, $p < 0.001$). Dans les deux cas,

ces résultats sont dus à la variable « longueur » qui est positivement corrélée avec le nombre de bris.

Tableau 4. Analyse des résultats des régressions multiples par type de sol (sable, till, tourbe et roche).

Sable	B'	$s_{B'}$	B	s_B	t(188)	p
Intercpt			2.995	2.421	1.237	0.217
LONG	0.072	0.070	0.455	0.442	1.031	0.304
DIAM	0.126	0.070	0.037	0.021	1.810	0.072
ANNEE	-0.105	0.070	-0.002	0.001	-1.494	0.137
PRESSION	0.123	0.071	0.005	0.003	1.746	0.082
R^2 = 0.042, F(4,199) = 2.20, p< 0.071						
Till	B'	$s_{B'}$	B	s_B	t(199)	p
Intercpt			1.630	1.809	0.901	0.369
LONG	0.268	0.074	0.863	0.240	3.603	0.000*
DIAM	0.032	0.073	0.006	0.013	0.431	0.667
ANNEE	-0.073	0.076	-0.001	0.001	-0.971	0.333
PRESS	0.011	0.074	0.000	0.002	0.147	0.883
R^2 = 0.071, F(4,199) = 3.40, p< 0.01*						
Roc	B'	$s_{B'}$	B	s_B	t(94)	p
Intercpt			-0.002	2.471	-0.001	0.999
LONG	0.422	0.092	1.061	0.231	4.599	0.000*
DIAM	0.090	0.092	0.020	0.020	0.976	0.332
ANNEE	-0.027	0.091	0.000	0.001	-0.301	0.764
PRESSION	0.222	0.091	0.008	0.003	2.447	0.016
R^2 = 0.239, F(4,94) = 3.40, p< 0.001*						
Tourbe	B'	$s_{B'}$	B	s_B	t(9)	P
Intercpt			37.189	45.867	0.811	0.438
LONG	0.112	0.306	0.538	1.467	0.367	0.722
DIAM	-0.376	0.298	-0.117	0.092	-1.262	0.239
ANNEE	-0.254	0.321	-0.018	0.023	-0.790	0.450
PRESS	0.013	0.311	0.001	0.016	0.042	0.968
R^2=0.260, F(4,9) = 0.770, p<0.57						

Les valeurs de t confirment que les coefficients de régressions associés à cette variable indépendante sont significativement différents de zéro pour le till (t = 3 .6, p <0.0001) et pour le roc (t = 4.6, p < 0.0001). Dans les deux cas, toutes les autres variables indépendantes ne permettent pas d'expliquer la variance du nombre de bris.

Les figures 3 à 6 confirment l'interprétation du tableau 4. Les régressions entre valeurs prédites et observées confirment le peu de structure entre les variables indépendantes et le taux de bris. La presque totalité des valeurs observées non nulles se retrouvent à l'extérieur de l'intervalle de confiance pour chaque type de sol. On peut donc confirmer visuellement que l'approche régressive ne permet pas de prédire le taux de bris à partir des variables quantitatives de la base de données.

Il faut aussi noter qu'étant donné que la région d'Ahuntsic a un historique de taux de bris plus long (24 ans), une analyse par régression multiple a aussi été tentée sur l'ensemble des données provenant de cette région. Les résultats n'étaient pas plus concluants que ceux présentés ici.

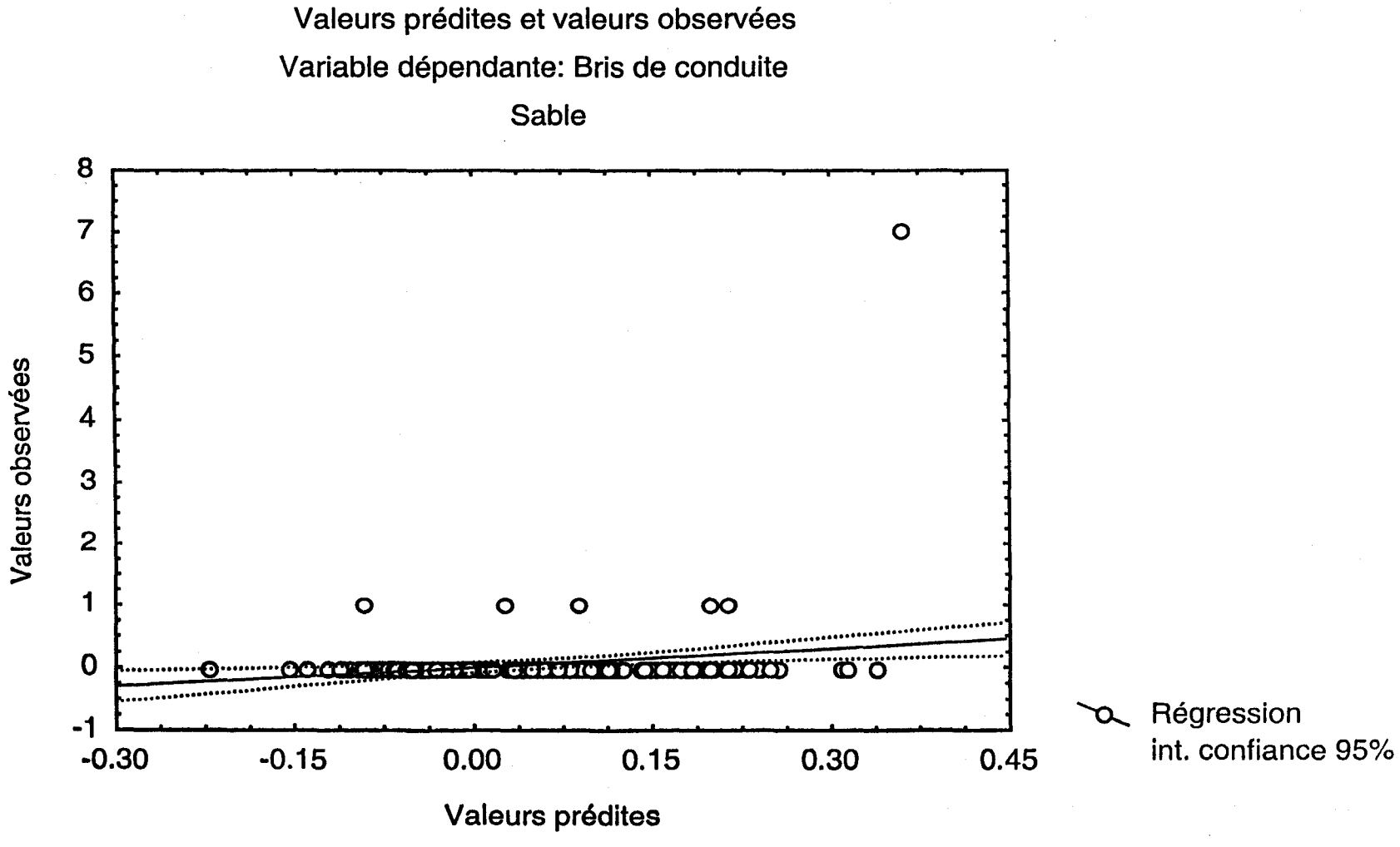


Figure 3 Relation entre les valeurs observées et les valeurs prédites, selon la régression multiple présentée au tableau 4 pour le sable.

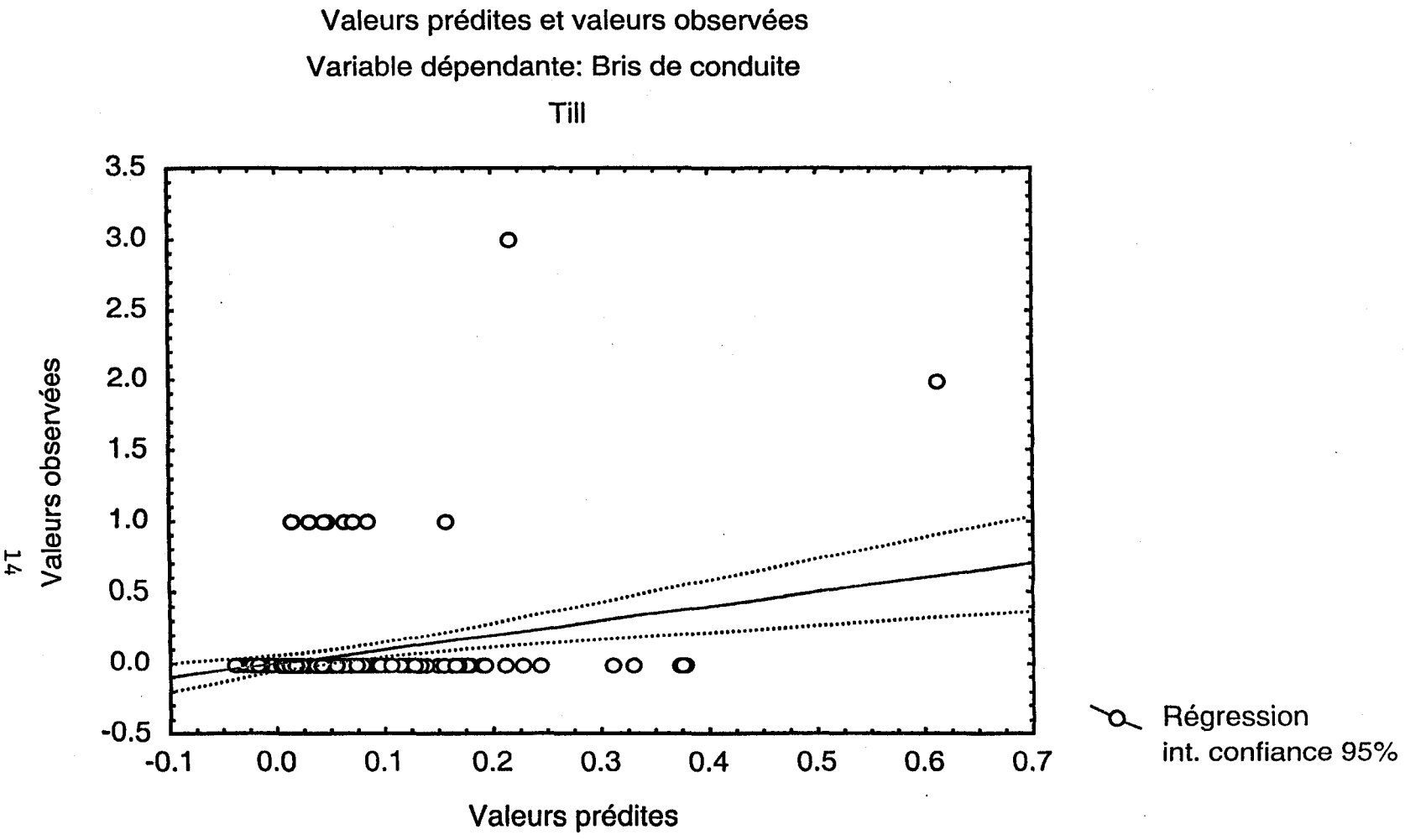


Figure 4 Relation entre les valeurs observées et les valeurs prédites, selon la régression multiple présentée au tableau 4 pour le till.

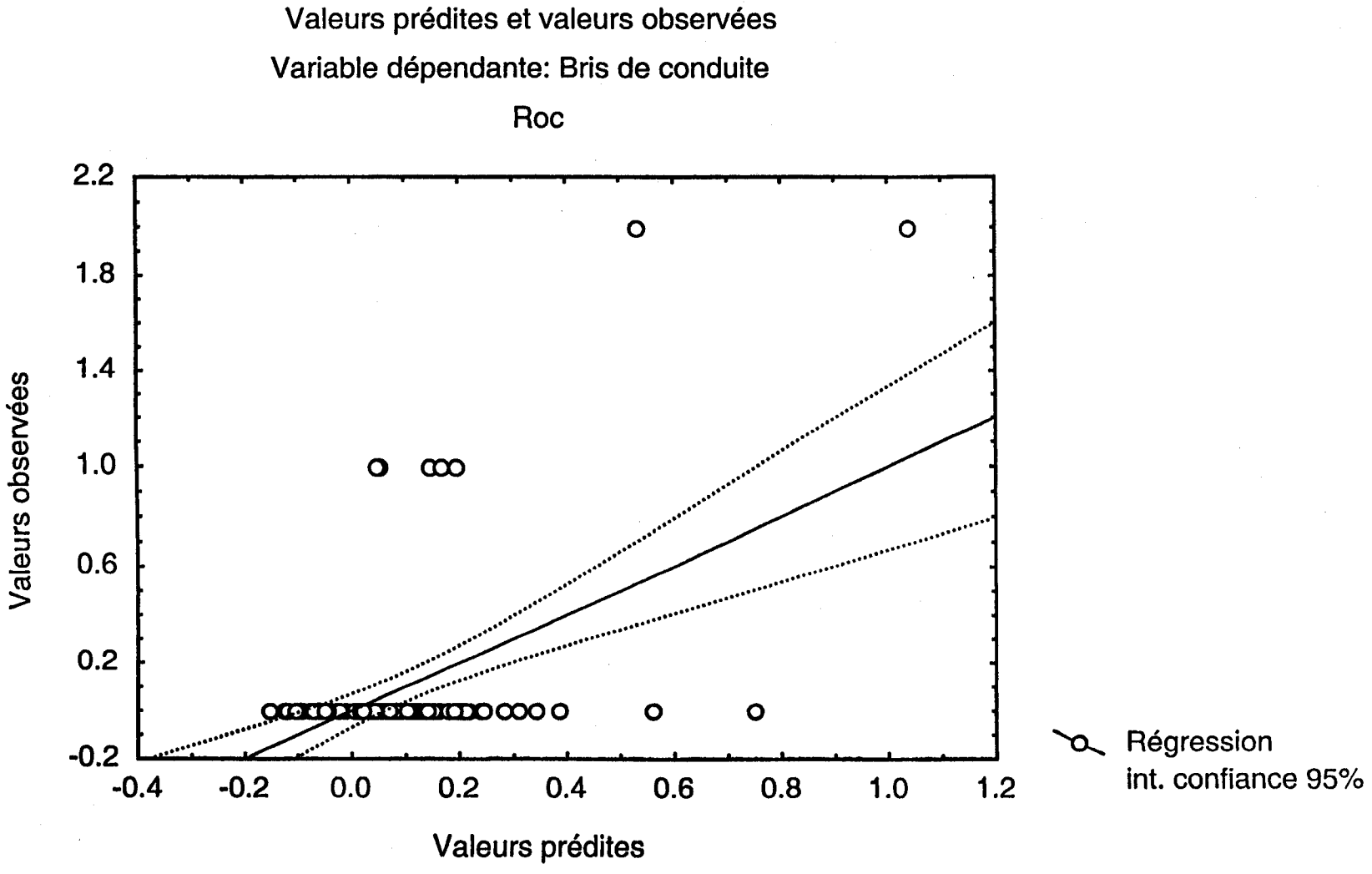


Figure 5 Relation entre les valeurs observées et les valeurs prédites, selon la régression multiple présentée au tableau 4 pour le roc.

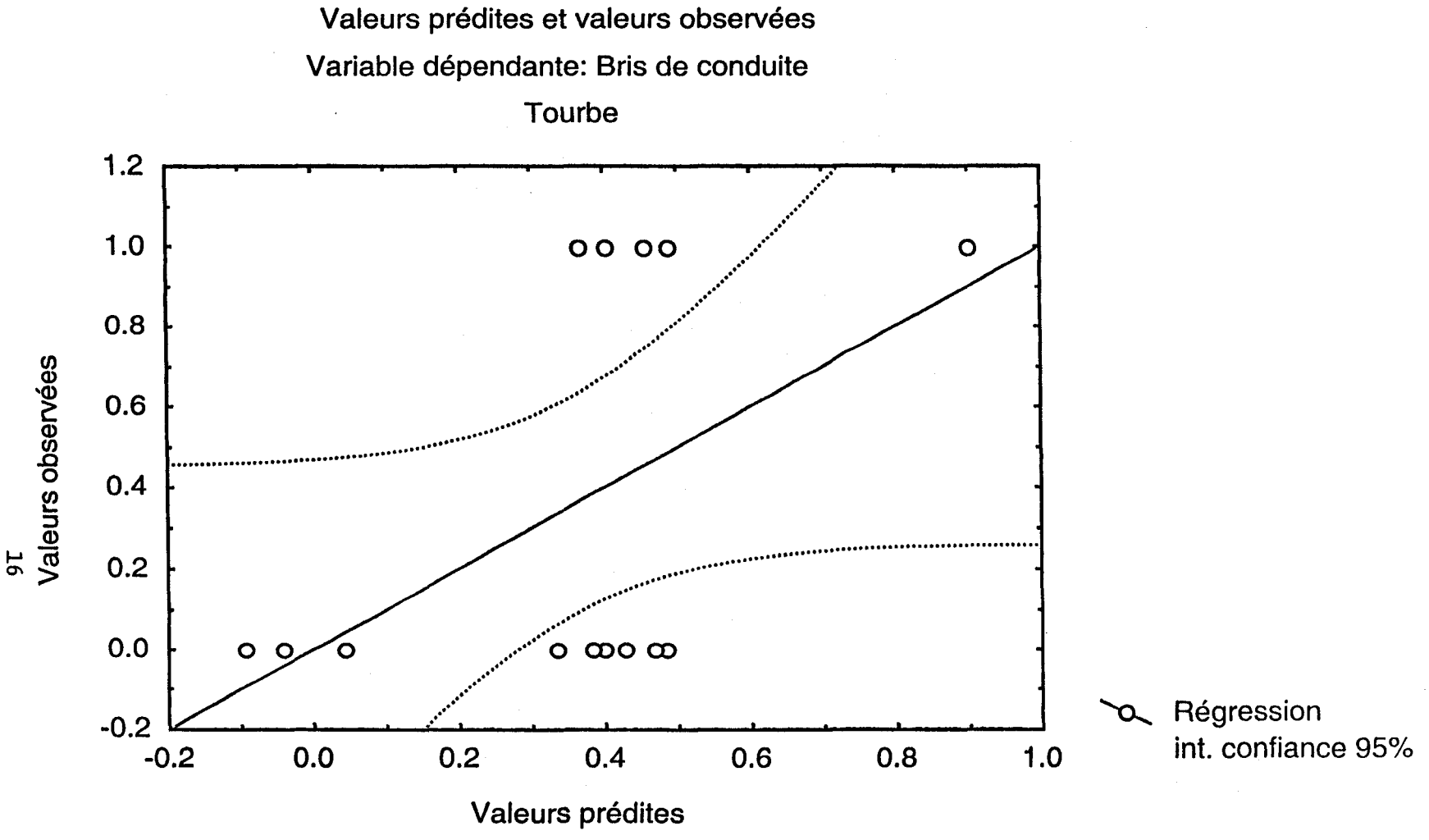


Figure 6 Relation entre les valeurs observées et les valeurs prédites, selon la régression multiple présentée au tableau 4 pour la tourbe.

4.0 ANALYSE MULTIVARIÉE

L'objectif de l'utilisation de l'analyse multivariée était dans un premier temps d'identifier la possibilité d'utiliser des méthodes multivariées pour définir des regroupements de tronçons sur la base d'un ensemble de variables caractéristiques du taux de réparation de bris (diamètre des conduites, pression, période de construction, type de sol). Par une analyse multivariée, on vise donc à définir la structure de l'ensemble des données disponibles, effectuer une classification des tronçons sur la base des variables considérées globalement et confronter les regroupements obtenus par l'analyse multivariée avec les regroupements effectués par la firme BPR d'une manière intuitive sur la base des regroupements naturels et des données disponibles.

4.1 MÉTHODOLOGIE

L'examen des données disponibles a révélé que les méthodes traditionnelles telles que l'analyse en composantes principales et les méthodes traditionnelles de classification ne pouvaient s'appliquer directement aux données en raison des types de variables mesurées qui sont soit de type nominale (type de sol), soit de type ordinale (classes de pression, diamètres, période de construction).

Pour effectuer l'analyse multivariée, nous avons opté pour la méthode d'analyse factorielle des correspondances (AFC) suivie d'une classification sur les facteurs issus de l'AFC. La méthode d'analyse des correspondances, initialement développée en France au début des années 1960 par Benzécri (Lebart *et al.* 1984), permet notamment de décrire un tableau de fréquences ou un tableau de données de type présence / absence. La méthode considère l'ensemble des colonnes du tableau ou l'ensemble des lignes du tableau comme des vecteurs points et permet de réduire les dimensions du tableau en projetant ces ensembles de points dans un espace multidimensionnel dans lequel les axes sont orthogonaux et expliquent de manière décroissante la variabilité du tableau. La détermination de cet espace factoriel revient à faire passer une droite à travers un nuage de points dans le sens de l'élongation maximum du nuage de points, ensuite une seconde droite perpendiculaire à la première dans le sens de l'élongation maximum secondaire, et ainsi de suite. On obtient donc une représentation graphique dans le plan des axes pris deux à deux (1; 2), (1; 3), (2; 3), ..., dans laquelle les vecteurs colonnes du tableau (ou vecteurs lignes du tableau) s'interprètent en fonction de leur distance entre eux et par rapport à l'origine des axes. Deux points très proches

indiquent une variation identique, tandis que deux points opposés dans le plan indiquent une variation identique, mais inverse.

En raison des propriétés de symétrie du tableau analysé, on peut faire un lien (ou une correspondance) entre les points vecteurs-colonnes (variables) et une classe de points vecteurs-lignes (tronçons) à partir de leur proximité dans l'espace factoriel et caractériser cette classe de points par la ou les variable(s) responsable(s) du regroupement.

Dans l'analyse des résultats de l'AFC pour de grands tableaux de données, il devient pratiquement impossible d'examiner la représentation de chacun des points-lignes dans les plans de l'espace factoriel pris deux à deux. Une classification des points lignes du tableau (tronçons) peut alors être effectuée par une méthode de classification automatique. L'analyse des résultats concernant les points lignes se limite alors à l'interprétation des centres de gravité des classes obtenues. Parmi les méthodes de classification disponibles, on peut considérer la classification ascendante hiérarchique (CAH) (Lebart *et al.* 1984). Cette méthode procède au calcul des distances entre chaque paire de points. La distance utilisée est la distance euclidienne puisque la classification est effectuée sur les coordonnées factorielles des tronçons. La méthode agrège chaque paire de points qu'elle remplace par leur centre de gravité. Au fur et à mesure que progresse l'agrégation, il faut agréger des groupes de points et définir un critère d'agrégation. Le critère utilisé est celui de Ward, lequel consiste à agréger les groupes de façon à maximiser la variance inter-groupe.

Le logiciel utilisé pour l'application de ces deux méthodes est le logiciel Statistica. En ce qui concerne la CAH, le logiciel accepte un maximum de 300 individus à agréger. Pour classer un plus grand nombre d'individus, Il a fallu recourir à une méthode plus puissante, soit la méthode des Nuées Dynamiques. Cette méthode procède à partir de noyaux d'agrégation que l'on définit au départ. Les nuages de points se forment en agrégeant les points voisins; au fur et à mesure que les nuages se forment, les centres de gravité se déplacent. Cette méthode permet de classer de grands ensembles de points en utilisant les mêmes définitions de distances et les mêmes critères de partition qu'avec la CAH. Il faut cependant définir au départ le nombre de classes qu'on veut obtenir alors qu'avec la CAH, on peut déduire le nombre de classes à retenir en examinant la structure de l'arbre hiérarchique.

4.2 RÉSULTATS DE L'ANALYSE DES CORRESPONDANCES

4.2.1 Tableau analysé

L'analyse multivariée d'un tableau couvrant toute la population des tronçons de la ville de Montréal était difficilement envisageable en raison de la lourdeur et de la difficulté découlant de l'analyse d'un tableau d'aussi grande taille (18174 lignes). L'analyse d'un échantillon représentatif de la population a été jugée préférable, car elle pouvait conduire aux mêmes résultats sans avoir à affronter les difficultés découlant de l'analyse d'un tableau aussi gigantesque. Le tableau qui a été analysé par l'AFC comprend les données de l'échantillon de 701 tronçons retenus pour l'analyse de régression multiple (cf. section 3.1). Rappelons que cet échantillon couvre toutes les régions de la ville de Montréal (y compris Ahuntsic) et que la représentativité de cet échantillon en ce qui concerne le taux de réparation de bris pour chacune des années 1993 à 1996 a été vérifié au moyen d'un test.

Les variables considérées dans l'analyse multivariée sont le diamètre, la pression, l'année de construction, le type de sol et le taux de bris. Afin de rendre le tableau analysable par l'AFC, il a fallu transformer les variables de façon à ramener l'information disponible en un tableau de fréquences. C'est pourquoi, on a procédé aux transformations suivantes :

- Les variables diamètre, pression, année et type de sol ont été éclatées en autant de catégories que ces variables contenait;
- Pour chacune des nouvelles variables obtenus par l'éclatement, on a codifié l'information par un 1 ou un 0 selon que pour un tronçon la catégorie est présente ou absente;
- Les catégories qui ont été retenues pour l'analyse sont les suivantes:

    - Diamètre : 6, 8, 10, 12 po
    - Pression : 30-39, 40-49, 50-59, 60-69, 70-79, 80-89, 90 et plus
    - Année : avant 1949, 1949 à 1969, 1970 et plus
    - Type de sol : argile (AR), till (TI), sable (SA), tourbe (TO), roc (RO)

En raison du nombre trop faible de tronçons de diamètre 2 ou 4 et de pression 90, on les a regroupées avec les catégories voisines. Les années ont été regroupées en trois périodes: avant 1949, de 1949 à 1969 et après 1969, comme dans les études de Kleiner et de BPR.

Le tableau comprenait donc 19 colonnes de données binaires contenant de l'information sur des variables "statiques" du réseau d'aqueduc. À ce tableau, on a adjoint des colonnes contenant de l'information sur les bris d'aqueduc pour les années 1993 à 1996. Une colonne appelée "bris" était codifiée par un 1 ou un 0 selon qu'il y avait eu un bris dans le tronçon au cours des quatre années. Quatre autres colonnes ont été ajoutées au tableau, chacune correspondant au taux de bris moyen rapporté pour chacune des catégories de diamètre, pression, période et type de sol.

Le tableau ayant fait l'objet d'une analyse par l'AFC comprend donc 701 lignes et 24 colonnes.

4.2.2 Résultats de l'AFC

L'application de l'AFC a produit les résultats suivants.

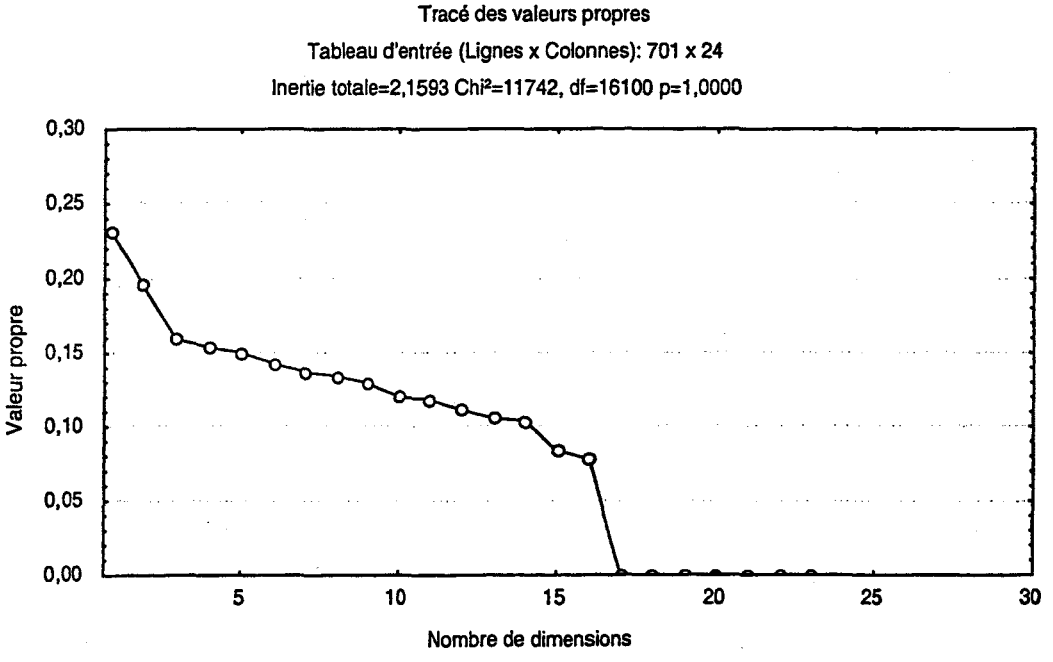


Figure 7. Tracé de la variance expliquée par les axes factoriels

L'AFC du tableau a permis d'expliquer les pourcentages de variance tels que représentés à la figure 7. Les pourcentages de variance expliqués selon les cinq premiers axes sont donnés ici:

Valeur propre :	0,232	0,196	0,160	0,154	0,150
Variance expliquée :	10,7	9,08	7,40	7,14	6,97
Variance cumulée :	10,7	19,8	27,2	34,5	41,3

Les pourcentages de variabilité peuvent paraître faibles, mais c'est généralement le cas lorsqu'on analyse un tableau binaire. Cette décroissance de la variabilité indique qu'après quatre axes, les principaux phénomènes sous-jacents au tableau sont représentés, mais qu'on pourrait considérer jusqu'à dix axes comme interprétables. Pour les fins de l'analyse, on a retenu les cinq premiers axes factoriels.

Les résultats de l'AFC en ce qui concerne les deux premiers axes sont reportés à l'annexe A. On y retrouve les coordonnées factorielles des points colonnes, ainsi que les calculs des contributions à l'inertie des axes et la qualité de représentation.

L'essentiel des résultats est reporté à la figure 8 qui montre la représentation des points colonnes dans l'espace des axes factoriels 1 et 2.

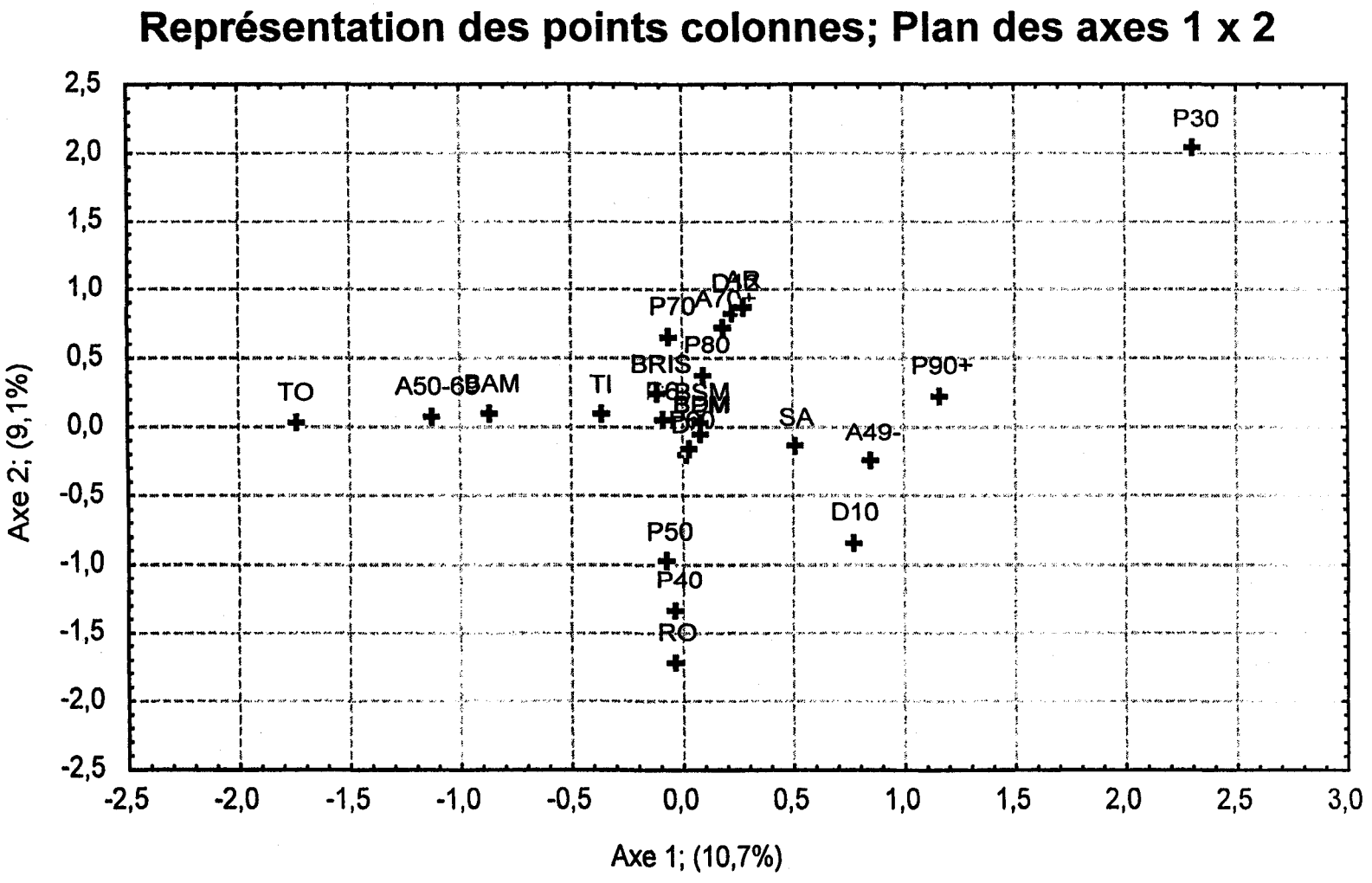


Figure 8 Représentation des points colonnes dans le plan des axes 1 et 2

L'interprétation de cette figure fait ressortir les points suivants :

- Dans le plan des axes 1 et 2, on retrouve une opposition entre les points représentant les différentes catégories des variables considérées, ce qui est normal compte tenu que chacun des vecteur-colonnes représentant ces catégories sont inversement reliés par la relation de présence / absence

- Les points qu'on retrouve au centre du nuage de points (près de l'origine des axes 1 et 2) sont mal représentés dans ce plan. Ces points auront un pouvoir déterminant pour les axes 3, 4 et plus.

- Certaines variables ressortent comme ayant un pouvoir plus déterminant au niveau de la structure du tableau; dans l'ordre ce sont les variables "pression", "type de sol" et "période", qui ont un pouvoir plus discriminant dans le plan des axes 1 et 2; la variable "diamètre" intervient plutôt sur le plan tracé par les axes 3 et plus.

- certains points se retrouvent projetés aux extrémités du nuage de points (P30, TO); cet éloignement dans le plan est influencé artificiellement par leur poids très faible dans l'analyse (22 et 14 présences sur 701 respectivement pour P30 et TO). En raison de leur faible poids, ils ne déséquilibrent pas l'analyse dans son ensemble.

- La variable "bris" se retrouve au centre du nuage de points indiquant un profil moyen par rapport à l'ensemble des autres variables.

- Les variables décrivant le bris moyen par catégorie se retrouvent également comme des variables avec un profil moyen, remplissant leur rôle de pondérer la structure du tableau.

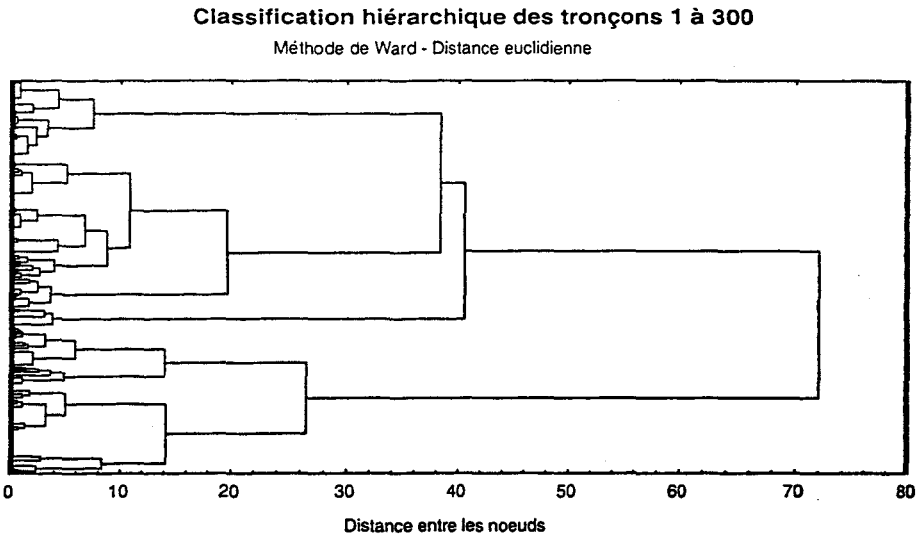
- On peut en conclure que, d'après la répartition des points colonnes dans le plan des axes 1 – 2, 1 - 3, 2 – 3 ..., que les points lignes du tableau se répartiront dans l'espace factoriel d'une manière qui reflétera la contribution de l'ensemble des variables à leurs caractéristiques.

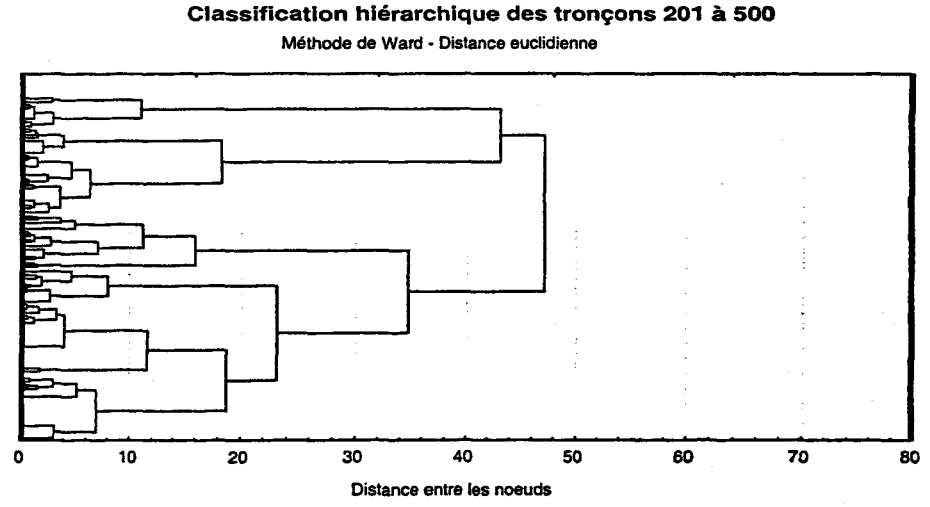
La phase d'analyse du tableau par l'AFC était une étape exploratoire qui avait pour but de dégager l'ensemble des interrelations entre les variables du tableau, de vérifier si les données étaient suffisamment concordantes permettant d'effectuer une analyse de classification des tronçons potentiellement interprétable. On peut présumer qu'en effectuant une classification sur les coordonnées factorielles des points lignes du

tableau, la classification résultante devrait être interprétable globalement en termes de sensibilité aux réparations et caractérisés par des catégories de variables ayant un pouvoir explicatif sur les taux de bris. En conservant seulement les cinq premiers axes factoriels, on effectue une sorte de filtrage du tableau de données initial afin de mettre en évidence les principales tendances qui se dégagent.

4.3 RÉSULTATS DE LA CLASSIFICATION ASCENDANTE HÉRARCHIQUE

La méthode de classification ascendante hiérarchique a été utilisée dans un premier temps. L'utilisation de cette méthode permet d'identifier si un nombre de groupes optimal ressortait dans l'analyse. Puisque le logiciel Statistica ne permettait que de classer un maximum de 300 individus, il a été décidé de procéder à trois analyses successives, soit les observations 1 à 300, les observations 200 à 500 et les observations 401 à 701. Cette façon de procéder ne posait pas de difficultés puisque les tronçons étaient disposés de façon aléatoire dans le tableau de données. Les résultats de cette classification sont présentés dans la figure 9.





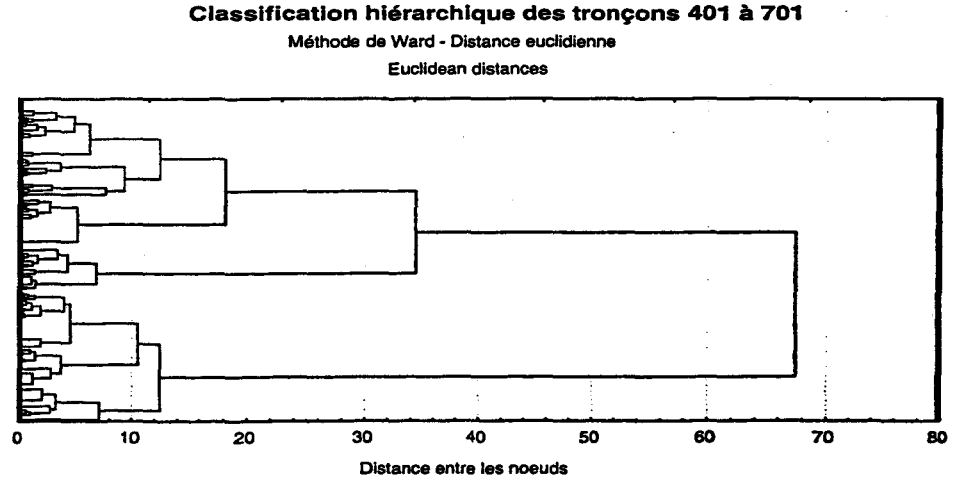


Figure 9 Classification ascendante hiérarchique des tronçons

L'examen des classifications hiérarchiques montre que dépendant des tronçons choisis, la structure de l'arbre hiérarchique change considérablement. Si on effectuait une coupure au niveau de la distance entre les nœuds au niveau égal à 30, on aurait 4, 4 et 3 classes de tronçons. C'est donc dire que 3 à 4 groupes de tronçons se dégagent de cette analyse. Par contre, si on voulait considérer un plus grand nombre de classes (8 par exemple), l'analyse serait instable.

Il ressort de cette analyse que trois à quatre groupes de tronçons posséderaient une certaine homogénéité. Il semble que la structure des données ne permet pas de faire des regroupements à un niveau plus fin.

4.4 RÉSULTATS DE LA CLASSIFICATION PAR LES NUÉES DYNAMIQUES

4.4.1 Comparaison des classifications obtenues par analyse multivariée avec celles obtenues par BPR

Afin de comparer les résultats avec les regroupements effectués par BPR, on a utilisé la méthode des Nuées Dynamiques pour délimiter huit classes de tronçons. Cette méthode de classification, conçue pour traiter de grands tableaux, permet de considérer les 701 tronçons à la fois et de définir le nombre de classes que l'on désire. Puisqu'on utilise les mêmes critères de distance et d'agrégation qu'avec la CAH, l'interprétation des classes obtenues peut être faite comme si on avait considéré la partie supérieure de la classification hiérarchique avec huit classes. Sur la figure 10, on montre l'histogramme de répartition des huit groupes obtenus par cette méthode de classification comparé avec le classement effectué par BPR.

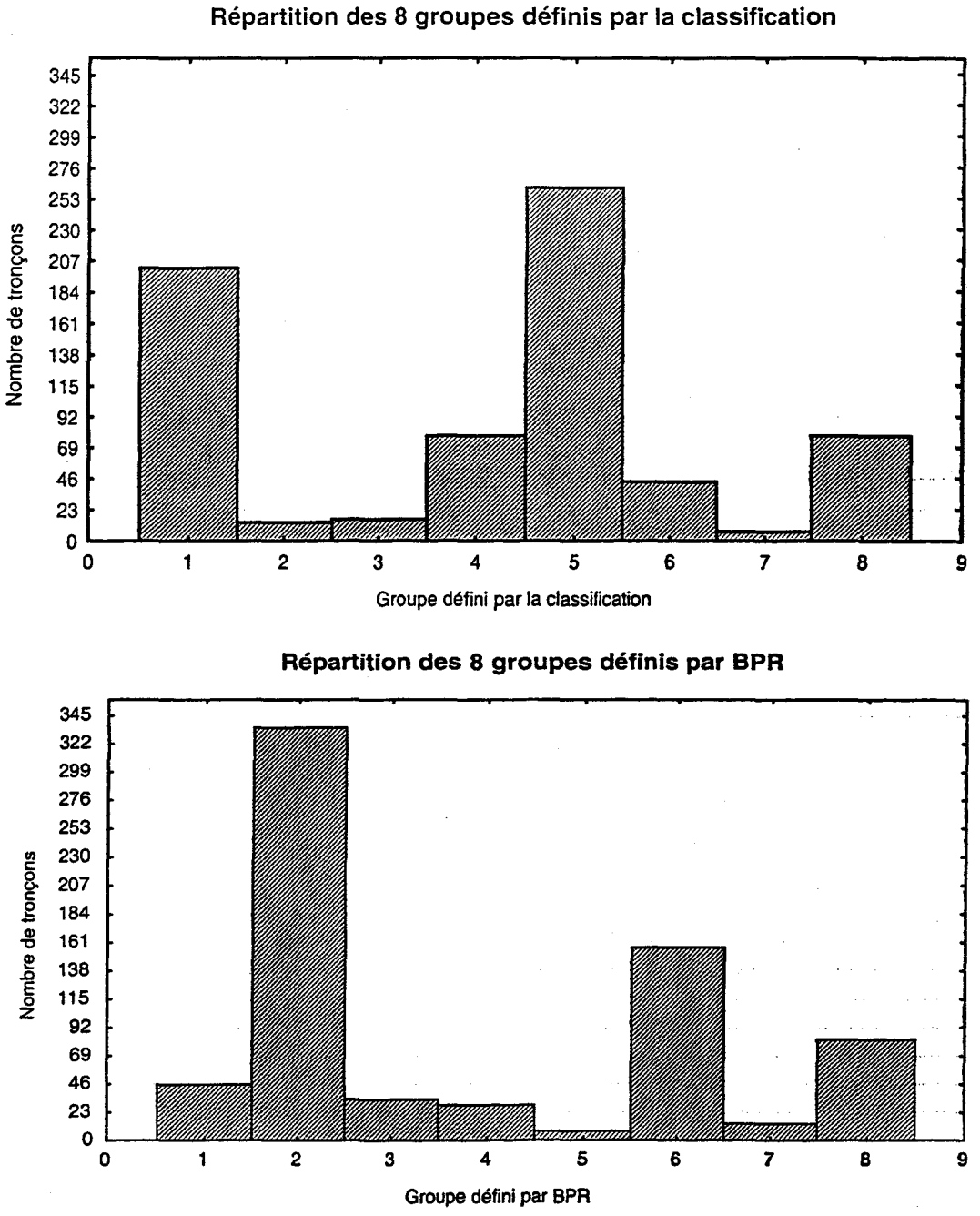


Figure 10 Comparaison des histogrammes des deux classifications

Les distributions des regroupements obtenus par classification automatique diffèrent beaucoup des groupements obtenus par BPR (Figure 10 et Tableau 5). Dans le groupe 2, qui est le plus important défini par BPR, on retrouve tous les tronçons avant 1949 pour tous les type de sol et pour toutes les pressions, mais pour les conduites de diamètre supérieur à 6 pouces.

Tableau 5. Comparaison des regroupements par classification avec ceux obtenus par BPR

Classification	Regroupement par BPR								
	1	2	3	4	5	6	7	8	Somme
1	9	52	1	20	2	85	11	21	201
2	0	13	0	0	0	1	0	0	14
3	0	10	4	0	1	1	0	1	17
4	6	43	25	0	0	0	1	4	79
5	26	164	2	5	4	43	1	16	261
6	2	27	0	1	0	11	0	2	43
7	0	5	0	0	0	0	0	3	8
8	2	21	1	3	0	16	0	35	78
Somme	45	335	33	29	7	157	13	82	701

Avec la classification automatique, on retrouve deux groupes importants (1 et 5) et deux autres groupes moins importants (4 et 8) qui ne correspondent pas du tout aux groupes définis par BPR. On peut donc retenir que la classification effectuée par analyse multivariée diffère complètement du classement effectué par BPR.

4.4.2 Caractérisation des classes de tronçons obtenues par analyse multivariée

Avec la classification par les Nuées Dynamiques, on pourrait montrer que les classes stables identifiées par la CAH se retrouvent parmi les classes 1, 4, 5 et 8.

Il est possible de caractériser chacun des huit groupes définis par la classification en fonction des points variables caractéristiques de leur regroupement (Figure 11 et Tableau 6)

Tableau 6. Caractérisation des huit classes de tronçons obtenues par analyse multivariée

Variable	Classification								
	1	2	3	4	5	6	7	8	Somme
D ≤ 6	32	0	0	5	42	1	0	1	81
D 8	129	11	11	61	175	9	1	48	445
D 10	3	1	4	5	6	29	0	0	48
D 12	37	2	2	8	38	4	7	29	127
AR	61	6	1	0	92	6	8	20	194
RO	7	0	9	79	3	0	0	2	100
SA	47	7	2	0	105	27	0	23	211
TI	73	1	5	0	61	10	0	32	182
TO	13	0	0	0	0	0	0	1	14
P 30-39	0	14	0	0	0	0	8	0	22
P 40-49	0	0	16	0	0	15	0	0	31
P 50-59	41	0	1	50	42	9	0	21	164
P 60-69	58	0	0	26	81	2	0	16	183
P 70-79	74	0	0	0	88	12	0	30	204
P 80-89	26	0	0	3	34	4	0	5	72
P 90+	2	0	0	0	16	1	0	6	25
A 49-	0	14	3	54	256	40	5	8	380
A49-69	201	0	14	20	0	3	0	1	239
A 70+	0	0	0	5	5	0	3	69	82

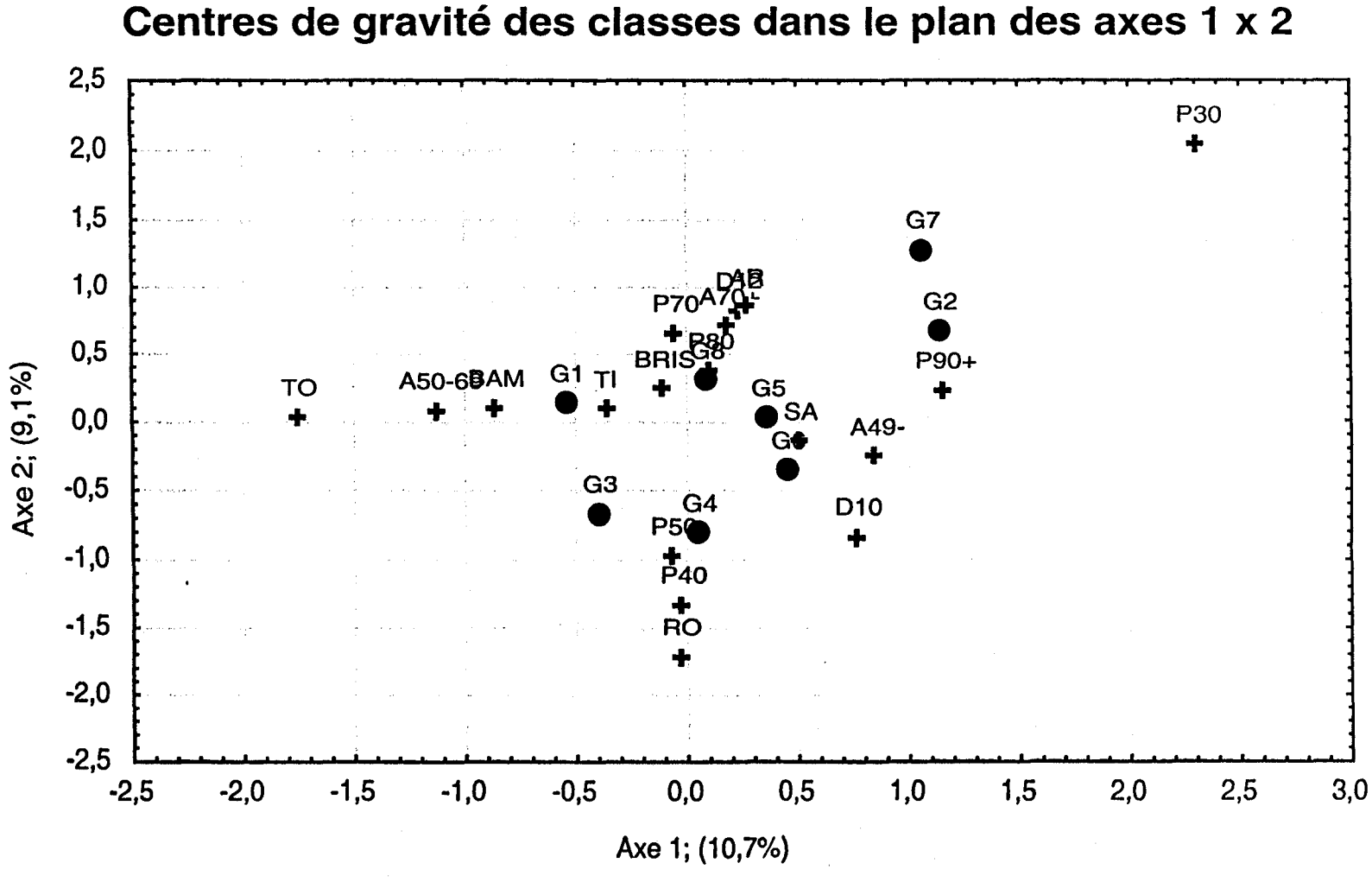


Figure 11 Représentation des centres de gravité des classes dans le plan des axes 1 et 2

On voit que les groupes 1, 5 et 8 sont caractérisés principalement par la période de construction (1 par A49-69); 5 par A49-; 8 par A70+). La classe 4 est principalement caractérisée par le type de sol (présence de roc) et un peu moins par la période (A49- et A49-69). Il semble donc que la période de construction soit un facteur discriminant pour la détermination d'au moins quatre zones homogènes.

En examinant les caractéristiques de chacune des huit classes et en interprétant les proximités dans l'espace factoriel, il est possible de caractériser ces classes plus finement. Par exemple, on peut constater que la présence de tourbière est une caractéristique importante pour le groupe 1. On peut aussi définir des caractéristiques pour les groupes 2, 3, 6 et 7. Le groupe 6 par exemple est principalement caractérisée par la présence de sable et par la période de construction avant 1949. Il faut toutefois se rappeler que les quatre groupes moins importants sont moins stables; il serait préférable de les valider par une autre méthode comme l'analyse de variance par exemple.

5.0 ANALYSE DE VARIANCE

5.1 INTRODUCTION

BPR a tenté de faire des regroupements de conduites à partir des variables indépendantes de la base de données (type de sol, année d'installation, diamètre et pression). Pour ce faire, deux critères ont été utilisés, soit la différence entre le taux de bris moyen de chaque groupe, et la longueur cumulative des tronçons dans un échantillon.

Afin d'asseoir ces regroupements sur une base statistique vérifiable, il a été convenu de vérifier si les moyennes calculées pour chaque groupe étaient significativement différentes. La méthode reconnue pour ce genre de vérification consiste à procéder à une analyse de variance (ANOVA).

5.2 MÉTHODOLOGIE

5.2.1 Approche générale

L'analyse de variance consiste à tester l'égalité simultanée des moyennes de k populations à l'aide de k échantillons de taille n_k :

$$H_0 : \mu_1 = \mu_2 = ... = \mu_k$$

Plutôt que de procéder à des tests directs sur la moyenne, on calcule d'abord ce qu'il est convenu d'appeler les « effets » de chaque population. Chaque observation x_{ij} (c'est-à-dire le taux de bris, dans notre cas) peut être vue comme la somme de trois termes :

$$x_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

où : $\mu = \frac{\sum_{i=1}^{k} \mu_i}{k}$ est la moyenne de toutes les populations (c'est à dire le taux moyen de bris pour tous les tronçons, dans notre cas) ;

ε_{ij} = la déviation ou différence entre la $j^{ième}$ valeur observée dans la population i et la moyenne μ_i de cette population (différence entre un taux de bris pour un tronçon donné et le taux de bris moyen de ce groupe).

α_i est l'effet de la population i.

On a la restriction suivante : $\sum_{i=1}^{k} \alpha_i = \sum_{i=1}^{k} (\mu_i - \mu) = 0$

Les moyennes de k populations ne seront pas différentes si l'hypothèse suivante sur les effets est respectée :

$$H_0 : \alpha_1 = \alpha_2 = ... = \alpha_k = 0$$

Si au moins une des moyennes est différente des autres ou de façon équivalente, si un des effets est non nul, alors l'hypothèse H_0 est infirmée.

La vérification de cette hypothèse passe par la comparaison de deux estimations indépendantes de la variance de la population. Ces deux estimations sont la somme du carré des erreurs SSE et la somme du carré des moyennes des groupes SSC. Ces deux valeurs sont définies comme suit, pour j observations provenant de k groupes :

$$SSC = \frac{n\sum_{i=1}^{k}(x_i - \bar{x})^2}{k-1}$$

$$SSE = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n}(x_{ij} - \bar{x})^2}{k(n-1)}$$

où : $n = \sum_{i=1}^{k} n_i$

La statistique F, permettant de confirmer ou d'infirmer l'hypothèse H_0 que les moyennes sont égales, est calculée par le rapport suivant :

$$F = \frac{SSC}{SSE}$$

Comme dans le cas de la régression multiple (section 3.2), la valeur de F doit dépasser la valeur critique $F(1-\alpha; \nu_1, \nu_2)$ pour infirmer l'hypothèse H_0. On associe à ce seuil de dépassement une valeur de p qui quantifie la certitude du résultat du test.

Puisque l'analyse de variance présuppose que les valeurs des variables utilisées ont une variance homogène. Cette hypothèse a été testée à l'aide du test de Levene (Levene, 1960). Dans bien des cas, le test a révélé que les variances des variables étaient significativement différentes. Cependant, l'analyse de variance est suffisamment robuste pour pouvoir détecter des moyennes significativement différentes même si cette l'hypothèse d'homogénéité des variances n'est pas respectée (Pfaffenberger et Patterson, 1987)

5.2.2 TEST *A POSTERIORI*

Si l'hypothèse de base H_0 est rejetée par l'ANOVA, il faut tenter de faire une analyse plus précise afin de déterminer laquelle ou lesquelles des moyennes sont différentes des autres.

Le test de Scheffé est une des ces méthodes. Pour ce test, il faut d'abord calculer C tels que :

$$C = \sum_{j=1}^{k} c_j \mu_j$$

où la somme des coefficients c_j souvent appelés « contrastes » doit être nulle.

L'hypothèse de base s'énonce alors comme suit :

$$H_0 : C = 0$$

La valeur critique de Scheffé est définie ainsi :

$$S = \sqrt{(k-1)F_\alpha(k-1, n-k)SSE\sum_{j=1}^{k}\frac{c_j^2}{n_j}}$$

où : k = nombre de groupes ;

F_α(k-1, n-k) = valeur de F pour k-1 degrés de libertés pour les groupes, n-k degrés de libertés pour la population;

c_i= coefficients des contrastes pour le groupe j ;

$$n = \sum_{j=1}^{k} n_j .$$

Si $|S_c| = \left|\sum_{j=1}^{k} c_j \overline{x_j}\right|$ > S, on rejette l'hypothèse de base. Ce test est ainsi répété entre chaque paire de groupes, afin de déterminer quelles moyennes testées comme identiques sont en effet différentes.

Des analyses de variance (ANOVA) à un seul critère ont été faites dans le cadre de cette étude. Les moyennes de la variable dépendante (taux de bris) ont été comparées successivement pour chaque critère (type de sol, diamètre, pression et année d'installation). Lorsque l'analyse de variance révèle des moyennes significativement différentes, on utilise le test de Scheffé pour établir quels sont les groupes dont les taux de bris moyens diffèrent significativement et qui, par conséquent, ne devraient pas être fusionnés. Si on regroupe des classes dont les moyennes ne sont pas différentes, on refait une ANOVA pour ce critère avec les nouveaux groupes, afin de s'assurer que le processus de regroupement produit toujours des classes dont les moyennes de taux de bris sont différentes statistiquement.

Il faut en plus souligner que les regroupements qui seront faits à partir d'ANOVA devront être les mêmes pour les 20 années de données d'Ahuntsic et les quatre années disponibles pour les autres régions. Les tests devront être fait séparément sur ces deux échantillons.

5.3 EXEMPLE D'APPLICATION DE L'ANALYSE DE VARIANCE : CRITÈRE DES DIAMÈTRES DE CONDUITES

Les ANOVA et tests de Scheffé ont été exécutés à l'aide du logiciel Statistica (Statsoft, 1994). Dans chacun des cas, les résultats présentés incluent le tableau des moyennes par groupe, avec variances et écart-types, les valeurs de F et p de l'analyse de variance, de même que les valeurs de p pour les tests de Scheffé. Un test qui démontre un effet (c'est-à-dire que la statistique infirme l'hypothèse de base pour un niveau de confiance $\alpha = 0.05$ et pour un p jugé significatif) est signalé dans les tableaux de résultats à l'aide d'un astérisque.

Les moyennes, écart-types et variances pour chaque classe de diamètres (tableau 7) permettent déjà de constater que le taux de bris moyen est plus faible pour les conduites de 8 po (0.73 bris/km) et 12 po (0.74 bris/km), que pour celles de 6 po (1.28 bris/km) et 10 po (1.06 bris/km).

Tableau 7. Résumé des moyennes, écart-types et variances par classe de diamètre

Diamètre (po)	N	Moyenne (bris/km)	Écart-type (bris/km)	Variance (bris/km)^2
6	1569	1.28	4.29	18.40
8	11881	0.73	3.71	13.79
10	1046	1.06	4.65	21.64
12	3678	0.74	3.54	12.55
Toutes les classes	18174	0.80	3.80	14.41

L'analyse de variance (tableau 6) permet de confirmer que certaines moyennes sont significativement différentes des autres ($F = 11.63$, $p < 0.000001$).

Tableau 8. Analyse de variance des taux de bris moyens par classe de diamètre

Statistique	Valeur
SSC	167.18
Degrés de libertés (k-1)	3
SSE	14.39
Degrés de liberté (n-k-1)	18170
F	11.63*
P	<0.000001*

Finalement, le test de Scheffé (tableau 9) nous permet de constater que les conduites de 6 po ont une moyenne différente des moyennes de deux autres classes de

diamètres (8 po, p = 0.000002 et 12 po, p = 0.00005). Les conduites de 8 po ont une moyenne significativement différente des conduites de 6 po seulement. Les conduites de 10 po ont une moyenne qui ne se distingue pas statistiquement des autres classes. Quant aux conduites de 12 po, il a déjà été mentionné que leur taux de bris moyen est significativement différent de celui des conduites de 6 po, mais il ne se distingue pas des autres classes de conduites.

Tableau 9. Valeurs de p pour les tests de Scheffé entre chaque classe de diamètre

Diamètres (po)	6	8	10	12
6		0.000002*	0.554	0.00005*
8	0.000002*		0.066	0.999
10	0.554	0.066		**0.113**
12	0.00005*	0.999	0.113	

Les résultats des tableaux 7, 8 et 9 montrent que certaines classes de diamètres pourraient être regroupées si on se base sur la moyenne des taux de bris. Par exemple, les conduites de 10 po et 12 po ont des moyennes qui ne se distinguent pas l'une de l'autre par le test de Scheffé.

Les conduites de 8 po, par contre ne devraient pas être regroupées avec celles de 6 po si on utilise les résultats de l'ANOVA et du test *a posteriori* comme critères de regroupement.

Une seconde analyse de variance a donc été testée avec deux classes de diamètres, soit la classe des conduites de 6 po, et une seconde classe qui englobe toutes les autres conduites.

Le tableau 10 montre que le taux de bris moyen pour la classe regroupée (0.76 bris/km pour les diamètres de 8 po à 12 po) demeure inférieur à celui des conduites de 6 po (1.28 bris/km). L'analyse de variance, dans ce cas, est équivalente à effectuer un test de Student puisqu'il y a seulement deux classes. Cette analyse (tableau 11) montre que les deux taux de bris moyens demeurent significativement différents ($F = 14.39$, $p < 0.000001$).

Tableau 10. Résumé des moyennes, écart-types et variances par classe de diamètre avec regroupement.

Diamètre (po)	N	Moyenne	Écart-type	Variance
6	1569	1.28	4.29	18.40
8-12	16605	0.755	3.74	14.01
Toutes les classes	18174	0.800	3.80	14.41

Tableau 11. Analyse de variance des taux de bris moyens par classe de diamètre avec regroupement

Statistique	Valeur
SSC	396.4
Degrés de libertés (k-1)	1
SSE	14.39
Degrés de liberté (n-k-1)	18172
F	14.39*
P	<0.000001*

La méthodologie proposée dans cette section permettra de procéder à des regroupements de tronçons selon un critère statistique objectif. L'analyse multivariée présentée à la page précédente permettra de dégager les classes de regroupement souhaitables étant donné la structure de la base de données. L'ANOVA pourra donc être faite à la lueur des résultats de l'AFC. On peut aussi souligner que cette méthode peut s'appliquer aux échantillons de tronçons avec bris seulement. Bien que les moyennes calculées seront différentes, les conclusions de l'ANOVA seront les mêmes, puisque les effets calculés sont différents à cause des valeurs non-nulles.

6.0 UTILISATION DE UN OU DEUX PARAMÈTRES DE CALAGE POUR LES COURBES EXPONENTIELLES

Une fois l'analyse de classification des tronçons par groupe terminée, on cherche à prédire l'augmentation du taux de bris à l'aide de l'équation exponentielle qui a été utilisée dans d'autres études pour la prédiction du taux de bris (ex. Kleiner 1998) :

$$N(t)_i = N(t_0)_i e^{A_i(t-t_0)}$$

où $N(t)_i$ = taux de bris par année pour le groupe i et l'année t;
$N(t_0)_i$ = taux de bris pour le groupe i durant la première année t_0;
A_i = coefficient du taux de bris pour le groupe i

Cette équation de régression peut être linéarisée par transformation logarithmique:

$$\log N(t) = \log(N(t_0)e^{A(t-t_0)}) = \log(N(t_0)) + A(t-t_0)$$

où : log ($N(t_0)$) = constante de l'équation de régression linéaire;
A = coefficient de la régression linéaire.

Les mêmes critères d'adéquation que ceux utilisés dans le cas des régressions linéaires peuvent donc être utilisés.

Kleiner (1998) a défini $N(t_0)$ comme étant le taux de bris initial pour un groupe homogène donné. Statistiquement, $\log(N(t_0))$ correspond à la constante de l'équation de régression linéaire qui est obtenue en linéarisant l'équation exponentielle. Cette valeur pourrait donc être traitée comme un paramètre de l'équation et ajustée en conséquence, plutôt qu'être considérée comme une valeur physique.

Il faut maintenant établir s'il est préférable de caler la courbe exponentielle avec un ou deux paramètres. En modifiant le nombre de paramètres, on modifie le nombre de degrés de liberté. On peut donc vérifier si l'augmentation du nombre de paramètres améliore l'adéquation de la régression en utilisant comme critère, le coefficient de détermination ajusté. Ce coefficient est calculé en divisant les numérateurs et les dénominateurs de R^2 par leurs degrés de liberté respectifs :

$$R^2_{ajus} = R^2 \frac{(n-1)}{(n-k-1)}$$

où : n = taille de l'échantillon;

k = nombre de paramètres.

Le coefficient de détermination ajusté permet ainsi d'évaluer et de comparer le pourcentage de la variance expliquée. Si l'ajout d'un paramètre permet une augmentation significative du R^2 ajusté, il est alors probable qu'il serait plus profitable de traiter $N(t_0)$ comme un paramètre plutôt que comme une constante ayant une signification physique.

7.0 CALCUL DES INTERVALLES DE CONFIANCE

Les intervalles de confiance sont habituellement calculés pour des régressions linéaires. Pour traiter une relation non-linéaire comme celle représentée par la courbe exponentielle, il faut d'abord linéariser l'équation (voir section 6). On a initialement l'équation suivante :

$$N(t) = N_o e^{A(t-t_0)}$$

On linéarise cette équation par transformation logarithmique :

$$\log N(t) = \log(N_0 e^{A(t-t_0)}) = \log(N_0) + A(t-t_0)$$

où : log (N_0) = constante de l'équation de régression linéaire;

A = coefficient de la régression linéaire.

Soit Y_i, l'estimation par régression linéaire de la vraie valeur (η_i) pour un X_i donné. On aurait, dans le cas qui nous concerne :

$$Y_i = \log N(t)$$
$$X_i = (t - t_o)$$

l'intervalle de confiance de la vraie valeur correspondant à X_i se définit comme suit :

$$Y_i - t_1 S\sqrt{\frac{1}{n} + \frac{(X_i - \overline{X})^2}{\sum (X_i - \overline{X})^2}} < \eta_i < Y_i + t_1 S\sqrt{\frac{1}{n} + \frac{(X_i - \overline{X})^2}{\sum (X_i - \overline{X})^2}}$$

où :

t_1 est la variable de Student à N-2 degrés de liberté correspondant à un niveau de signification fixé (5%), c'est-à-dire $t_{n-2}(\alpha/2)$;

S^2 est l'estimation de la variance σ^2 au point y_i.

Le calcul de l'intervalle de confiance nécessite donc que les valeurs de N(t) soient non nulles puisque Y_i = log N(t).

Il faudrait donc traiter le sous-échantillon des conduites ayant déjà eu un ou des bris séparément de celui contenant les tronçons avec un taux de bris nul. Ce traitement peut se faire dans un cadre probabiliste.

Pour chaque groupe homogène de conduites, on peut calculer la probabilité p_0 que le taux de bris soit nul :

$$P[N = 0] = p_0$$

Ce taux est simplement égal à la proportion de tronçons sans bris dans un groupe donné. La probabilité qu'un tronçon ait un taux de bris non nul est $1-p_0$. On s'intéresse alors à la probabilité du taux de bris N entier compris entre 0 et une valeur maximale donnée N_{max}, conditionnelle à ce que le taux soit non nul, que l'on note comme suit :

$$P[0 < N \leq N_{max} | N > 0] = p_x$$

Selon le théorème de Bayes on peut calculer cette probabilité de la manière suivante :

$$p_x = \frac{P[0 < N < N_{max}]}{P[N > 0]} = \frac{P[0 < N < N_{max}]}{1 - p_0}$$

Sur une période de plusieurs années, la proportion du taux de bris N ($0<N<N_{max}$) peut être calculé. On peut ensuite calculer la probabilité d'occurrence conditionnelle à N > 0. Il est alors possible de traiter les valeurs non nulles séparément, et de faire les calculs de régression non-linéaires et d'intervalles de confiance sur ces valeurs, connaissant leur probabilité d'occurrence.

8.0 CONCLUSIONS ET RECOMMANDATIONS

On peut dégager des analyses décrites dans ce rapport les conclusions suivantes :

- L'analyse des régressions multiples faite avec les données actuelles fournies ne permet pas la prédiction du taux de bris par tronçon en fonction des variables indépendantes quantitatives telles que la pression, l'âge de la conduite, le diamètre et la longueur du tronçon. Bien qu'ils ne soient pas présentés en détail dans ce rapport, les essais préliminaires limités à un échantillon constitué de tronçons avec bris non nuls n'ont pas été plus concluants. Cette vérification pourra être reprise en détail lors d'une prochaine étape du projet.

- La longueur du tronçon a été traitée comme une variable explicative dans l'analyse de régression. Il serait préférable que les tronçons dans la base de données soient de longueurs similaires, afin d'éviter le biais imposé par cette variable.

- L'analyse factorielle des correspondances (AFC) est un outil puissant d'analyse des données qui a permis d'identifier, à partir d'un échantillon représentatif de tous les tronçons de la ville de Montréal, la structure de la base de données et de faire ressortir des relations entre les différentes composantes de la banque de données.

- En combinant une analyse de classification aux résultats de l'AFC, il est possible d'identifier trois à quatre grandes classes de tronçons correspondant à des régions homogènes par rapport au taux de réparation de bris.

- Il est possible de caractériser chacune des zones homogènes en fonction des différentes variables considérées dans l'étude. La période de construction et le type de sol apparaissent comme les variables intervenant en premier lieu dans l'identification des zones homogènes.

- Les méthodes multivariées peuvent permettre d'identifier des sous-groupes correspondant à des régions homogènes. Il est cependant suggéré de valider ces groupes avec les résultats obtenus par une autre méthode comme l'analyse de variance.

- Dans l'hypothèse où on serait intéressé à identifier des zones homogènes à partir des seuls tronçons de conduite qui ont subi un bris, il serait opportun d'appliquer la même méthodologie (AFC suivie d'une classification multivariée sur les facteurs)

pour redéfinir la structure des données relatives aux bris et les interrelations globales entre les composantes qui pourraient modifier passablement la délimitation des zones homogènes.

- L'analyse de variance (ANOVA) permet de valider les regroupements à partir d'un critère objectif quantifiable, soit la différence statistiquement significative des taux de bris moyens. Il sera donc possible de tenter de reconstituer les groupes existants ou de modifier ces groupes à la lueur des résultats de l'ANOVA et les résultats de la classification par méthode multivariée.

- L'analyse de variance (ANOVA) permet de valider les regroupements à partir d'un critère objectif quantifiable, soit la différence statistiquement significative des taux de bris moyens. Il sera donc possible de tenter de reconstituer les groupes existants ou de modifier ces groupes à la lueur des résultats de l'ANOVA et les résultats de la classification par méthode multivariée.

- Au niveau des trois ou quatre grandes classes obtenues par classification automatique, il serait préférable d'accorder la priorité à une méthode multivariée plutôt qu'univariée, car cette dernière tient compte d'un ensemble d'interrelations à un niveau global. Par contre, au niveau plus fin (>4 classes), la délimitation des zones homogènes devient une question d'interprétation et de jugement (basé sur l'expérience et la connaissance du milieu) de l'ensemble des résultats des deux méthodes.

- Les courbes exponentielles de prédiction de taux de bris peuvent être calées avec un ou deux paramètres. Bien qu'un de ces coefficients ($N(t_o)$) puisse avoir une signification physique, les calculs réalisés par Kleiner (1994) à ce chapitre ne sont pas convaincants. On peut décider de caler ces courbes avec deux paramètres si, comme on s'y attend, l'ajout d'un paramètre permet d'expliquer plus de variabilité du taux de bris. Un critère objectif suggéré pour vérifier si cet objectif est atteint, est le calcul du coefficient de détermination ajusté.

- Le calcul des intervalles de confiance pour les courbes non linéaires de prédiction de taux de bris ne peut pas être fait directement. Il faut d'abord linéariser la régression. Cette étape nécessite que les valeurs nulles soient traitées séparément. Une approche probabiliste a été préconisée dans ce rapport, permettant ainsi le traitement des données non nulles, connaissant leur probabilité d'occurrence conditionnelle à $N(t) > 0$. Les courbes exponentielles de prédictions devraient alors

être développées avec les données historiques sur les tronçons avec un taux de bris > 0.

- Étant donné l'approche préconisée, les groupes qui seront ultimement suggérés devraient regrouper les conduites ayant un comportement similaire dans la même catégorie. La courbe exponentielle qui décrit ce groupe devrait s'appliquer à l'ensemble des tronçons de la catégorie, sans discrimination. L'incertitude associée à cette régression peut être quantifiée par l'intervalle de confiance. Ainsi, une approche plus conservatrice pour les conduites plus susceptibles aux bris pourrait consister à utiliser la limite supérieure de l'intervalle de confiance pour l'extrapolation du taux de bris.

- Il n'est cependant pas recommandé de tenter de transférer l'information de la courbe principale pour créer une courbe pour chaque tronçon, puisque cela serait contraire à la logique qui a permis la constitution de groupes homogènes et qu'il n'y a pas de fondement logique mathématique pour procéder à ce transfert. Une approche de type échantillonnage aléatoire stratifié à répartition optimale pourrait déterminer la proportion de tronçons à remplacer dans chaque groupe en fonction de la variabilité du taux de bris de l'effectif du groupe et des coûts de remplacement. Une fois cette proportion établie, on pourrait établir des priorités de remplacement à l'intérieur des groupes basés sur des critères physiques ou économiques.

L'ensemble des analyses statistiques décrites dans ce rapport a permis de conclure que la base de données pourrait contenir plus de variables discriminantes. Par exemple, il se peut que d'autres informations, telles que l'agressivité du sol, ou l'importance relative du secteur soient des variables qui puissent mieux expliquer la variance du taux de bris.

Une autre approche pourrait consister à reprendre l'analyse sous un angle probabiliste. On pourrait calculer la probabilité d'un bris de tronçon conditionnelle à un bris antérieur, par exemple. Ces pistes de solutions potentielles pourraient être examinées dans une phase subséquente de ce projet.

9.0 RÉFÉRENCES

Kleiner (1998). The conditions of the City of Montreal water distribution system. NRC/IRC paper 3/98. 26 p.

Lebart, L., A. Morineau et K. Warwick (1984). Multivariate Descriptive Statistical Analysis. John Wiley & Sons, 231 p.

Levene, H. (1960). Robust tests for the equality of variances, in Contributions to Probability and Statistics, Eds. I. Olkin & Paolo Alto, Stanford University Press : 278-292.

Pfaffenberger , R. et J.H. Patterson (1987). Statistical Methods for Business and Economics. 3rd edition, Irwin publishers, Illinois, 1246 p.

Statsoft (1994) Statistica for Windows. Vol 1-3.

ANNEXE A

Tableau des résultats de l'AFC

STAT. Column Coordinates and Contributions to Inertia

CORRESP. Input Table (Rows x Columns): 701 x 24

ANALYSIS Standardization: Row and column profiles

Column Name	Column Number	Coordin. Dim.1	Coordin. Dim.2	Mass	Quality	Relative Inertia	Inertia Dim.1
DIA_P6	1	-,09787	,05724	,014895	,001776	,049926	,000615
DIA_8	2	,01258	-,18600	,081833	,058878	,022370	,000056
DIA_10	3	,75789	-,82405	,008827	,087706	,058422	,021861
DIA_12	4	,22329	,84311	,023354	,160209	,051355	,005021
BRM_DIAD	5	,07501	-,03597	,132070	,260947	,001622	,003204
A49ETMOI	6	,84127	-,24312	,069880	,770649	,032202	,213240
A49_69	7	-1,13658	,09321	,043951	,810007	,032679	,244799
A70ETPLU	8	,17523	,72556	,015079	,070338	,055316	,001996
BRMOYPER	9	-,87145	,10951	,084060	,850376	,035315	,275246
AR	10	,26461	,86573	,035675	,304491	,044467	,010771
TI	11	-,36758	,10447	,033469	,052626	,043010	,019498
SA	12	,50360	-,13312	,038802	,114235	,042681	,042429
TO	13	-1,75147	,04421	,002575	,077131	,047450	,034052
RO	14	-,03764	-1,70109	,018389	,447670	,055076	,000112
BRMOYSOL	15	,07310	,04987	,129994	,192209	,002453	,002995
PSI30_39	16	2,30257	2,04513	,004046	,262700	,067644	,092483
PSI40_49	17	-,03987	-1,31623	,005701	,075404	,060713	,000039
PSI50_59	18	-,08073	-,96464	,030159	,277665	,047135	,000848
PSI60_69	19	,02199	-,15273	,033653	,008719	,042563	,000070
PSI70_79	20	-,06479	,67195	,037514	,183432	,043163	,000679
PSI80_89	21	,09240	,39153	,013240	,018669	,053154	,000487
PSI90_P	22	1,15075	,24321	,004597	,049356	,059675	,026249
BRMOYPSI	23	,06994	-,04268	,128860	,159067	,002519	,002718
BRIS_ABS	24	-,11462	,25259	,009379	,006807	,049091	,000531

STAT. Column Coordinates and Contributions to Inertia

CORRESP. Input Table (Rows x Columns): 701 x 24

ANALYSIS Standardization: Row and column profiles

Column Name	Cosine² Dim.1	Inertia Dim.2	Cosine² Dim.2
DIA_P6	,001323	,000249	,000453
DIA_8	,000268	,014433	,058610
DIA_10	,040191	,030557	,047515
DIA_12	,010501	,084632	,149708
BRM_DIAD	,212159	,000871	,048788
A49ETMOI	,711249	,021057	,059400
A49_69	,804595	,001947	,005411
A70ETPLU	,003876	,040469	,066461
BRMOYPER	,837157	,005139	,013219
AR	,026016	,136311	,278475
TI	,048693	,001862	,003933
SA	,106774	,003505	,007461
TO	,077082	,000026	,000049
RO	,000219	,271279	,447451
BRMOYSOL	,131166	,001648	,061044
PSI30_39	,146851	,086264	,115849
PSI40_49	,000069	,050348	,075334
PSI50_59	,001931	,143067	,275734
PSI60_69	,000177	,004002	,008542
PSI70_79	,001689	,086352	,181743
PSI80_89	,000985	,010347	,017684
PSI90_P	,047246	,001386	,002110
BRMOYPSI	,115905	,001197	,043161