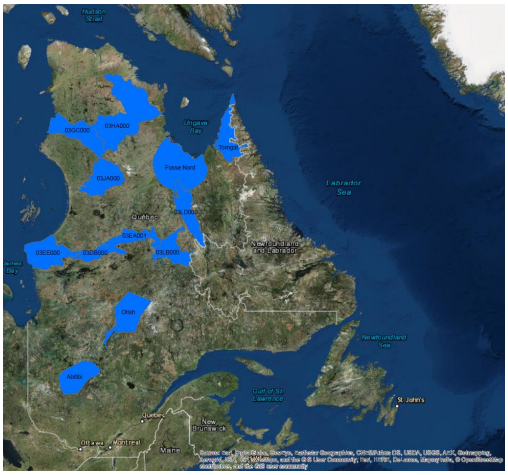

ANALYSE MULTIVARIÉE DES PARAMÈTRES MORPHOLOGIQUES DES LACS DE LA PROVINCE DE QUÉBEC

Par

Olivier Chimi Chiadjeu

Yves Gratton

André St-Hilaire

RAPPORT PRÉSENTÉ AU MINISTERE DES FORÊTS, DE LA FAUNE ET DES PARCS

OCTOBRE 2016

INRS CENTRE EAU TERRE ENVIRONNEMENT
UNIVERSITÉ DE RECHERCHE

Le rapport doit être cité comme suit.

Chimi Chiadjeu, O., Y. Gratton et A. St-Hilaire, 2016. Analyse multivariée des paramètres morphologiques des lacs de la province du Québec. Rapport No R1684, INRS-ETE, Québec (QC): vi + 23 p.

© 2016, Institut national de la recherche scientifique (INRS)

ISBN : 978-2-89146-875-6

RÉSUMÉ

Ce rapport décrit une classification statistique des lacs de la province du Québec à partir de paramètres morphologiques. Cette classification est effectuée dans le cadre du projet d'étude de l'habitat thermique des salmonidés dans les lacs de la province de Québec dans le contexte d'un réchauffement climatique appréhendé. Les études menées à l'aide du modèle MyLAKE sur quelques lacs montrent qu'une augmentation de la température entraînera une diminution de l'habitat thermique des salmonidés. L'utilisation de ce modèle nécessite un étalonnage et par conséquent des observations des lacs auxquels on s'intéresse. L'indisponibilité des observations rend donc impossible la généralisation de cette étude à l'ensemble des lacs de la province de Québec. Pour cette raison, nous avons tenté de regrouper un sous-ensemble des lacs du Québec en un nombre optimal de classes distinctes. Cette étude exploratoire nous a permis d'identifier les types de lacs de la province les plus fréquents et nous permettra d'orienter les choix des lacs à instrumenter. L'étalonnage du modèle effectué sur les lacs types de chaque classe permettra d'estimer le profil des températures de l'ensemble des lacs de la classe et indirectement les caractéristiques de l'habitat thermique pour cette classe. Disposant des données morphologiques (superficie, profondeur moyenne, altitude et développement de rivage) de centaines de milliers de lacs extraits d'un sous-ensemble de douze Unités de Découpage Hydrographique (UDH), nous démontrons qu'il est possible de regrouper les lacs du Québec en onze classes distinctes. Cette classification nous a permis d'identifier les lacs types présents dans les UDH. À l'aide d'un arbre de prédiction, nous démontrons qu'il est possible de prédire l'appartenance d'un lac à une classe à partir des paramètres morphologiques du lac présents dans la base de données. Ceci nous a conduits à identifier la médiane des paramètres de chaque classe comme la métrique qui caractérise le mieux la classe.

TABLE DES MATIÈRES

RÉSUMÉ III

TABLE DES MATIÈRES IV

LISTE DES TABLEAUX V

LISTE DES FIGURES VI

1. INTRODUCTION 1

1.1 CONTEXTE D'ÉTUDE 1
1.2 OBJECTIFS 2

2. MÉTHODOLOGIE 3

2.1 MÉTHODE DE CLASSIFICATION MISE EN ŒUVRE 3
2.2 ANALYSE EN COMPOSANTES PRINCIPALES 3
2.3 ANALYSE CANONIQUE DE CORRESPONDANCE 4
2.4 CLASSIFICATION PAR HIÉRARCHIE 5
2.5 VALIDATION 6

3. RÉSULTATS ET ANALYSES 7

3.1 DONNÉES 7
3.2 ANALYSE PRÉLIMINAIRE 8
3.3 ANALYSE DE CORRÉLATION CANONIQUE 11
3.4 ANALYSE EN COMPOSANTES PRINCIPALES 13
3.5 CLASSIFICATION ASCENDANTE PAR HIÉRARCHIE 14
3.6 VALIDATION DE LA CLASSIFICATION 20

4. BILAN ET PERSPECTIVE 22

5. RÉFÉRENCES 24

LISTE DES TABLEAUX

TABLEAU 1. NOMBRE INITIAL DE LACS ET NOMBRE DE LACS SÉLECTIONNÉS POUR CHACUNE DES DOUZE ZONES D'ÉTUDE 10
TABLEAU 2. STATISTIQUES SUR LES QUATRE VARIABLES MORPHOLOGIQUES : SUPERFICIE, PROFONDEUR MOYENNE, DÉVELOPPEMENT DU RIVAGE (DL) ET ALTITUDE. 10
TABLEAU 3. MÉDIANES DES PARAMÈTRES MORPHOLOGIQUES POUR CHAQUE CLASSE. 16

LISTE DES FIGURES

FIGURE 1. LOCALISATION DES DOUZE UNITÉS DE DÉCOUPAGE HYDROGRAPHIQUE (UDH)..... 7
FIGURE 2. SUPERFICIE DES LACS EN FONCTION DE LA PROFONDEUR MOYENNE. 9
FIGURE 3. COEFFICIENTS DE CORRÉLATION ENTRE LES VARIABLES MORPHOLOGIQUES ET MÉTÉOROLOGIQUES. 11
FIGURE 4. CORRÉLATIONS CANONIQUES AFFECTÉES AUX PRINCIPALES DIMENSIONS..... 12
FIGURE 5. PROJECTION DES VARIABLES SUR LES DEUX PRINCIPALES DIMENSIONS..... 12
FIGURE 6. INERTIE DES COMPOSANTES PRINCIPALES ET COURBE DE L'INERTIE CUMULÉE (BLEU)..... 13
FIGURE 7. REPRÉSENTATION DES INDIVIDUS SUR LES PRINCIPALES COMPOSANTES. 14
FIGURE 8. ÉVOLUTION DES CRITÈRES Q (PANNEAUX DU HAUT) ET SPRSQ (PANNEAUX DU BAS) POUR LES INERTIES INTRA-CLASSE (COLONNE DE GAUCHE) ET INTER-CLASSES (COLONNE DE DROITE), EN FONCTION DU NOMBRE DE CLASSES SÉLECTIONNÉES. 15
FIGURE 9. PROJECTION DES CLASSES SUR LES COMPOSANTES PRINCIPALES..... 16
FIGURE 10. DENDROGRAMME. LES ABSCISSES SONT LES NUMÉROS DES CLASSES TANDIS QUE LES ORDONNÉES SONT LES DISTANCES ENTRE LES GRAPPES BASÉES SUR LA MÉTHODE DE WARD [9]..... 17
FIGURE 11. POURCENTAGE DE LACS DANS CHAQUE CLASSE. 17
FIGURE 12. POURCENTAGE DES CLASSES DANS LES DOUZE ZONES D'ÉTUDE. 18
FIGURE 13. BOÎTES À MOUSTACHE DES PARAMÈTRES MORPHOLOGIQUES DE CHAQUE CLASSE. 19
FIGURE 14. VALIDATION INTRINSÈQUE DE LA CLASSIFICATION. 21

1. INTRODUCTION

Des études ont montré qu'une augmentation de la température ambiante entraînera une augmentation de la température de l'eau aussi bien dans les océans que dans les lacs [1] [2] [3]. Qu'adviendra-t-il alors des populations de certaines régions pour qui les poissons sont leur principal**e** source d'alimentation? La synthèse rédigée par D. J. Marcogliese (2001) souligne l'impact des changements climatiques sur l'écosystème aquatique d'eau douce et d'eau marine [4]. Pour assurer la pérennité des espèces en milieu aquatique, il faut s'assurer non seulement de la présence des éléments nutritifs et d'une qualité adéquate de l'eau, mais aussi d'une concentration d'oxygène dissous et d'une température de l'eau acceptables. Il est clair que l'augmentation de la température de l'eau ne peut pas être l'unique cause de la disparition d'une espèce mais elle demeure un facteur important. Une étude exhaustive de l'impact des changements de la température de l'eau sur une espèce devrait, en principe, prendre en compte l'environnement ainsi que l'interaction entre les différentes espèces présentes dans l'environnement auquel on s'intéresse. Cependant, une augmentation de la température ambiante pourrait entrainer la diminution, voir la disparition de certains habitats thermiques dans les lacs de la province du Québec [6].

1.1 Contexte d'étude

Ce projet s'inscrit dans le cadre des études effectuées à l'INRS-ETE dans le but de cartographier les habitats thermiques des salmonidés des lacs de la province de Québec, dans un contexte de réchauffement climatique. La simulation des profils verticaux de température des lacs est effectuée à l'aide du modèle numérique MyLAKE [7, 8]. Ce modèle prend en entrée les variables météorologiques et tient compte des paramètres morphologique**s** des lacs. La température de l'eau observée est utilisée pour l'étalonnage du modèle. Un nouvel étalonnage est nécessaire pour chaque lac d'intérêt. Cette phase d'étalonnage est très importante car elle permet d'ajuster les paramètres constant**s** du modèle afin qu'il puisse reproduire les températures de l'eau observées dans le passé. Le modèle est ensuite forcé à l'aide de scénarios climatiques

futurs afin de prédire les habitats thermiques dans le futur. Un regroupement ou une classification des lacs du Québec nous permettrait d'orienter la sélection des lacs types à instrumenter afin de minimiser les coûts et de mieux prédire les habitats thermiques futurs.

La classification statistique est une méthode d'analyse de données qui permet de regrouper des données complexes en un nombre restreint de classes simples en fonction de leurs caractéristiques ou en fonction d'indices de dissimilarité prédéfinis. Une classe est alors un ensemble d'objets qui ont les mêmes caractéristiques. En analyse de données, les catégories de méthodes de classification couramment utilisées sont : les classifications supervisées (ex. K-moyen) et non supervisées (ex. classification par arbre de décision) ainsi que les classifications non paramétriques (ex. Machine à vecteur de support) et paramétriques (ex. régression logistique) [6]. Le choix de la méthode de classification est fortement dépendant de l'application [9]. Les méthodes de classification par partitionnement sont des méthodes qui nécessitent une initialisation du nombre de classes et/ou de leurs centres (de masse) qui peuvent varier pendant la classification. Cependant, il faut fixer *a priori* le nombre de classes et choisir les centres initiaux des classes. Cette méthode dépend donc fortement du choix du centre des classes et du nombre de classes. Bien qu'elle soit bien adaptée à des données numériques, elle est moins flexible que la classification par arbre de décision (par hiérarchie). Les principales limites de la méthode de classification hiérarchique sont le coût en temps de calcul et l'impossibilité d'optimiser les classes après la classification. Les méthodes mixtes regroupent à la fois des méthodes par partitionnement et des méthodes hiérarchiques, et parfois aussi des méthodes d'analyses multivariées, afin de pallier aux carences liées à l'utilisation indépendante de chacune des différentes méthodes.

1.2 Objectifs

L'objectif de cette étude est de proposer une classification des lacs pour les différentes zones géographiques de la province de Québec à partir des données morphologiques disponibles dans les banques de données gouvernementales. Pour chaque classe que nous définirons, nous proposerons des paramètres permettant de la caractériser. Nous déterminerons les lacs types de la province et les lacs types des différentes sous-régions.

2. MÉTHODOLOGIE

2.1 Méthode de classification mise en œuvre

Comme nous l'avons souligné en introduction, il existe plusieurs méthodes de classification. Pour un meilleur résultat et afin de pallier aux inconvénients liés à l'utilisation d'une seule méthode, les études récentes encouragent l'utilisation des méthodes dites mixtes. Une méthode mixte est une méthode qui regroupe plusieurs algorithmes de classification. Dans cette étude, l'algorithme de classification comprend trois analyses différentes:

1- Analyse préliminaire des données, incluant une Analyse Canonique de Correspondance (ACC) [9]
2- Analyse en composantes principales (ACP) [9, 10]
3- Classification Ascendante Hiérarchique (CAH) [11]

À la première étape, nous avons mené une analyse préliminaire afin d'extraire les valeurs extrêmes des paramètres morphologiques. Cette étape nous a permis de fixer les seuils pour les différentes variables caractérisant les lacs. Au cours de cette étape, à l'aide d'une analyse canonique de correspondance (ACC), nous avons vérifié la pertinence de l'utilisation des variables météorologiques pour la classification des lacs. La seconde étape nous a permis non seulement de réduire la dimension des données, mais aussi de définir un sous-espace dans lequel nos données seront facilement représentées. En plus de contenir les informations contenues dans l'espace des données initiales, les variables projetées sur ce sous-espace sont faiblement corrélées. La troisième étape permet d'effectuer la classification Ascendante Hiérarchique proprement dite. Nous avons effectué la CAH sur les variables projetées dans le sous-espace obtenu après l'ACP. Dans la suite**,** nous allons décrire les différentes étapes de l'algorithme de classification mise en œuvre.

2.2 Analyse en Composantes Principales

Il existe plusieurs livres et articles qui traitent de l'ACP. Nous ne nous attarderons pas ici

sur les équations mathématiques de l'ACP. Pour plus d'information sur l'ACP le livre de G. Saporta [9] est suggéré. Lorsqu'on dispose d'une base de données de grande taille et multidimensionnelle (supérieure à trois dimensions), il est difficile, voire impossible, de les représenter ou les visualiser facilement. L'ACP est une méthode d'analyse multivariée qui permet d'obtenir une représentation des données dans un sous-espace de dimensions inférieures à celles de l'espace original. Les axes du sous-espace ainsi formés sont appelés composantes principales. Le sous-espace est déterminé suivant deux principaux critères: la moyenne des carrés des distances entre l'axe de projection et les individus doit être la plus grande possible et les composantes principales doivent être orthogonales et non-corrélées. La variance expliquée est aussi appelée l'inertie expliquée par chaque composante. Ainsi, la première composante est celle qui a la plus grande inertie, i.e. la direction qui explique le plus grand nombre de variables. La deuxième composante est celle qui possède la seconde plus grande inertie et ainsi de suite. Lorsqu'on sélectionne les composantes qui expliquent par exemple 90% de l'inertie, on perd peu ou presque pas d'information, mais on réduit de beaucoup la dimension de l'espace des données.

2.3 Analyse canonique de correspondance

L'analyse canonique de correspondance (voir [9]) permet de montrer qu'il existe des éventuelles corrélations entre deux groupes de variables quantitatives ou bien de vérifier s'ils décrivent un même phénomène. Si ce n'est pas le cas, on peut ignorer l'un des deux groupes de variables. Dans cette étude, le premier groupe est composé des variables morphologiques et le second groupe des variables météorologiques. Le principe de l'analyse est le suivant : à la première étape, on détermine le couple de variables canoniques qui maximise le coefficient de détermination des composantes canoniques des différents groupes sous la contrainte que leur variance soit égale à 1. À la deuxième étape on détermine le couple de variables ayant le second plus important coefficient de détermination et ainsi de suite. À chaque couple de variables canoniques, on associe un coefficient de corrélation canonique. Les variables canoniques à l'intérieur de chaque groupe sont orthogonales. L'analyse est effectuée sur les composantes ayant les corrélations canoniques les plus importantes.

2.4 Classification par Hiérarchie

La classification par hiérarchie peut se faire de façon ascendante ou descendante. Dans cette étude nous allons porter notre attention sur la Classification Ascendance Hiérarchique (CAH) [11]. La CAH appartient au groupe des méthodes issues de l'agrégation. Le principe est le suivant. Initialement, on suppose que chaque individu représente une classe, ensuite on regroupe les individus entre eux en fonction de l'indice de dissimilarité choisi jusqu'à former le nombre de classes souhaité. Bien que coûteux en temps de calcul, cette méthode permet, grâce à la lecture de l'arbre (appelé un dendrogramme), de déterminer le nombre optimal de classes (Nopt). Nopt est ici le nombre de classes qui permet statistiquement de mieux différencier les individus de classes différentes et de mieux regrouper les individus d'une même classe. Le choix du nombre de classes se fait généralement visuellement à partir du dendrogramme ou à l'aide de critères statistiques. Il existe plusieurs critères permettant de déterminer de façon automatique le nombre optimal. Dans cette étude, nous avons utilisé la statistique « Semi-partial R-squared » (SPRSQ) et la proportion de la variance expliquée par les classes (Q) définies par :

$$SPRSQ = \frac{\Delta_{Inertie\ inter}}{(Inertie\ inter + Inertie\ intra)} \quad (1)$$

$$Q = \frac{Inertie\ inter}{Inertie\ inter + Inertie\ intra} \quad (2)$$

$\Delta_{Inertie\ inter}$ est la différence d'inertie inter-classes, i.e. entre les centres de masses de chaque classe et le centre de masse de l'ensemble des données. Les inerties sont définies par :

$$InertIntra = \frac{1}{N}\sum_{i=1}^{Nc}\sum_{j=1}^{n_i} d^2(g_i, x_i) \quad (3)$$

$$InertInter = \frac{1}{N}\sum_{i=1}^{Nc} n_i * d^2(g_i, G) \quad (4)$$

où N est le nombre d'individu, N_c le nombre de classe, n_i est le nombre d'individus dans une classe i, g_i est le centre de la classe i, G le centre de gravité de l'ensemble des individus et d est la distance euclidienne. L'inertie inter-classes représente une mesure de distance entre les classes et l'inertie intra-classe la distance entre les éléments d'une classe et le centre de gravité de la classe. Lorsque le nombre de classes augmente, l'inertie inter-classes augmente et l'inertie intra-

classe diminue. Lorsque le nombre de classes augmente, Q augmente et tend vers 1. Quant à la statistique SPRSQ, dans l'ensemble elle décroit quand le nombre de classes augmente. Le nombre optimal de classes est la valeur pour laquelle on observe la dernière plus grande variation de Q. Si on utilise la SPRSQ, le nombre optimal de classes est le maximum qui suit une faible valeur de SPRSQ.

Parmi les méthodes de regroupement de classes utilisées en classification hiérarchique, nous avons utilisé la méthode de Ward [11]. La méthode de « Ward » est basée sur la minimisation de la variance intra-classe. Ici, un individu appartient à la classe dont la variance du groupe composé de l'individu et des autre individus de la dite classe est minimale. La méthode de « Ward » apparaît comme celle qui maximise l'inertie intra-classe. Dans cette étude on désire obtenir une classification basée sur la similitude les individus et non sur leur proximité. Donc, la méthode de Ward apparaît ici comme celle la mieux adaptée pour notre étude.

2.5 Validation

Dans cette section, nous allons démontrer, à l'aide d'un arbre de prédiction, que la classification proposée permet de prédire l'appartenance d'un lac à sa classe avec peu d'erreur. Une analyse discriminante sera aussi effectuée à l'aide de la métrique déjà utilisée pour caractériser les classes. L'arbre de prédiction sera utilisé pour mesurer la validité intrinsèque des individus par rapport à leur classe et par rapport aux autres classes. Par exemple, nous allons tester l'appartenance des lacs d'une classe (classe 1, par exemple) à elle-même et l'appartenance des lacs des autres classes (classe 2,.., classe 10) à celle-ci. Nous avons comptabilisé les vrais positifs (i.e. un lac appartient à une classe et l'arbre prédit qu'elle appartient à la classe), les faux positifs (i.e. un lac appartient à une classe et l'arbre prédit qu'elle n'appartient pas à la classe), les vrais négatifs (i.e. un lac n'appartient pas à une classe et l'arbre prédit qu'il n'appartient pas à la classe) et les faux négatifs (i.e. un lac n'appartient pas à une classe et l'arbre prédit qu'il appartient à la classe).

3. RÉSULTATS ET ANALYSES

3.1 Données

Les données utilisées ont été fournies par le Ministère des Forêts, de la Faune et des Parcs (MFFP). Ces données correspondent aux lacs situés dans douze zones correspondant à douze Unités de Découpage Hydrographique (UDH) de la province du Québec. La Figure 1 présente les douze zones d'étude.

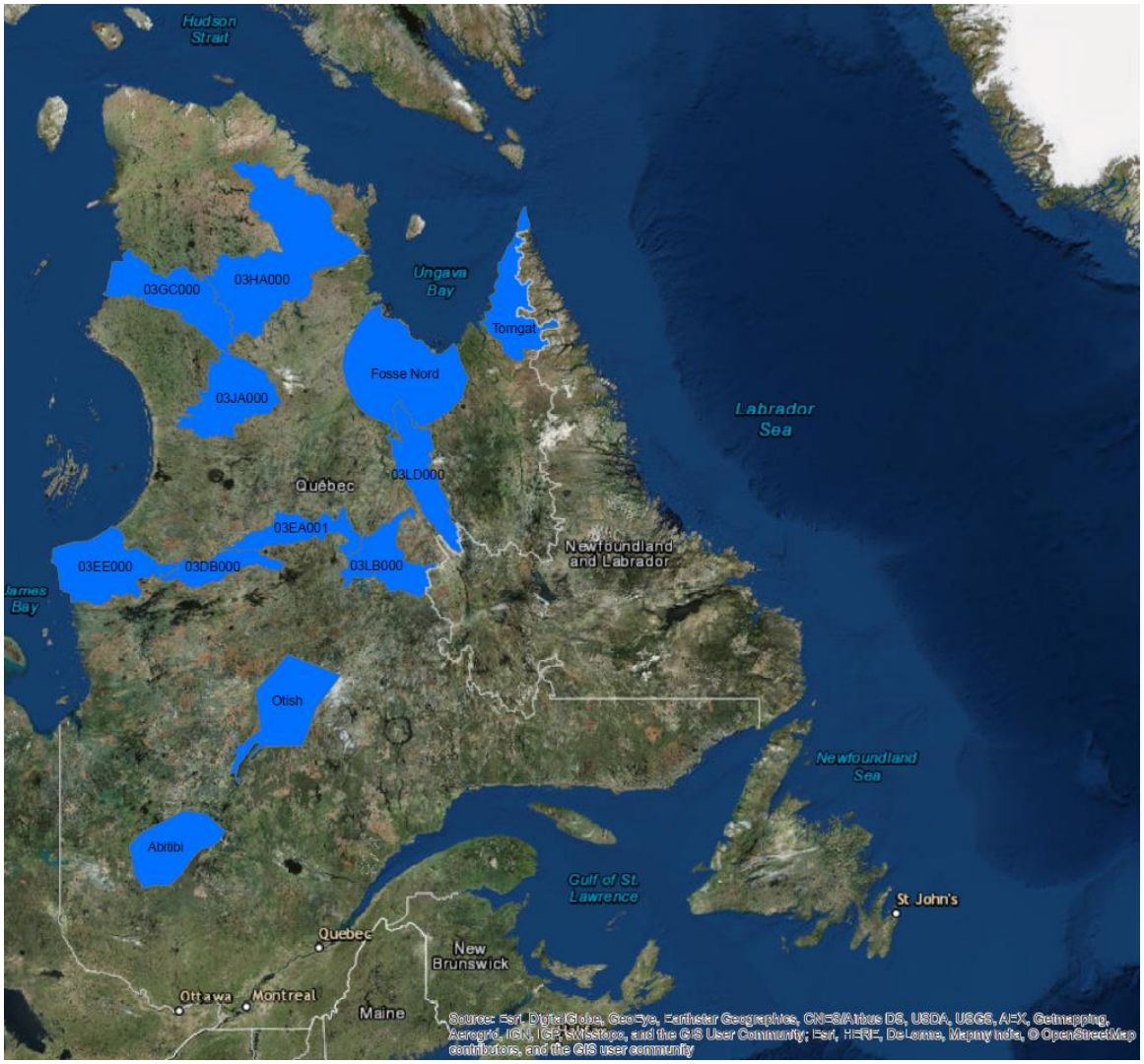

Figure 1. Localisation des douze Unités de Découpage Hydrographique (UDH).

Pour chaque lac, nous possédons les paramètres morphologiques suivants : la superficie du lac, la profondeur moyenne, le développement du littoral et l'altitude. Les superficies (Sup) des lacs sont estimées à partir des méthodes numériques de calcul des surfaces. L'altitude est définie par rapport au niveau de la mer (altitude 0) et au centre du lac. Le développement du littoral (DL) est le rapport entre le périmètre réel et le périmètre hypothétique du lac s'il était un cercle. Il est défini par :

$$DL = \frac{P}{2\sqrt{\pi * Sup}} \tag{5}$$

où P est le périmètre du lac. Quant aux profondeurs moyennes (Zmoy), elles proviennent de la Direction de l'expertise sur la biodiversité du MDDELCC. Elles sont estimées à l'aide de modèles régressifs développés à partir de la bathymétrie connue pour les lacs à relief doux (pente moyenne du bassin versant ≤ 10%) et à relief fort (pente moyenne du bassin versant > 10%). Les calculs reposent sur la superficie du lac (Sup, en km^2), la pente moyenne du bassin versant du plan d'eau (*S*, en %) et l'indice de développement du littoral (DL, en %) de la manière suivante :

$$Zmoy = \begin{cases} 2{,}718 * \left(1{,}431 * \left(0{,}238 * log(Sup)\right) * (0{,}104 * S) * (0{,}177 * DL)\right) * 1{,}086 \;\; si\; S \leq 10\% \\ 2{,}718 * \left(1{,}431 * \left(1{,}789 * log(Sup)\right) * (0{,}049 * S) * (0{,}214 * DL)\right) * 1{,}070 \;\; si\; S > 10\% \end{cases} \tag{6}$$

où la superficie utilisée pour le calcul de *S* est Sup *0,4.

Nous avons extrait, pour chaque lac, la moyenne des variables météorologiques suivantes sur une période allant de 1981 à 2010 : la température maximum (Tmax), la température minimum (Tmin), les précipitations (Precip) et la température moyenne (Tmoy). Nous avons utilisées les données de réanalyses distribuées par Environnement Canada (http://www.cccma.ec.gc.ca/french/data/data.shtml).

3.2 Analyse préliminaire

N'oublions pas que le but de notre classification est le regroupement des lacs qui ont les mêmes paramètres morphologiques et climatiques. Deux paramètres très importants dans le modèle de simulation du profil de température sont la superficie et la profondeur moyenne. Suite aux recommandations faites par le MFFP, nous avons sélectionné les lacs de superficie

supérieure à 10 ha. La Figure 2 représente la superficie de l'ensemble de ces lacs en fonction de leur profondeur moyenne.

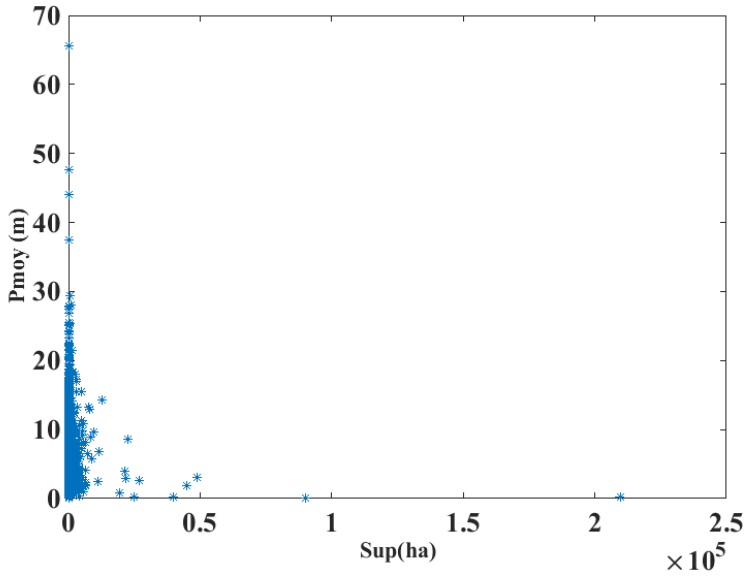

Figure 2. Superficie des lacs en fonction de la profondeur moyenne.

On peut conclure, à partir de cette figure, que les grands lacs ne sont forcément pas les plus profonds et vice versa. Nous avons deux lacs de grande superficie et de profondeur moyenne inférieure à 0,2 m et quatre lacs de profondeur moyenne supérieure à 30 m avec une superficie inférieure à 65 ha. Ces deux groupes de lacs seront considérés comme des classes à part entière, car bien que leurs morphologies ne soient pas représentatives de la morphologie d'un lac, les lacs qu'ils contiennent peuvent être une zone d'habitat dans une période de l'année. Néanmoins, les refuges thermiques dans les lacs ou plans d'eau de profondeurs moyennes inférieures à 0,2 m seront rapidement impactés par une augmentation et une diminution de la température. Ces lacs qu'on peut qualifier ici de particuliers seront retirés avant la classification par seuillage. En somme, nous allons classifier les lacs ayant une superficie comprise entre 10 et $0{,}5 \times 10^5$ ha, et une profondeur moyenne comprise entre 0,04 et 30 m. Dans le Tableau 1, nous avons énuméré les superficies, le nombre initial de lacs, le nombre de lacs sélectionnés et le pourcentage de lacs sélectionnés par zone d'étude. Sur les 621 928 lacs initiaux, nous avons sélectionné 27 846 lacs soit 4,47% des lacs disponibles dans notre base de données initiale. Bien que le pourcentage de lacs sélectionnés par zone soit faible, nous obtenons un nombre suffisant de lacs pour effectuer les analyses avec confiance. Le Tableau 2 présente les statistiques descriptives pour l'ensemble

des lacs utilisés dans l'analyse multivariée. On note une grande disparité en superficie et en altitude.

Numéro de zone	Superficies des zones (km^2)	Nombre initial de lacs	Nombre de lacs sélectionnés	Pourcentage de lacs sélectionnés (%)	Densité (Nombre de lacs sélectionnés / km^2)
Fosse Nord	36319	49854	4994	10,01	0,14
Abitibi	23951	12560	2155	17,15	0,09
Ostish	25718	11365	1019	8,97	0,04
Torngat	14072	45327	1515	3,34	0,11
03DB000	9192	15762	2144	13,60	0,23
03EA001	12414	40418	2349	5,81	0,16
03EE000	23302	39666	3639	9,17	0,16
03GC000	17955	65533	2037	3,10	0,11
03HA000	39353	65529	2174	3,31	0,05
03JA000	18599	167254	3968	2,37	0,21
03LB000	15314	35022	1483	4,23	0,10
03LD000	15912	37811	2005	5,30	0,12
Total		633946	29483	4,65	

Tableau 1. Nombre initial de lacs et nombre de lacs sélectionnés pour chacune des douze zones d'étude

	Moyenne	Médiane	Écart-Type	Minimum	Maximum
Superficie (ha)	85,11	22,80	649,16	10	$4,87x\ 10^4$
DL (%)	2,19	1,95	1,01	1,02	21,00
Zmoy (m)	4,17	4,07	2,31	0,04	29,00
Altitude (m)	27,47	236,21	150,84	0,00	983,00

Tableau 2. Statistiques sur les quatre variables morphologiques : superficie, profondeur moyenne, développement du rivage (DL) et altitude.

3.3 Analyse de Corrélation Canonique

La Figure 3 présente la matrice de corrélation entre les variables morphologiques et les variables météorologiques. On note une corrélation négative entre la latitude et la température maximale, la précipitation et la température moyenne et une corrélation positive entre l'altitude et la précipitation. Comme on pouvait s'y attendre, aux latitudes élevées on a de faibles températures et de faibles précipitations, tandis qu'aux faibles latitudes, les températures sont élevées. Pour des altitudes élevées, de fortes précipitations sont observées alors que pour les faibles altitudes, c'est l'inverse.

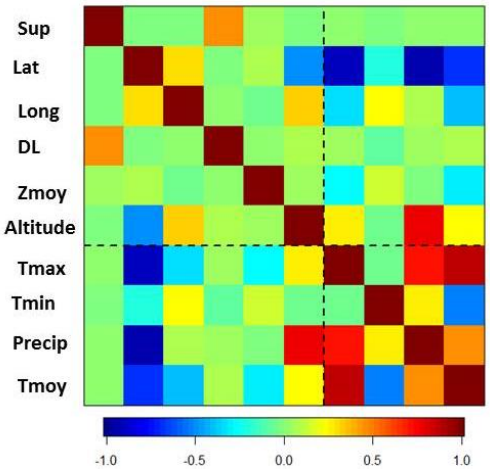

Figure 3. Coefficients de corrélation entre les variables morphologiques et météorologiques.

La Figure 4 représente les corrélations canoniques en ordre décroissant entre les différents couples de variables canoniques (dimensions) qui maximisent le coefficient de détermination. À partir de ce graphique, nous avons sélectionné les trois premières dimensions pour la suite de notre analyse. La Figure 5 représente la projection des variables météorologiques et morphologiques sur les deux premières dimensions. Sur cette figure, deux groupes de variables sont clairement distinguées. La première dimension représente bien la précipitation, la latitude, et la température maximale. C'est encore elle qui oppose la latitude à la précipitation et à la température maximale. Sur la Figure 3 on note qu'il n'y a pas de corrélation entre les variables morphologiques et météorologiques (Profondeur moyenne, Développement du littoral et surface). La Figure 5 présente deux groupes distincts de variables.

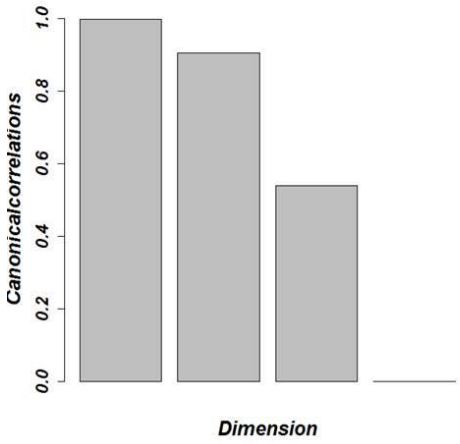

Figure 4. Corrélations canoniques affectées aux principales dimensions.

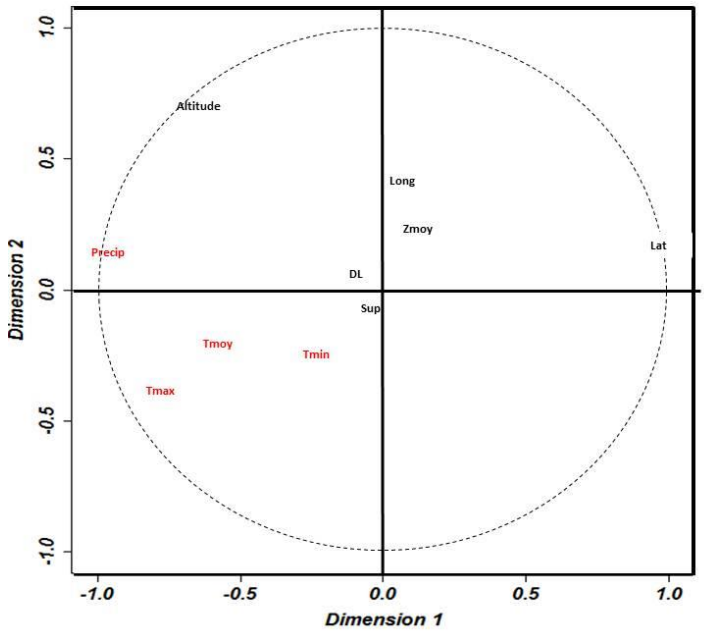

Figure 5. Projection des variables sur les deux principales dimensions.

L'ACC permet de mettre en évidence les corrélations entre deux groupes de variables. Elle permet aussi de décelée d'éventuelle relation entre les deux groupes de variables. Sur la figure 5 on note deux groupes distincts de variable : climatologie et morphologie. Ceci permet de conclure qu'il n'y a pas de relation entre la climatologie et la morphologie donc, la climatologie ne peut remplacer les variables décrivant la morphologiques des lacs. Dans la suite, la classification se fera uniquement avec le groupe des variables morphologiques.

3.4 Analyse en Composantes Principales

La Figure 6 présente les inerties (ou variance) expliquées pour les différentes composantes. Les trois premières composantes regroupent 80% de l'information. L'espace de représentation composé des trois premières composantes sélectionnées est non seulement interprétable visuellement mais, il ne produit pas une déformation trop prononcée des données (i.e. on n'a presque pas de perte d'information).

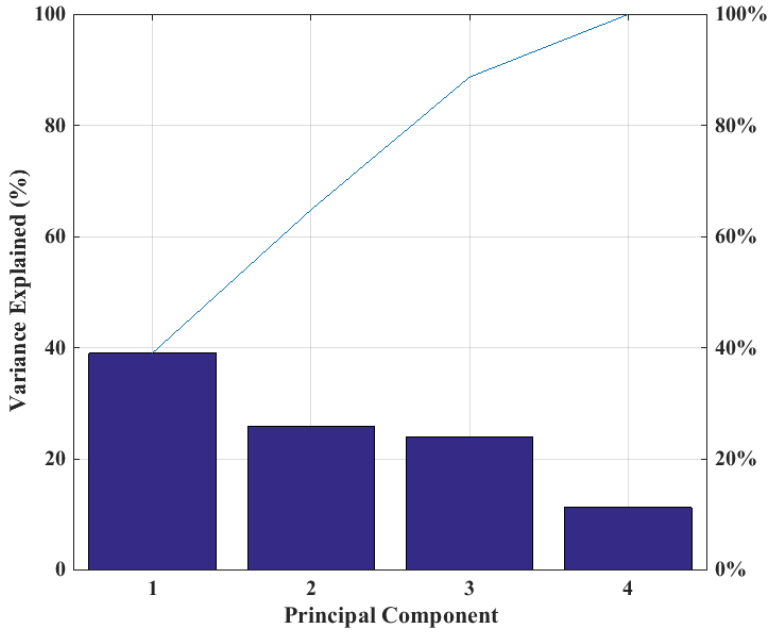

Figure 6. Inertie des composantes principales et courbe de l'inertie cumulée (bleu)

En retenant les trois premières composantes, on réduit la dimension des données tout en conservant plus de 80% de l'information. La Figure 7 présente la projection des données sur les trois principales composantes. La première composante décrit correctement la superficie et le développement de rivage. La deuxième composante décrit la profondeur moyenne et l'altitude. La troisième composante permet de différencier la profondeur moyenne de l'altitude. Il y a une forte corrélation entre la superficie et le développement de rivage et entre l'altitude et la profondeur moyenne. On peut déduire ici que les lacs de faible profondeur sont situés en haute altitude et inversement. Nous ne pouvons pas identifier les différentes classes à l'aide d'uniquement une ACP. Une CAH sera effectuée sur les trois premières composantes qui expliquent 80% de l'inertie des variables.

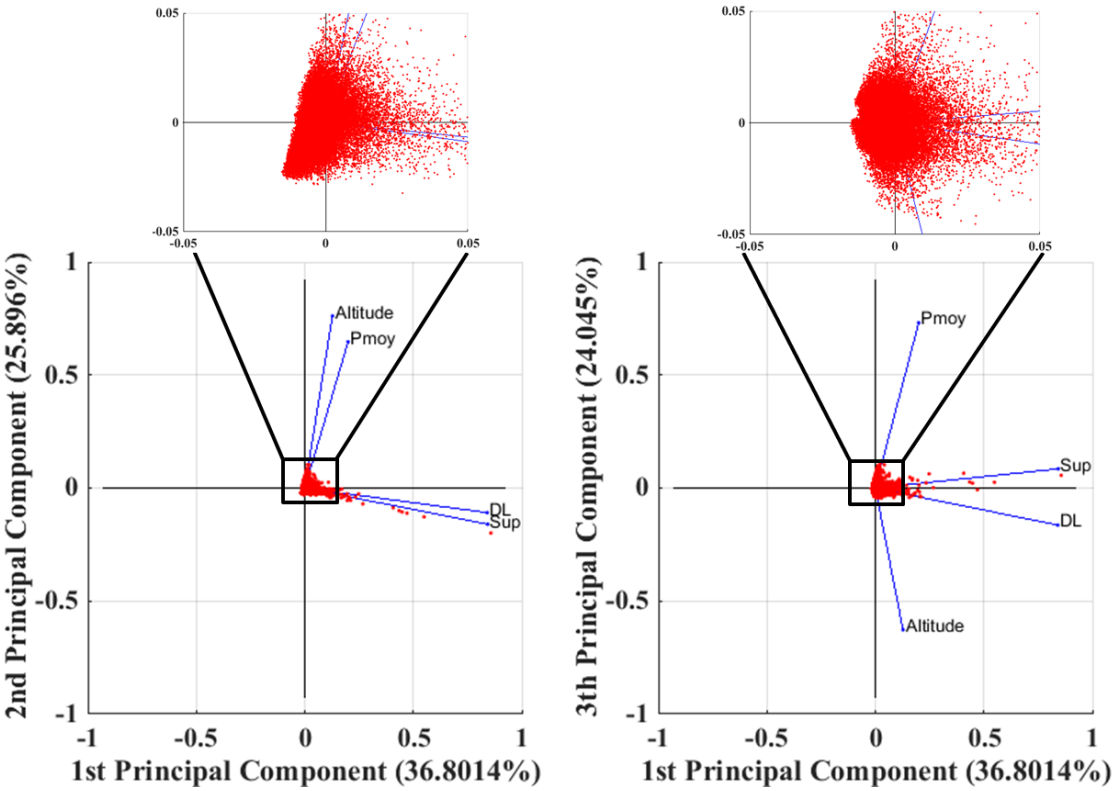

Figure 7. Représentation des individus sur les principales composantes.

3.5 Classification Ascendante par Hiérarchie

À l'aide de l'ACP, on a vu qu'on pouvait représenter nos données dans un espace à trois dimensions sans une déformation trop prononcée. La CAH est effectuée sur les données projetées sur ces dimensions. Comme il a été mentionné plus haut, pour déterminer le nombre optimal de classe (Nopt), nous avons fait varier le nombre de classes à retenir de 2 à 30. Nous avons évalué les grandeurs suivantes : l'inertie inter et intra-classe et calculé Q et SPRSQ. La Figure 8 présente l'ensemble des grandeurs obtenues en fonction du nombre de classes retenues.

Sur la courbe de l'évolution de Q en fonction du nombre de classes (les deux panneaux du haut), on note quatre paliers plus ou moins importants. Chaque palier correspond à un pic du SPRSQ qui est lui-même précédé d'une faible valeur du SPRSQ. Il faut faire un compromis entre le choix des critères et le nombre de classes. Avec onze classes, Q vaut 0,85. Au-delà de onze classes, Q augmente de 0,08 ce qui est peu significatif. La valeur du SPRSQ obtenue pour onze classes correspond au dernier plus grand pic observé sur la Figure 8. Nous fixerons donc le

nombre de classes à onze.

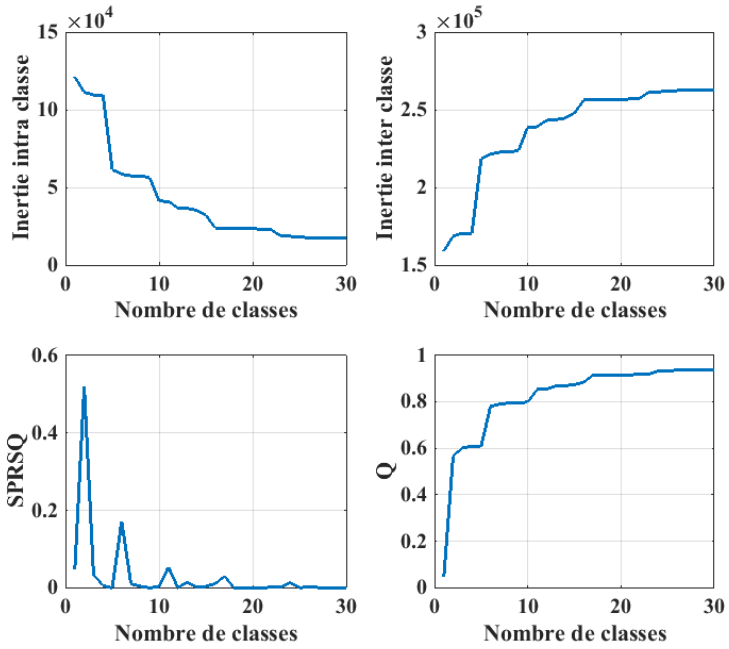

Figure 8. Évolution des critères Q (panneaux du haut) et SPRSQ (panneaux du bas) pour les inerties intra-classe (colonne de gauche) et inter-classes (colonne de droite), en fonction du nombre de classes sélectionnées.

Sur la Figure 9 nous avons projeté les onze différentes classes retenues sur les composantes principales de l'ACP. Dans le Tableau 3, nous présentons les médianes des variables morphologiques de chacune des classes. À l'aide de la projection, on peut déduire que la classe 11 représente les lacs à grands développements du rivage et à grandes superficies. Ensuite suivent les classes 6 et 5. Ces résultats se confirment lorsqu'on observe la médiane des paramètres de chaque classe dans le Tableau 3. En dehors de ces trois classes qui se démarquent par leurs superficies, le découpage des autres classes est basé sur la combinaison des autres paramètres morphologiques.

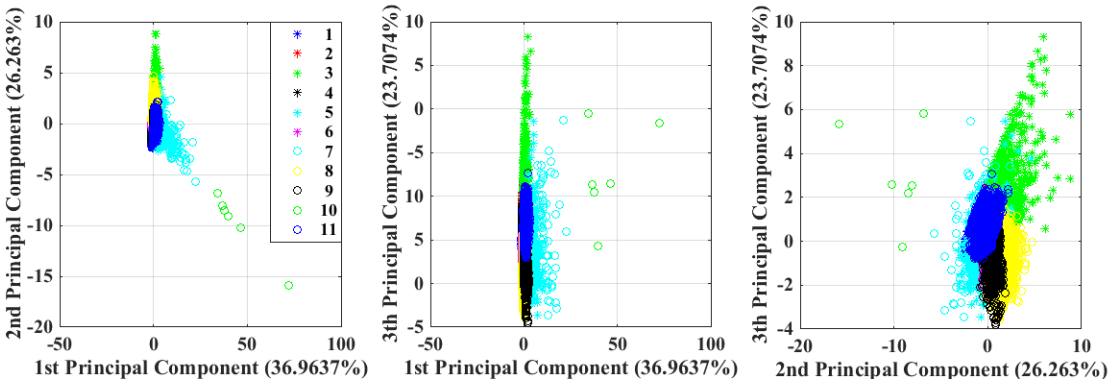

Figure 9. Projection des classes sur les composantes principales.

	Sup (Ha)	DL	Pmoy (m)	Altitude (m)
Classe 1	17,49	1,94	0,85	198,00
Classe 2	27,28	1,75	5,92	203,00
Classe 3	70,50	1,88	10,71	263,76
Classe 4	16,91	1,77	3,87	186,08
Classe 5	364,13	5,17	4,32	237,79
Classe 6	17,47	1,75	3,88	399,90
Classe 7	4 891,58	7,99	5,40	230,00
Classe 8	23,22	1,91	4,72	550,39
Classe 9	42,46	2,89	4,10	407,17
Classe 10	22 199,92	15,21	2,81	164,01
Classe 11	73,99	3,10	4,90	207,00

Tableau 3. Médianes des paramètres morphologiques pour chaque classe.

La Figure 10 représente le dendrogramme obtenue pour les onze classes. À titre d'exemple, parcourant le dendrogramme du haut vers le bas, on peut répartir nos données en deux groupes : le groupe composé des classes {1, 2, 3, 4, 11} et le groupe composé des classes {5, 6, 7, 8, 9, 10}. La Figure 11 présente le pourcentage des lacs dans chaque classe à partir des données disponibles.

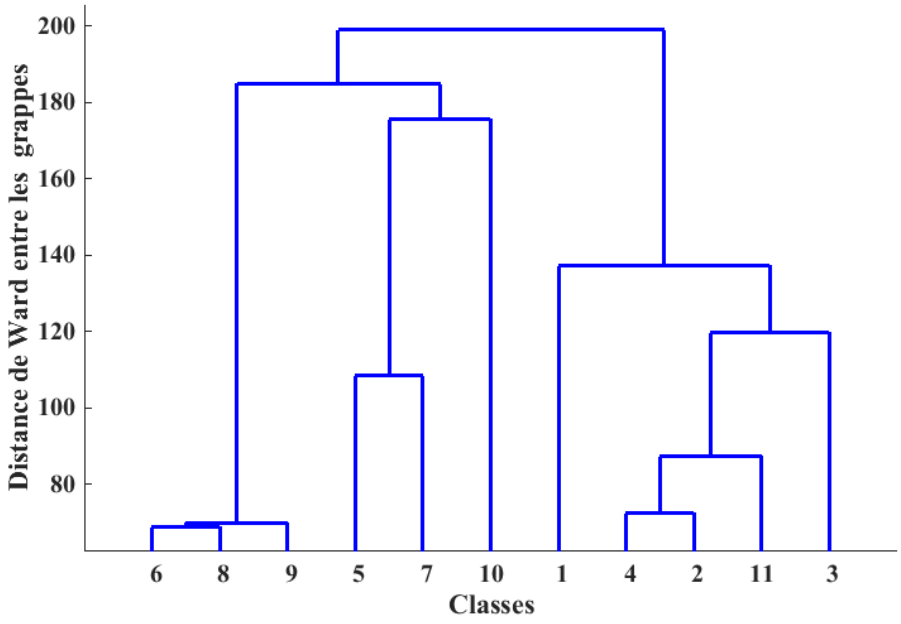

Figure 10. Dendrogramme. Les abscisses sont les numéros des classes tandis que les ordonnées sont les distances entre les grappes basées sur la méthode de Ward [9].

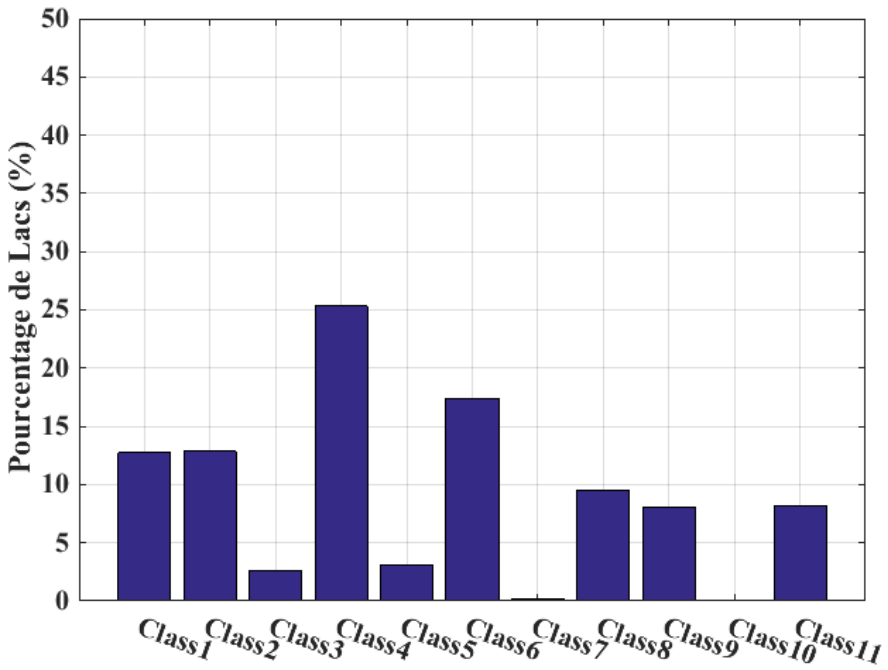

Figure 11. Pourcentage de lacs dans chaque classe.

La classe 4 est la classe la plus représentée sur l'ensemble des données disponibles. Ceci nous permet de conclure que, dans notre base de données, nous avons plus de lacs ayant en médiane une superficie autour de 16,91 ha et une profondeur moyenne autour de 3,87 m. Les grands lacs (classes 7 et 10) sont peu présents. Ils représentent environ 0,22% des lacs de la base de données. Le but principal de cette étude est de déterminer le lac type pour chaque zone d'étude, la Figure 12 présente donc le pourcentage de lacs de chaque classe dans les zones d'étude. Le lac type d'une zone d'étude est défini comme la classe de lacs la plus présente dans la zone d'étude.

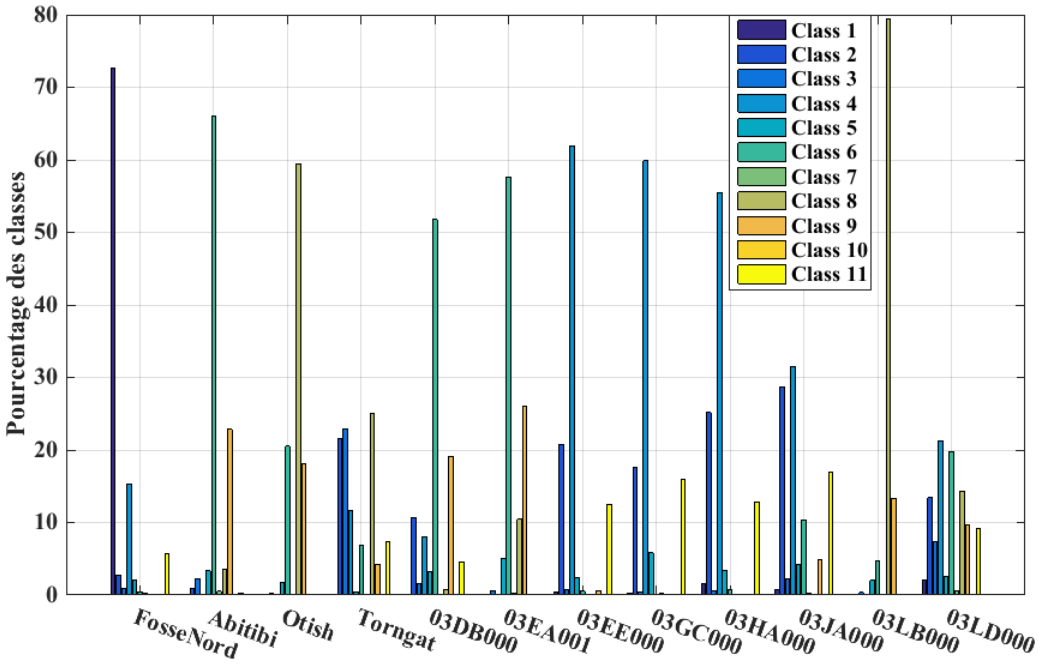

Figure 12. Pourcentage des classes dans les douze zones d'étude.

Par exemple dans la Fosse Nord, la classe 1 représente le lac type. Dans d'autres zones, on peut facilement identifier le lac type, mais pour certaines zones comme la Torngat, la 03JA000 et la 03LD000, on ne peut pas déterminer avec certitude un seul lac type. Dans ces zones, on a deux ou trois types de lacs qui sont plus ou moins également représentés. Sur la Figure 13, nous avons représenté les boîtes à moustache des différents paramètres morphologiques de chacune des classes.

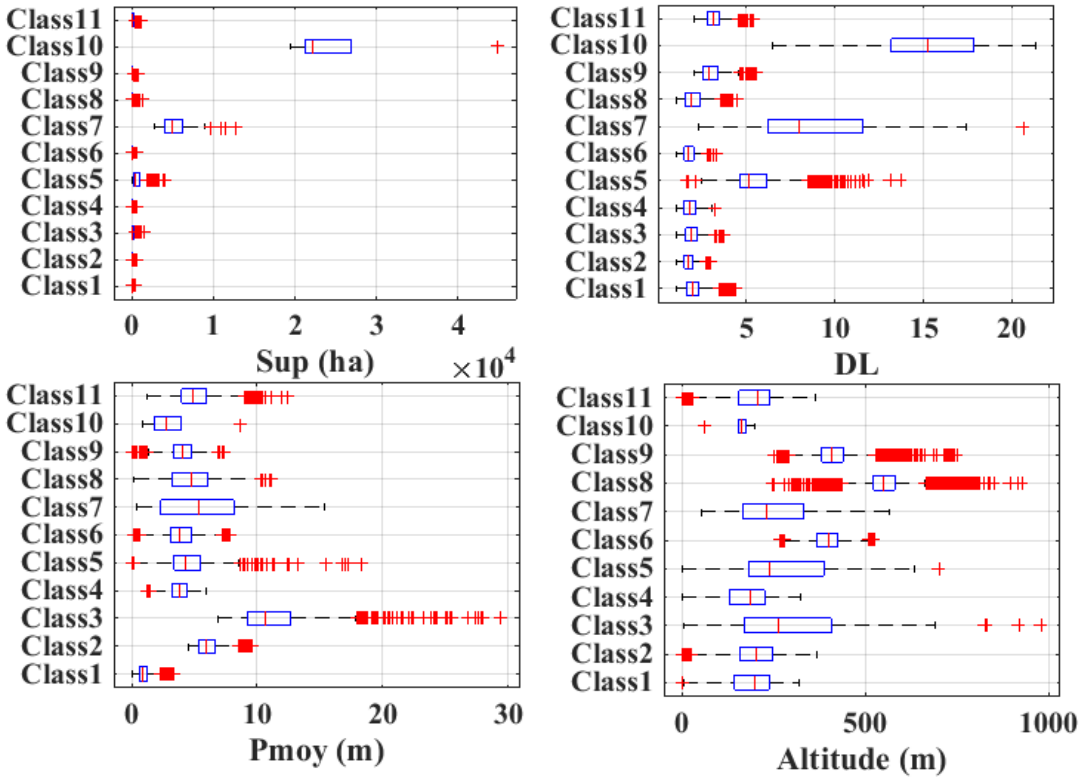

Figure 13. Boîtes à moustache des paramètres morphologiques de chaque classe.

Les paramètres permettant de distinguer les classes sont la superficie, la profondeur moyenne, le développement du rivage et l'altitude. En dehors des lacs des classes 7 et 10 qui ont une grande superficie, la superficie des lacs des autres classes se retrouve presque toujours dans le même intervalle. Ce paramètre permet aussi de distinguer les classes 7 et 10 entre elles et d'avec les autres classes. Les classes 1, 2 et 3 ont des développements de rivage presque identiques mais leurs profondeurs moyennes permettent de les distinguer. Le développement du rivage permet également de distinguer les classes 5, 7 et 10. La profondeur moyenne permet de distinguer la classe 3 des autres classes. Quant aux autres classes, comme on l'a noté pendant l'ACP, une combinaison entre le développement de rivage, la profondeur moyenne et l'altitude permet de les distinguer des autres classes.

3.6 Validation de la classification

L'objectif de la validation est de montrer d'une part que la métrique (médiane) utilisée est représentative des classes et, d'autre part, qu'à partir de notre classification, on peut prédire l'appartenance d'un lac à une classe à partir de ses paramètres morphologiques. La classification nous a permis de labéliser les lacs (d'attribuer à chaque lac une classe). Pour valider la classification, nous avons mené une analyse discriminante. Ceci nous a permis de définir les critères statistiques permettant de décrire et de prédire l'appartenance d'un lac à sa classe. A l'aide du model de prédiction obtenu, nous avons testé l'appartenance des lacs classifiés à leurs propre classe et aux autres classes (validation intrinsèque).

La Figure 14 représente le pourcentage de vrais positifs (i.e. un lac appartient à une classe et l'arbre prédit qu'elle appartient à la classe), les faux positifs (i.e. un lac appartient à une classe et l'arbre prédit qu'elle n'appartient pas à la classe), les vrais négatifs (i.e. un lac n'appartient pas à une classe et l'arbre prédit qu'il n'appartient pas à la classe) et les faux négatifs (i.e. un lac n'appartient pas à une classe et l'arbre prédit qu'il appartient à la classe).

Pour l'ensemble des classes les taux de vrais positifs et vrais négatifs sont nettement supérieurs aux taux de faux positifs et faux négatifs. Supposons qu'on souhaite avoir moins de 20% de taux de faux positifs et de faux négatifs. Soit une classe quelconque : si on suppose que d'affirmer qu'un lac appartient à cette classe alors qu'il n'y appartient pas est plus dangereux que d'affirmer qu'un lac n'appartient pas cette classe alors qu'il y appartient, notre classification répondra alors à nos attentes. Dans le cas contraire, on peut émettre des réserves sur la qualité de classification obtenue car les taux de faux positifs sont supérieurs à 20% pour les classes 2, 3, 9, et 11. Ceci est dû à la forte dispersion de leurs paramètres morphologiques et au nombre élevé de lacs ayant potentiellement des valeurs aberrantes (Figure 13). On pourrait s'attendre aussi à ce que les classes 7 et 10 aient un taux élevé de faux positifs car elles ont une faible population. Pour ces deux classes un seuil sur la superficie suffirait pour les identifier.

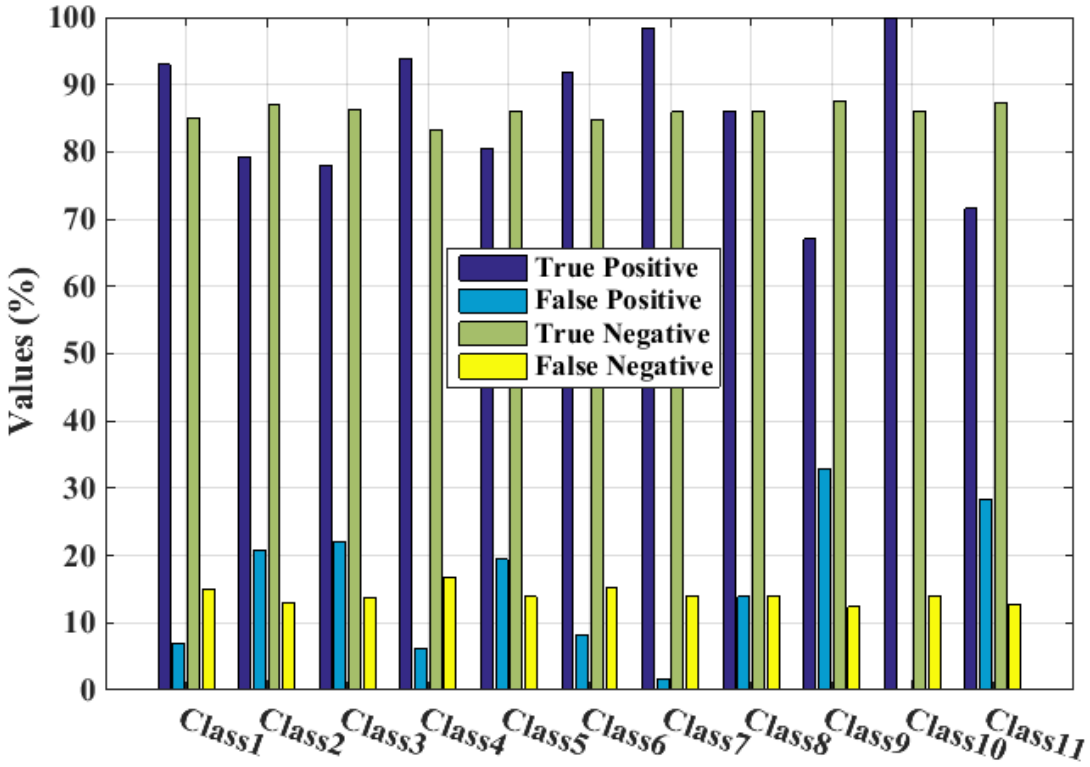

Figure 14. Validation intrinsèque de la classification.

Supposons maintenant que la médiane est la métrique qui caractérise la morphologie d'une classe. La prédiction de l'appartenance des médianes des classes nous donne un taux de vrais positifs et de vrais négatifs de 100% ainsi qu'un taux de faux positifs et de faux négatifs de 0%. Ceci montre statistiquement que la médiane des paramètres morphologiques d'une classe peut servir comme une valeur qui représente la classe.

4. BILAN ET PERSPECTIVE

Une étude préliminaire menée à l'aide d'une analyse par corrélation canonique, nous a permis de conclure qu'il n'y a pas de lien significatif entre les paramètres météorologiques et morphologiques. Ce qui signifie que l'utilisation de données climatiques pour classifier les lacs est à proscrire. Nous avons confirmé que la classification faite sur les paramètres morphologiques donne les mêmes résultats que ceux obtenus avec les paramètres morphologiques et météorologiques. Dans cette étude, nous avons montré qu'une classification morphologique des lacs est possible avec les paramètres suivants : la surface, le développement de rivage, la profondeur moyenne et l'altitude. Une méthode mixte a été utilisée pour la classification. Nous avons combiné une analyse en composantes principales à la classification ascendante par hiérarchie. Cette méthode permet de palier aux lacunes liées à l'utilisation d'une seule méthode de classification. L'analyse en composantes principales nous a permis non seulement de définir un espace de représentation des données avec une faible distorsion mais aussi de réduire la dimension. Disposant d'une base de données sur douze zones d'étude réparties sur le territoire de la province de Québec, nous avons été en mesure de proposer des lacs types pour la toute la province en général et en particulier pour chaque zone d'étude. Après une analyse discriminante, la validation intrinsèque nous a permis de confirmer statistiquement la classification proposée. Elle a aussi permis d'identifier la variable qui représente le mieux les classes : la médiane.

Cette classification pourra aussi être utilisée comme outil pour réduire le nombre de lacs à instrumenter. On pourra identifier les lacs avec les paramètres morphologiques proches de la médiane de chaque classe ou, à l'aide du modèle de prédiction obtenu par l'analyse discriminante. Il est cependant indispensable d'évaluer l'erreur lié à l'utilisation d'un jeu de paramètres car les lacs appartenant à une classe n'ont pas tous les mêmes paramètres morphologiques. Il faudrait donc estimer l'incertitude sur le profil de température modélisé lorsqu'on utilise la médiane comme la métrique qui représente la classe en faisant varier les médianes. Il sera aussi possible utiliser les paramètres morphologiques des autres lacs, toujours avec la même calibration, pour évaluer l'erreur en comparant les profils de température simulées

et observées.

5. RÉFÉRENCES

[1] M. Hondzo et H. Stefan, «Regional water temperature characteristics of lakes subjected to climate change.,» *Climatic Change*, vol. 24, n° %13, pp. 187-211, 1993.

[2] H. Stefan, X. Fang et M. Hondzo, «Simulated climate change effects on year-round water temperatures in temperate zone lakes.,» *Climatic Change*, vol. 40, n° %13-4, pp. 547-576, 1998.

[3] T. Wood, S. Wheny, S. Piccolroaz et S. F. Girdner, «Simulation of deep ventilation in Crater Lake, Oregon, 1951–2099,» US Geological Survey., 2016.

[4] D. J. Marcogliese, «Implications of climate change for parasitism of animals in the aquatic environment.,» *Canadian Journal of Zoology*, vol. 79, n° %18, pp. 1331-1352, 2001.

[5] T. Saloranta, M. Forsius, M. Järvinen et L. Arvola, «). Impacts of projected climate change on the thermodynamics of a shallow and a deep lake in Finland: model simulations and Bayesian uncertainty analysis.,» *Hydrology Research* , vol. 40, n° %12-3, pp. 234-248, 2009.

[6] C. Bélanger, D. Huard, Y. Gratton, D. Jeong, A. St-Hilaire, J. Auclair et I. Laurion, «Impacts des changements climatiques sur l'habitat des salmonidés dans les lacs nordiques du Québec,» Rapport interne R1514, INRS-ETE, Québec (Qc), 167 p., 2013.

[7] T. Saloranta et T. Anderson, «MyLake – a multi-year lake simulation model suitable for uncertainty and sensitivity analysis simulations,» *Ecological Modelling*, vol. 207, n° 11, pp. 45-60, 2007.

[8] T. Saloranta, M. Forsius, M. Järvinen et L. Arvola, « Impacts of projected climate change on the thermodynamics of a shallow and a deep lake in Finland: model simulations and Bayesian uncertainty analysis», *Hydrology Research*, vol. 40, n° 12-3, pp. 234-248, 2009.

[9] G. Saporta, "Probabilités, analyse des données et statistique, Editions Technip, Paris, 656 p., 2011.

[10] M. Nelitz, C. Alexander, K. Wieckowski et P. Council, «Helping Pacific salmon survive the impact of climate change on freshwater habitats: Case Studies», 2007.

[11] J. Ward Jr, «Hierarchical grouping to optimize an objective function.,» *Journal of the American statistical association*, vol. 58, n° 1301, pp. 236--244, 1963.