# Multivariate Analysis of Canadian Water Quality Data

Geneviève Tardif

Thesis submitted to the Faculty of Graduate and Postdoctoral Studies

in partial fulfillment of the requirements for

the degree of Master of Science in Geography

Department of Geography,

University of Ottawa

Ottawa, Ontario, Canada

© Geneviève Tardif, Ottawa, Canada, 2015

## Abstract

Physical-chemical water quality data from lotic water monitoring sites across Canada were integrated into one dataset. Two overlapping matrices of data were analyzed with principal component analysis (PCA) and cluster analysis to uncover structure and patterns in the data. The first matrix (Matrix A) had 107 sites located throughout Canada, and the following water quality parameters: pH, specific conductance (SC), and total phosphorus (TP). The second matrix (Matrix B) included more variables: calcium (Ca), chloride (Cl), total alkalinity (T_ALK), dissolved oxygen (DO), water temperature (WT), pH, SC and TP; for a subset of 42 sites. Landscape characteristics were calculated for each water quality monitoring site and their importance in explaining water quality data was examined through redundancy analysis.

The first principal components in the analyses of Matrix A and B were most correlated with SC, suggesting this parameter is the most representative of water quality variance at the scale of Canada. Overlaying cluster analysis results on PCA information proved an excellent mean to identify the major water characteristics defining each group; mapping cluster analysis group membership provided information on their spatial distribution and was found informative with regards to the probable environmental influences on each group. Redundancy analyses produced significant predictive models of water quality demonstrating that landscape characteristics are determinant factors in water quality at the country scale. The proportion of cropland and the mean annual total precipitation in the drainage area were the landscape variables with the most variance explained.

Assembling a consistent dataset of water quality data from monitoring locations throughout Canada proved difficult due to the unevenness of the monitoring programs in place. It is therefore recommended that a standard for the monitoring of a minimum core set of water quality variable be implemented throughout the country to support future nation-wide analysis of water quality data.

## Résumé

Des données physico-chimiques de qualité de l'eau, provenant d'emplacements de suivi des eaux lotiques situés à travers le Canada, ont été intégrées dans un ensemble de données. Deux matrices de données se chevauchant ont été étudiées à l'aide d'analyses en composantes principales (ACP) et d'analyses typologiques, afin de mettre au jour la présence d'une structure et de répétitions dans la disposition des données. La première matrice comprenait 107 sites établis à travers le Canada, et les paramètres de qualité de l'eau suivants : pH, conductance spécifique (SC), et phosphore total (TP). La deuxième matrice incluait plus de variables : calcium (Ca), chlorure (Cl), alcalinité totale (T_ALK), oxygène dissous (DO), température de l'eau (WT), pH, SC et TP; pour un sous-ensemble de 42 sites. Des caractéristiques paysagères ont été calculées pour chaque site étudié, et leur importance dans l'explication des données de qualité de l'eau a été examinée par des analyses de redondance.

Les premières composantes principales, produites par les analyses ACP sur chacune des matrices, étaient corrélées plus fortement avec la conductance spécifique, suggérant que ce paramètre est le plus représentatif de la variance dans la qualité de l'eau à l'échelle du Canada. La superposition des résultats des analyses typologiques sur les ceux de l'ACP s'est révélée un excellent moyen pour identifier les caractéristiques de qualité de l'eau majeures définissant chaque groupement; la représentation des groupements sur carte a fourni de l'information sur leurs distributions spatiales et s'est avérée instructive en ce qui a trait aux influences environnementales touchant chaque groupe. Les analyses de redondance ont produit des modèles significatifs de prévision de la qualité de l'eau, démontrant que les caractéristiques paysagères constituent des facteurs déterminants de la qualité de l'eau à l'échelle du pays. La proportion de terre cultivée et la moyenne des précipitations totales annuelles dans les bassins versants étaient les variables paysagères expliquant le plus de variance.

Assembler des données provenant de sites de suivi de la qualité de l'eau établis à travers le Canada en un ensemble uniforme s'est avéré difficile à cause de l'hétérogénéité des programmes de suivi en place. Il est donc recommandé qu'un standard soit établi à travers le pays, pour le suivi minimum de

certaines variables clés, afin de supporter les analyses de données de qualité de l'eau à l'échelle nationale.

## Acknowledgments

I would like to thank the individuals from the monitoring agencies that provided the water quality data necessary to this study and answered my questions about the data, specifically, Denis Parent, Environment Canada Water Quality Monitoring and Surveillance; Doreen LeClair, Alberta Environment and Sustainable Resource Development; Elise Watchorn, Manitoba Conservation and Water Stewardship Water Quality Management Section; Aaron Todd, Ontario Ministry of the Environment and Climate Change; Serge Hébert, Québec, Ministère du Développement durable, Environnement, et Lutte contre les changements climatiques; and Pam Minifie, Saskatchewan Water Security Agency. Thanks are extended to individuals from organisations that provided me with spatial datasets used in this study, namely, Judy Kwan, Environment Canada Meteorological Service of Canada; Lisa Koponen and Dr Jon Pasher, Environment Canada Landscape Science and Technology; and Mark Henry, Statistics Canada Environment, Energy, and Transportation Statistics Division.

I am extremely grateful to my thesis director, Professor Konrad Gajewski, for his immense support and guidance throughout the course of this project, and for his leadership, stewardship and mentoring in relation to environmental data analysis and the entire research process. I would also like to acknowledge the other members of my thesis committee, Professor Michael Sawada and Professor Denis Lacelle, for their scientific input, thoughtful questions and expert advice.

Finally, special thanks to my workplace supervisor at Environment Canada, Denis Parent, for his generous support and counsel throughout the duration and production of my thesis, and for having inspired my keen interest in water quality monitoring.

## Contents

## List of Figures

## List of Tables

## 1.0 Introduction

With its unique physical and chemical properties, water is often being referred to as the universal solvent. Water is also omnipresent and in constant circulation in our environment. For these reasons, the quality of fresh-water becomes a most reliable testimony of the nature of the environment where it travelled: from the bedrock, to the soil, its vegetation, and to the human use of land. Around the world, governments monitor the quality of ground and surface water to ensure its suitability for certain uses, such as those identified by the Canadian Council of Ministers of the Environment (CCME) that form the basis upon which Canadian water quality guidelines are established: community water supplies, protection of aquatic life, recreational uses, and agricultural uses (Canadian Council of Ministers of the Environment, 2014a). However, governments also monitor water quality to be able to fulfill research needs, and to promote the continual growth of the body of knowledge regarding the science of water quality interactions with environmental and anthropogenic influences. To that end, large numbers of water quality parameters are monitored regularly, and provide the necessary information for better decision to be made for the protection of water resources.

## 1.1 Problematic

Canada being the second largest country in the world, and extending across ten distinct ecological regions (level I, Commission for Environmental Cooperation, 1997) makes the task of water quality monitoring on its territory a most challenging one. In addition, the responsibility over water quality monitoring in Canada is shared between the federal, provincial and territorial governments, resulting in several distinct water quality programs in place across the country, monitoring for different sets of parameters and at varying frequencies. For these reasons, among others, nation-wide water quality studies are scarce. However, the need for a holistic account of water quality across Canada exists, especially for water resources managers of the federal government at Environment Canada who are attempting to maintain a national monitoring network of

highest quality and delivering water quality information where it is most needed. This thesis is a study of Canadian lotic water quality data, with the main objective to uncover the structure and pattern in a national dataset of water quality data, and with a secondary objective to examine relationships between water quality and landscape characteristics.

Three themes will be explored in a preliminary literature review. The first will review previous studies that have attempted a multivariate analysis of water quality data as a mean to reduce the dimensionality of complex datasets and uncover information. The second theme will relate to spatial analyses, as spatial variation information may provide further insight into water quality data patterns. The third theme will concern water quality in relation to landscape characteristics.

The main portion of this thesis, written in article format, will be a multivariate analysis of the water quality across Canada. To perform this analysis it was necessary to create a dataset from several sources, and the development of this national dataset will be described below. Finally the development and analysis of a database of landscape and climatic characteristics will be described. These are related to the water quality data using other multivariate statistical analyses.

## 1.2 Literature Review

### 1.2.1 Multivariate Data Analyses in Water Quality Datasets

One goal of multivariate data analyses is to summarize complex multi-variables datasets by identifying relationships between variables and cases in the dataset. Essentially, multivariate methods generate hypotheses but do not test them (Rencher, 2002). Several studies were published where multivariate analyses were performed on water quality data from a defined watershed, for example, Bu *et al* (2010), Shrestha and Kazama (2007), Singh *et al* (2004), Varol and Sen (2009), Zhang *et al* (2009), and Zhao and Cui (2008). They are fairly similar in that they all have the objective

to evaluate temporal and spatial variations in water quality data in their respective study basin, and all make use of a selection of the following multivariate methods.

#### *1.2.1.1 Cluster Analysis*

Cluster analysis is a numerical method to uncover structure in the data in such a way that the individuals of the same cluster will resemble each other more than they do those of another cluster (Everitt, 2011). Many clustering methods exist. Hierarchical techniques are perhaps the most popular and are commonly represented by the dendogram. In the hierarchical agglomerative method, the individuals are successively joined to their closest counterpart(s), based on a defined measure of proximity, into a decreasing number of clusters. In the divisive method, the initial cluster contains all individual and it is successively divided into an increasing number of clusters. Several cluster optimizing techniques exist depending on the type of data, for example continuous, binary, structured (Everitt, 2011).

#### *1.2.1.2 Discriminant Analysis*

Discriminant analysis is often compared to cluster analysis, with the difference that in discriminant analysis the number of groups (or clusters) is defined by the researcher *a priori*. The groups are separated by linear functions of the variables, called the discriminant functions (Rencher 2002). Users may eliminate redundant variables from the process with *forward selection*, a technique which starts with the variable that best separates the group and then adds one variable at a time, the one added is that which provides the maximal additional separation of the group. Alternatively *backward elimination* starts with all variables, and eliminates one variable at a time, the one that contributes the least to group separation. In stepwise selection, the two techniques are combined as variables are added one at a time, and at each step, variables are re-examined to determine if any have become redundant and can be dropped.

#### *1.2.1.3 Principal Component Analysis*

Principal component analysis (PCA) maximizes the variance from linear combinations of the original variables (Rencher, 2002). The first principal component is the linear combination of the variables that explains the most variance; the second principal component will be the linear combination of the variables that explains the most variance orthogonally from the first component; and so on. The e*igenvalues* quantify the variance explained by each principal component. The loadings correspond to the correlation between the original variables and each of the principal components. Scores are the third output from the analysis and are essentially the values of the components for each case.

#### *1.2.1.4 Factor Analysis*

In factor analysis, it is the original variables which are linear combinations of a few random variables called factors (Rencher, 2002). These factors cannot be measured or observed; they are hypothetical. They are meant to reduce the dimensionality of the original dataset, taking into account the existing correlation between the variables. Loadings come into play in the weighting of the factors for each variable. Rotation of the loadings may be applied in order to facilitate the interpretation of the factors. *Varimax rotation* is one specific rotation technique where the user attempts to maximize the variance of the squared loadings for every column of the rotated loadings matrix, making the loadings either small or large to separate them apart (for more information see Rencher, 2002).

In Canada, factor analysis was used in a study on the hydrochemistry of surface waters in the Mackenzie River drainage basin (Reeder *et al*, 1972). Variables included sodium, potassium, magnesium, calcium, strontium, boron, iron, uranium, silica, fluorine, chlorine, bicarbonate, sulfate, and organophosphate. Associations between sites were analyzed in Q mode analysis, and R mode analysis described the dependence among variables. The study identified bedrock geology as the

main environmental factor explaining the inorganic composition of surface waters in the Mackenzie River drainage basin.

Shrestha and Kazama (2007) used these four techniques to study the spatial and temporal variability of water quality in the Fuji River basin, Japan. The water quality parameters studied were discharge, temperature, dissolved oxygen, biochemical oxygen demand, chemical oxygen demand, pH, total suspended solids, electrical conductivity, total coliforms, nitrate nitrogen, ammonical nitrogen, and inorganic dissolved phosphorus. Hierarchical agglomerative clustering using Ward's method with squared Euclidean distances as a measure of similarity was used to segregate sites into three groups which successfully correspond to level of pollution. Discriminant analysis was then used to evaluate temporal and spatial variations in the data. Four temporal groups were identified and the authors found that the backward stepwise mode correctly assigned water quality data to appropriate season with fewer parameters than with the forward stepwise mode. A similar exercise was performed with sites locations, defining three groups; again, the backward stepwise method provided better performance in correctly assigning stations to level of pollution groups with fewer parameters. A PCA was performed on each group defined by the initial cluster analysis, and provided a good understanding of which parameters were explaining the most variation in each group. In the less polluted sites, they were parameters related to discharge, temperature and organic pollution. In the moderately polluted sites, parameters related to organic pollution from domestic wastewater and nutrients from agriculture were explaining the most variation. In highly polluted sites, parameters related to organic pollution and nutrients from domestic and industrial wastewater were leading the variations. Similar analyses were conducted in the Jinshui basin, China (Bu *et al,* 2010), the Luan River, China (Zhao and Cui, 2008), the Gomti River basin, India (Singh *et al,* 2004), the Behrimaz basin, Turkey (Varol and Sen, 2009), and the Xiangjiang basin, China (Zhang *et al,* 2009).

Simoneau (1985) used PCA and cluster analysis to study the spatial variability of 134 river monitoring sites in Québec, Canada. The dataset included twelve water quality variables: calcium, magnesium, chloride, sulfate, iron, total nitrogen, total phosphorus, total organic carbon, tannins and

lignins, turbidity, alkalinity and pH. The cluster analysis produced six groups, and group membership was overlaid on the PCA scores plot. This provided the necessary information to identify the dominant water quality characteristics associated with site locations and group membership, and to identify the apparent environmental influences explaining water quality in each group.

### 1.2.2 Spatial Analyses

Several methods are available to use the spatial information in the data to improve understanding of the patterns in the data. The following techniques integrate the spatial structure of the datasets into their analyses of environmental data.

#### *1.2.2.1 Semi-variogram*

The spatial structure of the data may influence the results obtained from a statistical analysis and this is especially true when the data are spatially autocorreled, i.e. values show similarity (positive autocorrelation), or dissimilarity (negative autocorrelation) according to the geographic distance separating them. Spatial autocorrelation can be measured with Moran's *I* coefficient, which is somewhat similar to Pearson's correlation coefficient with a spatial dimension (Legendre and Fortin, 1989). The semi-variogram is used to graph the semi-variance in the data as a function of distance. Theoretical semi-variograms are modelled or "fitted" from the experimental semi-variogram in order to produce the parameters describing the relationship and to be able to use them for comparison or in interpolation methods (e.g. kriging). The parameters are:

- the range: distance where the semi-variance stops increasing;
- the sill: ordinate value where the semi-variance stops increasing; and
- the nugget: refers to the type of non-linear function that best "fit" mostly the short distances part of the semi-variogram (Legendre and Fortin, 1989).

#### *1.2.2.2 Spatial PCA*

Arslan (2009, 2013) proposed a method for integrating information on the spatial structure of a dataset into principal component analysis. His method is described with a case study where the water quality data from the Porsuk River in Turkey was to be characterized. The water quality parameters analyzed were discharge, temperature, dissolved oxygen, biochemical oxygen demand, ammonical nitrogen, nitrite as nitrogen, nitrate as nitrogen, and orthophosphate. Whereas classical PCA is computed from a dataset's covariance (or correlation) matrix, the spatially weighted PCA was performed on a spatial covariance matrix, which was essentially the covariance matrix weighted with a spatial weight matrix. The spatial weight matrix itself was computed from values of semi-variogram modeling performed for each variable. As a result, the spatial covariance matrix was said to reflect the correlation among variables taking into account their spatial associations. While the computation method seemed reasonable, Arslan's articles presented no obvious demonstration that the characterization of the water quality data from the Porsuk River was enhanced with the spatially weighted PCA compared to the classical PCA.

### 1.2.3 Relating Environmental Data and Landscape Features

#### *1.2.3.1 Multivariate Canonical Analysis*

Canonical correlation analysis encompasses multivariate analysis methods that determine the linear relationship between two sets of multivariate data. Canonical correspondence analysis is the approach using correspondence analysis, a multivariate ordination method that summarizes the information from a dataset of counts per categorical variables (Rencher, 2002). These methods are therefore popular in the analysis of abundance data. In Stanfield and Kilgour (2006), correspondence analysis was first used to describe "synthetic" uncorrelated variables with the most variation in taxon abundance for fish and benthos (separately) from original variable sets. Then, canonical correspondence analysis was conducted to illustrate both relationships between taxa compositions

from fish and benthos with landscape features. These analyses were followed by regression analyses, described in section 1.2.3.2.

Redundancy analysis is another canonical ordination method that combines regression and principal component analysis (PCA). It describes the variance in a response dataset constrained by an explanatory dataset. Essentially, it performs a multi-variate regression of the response data on the explanatory data, followed by a PCA on the fitted response values (Borcard *et al*, 2011). It therefore shares common properties with PCA which can accommodate the analysis of multi-dimensional data sets. The PCA produces eigenvalues and eigenvectors, which are used to produce a canonical ordination. The resulting axes are orthogonal to one another, and are linear combination of the explanatory variable (Borcard *et al*, 2011). In a case study from Zimbabwe, Mazvmavi *et al* (2005) used redundancy analysis to identify landscape characteristics that explain variance in river flow statistics in 52 basins. The landscape characteristics analyzed included the mean annual precipitation, average annual potential evapotranspiration, drainage density, median slope, and proportions of catchment area composed of selected geology classes, and land cover classes. Response river flow statistics were the mean annual flow, coefficient of variation of annual flow, base flow index, average number of days per year with no flow, and daily flows with 0.90, 0.70 and 0.50 exceedance probabilities. Mean annual precipitation was found to be the landscape characteristic most correlated with variations in river flow statistics, alone explaining 50% of the variations.

#### *1.2.3.2 Regression and Scale Issues*

Moerke and Lamberti (2006) used regression to study scale-dependent influences on water quality, habitat and fish assemblages in the Kalamazoo River basin, Michigan, USA. Their dependant and independent variables spanned four spatial scales: watershed, total upstream corridor with 100 meters riparian buffer, local (200 meters) upstream corridor with 50 meters riparian buffer and within-stream. Metrics analyzed at the watershed scale were the watershed area, low flow yield, elevation, distance to the main stream, road density, and factors of soil composition, surficial geology and land

cover. Analyses of the stream corridor included metrics of percentages of land cover classes within the corridor. Metrics used in within-stream analyses included dissolved oxygen, specific conductance, turbidity, temperature, nitrate nitrogen, ammonium nitrogen, soluble reactive phosphorus, discharge, and factors describing the habitat, canopy, and stream banks. Variations in fish assemblages in response to explanatory landscape factors were evaluated through canonical correspondence analysis and relationships between landscape variables and water quality were determined with stepwise multiple regression and forward selection of variables. Water quality measures were best described by broad scale metrics (watershed and total stream corridor) whereas fish assemblages were best described at the local spatial scale. Authors also determined that regression models successfully explained up to 74% of the total variation in water quality metrics.

Multi-scale regression analysis was also used to study the relationship between climate variables and runoff trends in the Athabasca River Basin, Canada (Peters *et al,* 2013). The Athabasca River Basin was subdivided into three sections: lower, middle and upper Athabasca River, and hydrometric (flow) and climatic (temperature and precipitation) data and runoff indices were analyzed for four time periods (1958-1976, 1977-2009, 1958-2009 and 1913-2009) for trends. The rank-based non-parametric correlation test, Kendall Tau, was used to assess strength of association between climate variables and runoff trends and variability. There were important differences in results depending on the sub basin and time range. Nonetheless, multiple non-linear regression analysis indicated that variations in precipitation explained greater than 67% of the annual median/mean lower basin (the area of most concern) runoff variability since 1958.

Stanfield and Kilgour (2006) studied the relationship between percent impervious cover and biophysical properties in tributary streams to Lake Ontario, Canada. Backward stepwise regression was used to model the relationship between biophysical properties (fish and benthos assemblage, instream habitat and temperature), and landscape features (drainage area, slope, base flow index and percent impervious cover). Whereas in their first model, all landscape features were included, in the second, the percent impervious cover was excluded, and the residuals from that second model

were used in a third model to relate to percent impervious cover. The data had been separated into a calibration set and a validation set before analysis to allow subsequent validation of the model. They found their model relatively robust and concluded that under a 10% impervious cover threshold, variation in percent impervious cover changed in an incremental way fish, benthos, temperature and the percent stable banks. Above 10% impervious cover, fish and benthos, both consisting mainly of warm water or tolerant assemblages, were clearly altered. There was, however, no apparent relationship with percent impervious cover.

#### *1.2.3.3 Spatial Regression*

Chang (2008) used spatial regression to identify relationships between water quality data and landscape factors in the Han River basin, South Korea. Spatial regression is said to incorporate spatial dependence to the regression, either as an error-based model, where error terms would be correlated across different spatial units, or as a lag-based model, where the dependent parameters would be influenced by the independent parameters in adjacent places. Spatial lag would correspond to a capacity from upstream sites to predict the water quality at downstream sites. In Chang's study, most spatial regressions were lag based. The author also performed ordinary least square regression in order to compare both sets of regression results. The water quality data consisted of temperature, pH, dissolved oxygen, biochemical oxygen demand, chemical oxygen demand, suspended sediments, total phosphorus and total nitrogen from 118 monitoring sites. For the regression analyses, the mean annual water quality averaged over a three year period was used, in order to minimize the potential influence of inter-annual hydroclimatic variability. Landscape factors consisted of percent landcover of each of seven types, percent soil of each of three types based on level of drainage, mean elevation, and standard deviation of slope. Regressions were conducted at two scales: the full drainage area of the monitoring site, and a 100 meters riparian buffer zone around the river segment upstream from the monitoring site.

Chang concluded that spatial regression models better explained variations of water quality, especially for the cases where Moran's *I* value for water quality parameters were high; in those cases, the spatial regression explained more than 10% additional variations in water quality when compared to ordinary regression. Regarding scale, dissolved oxygen was better explained at the whole basin scale while oxygen demand, suspended sediments and nutrients were better explained at the 100 meters buffer riparian scale. Whereas landcover was the primary influence at the whole basin scale, the relative importance of topography and soil characteristics increased at the riparian scale.

#### *1.2.3.4 Geographically Weighted Regression*

Tu (2011) studied the relationship between water quality data and land use characteristics at different levels of urbanization in eastern Massachusetts, USA, using geographically weighted regression (GWR). GWR produces an array of geographically located nonparametric regression coefficients relating a dependent and an independent variable. The regression coefficient for each value of the dependent variable is calculated with independent values of all neighbouring cases, with close-by values contributing more weight than values farther away, as the weight of contribution to the calculation of the regression coefficients gradually declines with increasing distance from the local regression point (Brunsdon et al, 1998). The water quality variables included were specific conductance, dissolved solid, calcium, magnesium, sodium, potassium, chloride, ammonia nitrogen, nitrite nitrogen, ammonia plus organic nitrogen, nitrate plus nitrite nitrogen, phosphorus, and orthophosphate phosphorus. The explanatory variables were the percentages of six land use types in the drainage area for each site. GWR models were estimated between one water quality variable and one land use feature at a time. Values of *t*-test and $R^2$ were produced and demonstrated a number of significant relationships which varied spatially. For example, dissolved solids showed a more significant positive relationship with percentages of commercial land when further away from Boston; specific conductance showed significant positive relationship away from Boston and significant negative relationship closer to Boston. Each regression coefficient were also plotted against the

combined percentage of developed land to further demonstrate the varying relationships between water quality data and land uses according to the urbanization gradient.

#### *1.2.3.5 Regression Trees*

A regression tree analysis (also referred to as CART; classification and regression tree) is a binary recursive partitioning method where a dataset is successively split into increasingly homogeneous subsets (each split selected is the one reduces the deviance the most), producing a class of tree-based models (Qian, 2010). CART is an exploratory tool to model interactions within explanatory variables providing good predictive and classification capabilities. It may also serve for the identification of variables influencing significantly the variability of response variables (Qian, 2010).

In a recent report from the USGS (Harden *et al*, 2013), regression tree analysis was used to model the relationship between landscape features and median annual nutrient (nitrate, total nitrogen and total phosphorus) yields in central and eastern North Carolina, USA. The 48 monitoring sites in this study were classified into one of seven land cover classes: undeveloped, low agricultural, high agricultural, low urban, high urban, mixed, and high point source flow. Percentages of each of these landcover classes were calculated for the drainage area of each site, as were the hydrologic soil drainage groups, the amount of precipitation, and the number of wastewater discharge facilities and cattle and swine permits. An ANOVA was performed to check for significant differences in annual nutrient yields when sites were grouped by their land use category. Tests showed total phosphorus provided the most significant differences between pairs of land use categories. Four regression tree models were then developed individually for each variable: 1) for all 48 sites; 2) for sites where point source flow contribution to stream flow was less or equal to 10%; 3) for sites from model 2 and with a basin area of less than 1,000 $mi^2$; and 4) for sites from model 2 and with a basin area of less than 100 $mi^2$. Models were interpreted individually and then in unison, allowing the authors to conclude

that, within the studied landscape feature variables, percentage of forested area was the predominant predictor variable for total nitrogen and total phosphorus.

Based on the literature review, multivariate statistical analysis methods were selected to study the physical-chemical surface water quality data from sites across Canada. In the following research article, a principal component analysis (PCA) will identify parameters that explain the greatest variance in the Canadian water quality data. Then, a cluster analysis will define groups of sites according to their water quality similarities, and this information combined with PCA information will expose the important water quality characteristics defining each group. Finally, relationships between landscape characteristics and the variance in water quality data will be described through redundancy analysis. Methods analysing spatial variance information such as spatial regression or geographically weighted regression were not retained for this study as it was expected they would not perform well at the scale of the study (Canada), especially given the irregular distribution of the data.

# 2.0 Multivariate Analysis of Canadian Water Quality Data

## 2.1 Introduction

Water quality refers to the physical, chemical and biological characteristics of water. For decades, water resources managers around the world have monitored the water their people depend upon, in order to ensure safe drinking water and a functioning ecosystem. Where safety is of concern, physical-chemical water quality data (hereinafter water quality data) are compared against guideline thresholds that indicate the suitability of the water for a particular use. This is done on a variable by variable basis. It is a different and more complex process when multivariate water quality data needs to be interpreted for reasons such as cause and effect research studies or planning a water monitoring network. One complexity arises from the very large amount of data generated by the numerous parameters measured in water quality monitoring programs. The use of multivariate analysis methods, such as Principal Component Analysis (PCA) as a means to reduce the dimensionality of datasets and identify relationships between variables, and cluster analysis as a means to classify sites based on similarities, has proven helpful in a number of studies (e.g., Bu *et al,* 2010; Ribeiro *et al*, 2014; Shrestha and Kazama, 2007; Simoneau, 1985; Singh *et al,* 2004; Varol and Sen, 2009; Zhang *et al* 2009; Zhao and Cui, 2008). In these studies, the structure in the data uncovered by the analysis allowed researchers to better identify external factors having a major influence on the variance in water quality data.

External factors that influence water quality include the geology, climate, topography, aspects of the drainage network, vegetation, land use and of course, sources of pollution. As such, knowledge of regional landscape characteristic allows more informed interpretation of water quality data. Due to geographic information systems and to the increasing availability and quality of spatial data on the internet, landscape information can be easily collected and computed, and may then be related quantitatively to water quality data. Redundancy analysis is a multivariate method that allows the study of such relationships; it is a canonical ordination method that provides information on the

structure in a set of response variables, in this case water quality data, as explained by a set of constraining variables (landscape characteristics). For example, this method has proven useful in a case study from Zimbabwe to identify the most important basin characteristic in the determination of river flow statistics; mean annual precipitation was explaining 50% of the variations in flow statistic (Mazvimavi *et al*, 2005).

Efficient water quality monitoring programs are targeted to, and dictated by regionally-specific conditions and objectives. These attributes define all aspects of the water quality monitoring programs including the sampling locations, sampling frequency and suit of parameters to measure. As a consequence, different monitoring programs generate different water quality databases. For example, monitoring agencies concerned with intense agricultural activity on their lands may maintain a water quality datasets with bimonthly measurements for a lengthy set of nutrient and pesticide parameters, whereas programs monitoring undisturbed locations may only include in their datasets seasonal data for selected physical measures, majors ions and metals. The heterogeneity of water quality databases becomes a challenge in large-scale studies. As multiple water quality datasets are integrated for analysis, it may be necessary to select a subset of parameters and retain only those present in a majority of datasets.

Canada is a federation where responsibilities over most resources management are shared between the federal government and the provinces and territories. This is the case for the management of water quality, as stipulated under the *Canada Water Act* (Environment Canada, 2010). As specified in the *Department of the Environment Act,* national leadership in matters of water management is to be fulfilled by the Minister of the Environment (Environment Canada, 2014). In this context, the Water Quality Monitoring and Surveillance division delivers the water quality programs of Environment Canada in collaboration with over 20 partners, both federal and provincial, and under federal-provincial-territorial agreements (Environment Canada, 2012). After a recommendation by the Office of the Auditor General of Canada in 2010 (Office of the Auditor General of Canada, 2010), the Water Quality Monitoring and Surveillance division established risk-based methods to provide for the

systematic selection of monitoring sites across Canada where the monitoring needs, based on risk to water, are the greatest (Environment Canada, 2013a). Areas of higher risk are then targeted, i.e. areas where there is elevated environmental pressure, or where the ecosystem is more vulnerable or valuable (e.g. providing habitat to endangered species). As the Canadian water quality monitoring network is considered under a risk-based light, a parallel question surfaces: what can we learn about the monitoring network by looking at the Canadian water quality database itself, in a holistic manner? A national-level analysis, to determine patterns and relationships and how they vary spatially, may provide useful supplementary information to water resources managers concerned with the just and efficient allocation of monitoring resources. This is the object of this study.

The goal of this study is to use multivariate statistical analysis to generate large scale knowledge on the spatial variance of physical-chemical lotic fresh-water quality data across Canada. Water quality data from various monitoring programs across Canada will be integrated into one dataset, and averaged across time to yield one value for each parameter at each site. Analyses will be conducted in parallel on two water quality data matrices, one including more sites but few parameters, the other including more parameters but fewer sites. A principal component analysis (PCA) will identify parameters that explain the greatest variance in the water quality data. A cluster analysis will define groups of sites according to their water quality similarities, and this information superimposed on PCA information will expose the important water quality characteristics defining each group. Finally, landscape characteristics of drainage areas associated with each monitoring site will be calculated and their importance in explaining the variance in water quality data will be described through redundancy analysis.

## 2.2 Data and Methodology

### 2.2.1 Physical-Chemical Water Quality Data

Provincial and federal agencies responsible for surface fresh-water quality monitoring in Canada each have a network of designated monitoring locations they visit on a regular basis to take samples for physical-chemical analyses. The frequency of sampling events varies from weekly, for special study sites, to biannually in very remote locations. Depending on the sampling location, samples may be collected from the shore, from a bridge, by wading in the river, through ice, etc. Standardized protocols exist for most sampling situations (e.g. Canadian Council of Ministers of the Environment, 2011) and are followed by sampling technicians to minimize sources of errors in the samples. Typically, water samples are collected in containers provided by the laboratories performing the chemical analyses, and are transported back to the laboratories within specified hold time. A number of parameters that are more susceptible to change during transport from the field to the laboratory may be measured *in situ* using a hand-held meter; these include temperature, dissolved oxygen, pH and specific conductance.

Laboratories selected to perform chemical analyses must be accredited for the analysis of parameters of interest with the Canadian Association for Laboratory Accreditation or the Standards Council of Canada. Thus, they have demonstrated their competency for the analysis of these parameters through successful proficiency testing, the use of standard methods and an efficient quality assurance management program. Laboratory physical and chemical analyses are performed for selected parameters dependant on the monitoring location and program. Physical parameters monitored may include temperature, specific conductance, turbidity and total suspended solids. Inorganic parameters measured may include pH, dissolved solids, alkalinity, major ions (calcium, magnesium, sodium, potassium, and bicarbonate) and trace metals (zinc, arsenic, lead, nickel, iron, etc.). Nutrients monitored usually include total phosphorus and may include a number of nitrogen forms (e.g., total nitrogen, nitrate, nitrite, ammonia, etc). Other parameters monitored include

dissolved oxygen, sanitary measures such as total coliforms, and fecal coliforms, and organics measures, which include total organic carbon and pesticides compounds, such as 2,4-D and glyphosate.

In Canada, water level and discharge information is collected and produced by the Water Survey of Canada, of Environment Canada. This agency has an extensive network of approximately 2500 gage stations across the country (Environment Canada, 2013b). Several water quality monitoring stations are therefore located in proximity to water gage stations, or co-located with them. Water quantity information is often useful for the interpretation of water quality data, because it provides contextual information on the status of the river when the sample was taken (e.g. base flow, or during or after a storm event, etc), and on the hydrologic regime of the river. Also, at gages where water discharge information is generated from measured levels, this information enables the calculation of loading estimates from measured pollutants concentrations.

For this study, water quality data were gathered for river monitoring sites located within five kilometres of a water gage stations (on the same stream branch) generating discharge data. This criterion was established in order to allow the integration of discharge information in the analysis of the water quality data. Data collected between 2008 and 2012, inclusively, were retained. Data were first sought from Environment Canada's Water Quality Monitoring and Surveillance division, which provided data collected from federal monitoring sites and from sites managed under shared federal-provincial monitoring agreements. In provinces where federally mandated sites were too sparse, additional water quality data were sought from provincial departments (Table 1).

**Table 1.** Water quality data sources

| |
|---|
| Alberta Environment and Sustainable Resource Development (2014) |
| Environmental Canada, Water Quality Monitoring and Surveillance (2014); *Includes data from Environment Canada Federal Water Quality Monitoring Program and data collected as part of Federal-Provincial Water Quality Monitoring Agreements, including the following organizations: |
| British Columbia Ministry of the Environment; |
| New Brunswick Department of Environment; |
| Newfoundland and Labrador Department of Environment and Conservation; |
| Nova Scotia Environment; |
| Prince Edward Island Department of Environment, Energy and Forestry. |
| Manitoba Conservation and Water Stewardship, Water Quality Management Section (2014) |
| Ontario Ministry of the Environment and Climate Change (n.d.) |
| Québec, Ministère du Développement durable, Environnement, et Lutte contre les changements climatiques (2014) |
| Saskatchewan Water Security Agency (2014) |

The variable names for water quality parameters varied from one jurisdiction dataset to another (e.g. ammonia vs ammonia dissolved, total phosphorus vs phosphorus unfiltered total). Therefore parameter names were standardized prior to data integration. During integration, only data for parameters monitored at a majority of sites were retained. This greatly reduced the data available since many parameters are monitored only in part of the country. Only datasets and parameters spanning at least a three-year period within 2008 to 2012 were retained. In cases where pH had two measures for the same sample, one taken *in situ* during sample collection and one from the laboratory, the *in situ* value was retained, because samples rarely get to the laboratory within the two-hour time limit recommended for laboratory pH measurement (Environmental Protection Agency, 2012), which is very short due to the high susceptibility of pH to change during transport. The opposite was done with specific conductance (the conductivity at 25$^{o}$C) as it is a more stable parameter with a holding time of 28 days (Environmental Protection Agency, 2012b).

The number of samples per parameter differed considerably between sites, reflecting the various sampling frequencies established by the different monitoring programs. Samples counts also differed by year, for example some sites had limited monitoring done over a couple of year before engaging in a regular monthly sampling schedule. The distribution of sample counts over the years was adjusted in cases where the count doubled or more, between different years. In these cases, a number of samples from the over-sampled years were randomly removed from the dataset. Lastly, samples counts varied importantly from season to season. Most sites had limited, or sometimes no sample measures available over the winter months. For that reason, measurements from the first quarter of the year (January, February and March) were removed from all analysis. Then, to preserve the effect of the seasonality (except winter) on water quality data, data were averaged by quarter of the year, over all years. These quarterly averages, calculated from different numbers of samples, presented an uneven variance that could be a source of error when analyzing quarterly values. Therefore the quarterly averages were averaged again to produce one overall average value per parameter per site, and this formed the dataset on which analyses were performed.

Concentrations lower than the detection limit, also called “censored data”, were substituted by half the value of the detection limit. This method is widely used to estimate summary statistics on datasets which include censored data. Helsel and Hirsch (2002), recommends against its use, in favor of methods that estimate values for the censored data based on data characteristics, including number of samples and distribution. However, these data characteristics varied in this study across sites and quarters of the year, suggesting the use of different methods to estimate values for censored data. Thus, the simpler method of substituting by half the value of the detection limit was selected as it could be applied consistently over the entire dataset.

The parameters monitored varied greatly between sites, however, the multivariate analyses used do not accommodate missing data. Therefore, two data matrices were created for parallel analysis. Matrix A included a greater number of sites (107), but a limited number of parameters:

specific conductance (SC), pH, and total phosphorus (TP). Matrix B had only 42 sites but included data for eight parameters: calcium (Ca), chloride (Cl), total alkalinity (T_ALK), dissolved oxygen (DO), water temperature (WT), specific conductance (SC), pH and total phosphorus (TP).

### 2.2.2 Landscape Characteristics Data

A set of landscape characteristics data was assembled in order to explore their relationship with water quality and to study the continental-scale spatial variation in water quality data. Given the exploratory nature of this phase of the study, only easily-retrievable landscape characteristics data were assembled.

Georeferenced drainage areas for every site were assembled as vector polygons in the geographic information system (GIS) ArcGIS (ESRI, 2012). These were provided by Environment Canada Water Quality Monitoring and Surveillance (2014), by Environment Canada Meteorological Service of Canada (2014), and in part adapted from the sub-sub-drainage-area framework (SSDA) of the Water Survey of Canada, provided in the Atlas of Canada 1,000,000 National Frameworks Data, Hydrology dataset (Natural Resources Canada, 2009), or adapted from the USGS Watershed Boundary Dataset framework (USGS, 2014). In cases where the drainage areas had to be modified, digital elevation models (Geobase, 2008-2012), and river networks from the National Hydrographic Network (Geobase, 2011-2014), were used to manually trace the site specific drainage area extents.

The georeferenced drainage areas were used to calculate the following site landscape characteristics (Table 2; for units and data sources see Table): proportion of cropland (Crop), population density (Pop), density of facilities registered in a pollutant release and transfer register (PtSo), dams density (Dams), road density (Road), mean annual average temperature (Temp), mean annual total precipitation (Prec), lake and reservoir proportion (Lake), stream density (Strm), and circularity ratio (Circ; see below). In addition, the following site information was compiled: average stream discharge at the site (Q), latitude (Lat) and longitude (Long).

Climate variables were calculated from raster datasets of annual average temperature and annual total precipitation, obtained by processing historical temperature and precipitation data for years 1961 to 1990, from Environment Canada Canadian Centre for Climate Modelling and Analysis (n.d.). Although this time period does not correspond to the time span for the water quality data (2008 to 2012), the spatial patterns of climate normals tend to be correlated from one period to another, so in the analyses performed in this study, the resulting ordinations should be similar.

Spatial data preparation, including projection transformations, data selection, merging, raster treatment, and geometry calculations, were processed using the ArcGIS software (ESRI, 2012). The circularity ratio *Circ*, calculated from the area *A* and the perimeter *P* of the drainage area polygons, is the area *A* divided by the area of a circle of perimeter *P* (Allaby et Allaby, 1999), and was calculated according to the formula:

$$Circ = \frac{4\pi A}{P^2}$$

Calculations of proportions and densities were processed in the Geospatial Modelling Environment (GME) software (Beyer, 2009-2012). As opposed to other GIS software, GME performs calculations on one polygon at a time. This method requires more processing on the part of the software but it avoids the introduction of errors in calculations where multiple polygons overlap (which is applicable in this study as some drainage areas are included within greater drainage areas).

Site latitude and longitude were provided along with the water quality data. Average stream flow was calculated from water gage sites daily mean discharge data, obtained from the Water Survey of Canada HYDAT database (Environment Canada, n.d.). To be consistent with the time span of the water quality data, daily mean water discharge data between 2008 and 2012 inclusively were retained, or data for the most recent five years where data between 2008 and 2012 were not available. Similarly to the treatment of water quality data, daily discharges from the first quarter of the

year were excluded, and the remaining data were first summarized as quarter averages, and those values were averaged again to produce one overall summary value per site.

**Table 2.** Landscape characteristics data, units and data sources

| Variable | Description | Unit | Data source(s) |
|---|---|---|---|
| Crop | Area classified as cropland / total drainage area | unitless | Land Cover 2010 (Commission for Environmental Cooperation, 2012) |
| Pop | Number of people / total drainage area | people / $km^2$ | Population Census 2011 (Statistics Canada, 2011); County totals: Vintage 2011, Census year 2010 (United States Census Bureau, 2012); 2010 Census Tracts (United States Census Bureau, n.d.(a)) |
| PtSo | Number of facilities registered in pollutant release and transfer register / total drainage area | facilities / $km^2$ | Pollutant release and transfer register (Environment Canada, 2009; Environmental Protection agency, 2009) |
| Dams | Number of dams / total drainage area | dams / $km^2$ | Atlas of Canada 1,000,000 national Frameworks Data, Hydrology – Dams (Natural Resources Canada, 2003); CanVec Feature catalogue, edition 1.1.2 (Natural Resources Canada, 2006) Major Dams of the United States (U.S. Army Corps of Engineers, 2005); NHD Dam Events (USGS, 2014) |
| Road | Length of road / total drainage area | km / $km^2$ | Road network file census year 2011 (Statistics Canada, 2011); United States Census Bureau Road Network, census 2010 (U.S. Census Bureau (n.d.(b)) |
| Temp | Drainage area mean annual average temperature | ºC | CCCma-CanESM2 modelled, historical, 1961-1990, near surface temperature averages, NAM-44 and ARC-22 (Environment Canada Centre for Climate Modelling and Analysis, n.d.) |
| Prec | Drainage area mean annual total precipitation | mm | CCCma-CanESM2 modelled, historical, 1961-1990, precipitation totals, NAM-44 and ARC-22 (Environment Canada Centre for Climate Modelling and Analysis, n.d.) |
| Lake | Area of lake and reservoir / total drainage area | unitless | Atlas of Canada 1,000,000 National Frameworks Data, Hydrology, Version 6.0. (Natural Resources Canada, 2009) |
| Strm | Length of stream / total drainage area | km / $km^2$ | Atlas of Canada 1,000,000 National Frameworks Data, Hydrology, Version 6.0. (Natural Resources Canada, 2009) |
| Circ | Drainage area circularity ratio | unitless | ArcMap 10.1 calculate geometry, Canadian Lambert Conformal Conic projection (ERSI, 2012) |
| Q | Average stream discharge at the site (excluding months of January, February and March) | $m^3$ / second | HYDAT (Environment Canada, n.d.) |

| | | | |
|---|---|---|---|
| Lat | Latitude coordinate of the monitoring site (North American Datum 1983) | decimal degrees | Monitoring agencies (Table 1) |
| Long | Longitude coordinate of the monitoring site (North American Datum 1983) | decimal degrees | Monitoring agencies (Table 1) |

### 2.2.3 Statistical Methods

All statistical analyses were performed with the R software, version 3.0.2 (The R Foundation for Statistical Computing, 2013), including packages vegan (Oksanen *et al*, 2013), and packfor (Dray, 2013).

The two data matrices were each analyzed using principal component analysis (PCA) and cluster analysis to uncover structures and patterns in the data. The PCA and the cluster analysis are referred to as unconstrained analyses because they operate solely on one multivariate dataset; there are no independent data. PCA is an ordination method that uses the covariance or the correlation matrix of the data to produce the principal components, which are linear combinations of the original variables (Rencher, 2002). The correlation matrix was used in this study because the data for the water quality parameters were of greatly different magnitudes. The first component of a PCA is the linear combination of the variables that explains the most variance in the original data; the second component is the linear combination of the variables that explains the most variance orthogonally from the first component; and so on. As a result, a fewer number of components can explain most of the variance from all variables included in the original dataset; this is the purpose of ordination methods. The variances explained by each component are indicated by the eigenvalues. The correlation between the original variables and each of the components are referred to as the loadings. The values of the components for each case, here, the monitoring sites, are the scores.

Cluster analysis summarizes the structure in the data based on similarity, so that individuals (in this case, the monitoring sites) of the same cluster resemble each other more than they do those

of another cluster. There are various clustering methods and measures of similarity. In this study, the Euclidean distance was selected to measure the similarity between water quality data. Ward's hierarchical agglomerative method, based on least increase in sum of squared distances (dispersion) in clusters, was the selected clustering method applied to the standardized water quality data. This is consistent with cluster analysis methods performed in other water quality studies (Shrestha and Kazama (2007), Simoneau (1985), Varol and Sen (2009), and Zhang et al (2008)). Cluster analysis produces a dendrogram depicting the agglomeration structure. A choice needs to be made on the number of clusters to retain. The graph of fusion level values (Borcard, 2011), presenting the dissimilarity values where two branches of the dendrogram merge, was used to visually select clustering levels, aiming to obtain a large enough number of groups to interpret without over-reducing the dissimilarity value. For both the PCA and the cluster analysis, the standard functions in R (The R Foundation for Statistical Computing, 2013) were used.

Simoneau (1985) used PCA and cluster analysis on data from a river monitoring network of 134 sites in Quebec, with data for twelve variables. The cluster analysis classified the sites in six groups. As they were overlaid on the PCA scores figure, the water quality aspects dominating each group were identified, as well as the environmental characteristics explaining them.

Redundancy analysis, a constrained ordination method, was used to further explore the spatial variance in water quality data and its relationship with drainage area landscape characteristics. Redundancy analysis is used to describe the variance in a response dataset (here the water quality data) explained, or constrained, by an explanatory dataset (the landscape characteristics). Essentially, it performs a multi-variate regression of the response data on the explanatory data, then a PCA on the fitted response values (Borcard *et al*, 2011). This PCA produces canonical eigenvalues and eigenvectors, which are used to produce an ordination. The resulting axes are orthogonal to one another, and are linear combination of the explanatory variable. Where canonical eigenvectors are multiplied with the fitted response values (as opposed to the original response values), as performed in this study, resulting sites scores are referred to as "sites

constraints" and are also linear combinations of the explanatory variables (Borcard *et al*, 2011). In a case study from Zimbabwe, Mazvmavi *et al* (2005) used redundancy analysis to identify landscape characteristics that explain variance in river flow statistics. The importance of identified landscape characteristics was also estimated with percentages of variance explained.

Redundancy analysis produces an $R^2$ value expressed as a proportion of the variance in the response data explained. Similarly to the $R^2$ value in a multiple regression, any variable included in the explanatory dataset increases the $R^2$, regardless of its relation, existent or not, to the response dataset (Borcard *et al*, 2011). Ezekiel's formula produces a $R^2$ adjusted according to the size of the explanatory matrix, and solves this issue:

$$R^2_{adj} = 1 - \frac{n-1}{n-m-1}(1-R^2)$$

where $n$ is the number of cases and $m$ is the number of explanatory variables (Borcard *et al*, 2011).

The significance of redundancy analysis results may be tested with permutation tests (Borcard *et al*, 2011). These tests create a reference distribution of a selected statistic by running a large number of times the analysis against a randomly generated dataset. Then, the true statistic under the real dataset is compared to the reference distribution. The null hypothesis is rejected if the *p* value is equal to or smaller than the predefined significance level (0.05 in this study). The *p* value is the proportion of the permuted values equal to or larger that the true value. For testing redundancy analysis results, the *pseudo-F* test statistic is calculated:

$$F = \frac{SS(\hat{Y})/m}{RSS/(n-m-1)}$$

where $n$ is the number of cases, $m$ is the number of canonical eigenvalues, $SS(\hat{Y})$ is the sum of squares of the fitted values and *RRS* is the residual sum of squares (Borcard *et al*, 2011). The R package vegan (Oksanen *et al*, 2013), was used to run redundancy analyses and permutation tests of significance on water quality data in relation to landscape characteristics.

Forward selection of variables was used in the redundancy analysis to identify the landscape characteristics that explain the most variance in water quality data. Forward selection is the variable reduction method that is the most often applied in combination with redundancy analysis (Borcard *et al*, 2011). Starting from a null model, it selects, or adds one explanatory variable at a time; the one added being the one that increases the adjusted $R^2$ value the most, conditional to it being deemed significant at a predefined significance level (0.05 in this study). This goes on as long as the model under the latest added variable does not produce an adjusted $R^2$ above the one of the original model, in order to prevent the inclusion of too many variables (Borcard *et al*, 2011). The R package packfor (Dray, 2013) was used to apply the forward selection method in this study.

Variation partitioning is another method that may be used along with redundancy analysis to partition an explanatory dataset and to determine the importance of each partition in the explanation of the response dataset (Borcard *et al*, 2011). Since explanatory variables are not orthogonal, and are therefore correlated, it is assumed that there is some degree of overlap in how parts of an explanatory dataset explain the variance in a response dataset, i.e. they explain the same information. With an explanatory dataset partitioned in two, the variation partitioning method calculates the proportion of the variance in the response variables explained by (1) a shared portion of intercorrelated variables within each partition, (2) portions of each of the two partitions that are not intercorrelated, and (3) other factors not included in the analysis (Borcard *et al*, 2011). Variation partitioning was applied to the redundancy analyses of the water quality data matrices constrained by landscape characteristics partitioned between “natural” geographic variable (Temp, Prec, Lake, Strm, Circ, Q, Lat and Long) and anthropogenic variables (Crop, Pop, PtSo, Dams and Road), in order to compare how the variance in water quality data is explained by each of these partitions. The R package vegan (Oksanen *et al*, 2013), was used to apply the variation partitioning method.

## 2.3 Study Area

Matrix A sites (many sites, few variables) are distributed across the Canadian provinces and territories and include 2 sites in Yukon (YT), 21 in British Columbia (BC), 7 in Alberta (AB), 6 in Saskatchewan (SK), 7 in Manitoba (MB), 16 in Ontario (ON), 14 in Quebec (QC), 6 in New Brunswick (NB), 11 in Nova Scotia (NS), 3 in Prince Edward Island (PE), and 14 in Newfoundland and Labrador (NF) (Figure 1). The drainage area associated with each site varies from 2.8 $km^2$ (Leary's Brook, NF) to 286,840 $km^2$ (Red River, MB). The mean drainage area is 16,146 $km^2$, the median 1,068 $km^2$, and the standard deviation 43,206 $km^2$. Some site drainage areas are located within the drainage areas of other sites. The proportion of overlapping drainage areas to the sum of all drainage areas is 22.4%. Drainage areas extend over seven level I ecological regions (Commission for Environmental Cooperation, 1997). Taiga is the dominant eco-region in 6 site drainage areas; Northwestern Forested Mountains, 14; Marine West Coast Forests, 7; Northern Forests, 39; Great Plains, 9; North American Deserts, 3; and Eastern Temperate Forests, 29. Details on the physical, biological and anthropogenic characteristics of those ecological regions can be found in Commission for Environmental Cooperation (1997).

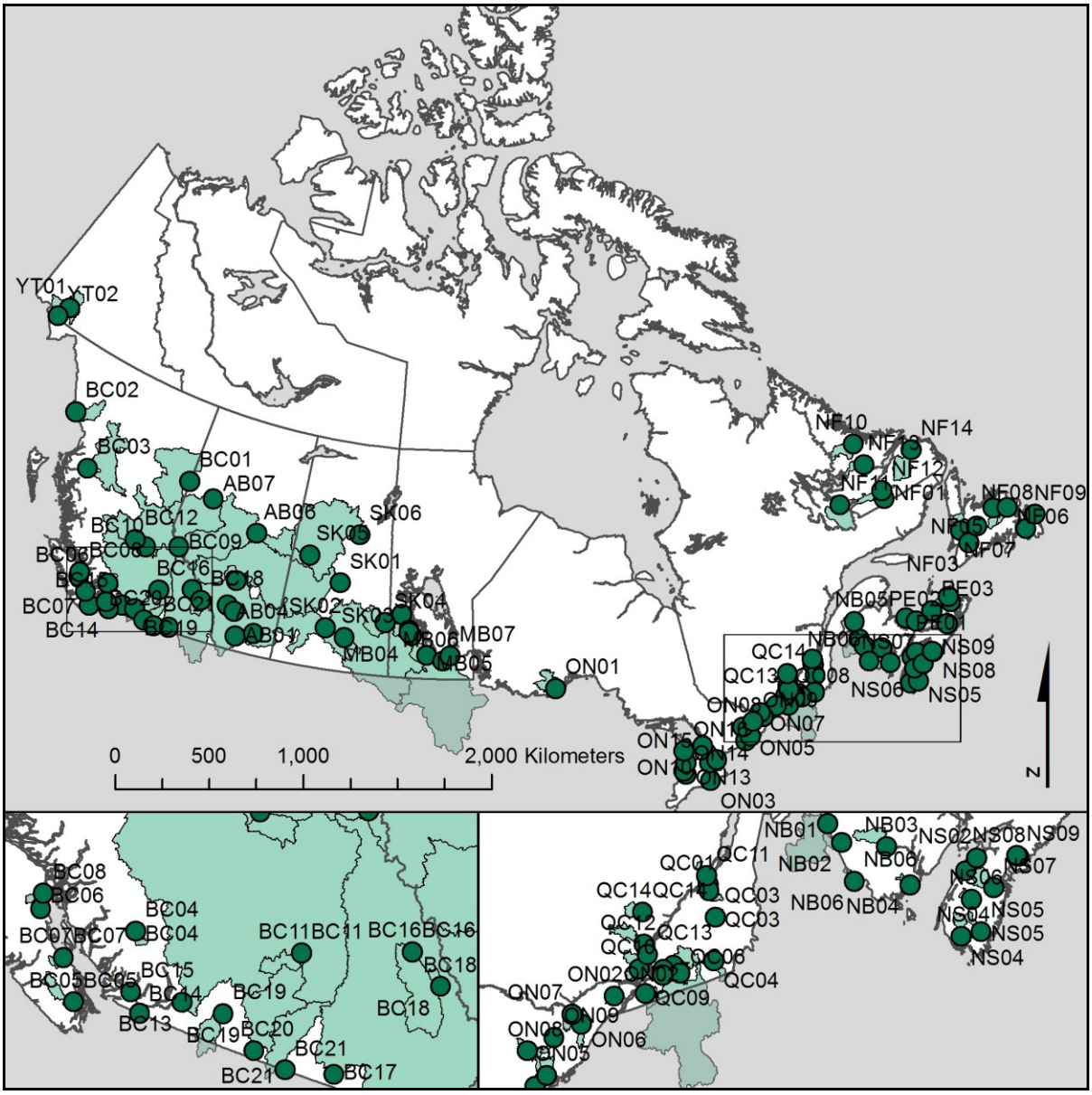



**Figure 1.** Matrix A site locations and drainage areas

Matrix B includes more variables but fewer sites and is therefore more geographically restricted. This matrix includes 5 sites in BC, 7 in AB, 6 in SK, 7 in MB, 13 in ON, 3 in NB, and 1 in NF (Figure 2). Drainage area size varies from 35 km$^2$ (North Alouette River, BC) to 286,840 km$^2$, the same as in Matrix A. The mean drainage area is 27,495 km$^2$, the median 1,652 km$^2$, and the standard deviation 60,010 km$^2$. The proportion of site drainage area overlap is 25.6%. Compared to Matrix A, Matrix B sites tend to be concentrated more in the central part of the country (the Prairies and ON) and their drainage areas tend to be larger. Dominant level I ecological regions (Commission

for Environmental Cooperation, 1997) in Matrix B drainage areas include: Northwestern Forested Mountains (4 sites), Marine West Coast Forests (3), Northern Forests (13), Great Plains (9), and Eastern Temperate Forests (13).

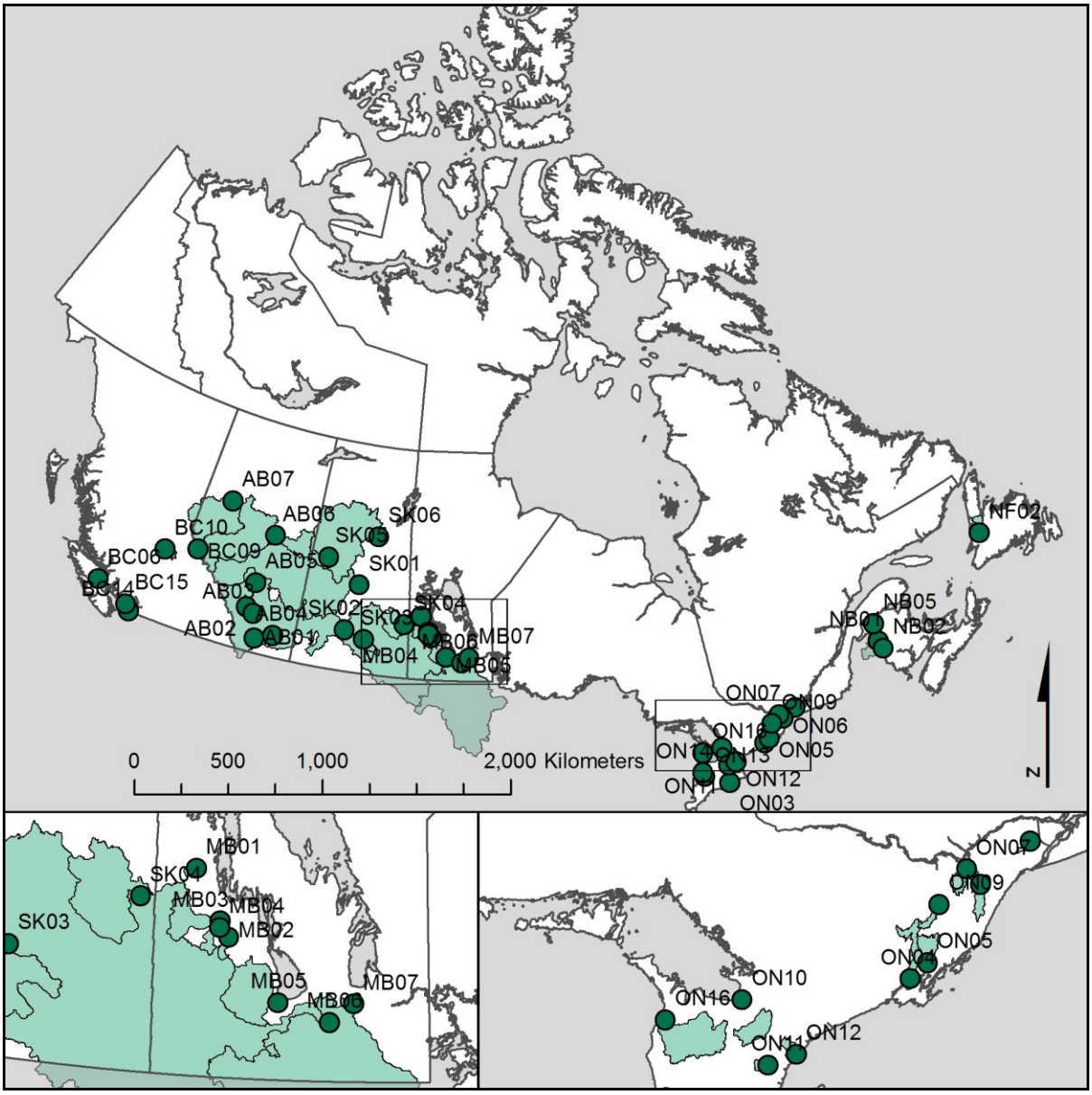



**Figure 2.** Matrix B site locations and drainage areas

## 2.4 Results

### 2.4.1 Water Quality Data Description

The water quality data described in this section are the site averages calculated as described in the methodology section. Over the entire dataset, values of pH vary between 4.44 (Roseway River, NS) and 8.60 (Qu'Appelle River below Qu'Appelle dam, SK), with fewer sites with lower pH values compared to sites with higher pH (Table 3 and Appendix B). Sites with lower pH are located in Nova Scotia (Appendix A) and reflect local conditions of elevated atmospheric acid deposition and a crystalline geology of poor buffering capacity. These sites are absent from the Matrix B dataset, where pH varies between 6.74 (North Alouette River, BC) and 8.60. Values of specific conductance (SC) vary between 13 (Little Mecatina River, NF) and 1210 µS/cm (Don River, ON) with fewer sites of high values located in Ontario, Manitoba and Saskatchewan. In the Matrix B subset, SC varies between 16 (North Alouette River, BC) and 1210 µS/cm. Total phosphorus (TP) measurements include a number of values below the detection limit and these were substituted by half the detection limit in the calculation of site average values. Resulting TP values range between 0.003 (Fraser River at Red Pass, BC and Exploits River, NF) and 0.4 mg/L (Red River, MB), with many more sites of lower TP (Appendix B). Within the Matrix B subset, TP varies between 0.003 (Fraser River at Red Pass, BC) and 0.4 mg/L, again with more sites of lower values.

Specific to the Matrix B dataset, calcium (Ca) values range from 1.7 (North Alouette River, BC) to 90.1 mg/L (Vermilion River, MB). Chloride (Cl) values vary between 0.6 (Fraser River at Red Pass, BC) and 249.0 mg/L (Don River, ON), with a greater number of sites of low Cl values (Appendix B). Values of dissolved oxygen (DO) vary between 8.3 (Assiniboine River, MB) and 13.5 mg/L (Qu'Appelle River below Qu'Appelle dam, SK). Total alkalinity (T_ALK) ranges between 4.1 (North Alouette River, BC) and 288.4 mg/L (Assiniboine River, SK). Finally, water temperature (WT) varies from 6.7 (Fraser River at Red Pass, BC) to 16.4 °C (Nanticoke Creek, ON).

SC, TP and Cl data were ln-transformed prior to analysis using unconstrained methods in order to normalize their distribution. Matrix A and Matrix B site average values are provided in Appendix A and density plots (kernel estimates) are provided in Appendix B.

**Table 3.** Water quality data summary statistics

| | | Count | Min | 5th percentile | Median | Mean | 95th percentile | Max | Standard Deviation |
|---|---|---|---|---|---|---|---|---|---|
| Matrix A | pH | 107 | 4.44 | 6.31 | 7.87 | 7.64 | 8.39 | 8.60 | 0.77 |
| | SC (µS/cm) | 107 | 13 | 18 | 224 | 280 | 715 | 1210 | 248 |
| | TP (mg/L) | 107 | 0.003 | 0.005 | 0.28 | 0.06 | 0.24 | 0.40 | 0.09 |
| Matrix B | Ca (mg/L) | 42 | 1.7 | 4.8 | 49.8 | 50.4 | 86.4 | 90.1 | 24.7 |
| | Cl (mg/L) | 42 | 0.6 | 1.2 | 9.9 | 20.9 | 61.4 | 249.0 | 39.9 |
| | DO (mg/L) | 42 | 8.3 | 8.7 | 10.7 | 10.6 | 12.9 | 13.5 | 1.3 |
| | T_ALK (mg/L) | 42 | 4.1 | 12.3 | 169.1 | 159.1 | 261.7 | 288.4 | 76.7 |
| | WT (°C) | 42 | 6.7 | 8.3 | 12.5 | 11.9 | 16.1 | 16.4 | 2.6 |
| | pH (mg/L) | 42 | 6.74 | 7.15 | 8.21 | 8.08 | 8.44 | 8.60 | 0.39 |
| | SC (µS/cm) | 42 | 16 | 45 | 405 | 441 | 986 | 1210 | 271 |
| | TP (mg/L) | 42 | 0.003 | 0.005 | 0.04 | 0.09 | 0.34 | 0.40 | 0.10 |

### 2.4.2 Unconstrained Analyses

#### *2.4.2.1 Matrix A*

A first principal components analysis (PCA) was performed on the correlation matrix of three variables (pH, SC and TP) and 107 sites (Matrix A). Two components were retained for analysis. The first component explains 77.5% of the variance in the original dataset and it is negatively correlated with all variables (Figure 3). The second component explains 16% of the variance and only TP and pH have noticeable loadings on it, in opposite directions. Only the first component has an eigenvalue greater than one, suggesting it is the only component explaining more variance than the original variables individually. Sites NS04-06 differ from others and have very low pH and SC (Figure 3).

Several sites in SK, MB and ON are positively correlated with TP and the opposite is true for several sites from NB and BC. Many sites in ON present higher scores on the second component (located upward in Figure 3) and have elevated pH values. Many sites from NF and NB have higher scores on the first component (located on the right side in Figure 3) and present low SC.

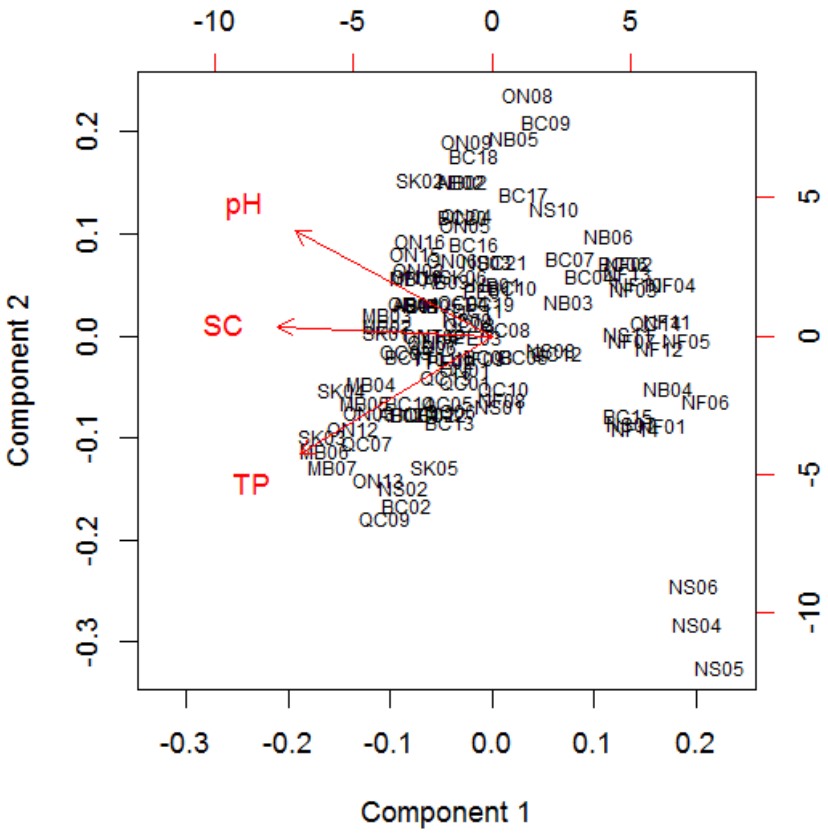



**Figure 3.** Principal component analysis biplot on Matrix A, presenting variable loadings (upper and right-hand side axes) and site scores (lower and left-hand side axes); Component 1 explains 77.5% of the variance and Component 2 explains 16.0%

The cluster analysis was cut to produce five groups, because the distance (dissimilarity) between cluster levels greatly reduced after that point (Appendix C). In Figure 4, cluster analysis groupings are identified on the plot of the PCA sites scores for the first and second components. A first group consists of sites with moderate SC, low TP, and a large range of pH values. Sites in the first group have scores close to zero on the first component and are positive on the second

component. Group 1 sites are found in ON, QC, the Maritimes and BC (Figure 5). The second and fifth groups, located on the right side along the first component axis, have low values of pH and SC. Sites in Group 2 are mostly located in the Atlantic region and on the Pacific coast. The three sites in group 5 are in NS and have low pH. The third and fourth groups, located on the left side along the first component axis, have generally higher values in all parameters. Sites of group 4, with negative scores on both components, have particularly high SC and TP data, and are mostly concentrated around Lake Winnipeg, although a few are found elsewhere. Group 3 sites, with negative scores on axis one and a wider range of scores on axis 2, have high values of SC and pH, and moderate values of TP. Group 3 sites are found across the country but are particularly dominant in the Prairies, and around the Great Lakes and the St. Lawrence River area. The most northern sites, in YK, belong to that group.

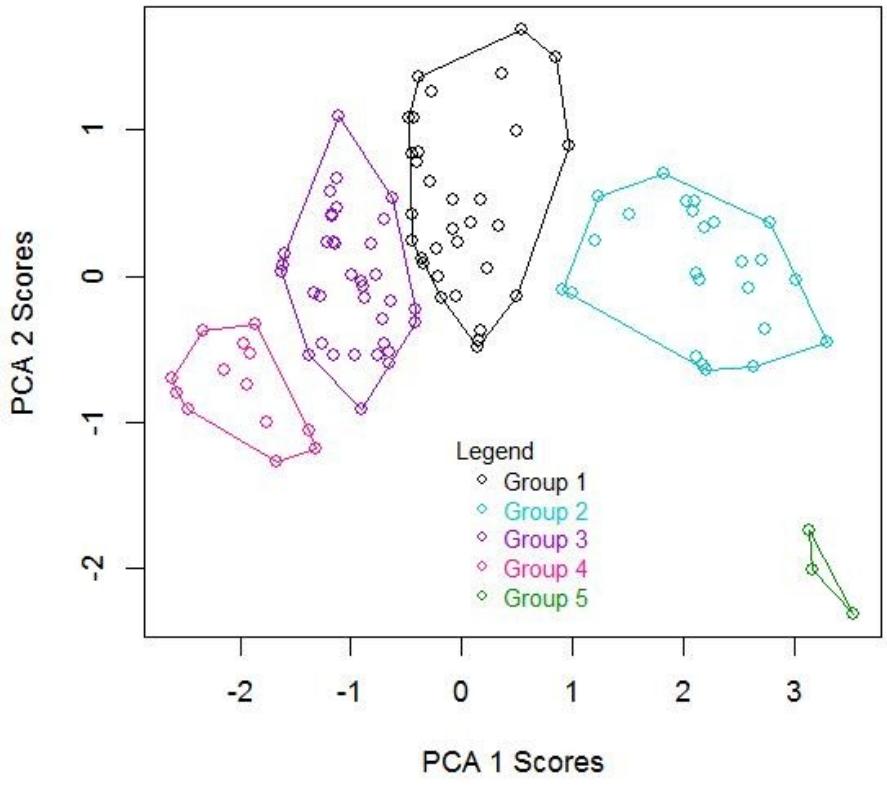



**Figure 4.** Scores on the first and second component of Matrix A, coded by cluster group membership

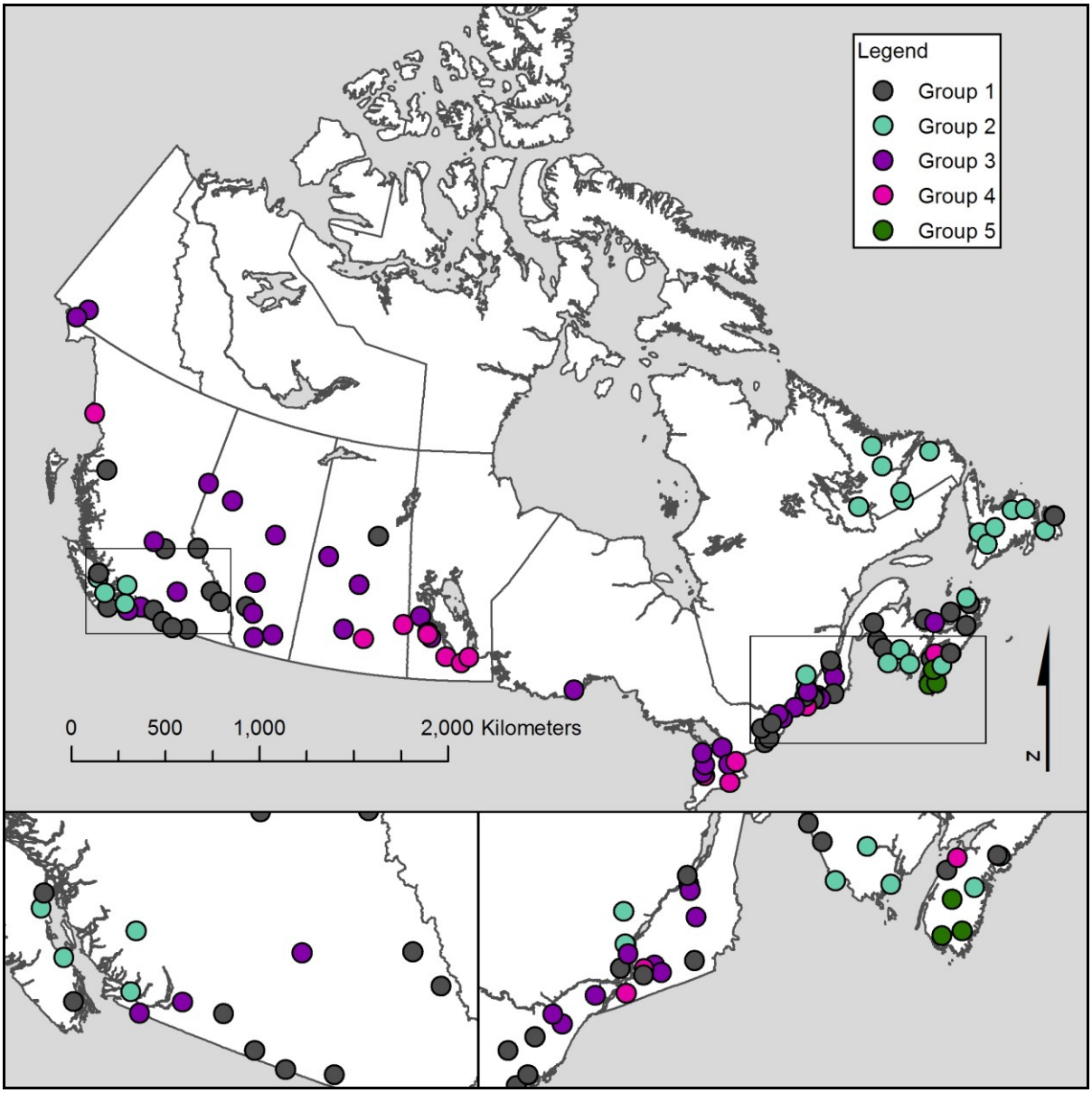



**Figure 5.** Location of sampling stations, coded by cluster membership (Matrix A)

*2.4.2.2 Matrix B*

A principal components analysis on Matrix B water quality data produces two components with an eigenvalue above 1 (Figure 6). The first component explains 56.4% of the variance in the original data, and the second component 16.2%; together, they explain 72.5% of the variance in the

original data. All water quality variables except WT and DO are similarly loaded (negative) on the first component. WT and DO are highly loaded (negative) on the second axis, and Cl is equally loaded (negative) in both components. All sites in MB have high scores (positive) on the second axis, negative scores on the first axis, and are negatively correlated with DO. Sites with highest positive scores on the first component are in BC and NF. Sites located in SK and ON present the highest negative scores on the second axis and are correlated with WT and DO.

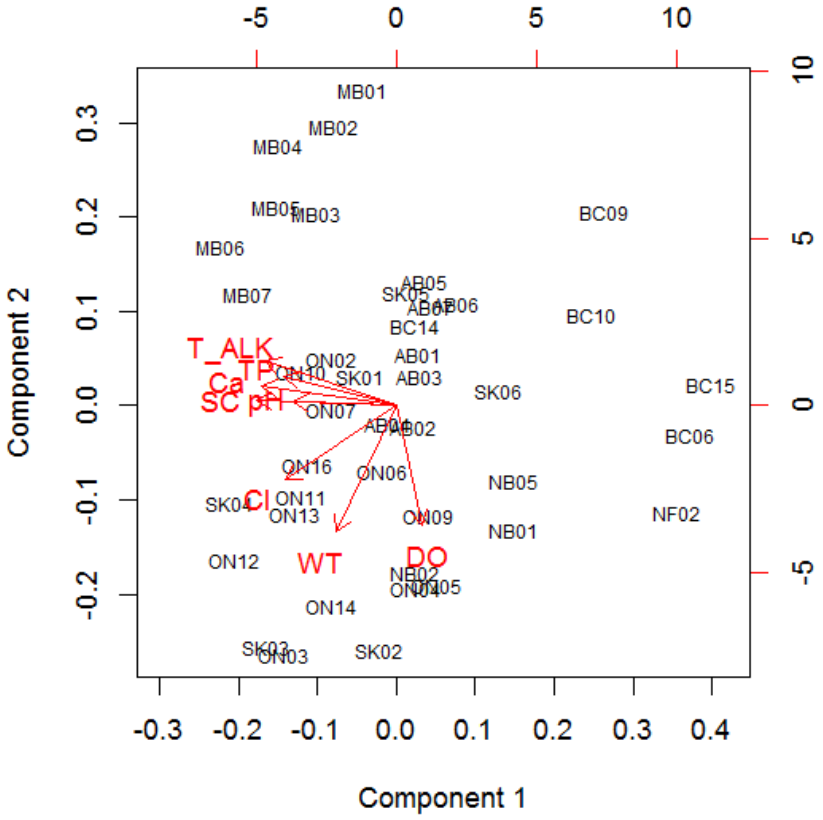



**Figure 6.** Principal component analysis biplot on Matrix B, presenting variable loadings (upper and right-hand side axes) and site scores (lower and left-hand side axes); Component 1 explains 56.4% of the variance and Component 2 explains 16.2%

The cluster analysis was cut to produce four groups; dissimilarity between cluster levels greatly decreased after that point (Appendix C). The first group contains sites with mostly positive scores on the first axis of the principal components analysis, but close to zero, and including a wide

range of scores on the second axis (Figure 7). The water quality data of Group 1 consists of moderate values, especially in Ca, SC, and T_ALK, and a large range of DO and WT values. Group 1 sites are mostly found in the Prairies; the 3 sites in NB are also in this group (Figure 8). The second and third groups, with negative scores on the first component, include higher values of Ca, SC and T_ALK. Group 3 sites have negative scores on the second axis, and have higher values of Cl, DO and WT. Sites of this group are located mostly in Southern ON, with a couple of sites in SK. Group 2 sites have with mostly positive scores on the second axis, and are located around Lake Winnipeg, as well as in Southern and Eastern ON. The fourth group, with positive scores on axis 1, are sites with lower values of Ca, Cl, pH, SC, T_ALK, TP and WT. The 5 sites belonging to that group are located in NF and BC.

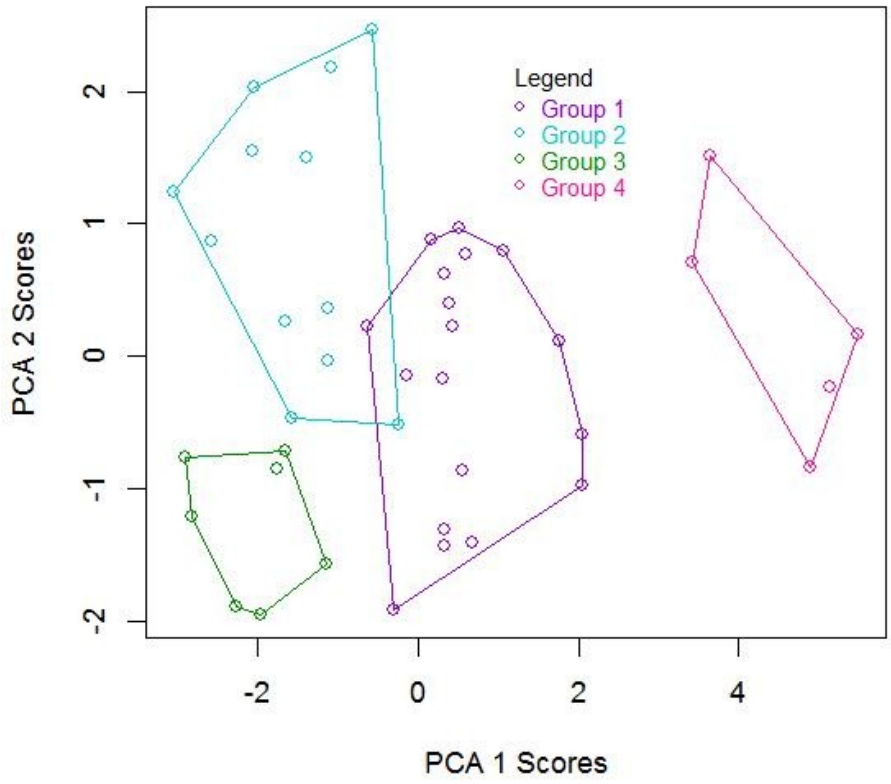



**Figure 7.** Scores on the first and second component of Matrix B, coded by cluster group membership

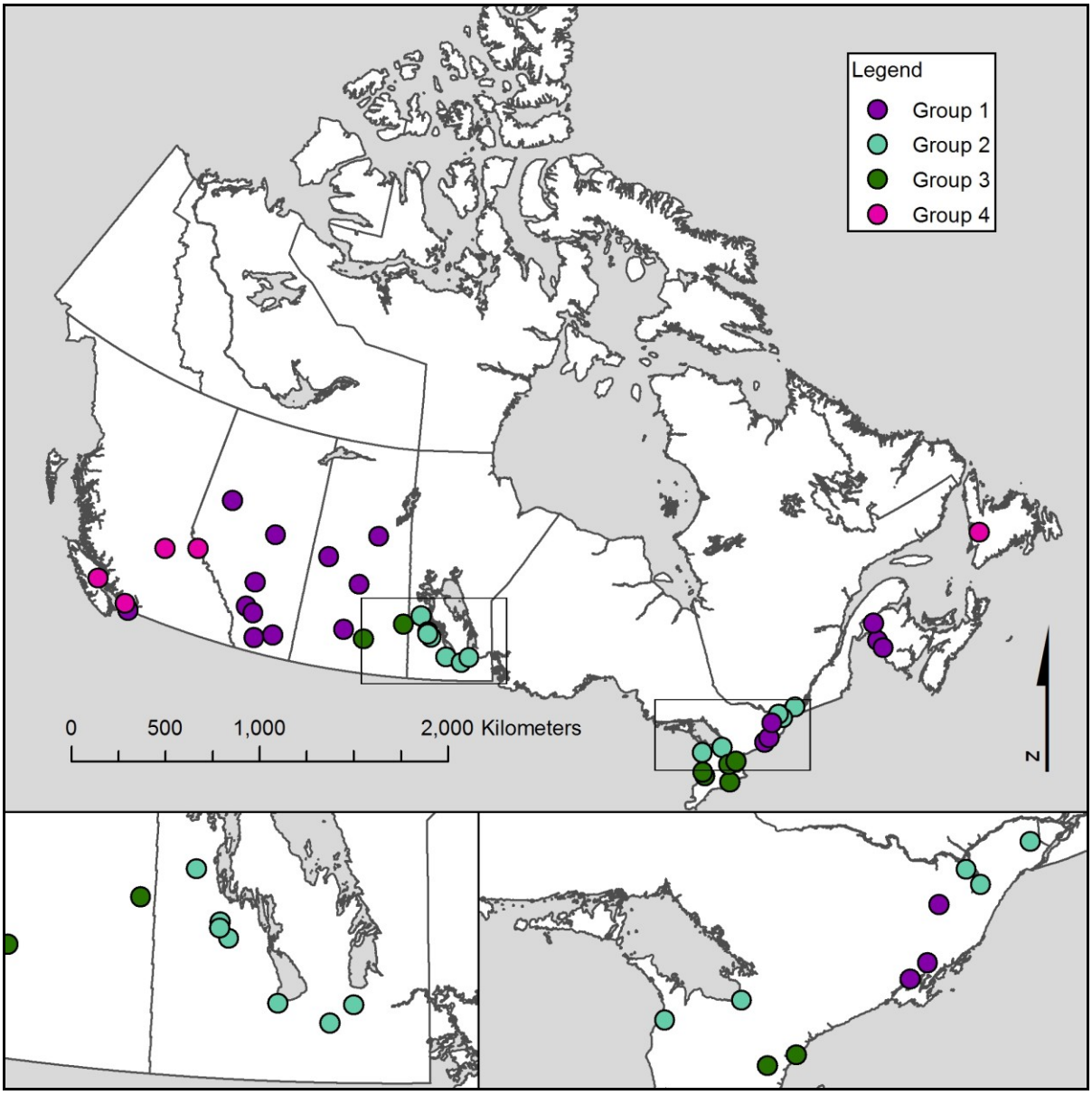



**Figure 8.** Location of sampling stations, coded by cluster membership (Matrix B)

There are similarities in how the cluster analysis of Matrix A and Matrix B data separate the sites. For example, in both analyses two sites in South and South-eastern SK are clustered in a group different than other sites in the province. Also, in ON, sites near Lake Huron are clustered in the same group as Ontario sites near the QC boundary, but different from sites located in between. However, clustering differences are also apparent throughout.

### 2.4.3 Landscape Characteristics Data Description

Landscape characteristics data were calculated for drainage areas associated with the water quality monitoring sites using readily available spatial datasets (see section Data and Methodology – Landscape Characteristics Data). Calculated values for the proportion of cropland (Crop) vary between 0 (several sites) and 0.99 (Wilmot River, PE) in Matrix A drainage areas, and between 0 and 0.98 (Ausable River, ON) in the Matrix B subset (Table 5). These surprisingly high maximum proportions do reflect actual high intensity cropland activity in those drainage areas. However, like most other calculated landscape characteristic data, they include an error associated with the resolution of the source spatial data; this error increases in importance as the scale at which calculations are made decreases. The Wilmot River and the Ausable River sites, having relatively small drainage areas in this study (45 $km^2$ and 114 $km^2$, respectively), contain less precise calculated Crop values. In both matrices, the population intensity data (Pop) vary between 0 and 3125 people/$km^2$ (Don River, ON). The number of facilities in the pollutant releases registry (per square kilometre; PtSo) ranges from 0 to 0.52 (Waterford River, NF) in Matrix A, and from 0 to 0.38 (Don River, ON) in Matrix B. Pop and PtSo data include a few sites of very high intensities compared to the low values of the majority of sites (Appendix D); however, they no longer appeared as outliers when they were ln-transformed. The number of dams per square kilometre (Dams) varies between 0 and 0.06 (de l'Achigan River, QC) in Matrix A, and between 0 and 0.04 (Oldman River, AB) in Matrix B. Values of road intensity (Road) range from 0 to 10.3 km/$km^2$ (Leary's Brook, NF) in Matrix A, and from 0 to 8.7 km/$km^2$ (Don River, ON) in Matrix B. Dams and Road data include more sites of lower intensity values (Appendix D), and were ln-transformed prior to analysis to normalize their data distribution.

Mean annual average temperatures (Temp) vary between -9.1 (Alsek River, YT) and 11.5°C (Annapolis River, NS) in Matrix A, and between -4.8 (Fraser River at Red Pass, BC) and 9.3°C (Nanticoke Creek, ON) in Matrix B (Table 5). Mean total annual precipitation (Prec) ranges between 432 (North Alouette River, BC) and 3251 mm (Dezadeas River, YT) in Matrix A, and between 484

(Beaver River, SK) and 3251 mm in Matrix B. The proportion of lakes and reservoirs in the drainage area (Lake) varies between 0 and 0.22 (St. Croix River, NB) in Matrix A, and 0 and 0.11 (Churchill River, SK) in Matrix B. The variables Prec and Lake include more sites of lower values (Appendix D) and were ln-transformed prior to analysis. Values for the density of streams (Strm) vary between 0 (Bras d'Henri River, QC; a small drainage area of 154 $km^2$ with no river segment captured at the scale of the river dataset used to calculate stream densities) and 0.51 $km/km^2$ (Waterford River, NF) in Matrix A, and between 0.04 (Assiniboine River, SK) and 0.37 $km/km^2$ (North Alouette River, BC) in Matrix B. The circularity ratio (Circ) varies between a minimum of 0.07 (Salmon River, ON) in both matrices, and a maximum of 0.63 (Leary's Brook, NF) and 0.54 (Tsolum River, BC) in Matrix A and B, respectively. The average daily discharge (excluding months of January, February and March) at the monitoring site (Q) varies between 0.4 (Bear River, PE) to 3202 $m^3$/second (Fraser River at Hope, BC) in Matrix A, and 1.2 (Ausable River, ON) to 669 $m^3$/second (Red River, MB) in Matrix B, and include a few outlier sites of very high values compared to a majority of sites with lower values (Appendix D); the Q data distribution was normalized by ln-transformation prior to analysis.

The most western site in the entire dataset is on the Alsek River, YT, the most eastern is on the Waterford River, NF, the most northern is on the Dezadeash River, YT, and the most southern is on the Nanticoke River, ON. Within the Matrix B subset, the most western site is on the Tsolum River, BC; the most eastern is on the Humber River, NF; and the most northern is on the Smoky River, AB; and the most southern is again the site on the Nanticoke River, ON.

**Table 4.** Summary statistics of landscape characteristics for sites used in the analysis of Matrix A and B (see text)

| | Matrix A | | | | Matrix B | | | |
|---|---|---|---|---|---|---|---|---|
| | Min | Max | Mean | Standard deviation | Min | Max | Mean | Standard deviation |
| Crop (unitless) | 0 | 0.99 | 0.22 | 0.29 | 0 | 0.98 | 0.36 | 0.30 |
| Pop (people/$km^2$) | 0 | 3125 | 110 | 433 | 0 | 3125 | 97 | 481 |
| PtSo (facilities/ $km^2$) | 0 | 0.52 | 0.01 | 0.06 | 0 | 0.38 | 0.01 | 0.06 |
| Dams (Dams/ | 0 | 0.06 | 0.01 | 0.01 | 0 | 0.04 | 0.01 | 0.01 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| $km^2$) | | | | | | | | |
| Road (km/ $km^2$) | 0 | 10.3 | 1.1 | 1.7 | 0.03 | 8.7 | 1.0 | 1.3 |
| Temp (ºC) | -9.1 | 11.5 | 4.0 | 4.2 | -4.8 | 9.3 | 4.3 | 3.2 |
| Prec (mm) | 432 | 3251 | 1015 | 389 | 484 | 3251 | 874 | 490 |
| Lake (unitless) | 0 | 0.22 | 0.03 | 0.04 | 0.00 | 0.11 | 0.02 | 0.02 |
| Strm (km/ $km^2$) | 0 | 0.51 | 0.15 | 0.09 | 0.04 | 0.37 | 0.12 | 0.07 |
| Circ (unitless) | 0.07 | 0.63 | 0.25 | 0.11 | 0.07 | 0.54 | 0.23 | 0.11 |
| Q ($m^3$/second) | 0.4 | 3202 | 158 | 437 | 1.2 | 669 | 82 | 147 |
| Lat (decimal degrees) | 42.8 | 60.8 | 48.6 | 3.8 | 42.8 | 55.7 | 48.9 | 3.7 |
| Long (decimal degrees) | -138.0 | -52.7 | -87.4 | 24.8 | -125.2 | -57.8 | -95.6 | 18.6 |

### 2.4.4 Constrained Analyses

#### *2.4.4.1 Matrix A*

A redundancy analysis of Matrix A water quality data found that landscape characteristics explain 54.0% (adjusted $R^2$) of the variance in the water quality data. Permutation tests determined that the redundancy model was significant at the 0.001 significance level. The first canonical axis is responsible for 83.1% of that 54.0% of variance explained; therefore the first ordination axis explains 44.9% of variance in the water quality data. All water quality variables (pH, SC and TP) are positively correlated with the first canonical axis (Figure 9). Prec and Crop are the landscape variables the most heavily loaded on this axis (negatively and positively, respectively). Several other landscape variables, including Pop, Road, PtSo, and Dams, are highly correlated within themselves, and with all three water quality variables, especially SC. Several sites located in NS, NF and NB, have negative high scores on the first axis and are correlated with Prec; sites of highest positive scores are located in SK and MB.

The second canonical axis is responsible for 9.7% of the variance in the water quality data explained by landscape characteristics (54.0%); therefore it explains 5.2% of the variance in the

water quality data. Long, Strm, Temp and Crop are the most highly loaded variables on this second axis (positive). The water quality variable TP is highly correlated with Crop and other variables indicating land use, including Pop, Road, PtSo, and Dams, and is negatively correlated with Lake. The variable pH is negatively correlated with Prec. A number of sites located in BC have negative high scores on the second axis, are positively correlated with Lat, and are negatively correlated with Temp. Two sites from PE have the highest positive scores on the second axis.

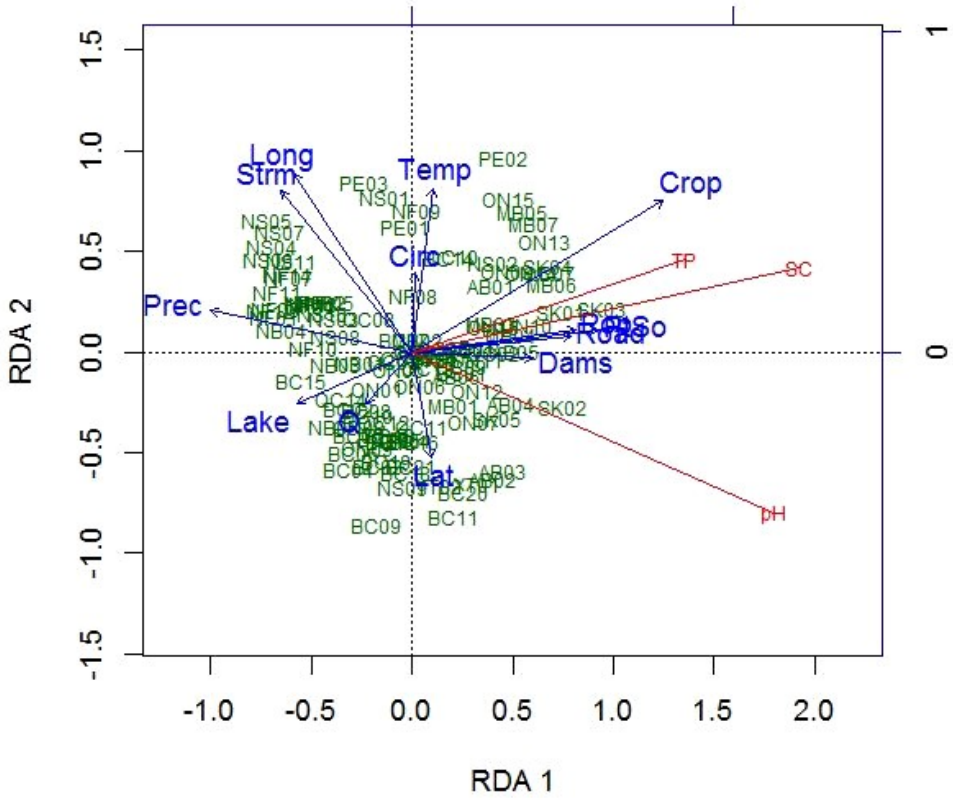



**Figure 9.** Redundancy analysis on Matrix A, presenting correlations with the first and second canonical axes (RDA 1 and RDA 2), for the response water quality variables (red arrows), the explanatory landscape variables (blue arrows), and the fitted sites scores (green); RDA 1 explains 44.9% of the variance in water quality and RDA 2 explains 5.2%

Forward selection was applied in order to attempt to identify landscape variables that explain the most variance in the water quality data. The method identifies only two variables, Crop and Prec;

these two produce a model with an adjusted $R^2$ above that of the full model. A model with only Crop and Prec explaining variance in water quality was also determined to be significant at the 0.001 level with permutation tests.

Crop and Prec may be correlated because agricultural activities are primarily practiced where climatic conditions are favourable. To further explore how different types of correlated landscape variables explain the variance in the water quality data, the landscape dataset was partitioned, i.e. its variables were assigned to two categories. One partition included "natural" variables, composed of Temp, Prec, Lake, Strm, Circ, Lat, Long and Q, and the other was composed of the anthropogenic variables Crop, Pop, PtSo, Road, and Dams. The variation partitioning method identifies 14.6% of the variance in water quality data as being explained by the natural portion (not correlated with the anthropogenic portion), 16.3% by anthropogenic portion (not correlated with the natural portion), and 35.5% explained by the shared portion of correlated natural and anthropogenic variables. Models with individual uncorrelated natural and anthropogenic portions were determined significant at the 0.001 level with permutation tests.

#### *2.4.4.2 Matrix B*

A redundancy analysis of Matrix B water quality indicates that 48.7% (adjusted $R^2$) of the variance in the water quality data is explained by landscape characteristics. The redundancy model was determined significant at the 0.001 level. The first canonical axis is responsible for 63.0% of the variance in water quality data explained by landscape characteristics (48.7%); therefore, the first axis explains 30.7% of the variance in the water quality data. All water quality variables but DO and WT are positively correlated with the first axis (Figure 10). As it was the case in the Matrix A analysis, Prec and Crop are the landscape variables with the highest loadings (negative and positive, respectively) on the first axis. Road and Pop also have important loadings (positive). Four sites located in BC and one in NF have very highly negative scores on the first axis, and are correlated with Strm, Lake, and discharge (Q), and negatively correlated with Crop and other anthropogenic

disturbance variables (Pop, Road, PtSo and Dams). Sites with highest positive scores on the first axis are located in SK, MB and ON.

The second canonical axis accounts for 15.3% of the variance in the water quality data explained by landscape characteristics; it explains 7.5% of the variance in the water quality data. WT is the only water quality variable correlated (positive) with axis 2. Lat (negative), and Long and Temp (positive), are the most highly loaded landscape variables on the second axis. Sites of highest negative scores on the second axis are located in MB and BC; sites located in ON and NB have the highest positive scores.

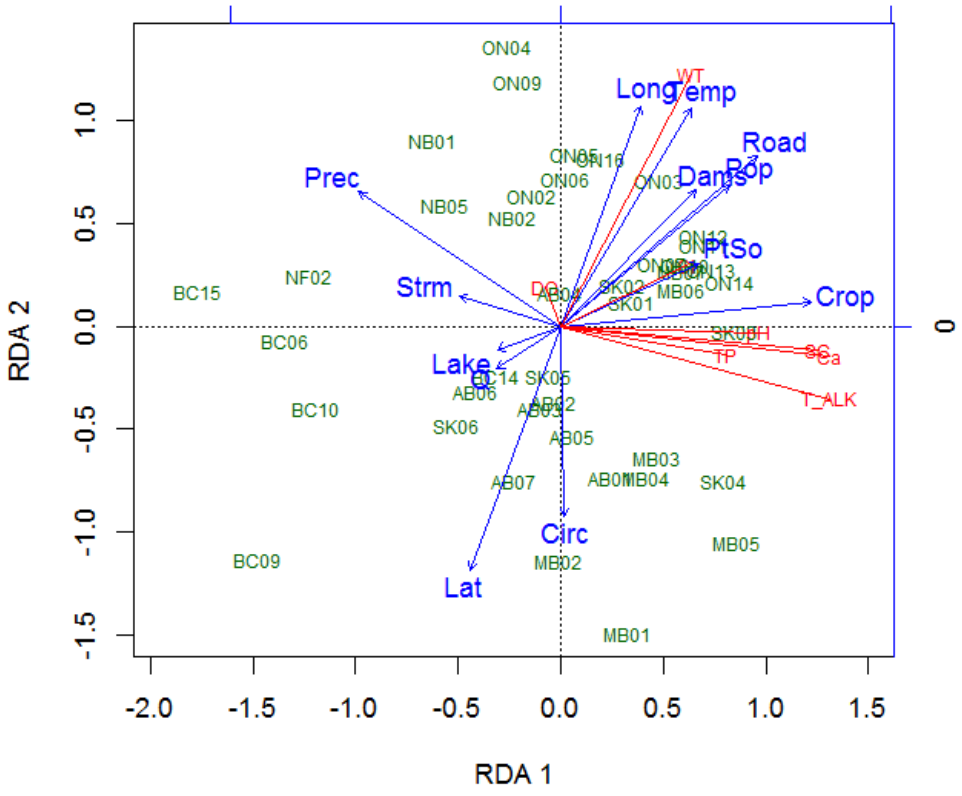



**Figure 10.** Redundancy analysis on Matrix B, presenting correlations with the first and second canonical axes (RDA 1 and RDA 2), for the response water quality variables (red arrows), the explanatory landscape variables (blue arrows), and the fitted sites scores (green); RDA 1 explains 30.7% of the variance in water quality and RDA 2 explains 7.5%

Forward selection identifies three variables, Crop, Prec, and Road as the most important landscape characteristics explaining the variance in the water quality. A model with only Crop, Prec and Road explaining variance in water quality was determined significant at the 0.001 level with permutation tests.

The Matrix B landscape dataset was partitioned into a set of natural variables and a set of anthropogenic variables. The variation partitioning determines that 18.0% of the variance in water quality data is explained by the natural portion (uncorrelated with the anthropogenic portion), 17.3% by the anthropogenic portion (uncorrelated with the natural portion), and 26.7% is explained by the shared portion of correlated natural and anthropogenic variables. Models with individual uncorrelated natural and anthropogenic portions explaining variance in water quality were determined significant at level 0.05 and 0.01 respectively.

## 2.5 Discussion

Water quality data from various monitoring programs across Canada were integrated into one dataset, and averaged across time to yield one value for each parameter at each monitoring site. Two overlapping matrices of water quality data were analyzed. The first matrix (Matrix A) had a geographic coverage consisting of 107 sites located throughout Canada, but few parameters; it included data for pH, specific conductance (SC) and total phosphorus (TP). The second matrix (Matrix B) included more water quality variables but only for a subset of 42 sites present in Matrix A; it included data for calcium (Ca), chloride (Cl), total alkalinity (T_ALK), dissolved oxygen (DO), water temperature (WT), and pH, SC and TP. In addition, several landscape characteristics were calculated for the drainage areas of each site: proportion of cropland (Crop), population density (Pop), density of facilities registered in a pollutant release and transfer register (PtSo), dams density (Dams), road density (Road), mean annual average temperature (Temp), mean annual total precipitation (Prec), lake and reservoir proportion (Lake), stream density (Strm), and circularity ratio (Circ). The following

site information was also compiled as landscape characteristics: average daily discharge at the site (Q), latitude (Lat) and longitude (Long). Both water quality matrices were analyzed with principal component analysis (PCA) and cluster analysis to uncover structure and patterns in the data, and they were analyzed in relation to landscape characteristic variables with redundancy analysis.

The PCA of the first matrix produced one significant principal component, which was most highly correlated with SC. The PCA on the second matrix produced two significant principal components, and again SC was the variable the most correlated with the first component. The fact that both matrices, together combining data for a large geographic coverage and for eight water quality parameters, produced this finding suggests that SC is the most representative parameter of water quality variance at the country scale. This is not surprising, as SC is a variable that is a general index of dissolved ions. It is also a convenient finding, as SC is easy to measure accurately, and automated monitoring instruments may be deployed in the river at selected locations and take regular (e.g. hourly) measurements of SC. Given a proper communication system in place, this data is accessible remotely by water quality monitoring officers. For example, the province of Newfoundland and Labrador has an extensive automated monitoring network (Government of Newfoundland and Labrador, 2014).

Identifying the group membership, derived from the cluster analysis, on the plot of the PCA scores proved to be a useful method to indicate the major water characteristics defining each group, and how similar or different the characteristics are within and between each group. For example, in the analysis of Matrix A data, sites of the fifth group (Figure 4) had very similar PCA scores attributable to very low pH, and these scores were most distant from those of the other groups of sites. According to this analysis, these sites, located in Nova Scotia, have a most distinct and localized set of water quality values.

Mapping the cluster analysis group membership provided information on the spatial distribution of sites of similar water quality and could be used to infer the probable environmental

influences on them. In the analysis of Matrix A, sites of the third and fourth group, with higher SC, were located in the Prairies and around the Great Lakes and St. Lawrence River (Figure 5). Several factors may be responsible for the elevated SC in this zone. First, the lower precipitation in the central portion of the country leads to a higher proportion of groundwater input to the river, and this is high in dissolved ions. Second, the sedimentary bedrock underlying most of the Prairies is more susceptible to weathering and dissolution, leading to more ions in the river water. Third, cropland activity, which is more important across this area than to the east or west increases the runoff of sediments to waterways, and this would also contribute to elevated SC. Sites of Groups 2 and 5, with lower SC, were located in regions with higher precipitation (Atlantic provinces and British Columbia), explaining, at least in part, the more dilute water. Finally, sites of Group 1, with moderate SC, were scattered across the country, and may then reflect environmental pressures occurring throughout Canada, such as urbanization. According to the landscape characteristics data, several sites in group 1 have higher population densities. This confirms the usefulness of cluster analysis, when used along with PCA and mapping, in identifying sites impacted by anthropogenic pressure, also demonstrated in similar studies (Shrestha and Kazama, 2007; Simoneau, 1985; Varol and Sen, 2009; and Zhang *et al* 2009).

The cluster analysis of Matrix B (wider spectrum of water quality parameters) produced more regionally defined cluster groups (Figure 8). Sites with higher SC were grouped according to values of DO and Cl; Group 2 had low DO concentrations (in MB and ON) associated with high levels of organic material or slow moving water, and Group 3 had higher Cl (in SK and Southern ON), which enters the water from the bedrock, agricultural activities or wastewaters (industries and municipalities). The addition of DO and Cl data to the analysis (including SC, pH and TP data), proved effective in uncovering supplementary structure in the water quality data. Further analyses including more water quality variables should be conducted nationally in order to identify other variable that provides additional information on the water quality variance. A monitoring standard of a core set of water quality variables needs to be implemented nation-wide to permit such future analysis. The establishment of a consistent national database of water quality data would greatly

facilitate any nation-wide future work. In the mean time, new analyses at smaller scale (e.g. major drainage areas, or ecological regions) would enable the inclusion of data from more parameters and sites, and would provide additional regionally significant information on water quality variance, while nonetheless relevant nationally.

Redundancy analyses on both matrices produced significant predictive models of water quality data and suggest that landscape characteristics are important influences to water quality at the country scale. In both analyses, the proportion of cropland and the mean annual total precipitation were the first and second landscape variables with the highest loadings on the first canonical axis, suggesting they are the most important landscape variables in explaining variance in water quality in Canada.

Precipitation may impact the quality of surface water several ways. Precipitation reduces the SC as it brings “new” low-conductance water to the rivers and consequently dilutes the base flow water that carries the ions from the bedrock and surficial geology. Precipitation runoff may carry pollutants and sediments from the land to the water. In areas where the air is polluted with acidic particles, precipitation leads to wet acid deposition which may reduce the pH in surface waters.

Agricultural croplands are a diffuse source of surface water pollution. They typically have higher erosion rates due to their lack of vegetation cover, and fertilizers and pesticides are commonly applied. Together, these factors cause cropland run-off to carry and leach sediments, nutrients and other pollutants to waterways, thus increasing turbidity, suspended sediments concentrations, nutrient concentrations (such as total phosphorus), SC, and other pollutants, in the receiving water. Cropland pollution may be alleviated by better management practices such as the implementation of nutrient management plans, crop rotation, and riparian buffers. However, as opposed to pollution point sources like wastewater treatment plants and industries where wastewater pollutants may be treated before being released, direct pollutant release treatment is not applicable with regards to cropland pollution, which may explain why it would be so correlated with variances of water quality.

Certain difficulties were encountered during the course of this study. Assembling a consistent dataset of water quality data across Canada was a challenge due to the unevenness of the monitoring programs in place. A majority of water quality parameters included in the datasets provided by the monitoring agencies were monitored only in parts of the country, which led to the strategy of analyzing the data in two ways, maximizing sites or variables. Nonetheless, data for several sites and parameters had to be excluded from the analysis. In the second part of the study, water quality was studied in relation to landscape characteristics; given the exploratory role of this part of the study, the landscape variables calculated were limited to those that could be generated from readily available spatial information. Therefore, the resolution of the spatial data used in landscape characteristics calculations was not always optimal, introducing a level of error in the calculated parameters that increased in the smaller size drainage areas. Certain landscape characteristics known to impact water quality were not included in this study but could provide important supplementary information. One would be bedrock information which may play a role in explaining elevated levels of SC in the Canadian Prairies, where easily weathered sedimentary rocks dominate the geology. Other variables of interest for further analysis would include terrain characteristics, and distance-based measures of anthropogenic disturbance (e.g. sum of releasing facilities weighted according to their distance to the water monitoring site).

In summary, despite a limited dataset, multivariate analyses uncovered structure in the water quality data from monitoring location across Canada, and have identified how landscape characteristic is related to this structure. Specific conductance (SC) was the water quality parameter the most representative of water quality variance across Canada. Redundancy analysis identified proportion of cropland area and mean total annual precipitation as the most important landscape characteristics explaining variance in water quality. In a context where water quality monitoring resources are sparse and need to be allocated most judiciously, these multivariate techniques may be able to provide simple models that allow the prediction of water quality characteristics from

unmonitored sites based on aspects of the watershed, although further studies would be needed to verify these results.

## 3.0 Conclusion

Multivariate analysis methods were used to summarize data and explore relationships between physical-chemical lotic water quality data in a Canada-wide study. Water quality data from sites across Canada were analyzed with principal component analysis (PCA), and cluster analysis, and their relationship with landscape characteristics were explored with redundancy analysis.

The water quality datasets analyzed were limited due to the heterogeneity of water quality monitoring programs in Canada. One water quality dataset had more sites (107) but included few parameters (specific conductance, pH and total phosphorus); the second dataset had fewer sites (42) and more parameters (calcium, chloride, total alkalinity, dissolved oxygen, water temperature, specific conductance, pH, and total phosphorus). The first principal component from both PCA analyses was correlated most highly with specific conductance, suggesting this parameter is the most representative of water quality variance at the country scale. Overlaying cluster analysis results on PCA information proved an excellent means to identify the major water characteristics defining each group; mapping cluster analysis group membership provided information on their spatial distributions. Redundancy analysis results produced significant predictive models of water quality demonstrating that landscape characteristics are determinant factors in water quality at the country scale. The proportion of cropland and the mean annual total precipitation were the first and second landscape variables with the highest loadings on the canonical axes of most variance explained.

A major hurdle in the analysis of nation-wide water quality data is the lack of commonality in the data generated by the various programs responsible for water quality monitoring in Canada. The integration of multiple datasets for analysis purposes is challenging and variables present across all datasets are extremely limited. This leads analysts to leave out of their studies data for several

parameters and sites, and conduct their analysis on fairly limited datasets. As such, one recommendation of this study is for the implementation of a nation-wide standard for the water quality monitoring of a core set of water quality variables, which was already proposed by the Canadian Council of Ministers of the Environment (CCME) in 2006. The CCME is an intergovernmental forum where working groups composed of experts from the federal, provincial and territorial governments, work together and take action on environmental matters of national and international concerns (CCME, 2014b). It is through the CCME water management committee that the protocols manual for water quality sampling in Canada (CCME, 2011) was put together, aiming to standardize procedures for field water quality sampling across jurisdictions. In 2006, this group published a document entitled *A Canada-wide Framework for Water Quality Monitoring* (CCME, 2006), calling for greater coordination among jurisdictions to support a Canada-wide network for monitoring sites of national, regional and local interest, and suggesting the monitoring of a minimum standard set of core key variables for each water uses (protection of aquatic life, industrial uses, agricultural uses, recreational uses and suitability of source water for drinking water supply). Core set of variables suggested for monitoring activities for the protection of aquatic life (which provides the data of most relevance in the study of interactions between water quality and the environment) were specific conductance, pH, dissolved oxygen, temperature, turbidity and nutrients. As multiple forms of nutrients exist, precise nutrient parameters should also be established in a minimum standard core set of monitoring variables. Total phosphorus would be one as it is already monitored throughout the country. Nitrogen is monitored in various forms across Canada. At least one form of nitrogen should also be selected and included in a standard core set of monitoring variables due to its important role in ecosystem productivity. A longer list of core monitoring variables could include major ions as well as selected metals. Ions are the controlling factors of the SC, and a more detailed study could indicate which ions are important sources of variance in different regions of Canada. Metals with detectable levels and set guidelines for the protection of aquatic life would be the most relevant in providing supplementary information. The establishment and implementation of a minimum standard core set of monitoring variables across Canada would be possible under the responsibility of the water management

committee of the CCME. In addition, the establishment of a nation-wide water quality database would greatly facilitate any future country-scale water quality studies in Canada.

Until such time as a consistent and national dataset becomes available, new multivariate analysis of water quality data should be conducted at smaller scales. Analysis at the scale of major drainage areas or ecological regions would enable the inclusion of data from more parameters and sites, and would provide more regionally significant information, while nonetheless relevant for a national monitoring network. Constrained analyses of water quality data and landscape features are promising and should be pursued. The inclusion of land cover information from 30-meter resolution satellite imagery, geology information, terrain, climate, and sources of anthropogenic pressures (such as population, point sources and roads) are expected to provide a reliable description of the role of landscape characteristic features in explaining water quality across Canada. This activity would enable the development of simple models allowing the prediction of water quality characteristics from unmonitored sites based on aspects of the watershed, and support decision making for water resources managers.

## References


Alberta Environment and Sustainable Resource Development. 2014. Surface Water Quality Data [Tabulate dataset]. Retrieved in March 2014 from http://esrd.alberta.ca/water/reports-data/surface-water-quality-data/default.aspx

Allaby, A., Allaby, M. 1999. A Dictionary of Earth Sciences. Retrieved in December 2014 from http://www.encyclopedia.com/doc/1O13-drainagebasinshapeindex.html

Arslan, O. 2009. A GIS-Based Spatial-Multivariate Statistical Analysis of Water Quality Data in the Porsuk River, Turkey. Water Qual. Res. J. Can. 44, 3: 279-293.

Arslan, O. 2013. Spatially Weighted Principal Component Analysis (PCA) Method for Water Quality Analysis. Water Resources. 40, 3: 315-324.

Beyer, H. L. 2009-2012. Geospatial Modelling Environment version 0.7.2.1 [Software]. Copyright (c) Hawthorne L. Beyer 2009-2012 Spatial Ecology LLC, http://www.spatialecology.com/gme/gmelicense.htm

Borcard, D., Gillet, F., Legendre, P. 2011. Numerical Ecology with R. Springer. New York. 306 pp.

Brunsdon, C., Fotheringham, S., Charlton, M. 1998. Geographically weighted regression – modeling spatial non-stationarity. The Statistician, 47: 431-443.

Bu, H. M., Tan, X., Li, S. Y., Zhang, Q. F. 2010. Water quality assessment of the Jinshui River (China) using multivariate statistical techniques. Environmental Earth Sciences, 60: 1631-1639.

Canadian Council of Ministers of the Environment. 2006. A Canada-wide Framework for Water Quality Monitoring. 25 pp. Retrieved from http://www.ccme.ca/en/resources/water/water_quality.html

Canadian Council of Ministers of the Environment. 2011. Protocols Manual for Water Quality Sampling in Canada. Retrieved in January 2015 from http://www.ccme.ca/en/resources/water/water_quality.html

Canadian Council of Ministers of the Environment. 2014a. Canadian Environmental Quality Guidelines. http://ceqg-rcqe.ccme.ca/en/index.html#void

Canadian Council of Ministers of the Environment. 2014b. About. http://www.ccme.ca/en/about/index.html

Chang, H. 2008. Spatial analysis of water quality trends in the Han River basin, South Korea. Water Research, 42: 3285-3304.

Commission for Environmental Cooperation. 1997. Ecological Regions of North America Toward a Common Perspective, Montreal, Quebec, 71 pp.

Commission for Environmental Cooperation. 2010. Land Cover [Grid dataset]. Retrieved in 2014 from http://www.cec.org/Page.asp?PageID=122&ContentID=25740&SiteNodeID=498&BL_ExpandID=

Dray, S. 2013. Forward Selection with permutation Version 0.0-8 [Software package]. Fetrieved from https://r-forge.r-project.org/R/?group_id=195

Environment Canada. n.d. Water Survey of Canada HYDAT Database [Tabulate dataset]. Retrieved in April 2014 from https://www.ec.gc.ca/rhc-wsc/default.asp?lang=En&n=9018B5EC-1

Environment Canada. 2009. National Pollutant Release Inventory [Tabulate dataset]. Retrieved from http://takingstock.cec.org/QueryBuilder.aspx?varlan=en-US#report=Facility|year=2009|chemicalsmedia=TotalReleaseTransfers|naics=3|menu=adv

Environment Canada. 2010. Shared Responsibility. Retrieved from http://www.ec.gc.ca/eau-water/default.asp?lang=En&n=035F6173-1

Environment Canada. 2012. Partners and Agreements. Retrieved from http://www.ec.gc.ca/eaudouce-fresh-water/default.asp?lang=En&n=0948A6DA-1

Environment Canada. 2013a. Canada Water Act Annual Report for April 2011 to March 2012. Retrieved from http://www.ec.gc.ca/eau-water/default.asp?lang=En&n=9E0D4BC1-1&offset=3&toc=show#X-2012103008545431

Environment Canada. 2013b. About the Water Survey of Canada. Retrieved in January 2015 from http://www.ec.gc.ca/rhc-wsc/default.asp?lang=En&n=EDA84EDA-1

Environment Canada. 2014. Federal Policy and Legislation. Retrieved from http://www.ec.gc.ca/eau-water/default.asp?lang=En&n=E05A7F81-1

Environment Canada Canadian Centre for Climate Modelling and Analysis. n.d. CanRCM4 Model Output / ARC-22 and NAM-44 CCCma-CanESM2 historical / monthly / atmosphere / average surface temperature and precipitation / 196101-197012, 197101-198012 and 198101-199012

[Data files]. Retrieved in 2014 from http://www.cccma.ec.gc.ca/data/canrcm/CanRCM4/index_cordex.shtml.

Environment Canada. Meteorological Service of Canada. 2014. Watershed Boundaries for Hydrometric Gauges operated by Water Survey of Canada [Shapefiles]. 401 Burrard St. Vancouver. British Columbia. V6C 3S5.

Environment Canada Water Quality Monitoring and Surveillance. 2014. Science and Technology/Water Science and Technology. Gatineau. Quebec. K1A 0H3.

Environmental Protection Agency. 2009. Release Chemical report [Tabulate dataset]. Retrieved from http://takingstock.cec.org/QueryBuilder.aspx?varlan=en-US#report=Facility|year=2009|chemicalsmedia=TotalReleaseTransfers|naics=3|menu=adv

Environmental Protection Agency. 2012. Water: Monitoring &Assessment; 5.4 pH. Retrieved in April 2015 from http://water.epa.gov/type/rsl/monitoring/vms54.cfm

Environmental Protection Agency. 2012b. Water: Monitoring &Assessment; 5.9 Conductivity. Retrieved in April 2015 from http://water.epa.gov/type/rsl/monitoring/vms59.cfm

Esri. 2012. ArcGIS 10.1 SP1 for Desktop [Software]. Copyright © 1999-2012 Esri Inc.

Everitt, B. 2011. Cluster Analysis. Wiley series in probability and statistics. 330 pp.

Geobase. 2008-2012. 50k_dem [Grid dataset]. Retrieved in 2014 from ftp://ftp2.cits.rncan.gc.ca/pub/geobase/official/cded/50k_dem/

Geobase. 2011-2014. National Hydro Network (NHN) [Geodatabase files]. Retrieved in 2014 from http://www.geobase.ca/geobase/en/data/nhn/index.html;jsessionid=FD3D75E2ADB6C38122E80EC1BB605E3B.geobase1

Government of Newfoundland and Labrador. 2014. Real Time Water Quality Monitoring Program. http://www.env.gov.nl.ca/env/waterres/rti/rtwq/index.html

Harden, S.L., Cuffney, T.F., Terziotti, Silvia, Kolb, K.R. 2013. Relation of watershed setting and stream nutrient yields at selected sites in central and eastern North Carolina, 1997–2008: U.S. Geological Survey Scientific Investigations Report 2013–5007. 47 pp. Retrieved from: http://pubs.usgs.gov/sir/2013/5007/

Helsel, D. R., Hirsch, R. M. 2002. Statistical Methods in Water Resources. USGS. 510 pp. Retrieved from: http://pubs.usgs.gov/twri/twri4a3/

Legendre, P., Fortin, M.-J. 1989. Spatial pattern and ecological analysis. Vegetation, 80: 107-138.

Manitoba Conservation and Water Stewardship. Water Quality Management Section. 2014. 123 Main Street, Suite 160. Winnipeg. Manitoba. R3C 1A5.

Mazvimavi, D., Burgers, S. L. G. E., Stein, A. 2005. Identification of basin characteristics influencing spatial variation of river flows. International Journal of Applied Earth Observation and Geoinformation, 8: 165-172. DOI: 10.1016/j.jag.2005.08.006.

Moerke, A. H., Lamberti, G. A. 2006. Scale-dependent influences on water quality, habitat, and fish communities in streams of the Kalamazoo River Basin, Michigan (USA). Aquat. Sci. 68: 193-205.

Natural Resources Canada. 2003. Atlas of Canada 1,000,000 National Frameworks Data, Hydrology – Dams [Shapefile]. Retrieved from http://geodiscover.cgdi.ca/wes/RecordSummaryPage.do?uuid=0D4E0553-0FEC-F9F6-F207-710FECA17DC9&recordLocale=en_US&view=summary&entryPoint=jsMap&mode=unmappable

Natural Resources Canada. 2006. pub/canvec/canada_fgdb/canvec_gdb_CA_HD.zip [Geodatabase]. Retrieved in 2014 from ftp://ftp2.cits.rncan.gc.ca/pub/canvec/. Documentation: Natural Resources Canada. 2011. CanVec Feature Catalogue, edition 1.1.2. Retrieved from ftp://ftp2.cits.rncan.gc.ca/pub/canvec/doc/

Natural Resources Canada. 2009. Atlas of Canada 1,000,000 National Frameworks Data, Hydrology, Version 6.0 [Shapefiles]. Retrieved in September 2012 from ftp://ftp2.cits.rncan.gc.ca/pub/geott/frameworkdata/drainage_areas/

Office of the Auditor General of Canada. 2010. 2010 Fall Report of the Commissioner of the Environment and Sustainable Development. Retrieved from http://www.oag-bvg.gc.ca/internet/English/parl_cesd_201012_02_e_34425.html#hd5f

Oksanen, J. Blanchet, F. G., Kindt, R., Legendre, P., Minchin, P. R., O'Hara, R. B., Simpson, G. L., Solymos, Stevens, M. H. H., Wagner, H. 2013. Vegan Version 2.0-10 [Software package]. http://cran.r-project.org/web/packages/vegan/index.html

Ontario Ministry of the Environment and Climate Change. n.d. pwqmn_raw_data_2002_2011.mdb [Tabulate dataset]. Retrieved in March 2014 from http://www.ene.gov.on.ca/environment/en/resources/collection/data_downloads/index.htm.

Peters, D. L., Atkinson, D., Monk, W. A., Tenenbaum, D. E., Baird, D. J. 2013. A multi-scale hydroclimatic analysis of runoff generation in the Athabasca River, western Canada. Hydrol. Process. 27: 1915-1934. DOI: 10.1002/hyp.9699

Quebec Ministère du Développement durable, Environnement, et Lutte contre les changements climatiques. 2014. 675 René-Lévesque Est, 29th floor. Quebec. Quebec. G1R 5V7

Qian, S.S. 2010. Environmental and ecological statistics with R. Chapman & Hall/CRC applied environmental statistics. United States of America. 421 pp.

R Foundation for Statistical Computing. 2013. [Software] http://www.r-project.org/

Reeder, S. W., Hitchon, B., Levinson, A. A. 1972. Hydrochemistry of the surface waters of the Mackenzie River basin, Canada-I. Factors controlling inorganic composition. Geochimica et Cosmochimica Acta. 36: 825-865.

Rencher, A. C. 2002. Methods of Multivariate Analysis, Second Edition. Wiley-Interscience. Canada. 708 pp.

Ribeiro, L., Kretschmer, N., Nascimento, J., Buxo, A., Rötting, T. S., Soto, G., Soto, M., Oyarzún, J., Maturana, H., Oyarzún, R. 2014. Water quality assessment of the mining-impacted Elqui River basin, Chile. Mine Water Environ. Springer-Verlag Berlin Heidelberg. DOI: 10.1007/s10230-014-0276-6

Saskatchewan Water Security Agency. 2014. 400 - 111 Fairford St. E. Moose Jaw. Saskatchewan. S6H 7X9.

Shrestha, S., Kazama, F. 2007. Assessment of surface water quality using multivariate statistical techniques: A case study of the Fuji River basin, Japan. Environmental Modelling & Software, 22: 464-475.

Simoneau, M. 1985. Spatial Variability in the Water Quality of Québec Rivers. Ministère de l'environnement du Québec, 17 pp.

Singh, K. P., Malik, A. Mohan, D., Sinha, S. 2004. Multivariate statistical techniques for the evaluation of spatial and temporal variations in water quality of Gomti River (India) – A case study. Water Research, 38: 3980-3992.

Stanfield, L.W., Kilgour, B.W. 2006. Effects of Percent Impervious Cover on Fish and Benthos Assemblages and Instream Habitats in Lake Ontario Tributaries. American Fisheries Society Symposium, 48: 577-599.

Statistics Canada. 2011. Road Network File, 2011 Census. Statistics Canada Catalogue no. 92-500-X [Shapefile]. Retrieved from http://www12.statcan.ca/census-recensement/2011/geo/RNF-FRR/index-eng.cfm

Statistics Canada. 2011. Population Census 2011 [Geodatabase]. 150 Tunney's Pasture Driveway. Ottawa. Ontario. K1A 0T6

Tu, J. 2011. Spatially varying relationships between land use and water quality across an urbanization gradient explored by geographically weighted regression. Applied Geography, 31: 376-392.

U.S. Army Corps of Engineers. 2005. National Atlas of the United States. Major Dams of the United States [Shapefile]. Retrieved from http://nationalatlas.gov/mld/dams00x.html

U.S. Census Bureau. n.d.(a). 2010 TIGER/Line Shapefiles United States Census Bureau Road Network, census 2010 [Shapefiles]. Retrieved from http://www.census.gov/cgi-bin/geo/shapefiles2010/main

U.S. Census Bureau. n.d.(b). 2010 TIGER/Line Census Tracts [Shapefile]. Retrieved from http://www.census.gov/geo/maps-data/data/tiger-data.html

United States Census Bureau. 2012. County totals: Vintage 2011; Census year 2010 [Tabulate dataset]. Retrieved from http://www.census.gov/popest/data/counties/totals/2011/index.html

USGS. n.d. NHDDamEvents_fgdb_6_28_2013.zip; WBD_National.zip [Geodatabase files]. Retrieved in 2013 from http://nhd.usgs.gov/

Varol, M. Sen, B. 2009. Assessment of surface water quality using multivariate statistical techniques: a case study of Behrimaz Stream, Turkey. Environmental Monitoring and Assessment, 159: 543-553.

Zhang, Q., Li, Z. W., Zeng, G. M., Li, J. B. Fang, Y. Yuan, Q. S., Wang, Y., Ye, F. 2009. Assessment of surface water quality using multivariate statistical techniques in red soil hilly region: a case study of Xiangjiang watershed, China. Environmental Monitoring and Assessment, 152 (1-4): 123-131.

Zhao, Z., Cui, F. 2008. Multivariate statistical analysis for the surface water quality of the Luan River, China. J Zhejiang Univ Sci A, 10, 1: 142-148.

## Appendix A. Average water quality data

*Calculations described in the Data and Methodology section

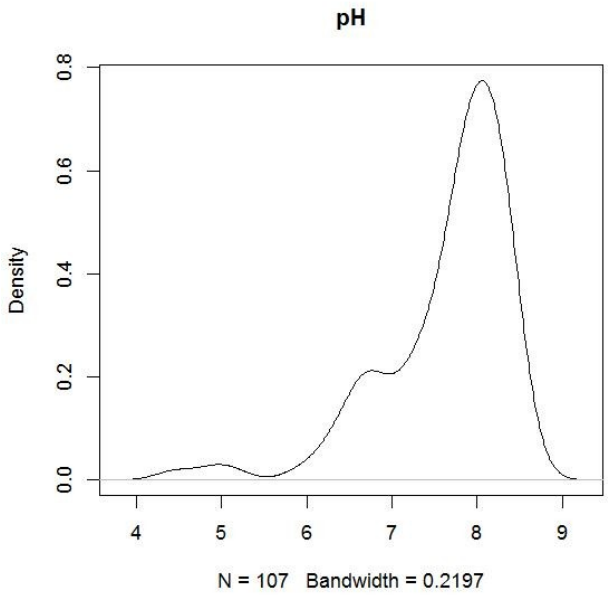
| Site ID | River Name | pH | SC (µS/cm) | TP (mg/L) | Ca (mg/L) | Cl (mg/L) | DO (mg/L) | T_ALK (mg/L) | WT (°C) |
|---|---|---|---|---|---|---|---|---|---|
| AB01 | Oldman River | 8.28 | 349.9 | 0.054 | 40.22 | 2.52 | 10.62 | 143.33 | 11.39 |
| AB02 | Elbow River | 8.21 | 397.9 | 0.012 | 53.03 | 14.38 | 11.4 | 148.06 | 9.7 |
| AB03 | Bow River below Carseland Weir | 8.02 | 378.4 | 0.028 | 46.94 | 14.34 | 11.04 | 136.94 | 8.95 |
| AB04 | Bow River at Ronalane | 8.22 | 418.4 | 0.05 | 47.67 | 11.84 | 10.89 | 140.56 | 11.36 |
| AB05 | Red Deer River | 8.03 | 363 | 0.037 | 50.94 | 3.09 | 10.79 | 171.11 | 8.77 |
| AB06 | Athabasca River | 8.01 | 289.5 | 0.059 | 36.77 | 2.61 | 10.8 | 120.72 | 9.04 |
| AB07 | Smoky River | 8.11 | 301.3 | 0.149 | 40.69 | 2.73 | 11.11 | 123.84 | 9.09 |
| BC06 | Tsolum River | 7.13 | 30.4 | 0.005 | 3.17 | 1.51 | 11.56 | 9.79 | 8.87 |
| BC09 | Fraser River at Red Pass | 7.89 | 139.4 | 0.003 | 17.26 | 0.59 | 10.03 | 53.63 | 6.66 |
| BC10 | Horsefly River | 7.79 | 99.4 | 0.019 | 14.29 | 0.6 | 11.1 | 45.37 | 8.5 |
| BC14 | Sumas River | 8.03 | 322.5 | 0.118 | 21.38 | 12.89 | 8.98 | 117.63 | 11.48 |
| BC15 | North Alouette River | 6.74 | 16.4 | 0.017 | 1.66 | 1.17 | 10.94 | 4.11 | 9.1 |
| MB01 | North Duck River | 8.27 | 491.7 | 0.04 | 72.78 | 1.69 | 9.62 | 261.69 | 7.22 |
| MB02 | Ochre River | 8.38 | 513.7 | 0.083 | 74.17 | 2.52 | 9.62 | 256.14 | 8.23 |
| MB03 | Valley River | 8.37 | 576 | 0.073 | 72.42 | 7 | 9.49 | 239.93 | 9.79 |
| MB04 | Vermilion River | 8.27 | 631.4 | 0.141 | 90.14 | 7.01 | 9.12 | 276.04 | 9.41 |
| MB05 | Whitemud River | 8.27 | 633.3 | 0.176 | 68.24 | 24.32 | 8.36 | 249 | 10.86 |
| MB06 | Assiniboine River | 8.39 | 895.7 | 0.38 | 80.65 | 27.11 | 8.33 | 260.88 | 12.85 |
| MB07 | Red River | 8.31 | 775.2 | 0.404 | 72.22 | 30.43 | 8.81 | 228 | 12.77 |
| NB01 | Aroostook River | 7.87 | 130.9 | 0.022 | 20.06 | 4.34 | 10.66 | 49.39 | 14.17 |
| NB02 | Big Presque Isle Stream | 8.28 | 303.9 | 0.013 | 53.62 | 9.86 | 12.14 | 127.9 | 13.13 |
| NB05 | Restigouche River | 8.1 | 175.9 | 0.006 | 31.89 | 1.32 | 10.68 | 83.72 | 13.94 |
| NF02 | Humber River | 6.94 | 41.9 | 0.004 | 4.27 | 3.68 | 11.81 | 10.53 | 9.74 |
| ON02 | Delisle River | 8.34 | 408.7 | 0.039 | 65.97 | 18.06 | 8.69 | 176.49 | 13.68 |
| ON03 | Nanticoke Creek | 8.25 | 548.1 | 0.189 | 70.54 | 35.65 | 12.29 | 187.25 | 16.35 |
| ON04 | Salmon River | 8.17 | 273.7 | 0.016 | 42.62 | 9.92 | 10.78 | 123.69 | 16.06 |
| ON05 | Napanee River | 8.23 | 224.4 | 0.019 | 32.16 | 8.79 | 10.62 | 98.61 | 16.05 |
| ON06 | Kemptville Creek | 8.05 | 373.8 | 0.023 | 45.15 | 14.05 | 9.88 | 179.61 | 14.63 |
| ON07 | Jock River | 7.78 | 561.1 | 0.042 | 61.82 | 50.27 | 8.99 | 197.48 | 13.71 |
| ON09 | Fall River | 8.55 | 214.6 | 0.011 | 26.64 | 11.39 | 9.07 | 88.39 | 16.1 |
| ON10 | Nottawasaga River | 8.24 | 536.3 | 0.037 | 74 | 29.66 | 9.11 | 225.1 | 13.43 |
| ON11 | Credit River | 8.06 | 693.4 | 0.041 | 73.32 | 74.77 | 11.24 | 222.86 | 12.21 |
| ON12 | Don River | 8.04 | 1210.2 | 0.18 | 87.67 | 249.5 | 11.37 | 188.87 | 13.21 |
| ON13 | Ausable River | 7.86 | 590.9 | 0.237 | 73.17 | 38.52 | 11.63 | 213.75 | 13.49 |
| ON14 | Bayfield River | 7.87 | 552.1 | 0.044 | 73.44 | 30.05 | 12.88 | 221.87 | 13.52 |

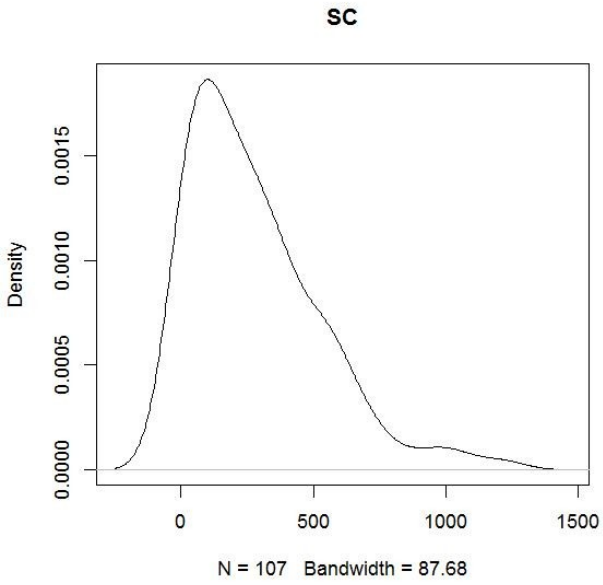
| Site ID | River Name | pH | SC (µS/cm) | TP (mg/L) | Ca (mg/L) | Cl (mg/L) | DO (mg/L) | T_ALK (mg/L) | WT (°C) |
|---|---|---|---|---|---|---|---|---|---|
| ON16 | Saugeen River | 8.31 | 557.4 | 0.028 | 74.74 | 13.17 | 10.12 | 231.03 | 15.3 |
| SK01 | North Saskatchewan River | 8.44 | 431.3 | 0.096 | 48.6 | 5.56 | 9.65 | 147.94 | 13.58 |
| SK02 | Qu'Appelle River below Qu'Appelle dam | 8.6 | 475.8 | 0.021 | 45.6 | 8.99 | 13.53 | 167.13 | 13.66 |
| SK03 | Qu'Appelle River at Lumsden | 8.43 | 991 | 0.346 | 57.61 | 61.97 | 12.99 | 182.06 | 14.25 |
| SK04 | Assiniboine River | 8.38 | 1038.1 | 0.185 | 86.67 | 21.42 | 12.27 | 288.42 | 13.71 |
| SK05 | Beaver River | 7.44 | 401.3 | 0.108 | 45.16 | 4 | 9.66 | 208.24 | 11.8 |
| SK06 | Churchill River | 8.22 | 154.3 | 0.033 | 16.04 | 2.47 | 10.22 | 76.09 | 11.8 |
| BC01 | Peace River | 8.12 | 195.4 | 0.145 | | | | | |
| BC02 | Iskut River | 7.93 | 174.9 | 0.323 | | | | | |
| BC03 | Skeena River | 7.86 | 89.1 | 0.047 | | | | | |
| BC04 | Cheakamus River | 7.34 | 41.1 | 0.008 | | | | | |
| BC05 | Cowichan River | 7.6 | 61.8 | 0.032 | | | | | |
| BC07 | Englishman River | 7.42 | 64.4 | 0.008 | | | | | |
| BC08 | Quinsam River | 7.69 | 104.5 | 0.027 | | | | | |
| BC11 | Salmon River | 8.17 | 357.5 | 0.085 | | | | | |
| BC12 | Fraser River at Marguerite | 7.97 | 139.8 | 0.114 | | | | | |
| BC13 | Fraser River at Hope | 7.92 | 119.3 | 0.116 | | | | | |
| BC16 | Columbia River at Nicholson | 8.04 | 239.4 | 0.018 | | | | | |
| BC17 | Columbia River at Birchbank | 7.9 | 130.3 | 0.008 | | | | | |
| BC18 | Kootenay River | 8.31 | 272.1 | 0.01 | | | | | |
| BC19 | Similkameen River | 7.89 | 127 | 0.028 | | | | | |
| BC20 | Okanagan River | 8.21 | 276.5 | 0.017 | | | | | |
| BC21 | Kettle River | 7.89 | 128.6 | 0.017 | | | | | |
| NB03 | Nashwaak River | 7.33 | 56.7 | 0.011 | | | | | |
| NB04 | Lepreau River | 6.26 | 23.1 | 0.006 | | | | | |
| NB06 | St. Croix River | 7.35 | 32.5 | 0.005 | | | | | |
| NF01 | Little Mecatina River | 6.42 | 13.1 | 0.011 | | | | | |
| NF03 | Lloyds River | 6.91 | 30.4 | 0.005 | | | | | |
| NF04 | Exploits River | 6.62 | 23.3 | 0.003 | | | | | |
| NF05 | Gander River | 6.27 | 21 | 0.004 | | | | | |
| NF06 | Terra Nova River | 5.93 | 16.8 | 0.004 | | | | | |
| NF07 | Northest River | 6.64 | 39.7 | 0.006 | | | | | |
| NF08 | Leary's Brook | 6.72 | 668.7 | 0.018 | | | | | |
| NF09 | Waterford River | 7.22 | 514.9 | 0.023 | | | | | |
| NF10 | Ugjoktok River | 7.01 | 21.9 | 0.005 | | | | | |
| NF11 | Atikonak River | 6.71 | 14.9 | 0.005 | | | | | |

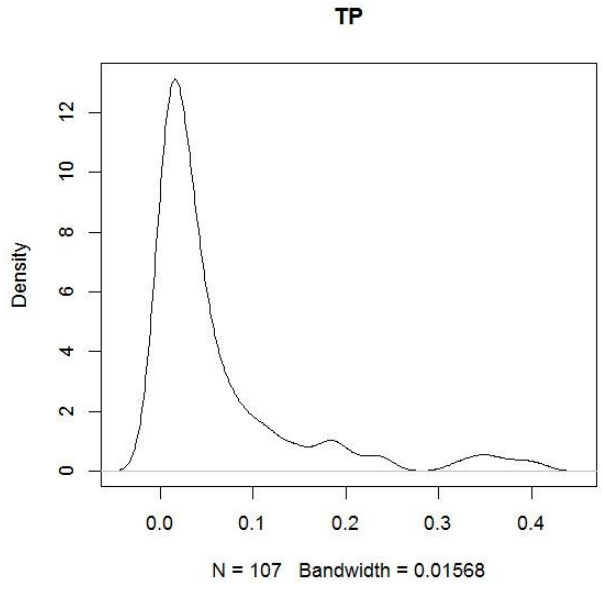
| Site ID | River Name | pH | SC (µS/cm) | TP (mg/L) | Ca (mg/L) | Cl (mg/L) | DO (mg/L) | T_ALK (mg/L) | WT (°C) |
|---|---|---|---|---|---|---|---|---|---|
| NF12 | Minipi River | 6.67 | 15.4 | 0.007 | | | | | |
| NF13 | Naskaupi River | 7.05 | 29.8 | 0.005 | | | | | |
| NF14 | Eagle River | 6.63 | 16 | 0.016 | | | | | |
| NS01 | Annapolis River | 7.39 | 123.1 | 0.045 | | | | | |
| NS02 | Cornwallis River | 7.82 | 303.6 | 0.233 | | | | | |
| NS03 | South River | 7.55 | 473 | 0.012 | | | | | |
| NS04 | Tusket River | 4.87 | 36.5 | 0.013 | | | | | |
| NS05 | Roseway River | 4.44 | 35.9 | 0.012 | | | | | |
| NS06 | Mersey River | 5.12 | 30.3 | 0.012 | | | | | |
| NS07 | Lahave River | 6.38 | 33.5 | 0.011 | | | | | |
| NS08 | Sackville River | 7.05 | 122.8 | 0.014 | | | | | |
| NS09 | Little Sackville River | 7.7 | 320 | 0.027 | | | | | |
| NS10 | Northest Margaree River | 7.43 | 159.9 | 0.005 | | | | | |
| NS11 | Cheticamp River | 6.68 | 40.8 | 0.006 | | | | | |
| ON01 | Kaministiquia River | 7.84 | 154.4 | 0.057 | | | | | |
| ON08 | Mississippi River | 8.47 | 74.6 | 0.005 | | | | | |
| ON15 | Maitland River | 8.3 | 572.5 | 0.032 | | | | | |
| PE01 | Mill River | 7.7 | 250 | 0.019 | | | | | |
| PE02 | Wilmot River | 7.74 | 325.5 | 0.047 | | | | | |
| PE03 | Bear River | 7.63 | 242.4 | 0.03 | | | | | |
| QC01 | Beaurivage River | 7.71 | 190.4 | 0.056 | | | | | |
| QC02 | Bras d'Henri River | 7.87 | 256.2 | 0.105 | | | | | |
| QC03 | Becancour | 8.18 | 406.9 | 0.083 | | | | | |
| QC04 | Coaticook River | 7.99 | 213.3 | 0.033 | | | | | |
| QC05 | Noire River | 7.77 | 224.1 | 0.079 | | | | | |
| QC06 | Yamask River | 7.6 | 291.8 | 0.069 | | | | | |
| QC07 | Des Hurons River | 8.04 | 724.7 | 0.205 | | | | | |
| QC08 | Richelieu River | 7.89 | 186.3 | 0.037 | | | | | |
| QC09 | Des Anglais River | 7.85 | 363.1 | 0.354 | | | | | |
| QC10 | Des Milles Iles River | 7.36 | 145.4 | 0.036 | | | | | |
| QC11 | Saint Charles River | 7.58 | 360.6 | 0.021 | | | | | |
| QC12 | L'Assomption River | 7.21 | 68.8 | 0.018 | | | | | |
| QC13 | De l'Achigan River | 7.79 | 271.2 | 0.062 | | | | | |
| QC14 | Du Loup River | 6.7 | 20.3 | 0.005 | | | | | |
| YT01 | Dezadeash River | 8.14 | 188.6 | 0.078 | | | | | |
| YT02 | Alsek River | 8.16 | 182.6 | 0.062 | | | | | |

## Appendix B. Water quality data density plots

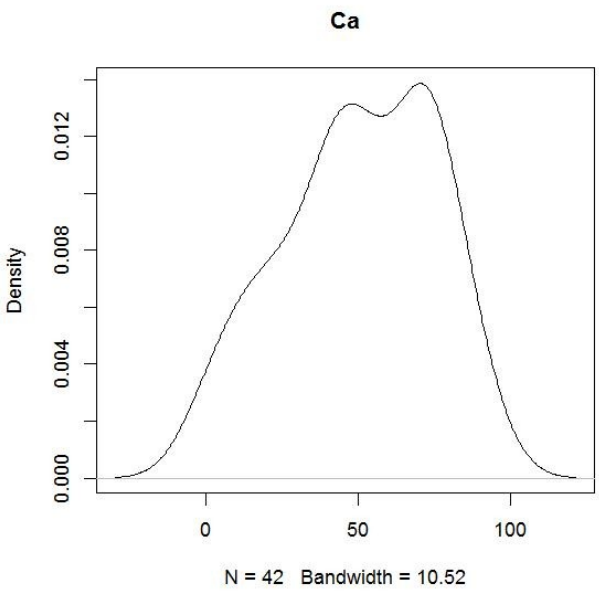
### Matrix A water quality data density plots before transformation

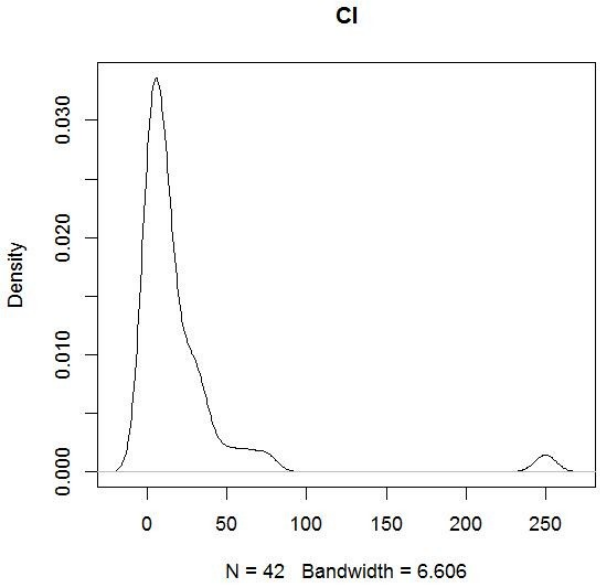


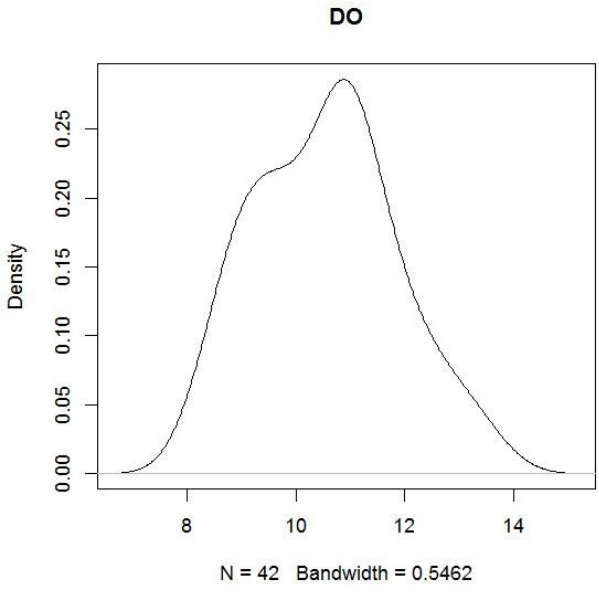



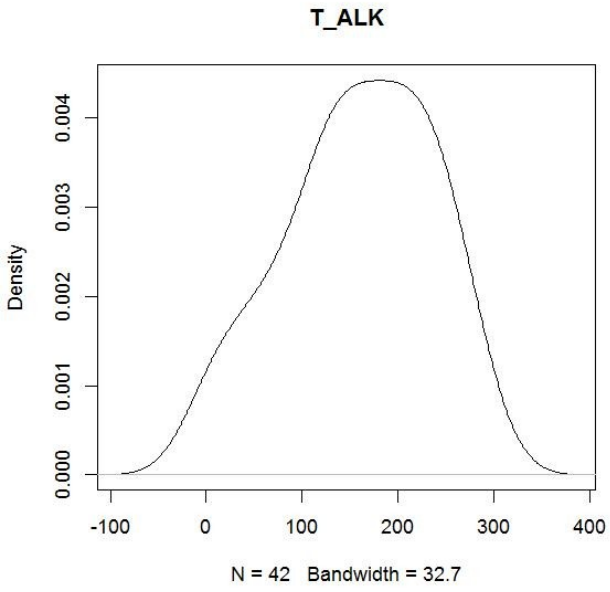



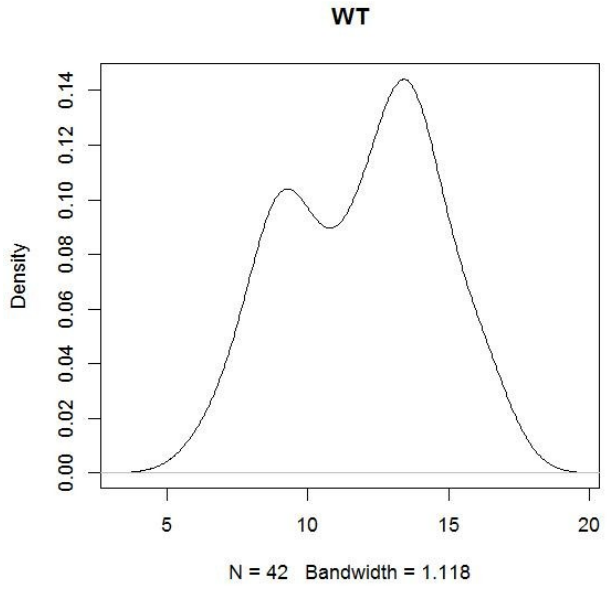
**Matrix B water quality data density plots before transformation**

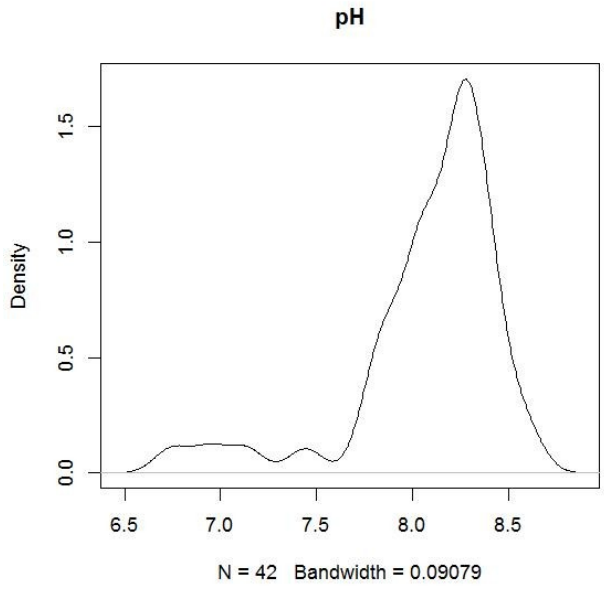


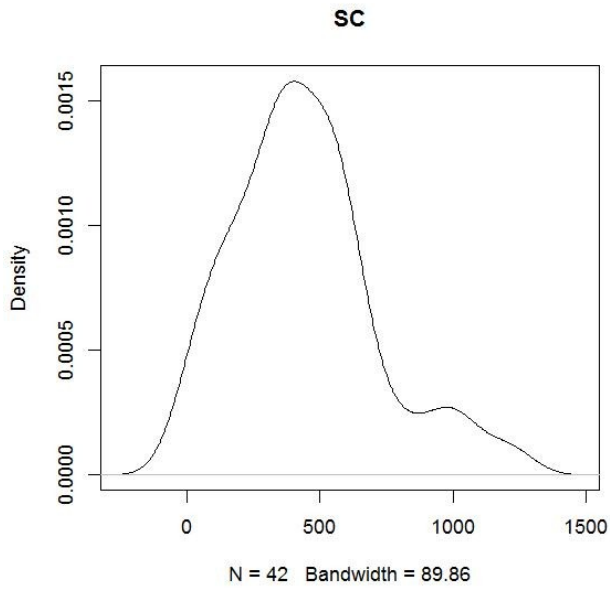



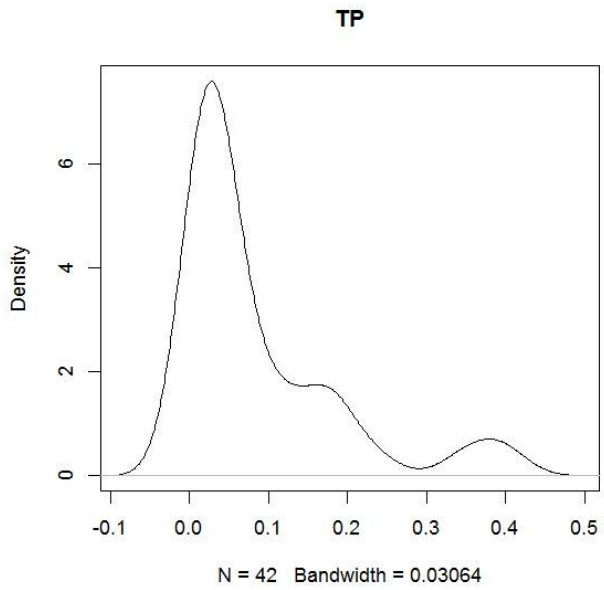




WT
Density
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
5
10
15
20
N = 42 Bandwidth = 1.118

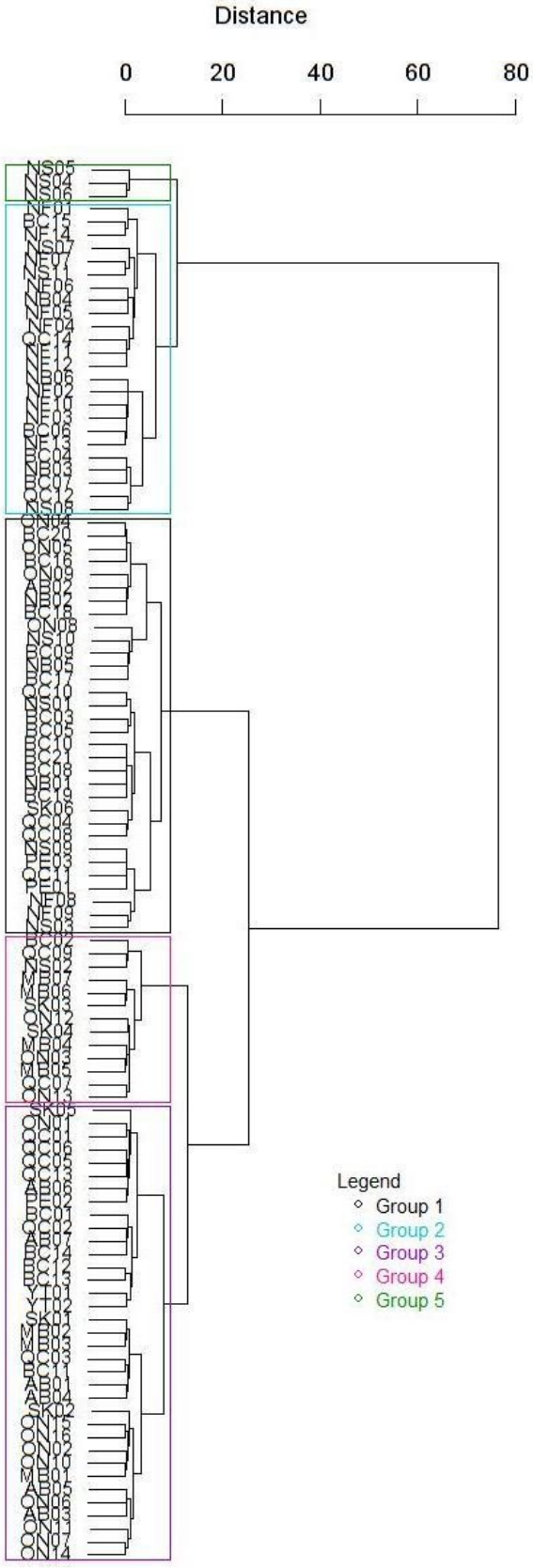

pH
Density
0.0
0.5
1.0
1.5
6.5
7.0
7.5
8.0
8.5
N = 42 Bandwidth = 0.09079


SC
Density
0.0000
0.0005
0.0010
0.0015
0
500
1000
1500
N = 42 Bandwidth = 89.86

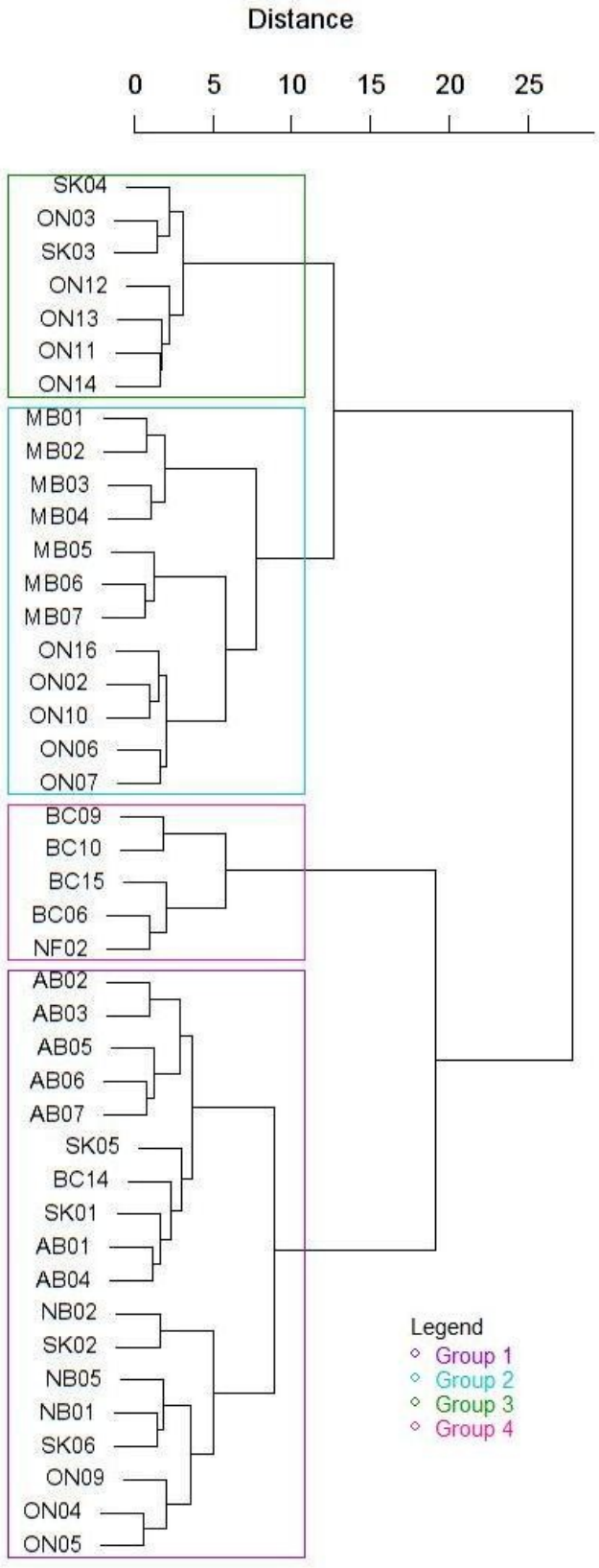

TP
Density
0
2
4
6
-0.1
0.0
0.1
0.2
0.3
0.4
0.5
N = 42 Bandwidth = 0.03064

## Appendix C. Cluster analysis dendrograms

**Matrix A cluster analysis dendrogram**

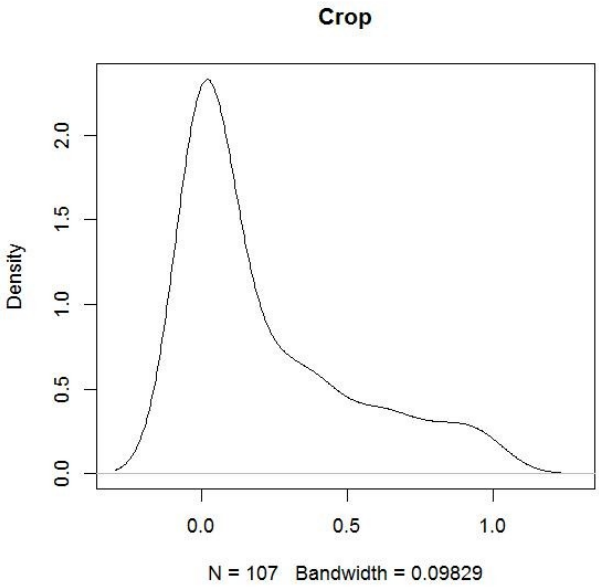


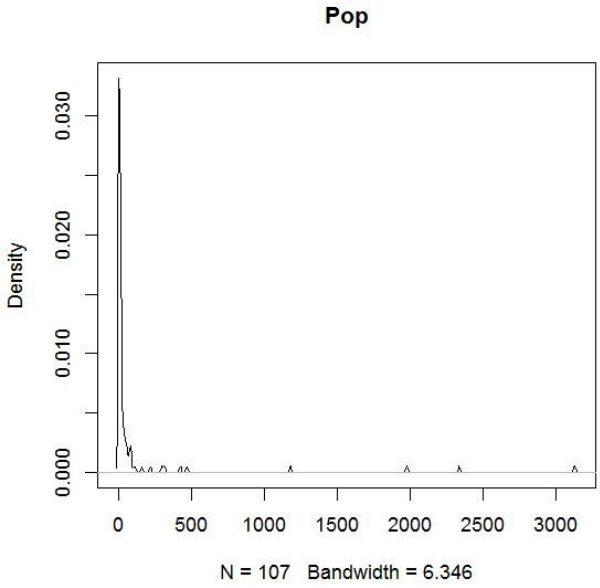
**Matrix B cluster analysis dendrogram**

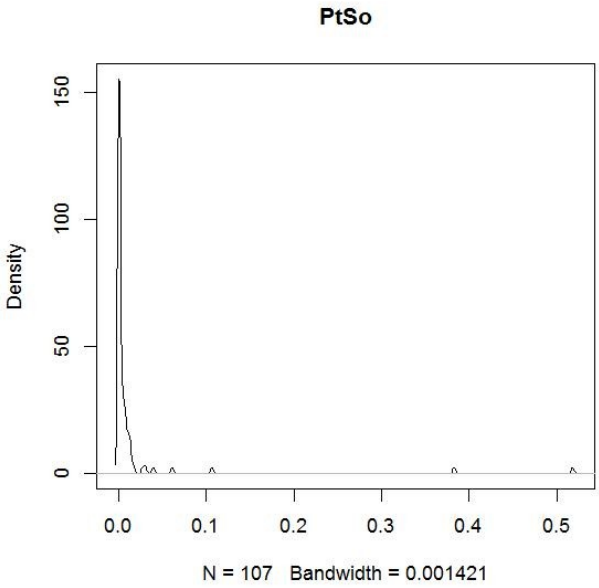


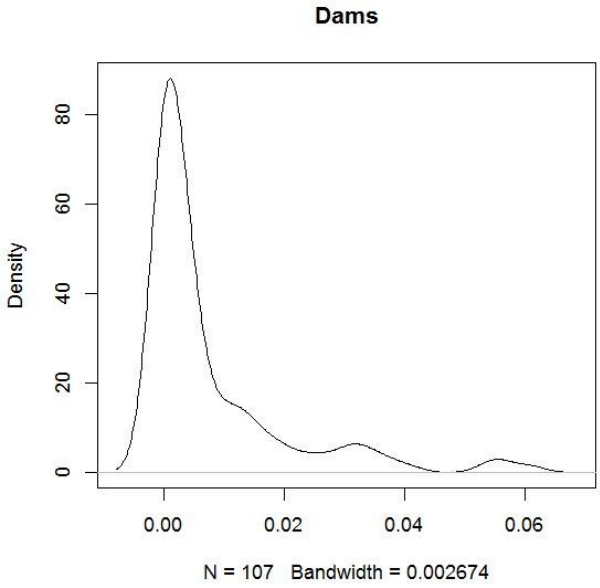
## Appendix D. Landscape characteristics data density plots

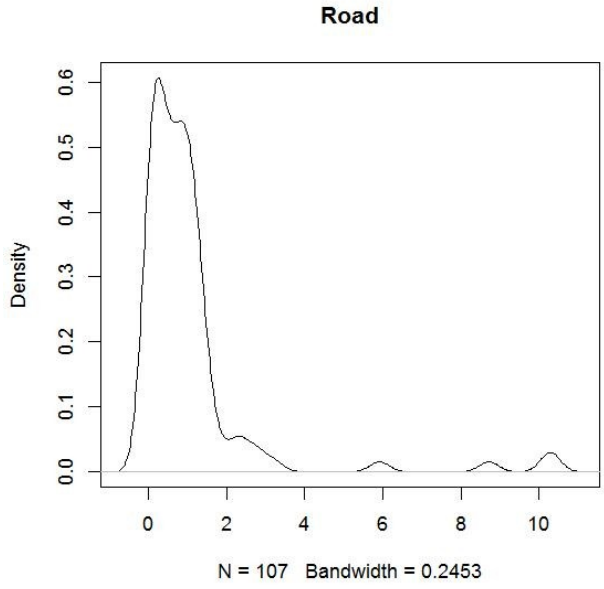
### Matrix A landscape characteristics data density plots before transformation

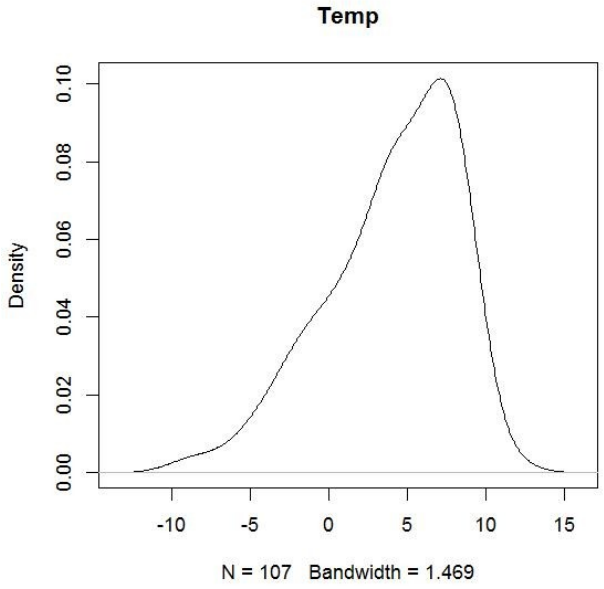


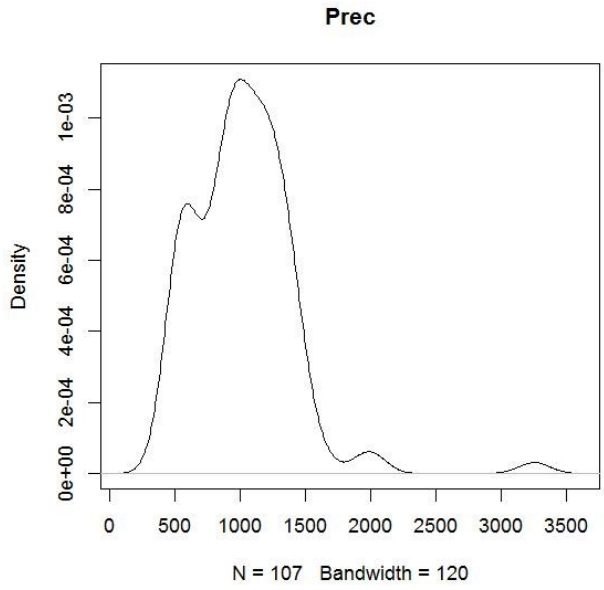



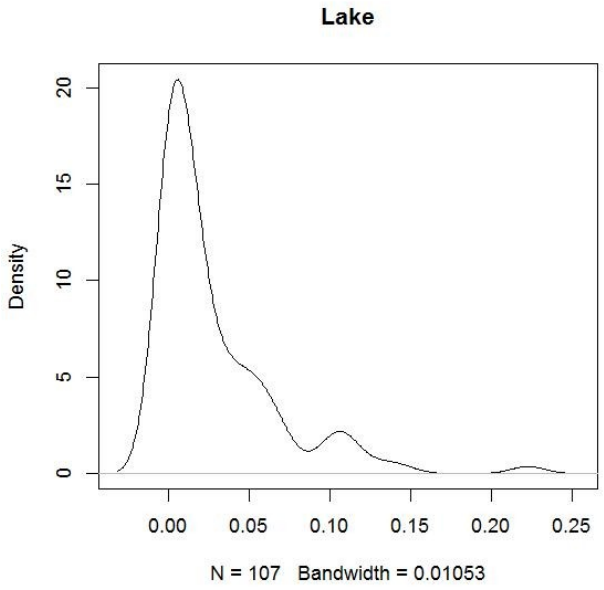



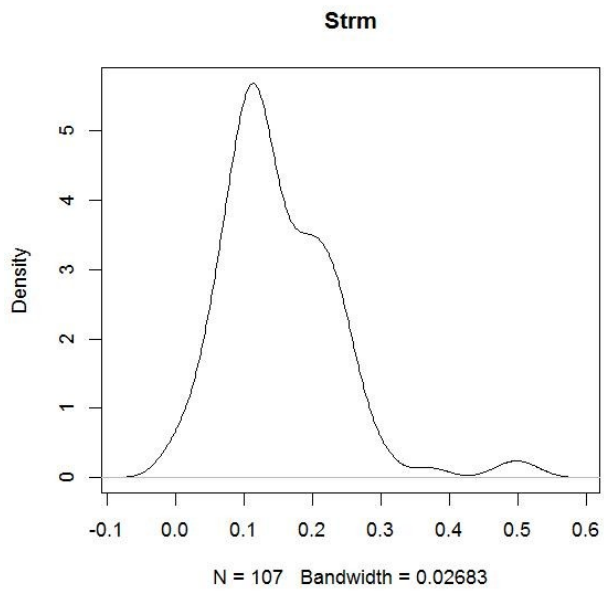



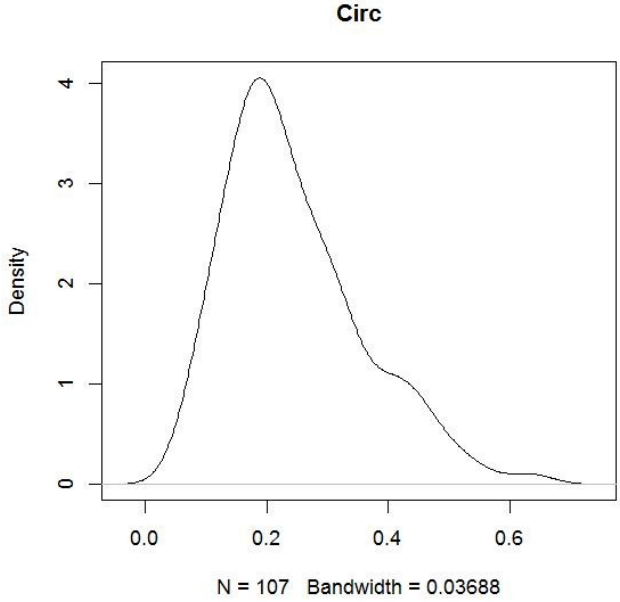

Road
Density
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0
2
4
6
8
10
N = 107 Bandwidth = 0.2453

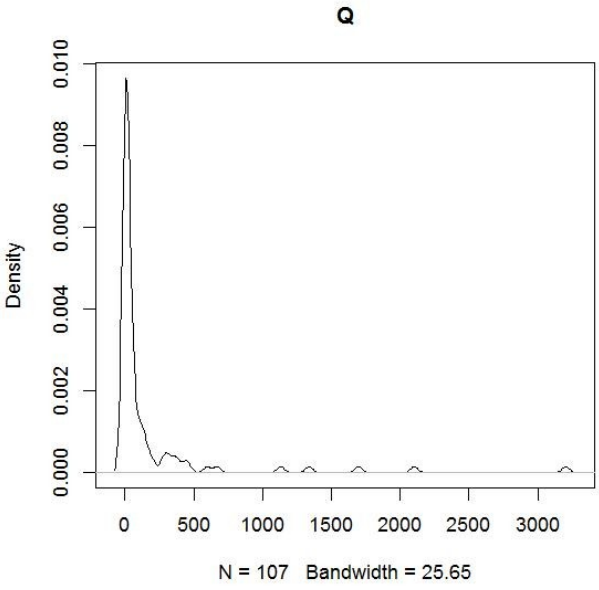

Temp
Density
0.00
0.02
0.04
0.06
0.08
0.10
-10
-5
0
5
10
15
N = 107 Bandwidth = 1.469

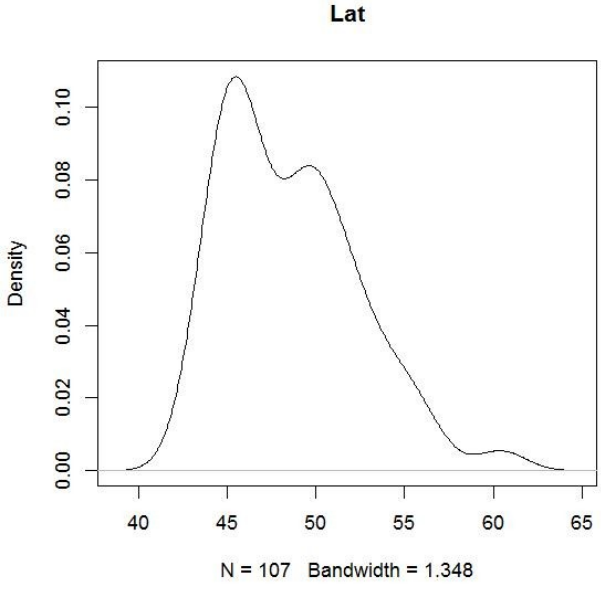

Prec
Density
0e+00
2e-04
4e-04
6e-04
8e-04
1e-03
0
500
1000
1500
2000
2500
3000
3500
N = 107 Bandwidth = 120

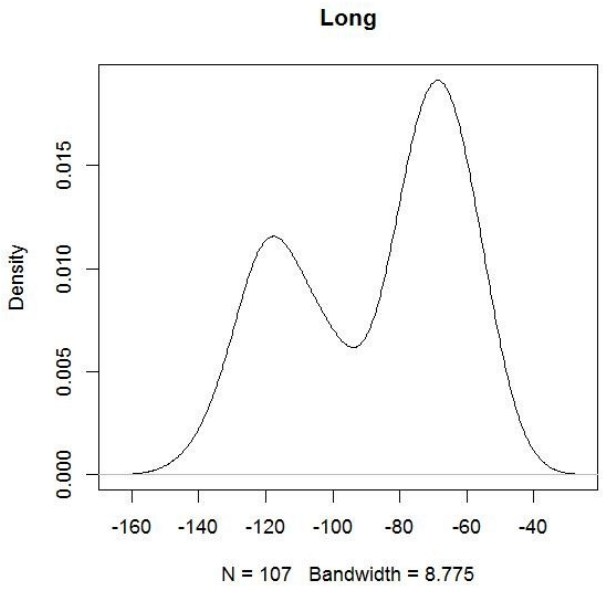

Lake
Density
0
5
10
15
20
0.00
0.05
0.10
0.15
0.20
0.25
N = 107 Bandwidth = 0.01053


Strm
Density
0
1
2
3
4
5
-0.1
0.0
0.1
0.2
0.3
0.4
0.5
0.6
N = 107 Bandwidth = 0.02683

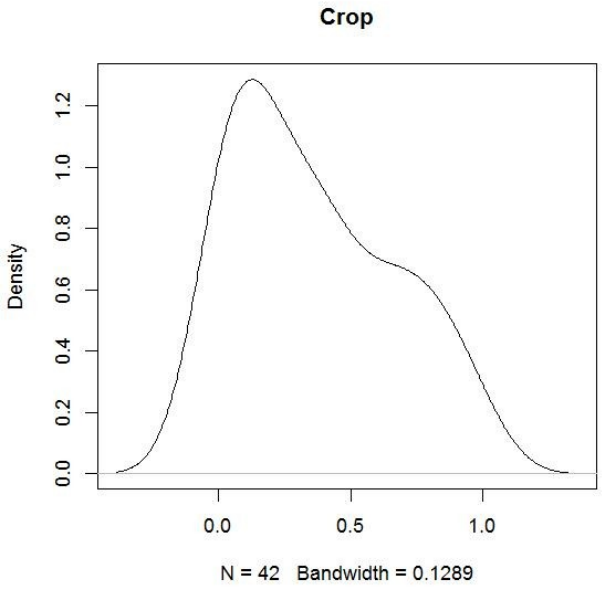

Circ
Density
0
1
2
3
4
0.0
0.2
0.4
0.6
N = 107 Bandwidth = 0.03688

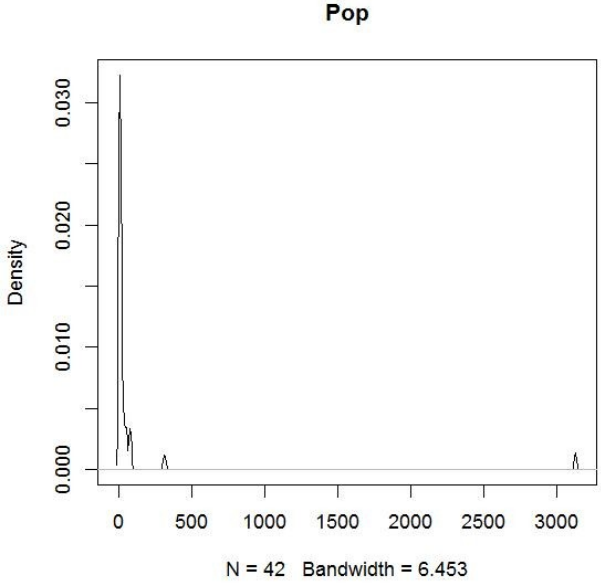

Q
Density
0.000
0.002
0.004
0.006
0.008
0.010
0
500
1000
1500
2000
2500
3000
N = 107 Bandwidth = 25.65

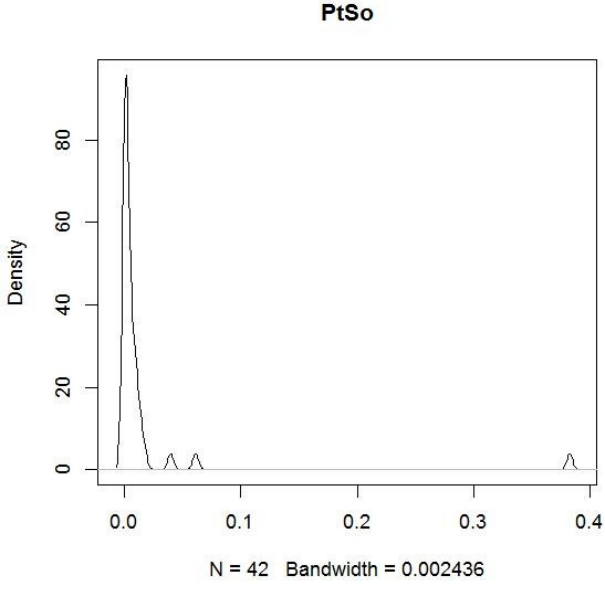

Lat
Density
0.00
0.02
0.04
0.06
0.08
0.10
40
45
50
55
60
65
N = 107 Bandwidth = 1.348

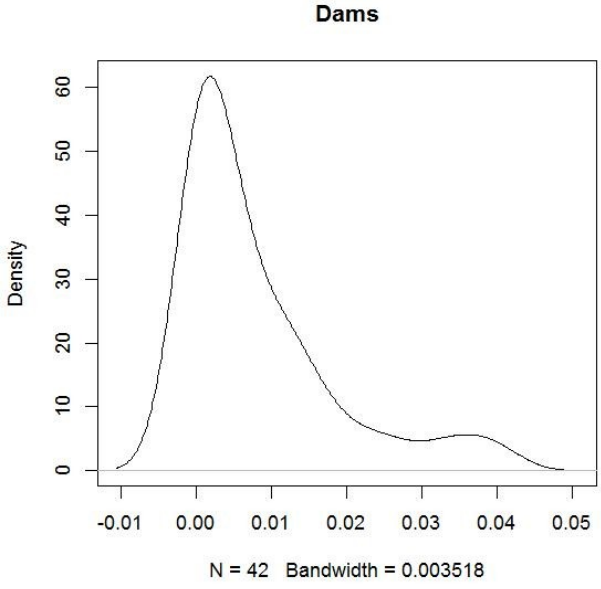

Long
Density
0.000
0.005
0.010
0.015
-160
-140
-120
-100
-80
-60
-40
N = 107 Bandwidth = 8.775

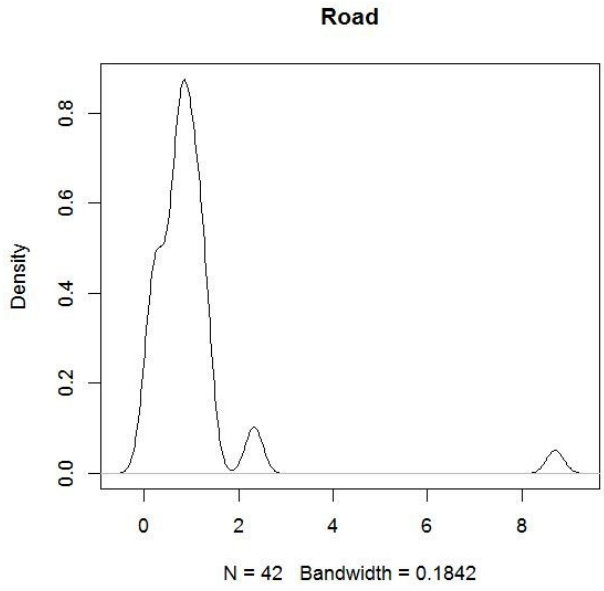
**Matrix B landscape characteristics data density plots before transformation**

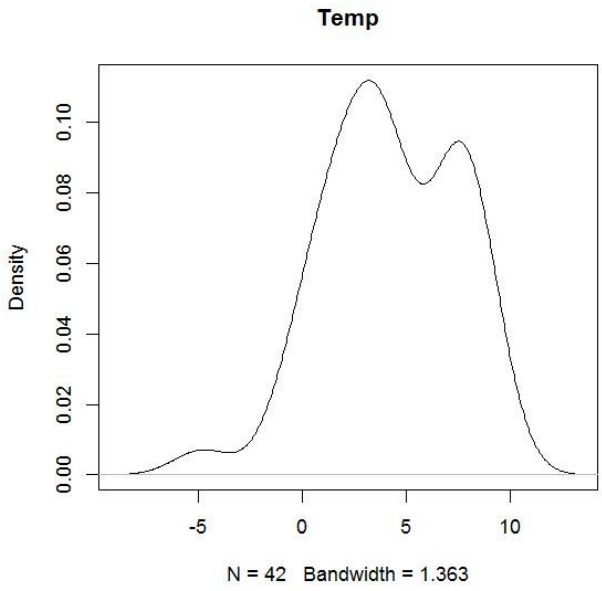


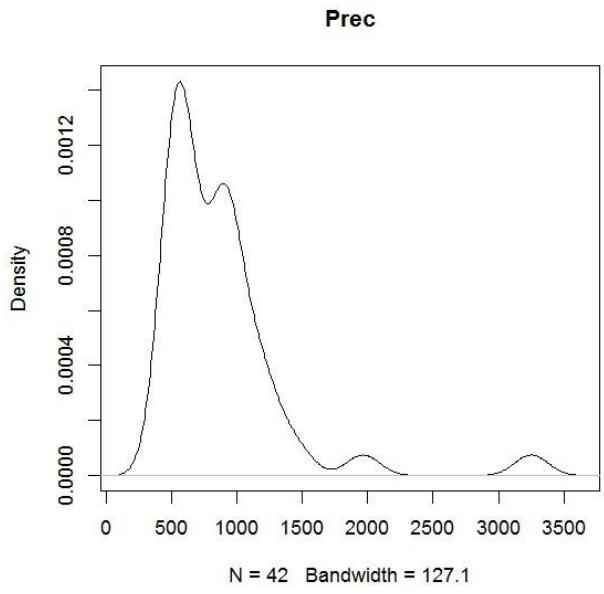



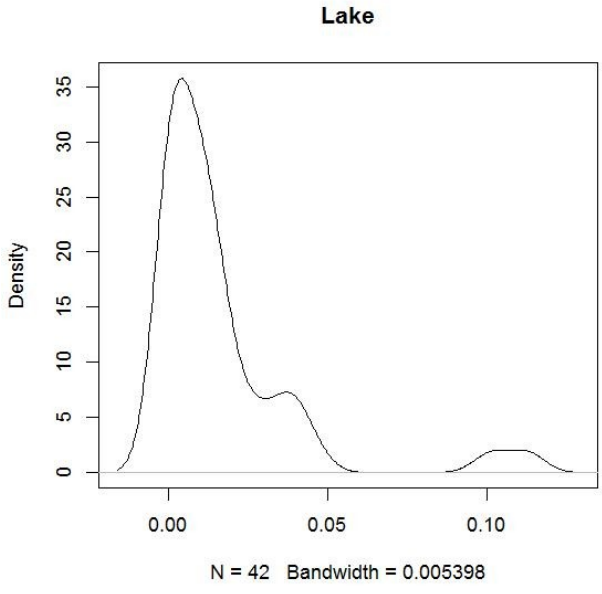



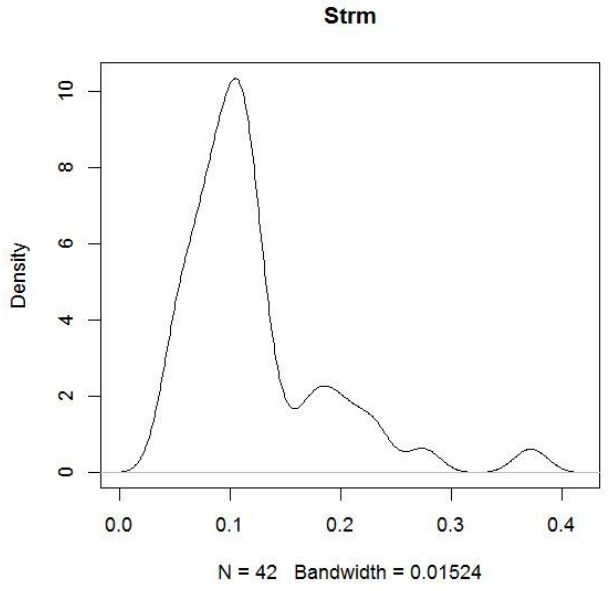



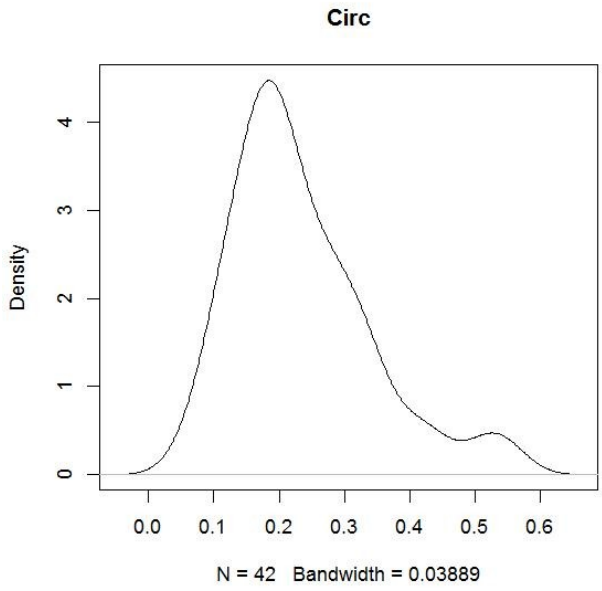

Road
Density
0.0
0.2
0.4
0.6
0.8
0
2
4
6
8
N = 42 Bandwidth = 0.1842

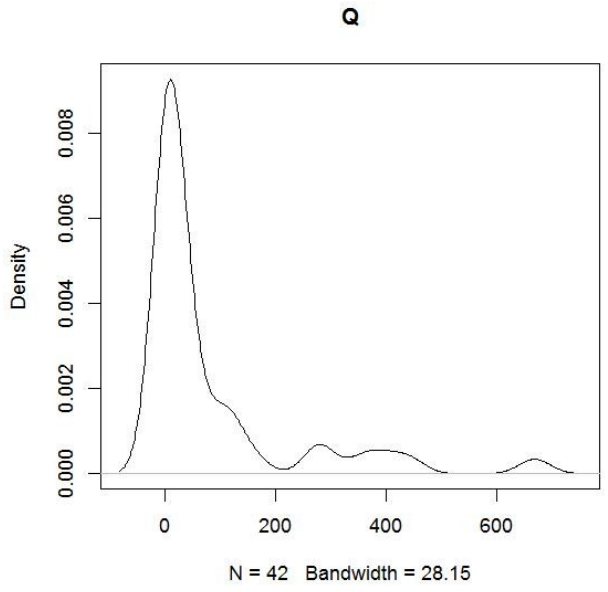

Temp
Density
0.00
0.02
0.04
0.06
0.08
0.10
-5
0
5
10
N = 42 Bandwidth = 1.363

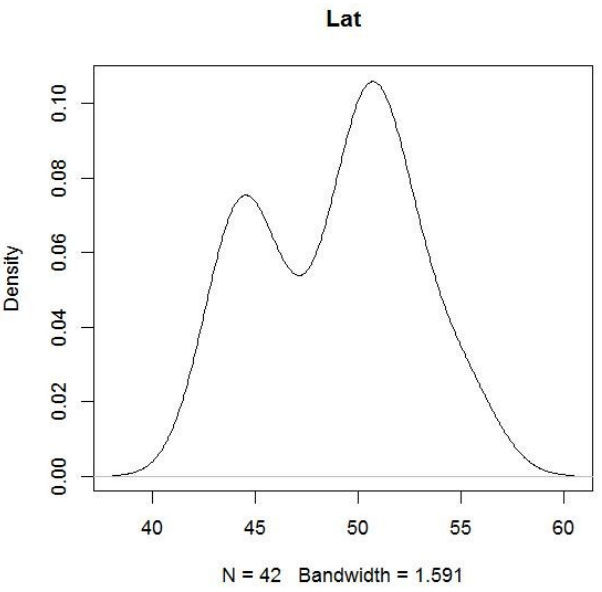

Prec
Density
0.0000
0.0004
0.0008
0.0012
0
500
1000
1500
2000
2500
3000
3500
N = 42 Bandwidth = 127.1

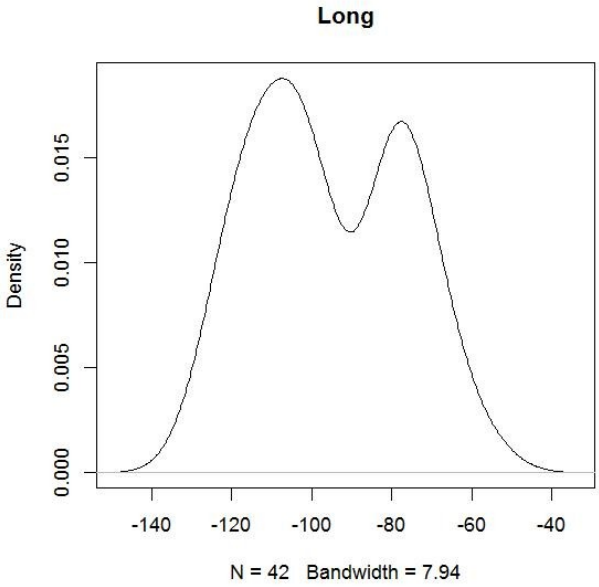

Lake
Density
0
5
10
15
20
25
30
35
0.00
0.05
0.10
N = 42 Bandwidth = 0.005398


Strm
Density
0
2
4
6
8
10
0.0
0.1
0.2
0.3
0.4
N = 42 Bandwidth = 0.01524


Circ
Density
0
1
2
3
4
0.0
0.1
0.2
0.3
0.4
0.5
0.6
N = 42 Bandwidth = 0.03889


Q
Density
0.000
0.002
0.004
0.006
0.008
0
200
400
600
N = 42 Bandwidth = 28.15


Lat
Density
0.00
0.02
0.04
0.06
0.08
0.10
40
45
50
55
60
N = 42 Bandwidth = 1.591


Long
Density
0.000
0.005
0.010
0.015
-140
-120
-100
-80
-60
-40
N = 42 Bandwidth = 7.94