THE EFFECTS OF AGRICULTURE ON CANADA'S MAJOR WATERSHEDS

THE EFFECTS OF AGRICULTURE ON CANADA'S MAJOR WATERSHEDS

By

DAN RAMUNNO, B.Sc.

A Thesis

Submitted to the School of Graduate Studies

In Partial Fulfillment of the Requirements

for the Degree

Master of Science

McMaster University

MASTER OF SCIENCE (2011)	McMaster University
(Statistics)	Hamilton, Ontario

TITLE:	The Effects of Agriculture on Canada's Major Watersheds
AUTHOR:	Dan Ramunno
SUPERVISOR:	Professor Abdel El-Shaarawi
NUMBER OF PAGES:	x, 61

Acknowledgments

I would like to thank my thesis supervisor Dr. Abdel El-Shaarawi for his support throught my Master's degree. His concrete directions and feedback greatly helped me develop the work of my thesis. He gave me many ideas and recommendations that I have implemented into my thesis, and his imput greatly aided in the computational and written aspects of my thesis. I would also like to thank my family for their support, especially through tough times. My immediate family members (my parents, my sister Sandra, my sister Lora and her husband Nipun) have been especially supportive throughout my Master's degree. In addition, I would also like to acknowledge the support that I received from my classmates and professors during my studies.

Abstract

Water contamination is one of the major environmental issues that negatively impacts water quality of watersheds. It negatively affects drinking water and aquatic wildlife, which can indirectly have negative effects on everyone's health. Many different institutions collected samples of water from four of Canada's major watersheds and counted the number of bacteria in each sample. The data used in this paper was taken from one of these institutions and was analysed to investigate if agricultural waste impacts the water quality of these four watersheds. It was found that the agricultural waste produced from nearby farms significantly impacts the water quality of three of these watersheds. Principal component analysis was also done on these data, and it was found that all of the data can be expressed in terms of one variable without losing very much information of the data. The bootstrap distributions of the principal component analysis parameters were estimated, and it was found that the sampling distributions of these parameters are stable. There was also evidence that the variables in the data are not normally distributed and not all the variables are independent.

Contents

List of Tables

List of Figures

1 Introduction

It is of interest to determine how waste produced from agricultural farmlands affects the water quality of four of Canada's major watersheds. To address this concern, water samples have been collected over several years and analysed for pathogens. The bacteria counts in these water samples were counted to assess the water quality of the watersheds. Water samples with higher bacteria counts strongly indicate that the corresponding watersheds contain more toxins and are more polluted, meaning that the water quality of these watersheds is lower. Some of this watershed data was collected by Edge *et al.* (2010), and a subset of their data was used in this paper for data analysis.

For each of the four watersheds, data was collected upstream and downstream of the sites where agricultural waste would enter the watersheds. If the amount of agricultural waste significantly impacts the water quality of the watersheds, one would expect the bacteria counts in the downstream locations to be higher than the bacteria counts in the upstream locations. However, if the amount of waste does not have a significant impact on water quality, one would expect the bacteria counts to be roughly the same upstream and downstream. There are ten different data sets used in this paper, each consisting of the Total coliform, Fecal coliform and E. coli natural log counts. One of the main objectives in this paper is to determine if agricultural waste significantly impacts the water quality of these four watersheds. This was mainly done by producing different box plots and scatter matrices of the data to investigate the trends in the data and to compare the bacteria counts amongst the watersheds.

The other objective of this thesis is to investigate if data reduction is possible through the use of principal component analysis on the watershed data sets. For each of the ten data sets, the three coefficients of each of the three principal components and the three corresponding standard deviations were calculated. The main purpose of these calculations was to derive three independent random variables that are linear combinations of the three original variables that would maximize sample variation in the data. The three coefficients of these three linear combinations would determine the amount of axis rotation needed to make the three principal components all orthogonal to each other with unit length. It is therefore legitimate to remove principal components because of their independence of each other. The standard deviations of the principal components were also calculated to determine which principal components should be used for data analysis. This technique is known as data reduction. It is used to express multivariate data sets in fewer dimensions without losing very much information in the data. This makes it easier to analyze the data since it can be very difficult to analyze and visualize data with many variables. In this paper, it has been shown how to express the data in each of the ten data sets in terms of one variable, which reduces the number of variables without losing very much information in the data. This data reduction technique was only used for three-dimensional datasets, which could be applied to datasets with any number of dimensions.

The data set corresponding to the bacteria counts collected from all of the locations downstream of the agricultural farmlands from all four watersheds, also known as the Ag data set, was further analysed by looking at the 12 approximate bootstrap distributions of the principal component analysis estimates computed on these data. The bootstrap distribution of each of the

12 principal component analysis estimates were estimated nonparametrically. Then two parametric bootstrap techniques were used to calculate the exact same 12 principal component analysis parameters. One of these parametric bootstrap techniques was based on a sample generated from the trivariate normal distribution with mean vector $[0, 0, 0]$ and covariance matrix $\mathbf{I}_3$ where $\mathbf{I}_3$ is the 3×3 identity matrix. The other technique was based on a sample generated from the trivariate normal distribution with mean vector $[0, 0, 0]$ and covariance matrix $\mathbf{R}$ where $\mathbf{R}$ is the correlation matrix of the Ag data set. The results of these three bootstrap techniques were compared to make inferences about the normality and independence of the three variables in the Ag data set.

2 Methodology

2.1 Principal Component Analysis

As discussed by Johnson and Wichern (2007), Hotelling (1936) and Jolliffe (2002), principal component analysis is a data reduction technique used to calculate principal components expressed as linear combinations of the variables in multivariate data. Using these linear combinations to transform the variables, the original coordinate system with axes $X_1, X_2, ..., X_p$ is rotated to obtain a new coordinate system with axes $Y_1, Y_2, ..., Y_p$, and the axes of this new coordinate system are positioned to maximize the variation in the data. The principal components are calculated using the covariance matrix $\mathbf{\Sigma}$ or correlation matrix $\Re$ in the case of population data observed from $X_1, X_2, ..., X_p$, which is very similar to the principal component calculations for sample data

(Jolliffe, 2002; Hotelling, 1936). These calculations do not require that the data are from a multinormal distribution, but having normality means that one can calculate principal components from a random sample to make inferences about the population principal components. In principal component analysis, the data is transformed from p variables with n observations into k new variables with n observations where $k \leq p$. Under this transformation, the new k variables are uncorrelated and independent if normality is assumed. This retains almost all of the variability of the original data, making it easier to visualize the data in fewer dimensions and to investigate the relationships amongst the variables in the data. For a given value of k, the new variables are selected to maximize the amount of explained variation for any set of k linear variables. (Johnson and Wichern, 2007; Hotelling, 1936).

2.1.1 Population Principal Components using the Covariance Matrix

Let's suppose that $\mathbf{X}' = [X_1, X_2, ..., X_p]$ is a random vector of variables that represent values observed from the population with covariance matrix $\mathbf{\Sigma}$ with eigenvalues $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p \geq 0$. If $Y_j = \mathbf{a}_j'\mathbf{X}$ with $\mathbf{a}_j' = [a_{j1}, a_{j2}, ..., a_{jp}]$ for $j = 1, 2, ..., p$, then $Var(Y_j) = \mathbf{a}_j'\mathbf{\Sigma}\mathbf{a}_j$ for $j = 1, 2, ..., p$ and $Cov(Y_j, Y_l) = \mathbf{a}_j'\mathbf{\Sigma}\mathbf{a}_l$ for $j, l = 1, 2, ..., p$ (Johnson and Wichern, 2007). In order to find the principal components, the linear combinations $Y_j, j = 1, 2, ..., p$, are constructed to be uncorrelated with maximal variance. The first principal component is $Y_1 = \mathbf{a}_1'\mathbf{X}$ and $\mathbf{a}_1'$ is computed by maximizing $Var(Y_1)$ under the constraint $\mathbf{a}_1'\mathbf{a}_1 = 1$ (Anderson, 1963; Jolliffe, 2002). The second principal component is $Y_2 = \mathbf{a}_2'\mathbf{X}$ and $\mathbf{a}_2'$ is computed by maximizing $Var(Y_2)$ under the constraints $\mathbf{a}_2'\mathbf{a}_2 = 1$ and $Cov(Y_1, Y_2) = 0$. For $j = 1, 2, ..., p$, the j^{th} principal

component is $Y_j = \mathbf{a}_j'\mathbf{X}$ and $\mathbf{a}_j'$ is computed by maximizing $Var(Y_j)$ given that $\mathbf{a}_j'\mathbf{a}_j = 1$ and $Cov(Y_j, Y_l) = 0$ for all $l < j$ (Anderson, 1963; Jolliffe, 2002).

Johnson and Wichern (2007), Anderson (1963) and Jolliffe (2002) have demonstrated that these formulas for the principal components can be expressed in terms of the eigenvalues and corresponding eigenvectors of the covariance matrix $\mathbf{\Sigma}$ with eigenvalues $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p \geq 0$. For $j = 1, 2, ..., p$, the j^{th} principal component can be expressed as $Y_j = \mathbf{e}_j'\mathbf{X}$ where $\mathbf{e}_j' = [e_{j1}, e_{j2}, ..., e_{jp}]$ is the normalized eigenvector corresponding to λ_j (Anderson, 1963; Jolliffe, 2002). It can also be shown that $Var(Y_j) = \mathbf{e}_j'\mathbf{\Sigma}\mathbf{e}_j = \lambda_j$ for $j = 1, 2, ..., p$ and $Cov(Y_j, Y_l) = \mathbf{e}_j'\mathbf{\Sigma}\mathbf{e}_l = 0$ for $j, l = 1, 2, ..., p; j \neq l$. This means that the principal components are all uncorrelated to each other and the variance of each principal component is equal to the corresponding eigenvalue of $\mathbf{\Sigma}$ (Anderson, 1963; Jolliffe, 2002). Johnson and Wichern (2007) used these results to show that $\sum_{j=1}^{p} Var(X_j) = \sum_{j=1}^{p} \lambda_j = \sum_{j=1}^{p} Var(Y_j)$. Based on this equality, the proportion of the total variation in the data explained by the k^{th} principal component is $\frac{\lambda_k}{\sum_{j=1}^{p} \lambda_j}, k = 1, 2, ..., p$. In practice, one can replace all p original variables with the first few principal components that explain 80-90% of the total variation without losing too much information of the original data (Johnson and Wichern, 2007).

Johnson and Wichern (2007) state that the element e_{jl} of the vector $\mathbf{e}_j$ and $Corr(Y_j, X_l) = \rho_{Y_j, X_l}$ are proportional to each other, and $|\rho_{Y_j, X_l}| = \frac{|e_{jl}|\sqrt{\lambda_j}}{\sigma_l}$ for $j, l = 1, 2, ..., p$ where $\sigma_l = \sqrt{Var(X_l)}$. Since this correlation coefficient measures the correlation between one principal component and one X variable at a time, it is usually better to use the elements of $\mathbf{e}_j$ to make conclusions based on the principal components instead of the correlation coefficients. Nevertheless, both of these methods usually give very similar results since large

elements of $\mathbf{e}_j$ in absolute value usually indicate that the correlation coefficients are large in absolute value (Johnson and Wichern, 2007).

2.1.2 Population Principal Components using the Correlation Matrix

Each of the X variables previously discussed may be standardized to obtain $Z_j = \frac{X_j - \mu_j}{\sigma_j}$, $j = 1, 2, ..., p$, where $\mu_j = E(X_j)$ and $\sigma_j = \sqrt{Var(X_j)}$. These standardized variables can be expressed as $\mathbf{Z} = (\mathbf{V}^{1/2})^{-1}(\mathbf{X} - \mathbf{b})$ where $\mathbf{b} = [\mu_1, \mu_2, ..., \mu_p]'$ and $\mathbf{V}^{1/2}$ is the diagonal matrix with diagonal entries $\sigma_1, \sigma_2, ..., \sigma_p$. It is well established that $E(\mathbf{Z}) = \mathbf{0}$ and $Cov(\mathbf{Z}) = \Re$ (Anderson, 1963; Hotelling, 1933; Hotelling, 1936; Johnson and Wichern, 2007). The principal components of $\mathbf{Z}$ are calculated in the exact same way as the principal components of $\mathbf{X}$ except the eigenvalues and eigenvectors are calculated from $\Re$ instead of $\mathbf{\Sigma}$. If $\mathbf{Z}' = [Z_1, Z_2, ..., Z_p]$ is a random vector representing the standardized population values with $Cov(\mathbf{Z}) = \Re$, then the j^{th} principal component of $\mathbf{Z}$ is $Y_j = \mathbf{e}_j'\mathbf{Z} = \mathbf{e}_j'(\mathbf{V}^{1/2})^{-1}(\mathbf{X} - \mathbf{b})$ for $j = 1, 2, ..., p$ where $\mathbf{b} = [\mu_1, \mu_2, ..., \mu_p]'$. It is important to note that using the covariance matrix usually results in different principal components than when using the correlation matrix. Also, $\sum_{j=1}^{p} Var(Y_j) = \sum_{j=1}^{p} Var(Z_j) = p$, and the correlation between Y_j and Z_l is $|\rho_{Y_j,Z_l}| = |e_{jl}|\sqrt{\lambda_j}$ for $j, l = 1, 2, ..., p$. Just like the principal components computed from $\mathbf{\Sigma}$, the eigenvalues of $\Re$ are $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p \geq 0$ (Johnson and Wichern, 2007). The proportion of the total variation in the data from the j^{th} principal component is $\frac{\lambda_j}{p}, j = 1, 2, ..., p$. Using $\Re$ instead of $\mathbf{\Sigma}$ to calculate the principal components is very useful when the values for each variable have completely different ranges or when the variables represent different types of measurements. Standardizing each variable means that each variable

has unit variance, and it ensures that the linear combinations representing the principal components are not heavily weighed on the variables with higher variances (Johnson and Wichern, 2007).

2.1.3 Sample Principal Components

Johnson and Wichern (2007) explain how to calculate the principal components of a multivariate random sample of random vectors $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n$ from a population with p variables $\mathbf{X}' = [X_1, X_2, ..., X_p]$ where $E(\mathbf{X}) = \mathbf{b}$ with $\mathbf{b} = [\mu_1, \mu_2, ..., \mu_p]'$ and $Cov(\mathbf{X}) = \mathbf{\Sigma}$. When doing multivariate analysis, it is useful to calculate the sample mean vector, covariance matrix and correlation matrix ($\overline{\mathbf{x}}$, $\mathbf{S}$ and $\mathbf{R}$ respectively) for statistical inferences. Johnson and Wichern (2007) have shown that the sample mean of all the $\mathbf{a}'\mathbf{x}_i$ values, $i = 1, 2, ..., n$, is $\mathbf{a}'\overline{\mathbf{x}}$ and the corresponding sample variance is $\mathbf{a}'\mathbf{S}\mathbf{a}$ where $\mathbf{a} = [a_1, a_2, ..., a_p]'$ is an arbitrary constant vector and $\mathbf{x}_i = [x_{i1}, x_{i2}, ..., x_{ip}]'$ corresponds to the i^{th} row vector of the $n \times p$ data matrix. They have also shown that the sample covariance of $(\mathbf{a}_1'\mathbf{x}_i, \mathbf{a}_2'\mathbf{x}_i)$ is $\mathbf{a}_1'\mathbf{S}\mathbf{a}_2$, $i = 1, 2, ..., n$, where $\mathbf{a}_1 = [a_{11}, a_{12}, ..., a_{1p}]'$ and $\mathbf{a}_2 = [a_{21}, a_{22}, ..., a_{2p}]'$ are both arbitrary constant vectors and $\mathbf{x}_i = [x_{i1}, x_{i2}, ..., x_{ip}]'$ is the same random vector that was previously defined (Johnson and Wichern, 2007).

Analogous to population principal components, $\hat{y}_j = \mathbf{a}_j'\mathbf{x}_i$ is the j^{th} sample principal component expressed as the linear combination of x_j variables for $j = 1, 2, ..., p$ and $i = 1, 2, ..., n$. Analogously, $\mathbf{a}_j$ is calculated so that $\hat{y}_j$ has maximal sample variance, $\mathbf{a}_j'\mathbf{a}_j = 1$ and the sample covariance is zero for all of the $(\mathbf{a}_j'\mathbf{x}_i, \mathbf{a}_l'\mathbf{x}_i)$ pairs for $l < j$. This ensures that the principal components are all independent of each other with unit length (Johnson and Wichern, 2007). Using these results, Johnson and Wichern (2007) have shown that the

maximum sample variance of the j^{th} sample principal component is $\hat{\lambda}_j$, which is the j^{th} largest eigenvalue of the sample covariance matrix $\mathbf{S}$ with corresponding normalized eigenvector $\hat{\mathbf{e}}_j$ (Johnson and Wichern, 2007). Thus, the j^{th} sample principal component is $\hat{y}_j = \hat{\mathbf{e}}_j'\mathbf{x}$, $j = 1, 2, ..., p$, where $\mathbf{x}$ represents an arbitrary data point sampled from the p population variables $\mathbf{X}' = [X_1, X_2, ..., X_p]$ and $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq ... \geq \hat{\lambda}_p \geq 0$. Based on previous results in this section, the j^{th} sample principal component has sample variance $\hat{\lambda}_j$ for $j = 1, 2, ..., p$ and the sample covariance of $\hat{y}_j$ and $\hat{y}_l$ is 0 for $j, l = 1, 2, ..., p; j \neq l$. It has also been shown that $\sum_{j=1}^{p} s_j = \sum_{j=1}^{p} \hat{\lambda}_j$ where s_j is the j^{th} diagonal element of $\mathbf{S}$ (Johnson and Wichern, 2007). Analogous to population principal components, the sample correlation between $\hat{y}_j$ and x_l is $r_{\hat{y}_j, x_l} = \frac{\hat{e}_{jl}\sqrt{\hat{\lambda}_j}}{s_l}$ for $i, l = 1, 2, ..., p$, where $\hat{e}_{jl}$ is the l^{th} element of $\hat{\mathbf{e}}_j$. Likewise, it is recommended to look at both the $\hat{e}_{jl}$ and $r_{\hat{y}_j, x_l}$ values when interpreting sample principal components (Johnson and Wichern, 2007).

The sample principal components may also be calculated using the correlation matrix $\mathbf{R}$, which normally results in different principal components than when using the covariance matrix $\mathbf{S}$. The $\mathbf{x}_i$ observation vectors, $i = 1, 2, ..., n$ can also be centered to have a sample mean of $\mathbf{0}$ without changing the sample covariance matrix $\mathbf{S}$ by using $\mathbf{x}_i - \overline{\mathbf{x}}$ instead of $\mathbf{x}_i$ (Johnson and Wichern, 2007). The j^{th} sample principal component using this transformed data is $\hat{y}_j = \hat{\mathbf{e}}_j'(\mathbf{x} - \overline{\mathbf{x}})$ for $j = 1, 2, ..., p$, where $\mathbf{x}$ is any arbitrary observation vector. It is easily seen that for each j^{th} sample principal component, if all of the $\hat{y}_{ij} = \hat{\mathbf{e}}_j'(\mathbf{x}_i - \overline{\mathbf{x}})$ values are calculated for $i = 1, 2, ..., n$, then these values have a sample mean of 0 and a sample variance of $\hat{\lambda}_j$ (Johnson and Wichern, 2007).

2.1.4 Sample Principal Components of Standardized Data

Just like with population principal components, the sample principal components of non-standardized data are greatly influenced by variables with higher sample variance or corresponding to values measured on a larger scale. To fix this issue, the data are standardized by converting each row vector of the $n \times p$ data matrix $\mathbf{X}$ into the transpose of $\mathbf{z}_i = \mathbf{D}^{-1/2}(\mathbf{x}_i - \overline{\mathbf{x}})$, $i = 1, 2, ..., n$; each j^{th} element of $\mathbf{z}_i$ is $\frac{x_{ij} - \overline{x}_j}{s_j}$ for $j = 1, 2, ...p$ and $\mathbf{D}$ is the diagonal matrix with diagonal elements $s_1^2, s_2^2, ..., s_p^2$ (Johnson and Wichern, 2007). As a result, the ij^{th} element of the standardized $n \times p$ data matrix $\mathbf{Z}$ is $z_{ij} = \frac{x_{ij} - \overline{x}_j}{s_j}$ for $i = 1, 2, ..., n$ and $j = 1, 2, ..., p$. The sample mean vector of the standardized data is $\overline{\mathbf{z}} = \mathbf{0}$ and the corresponding sample covariance matrix is $\mathbf{S_Z} = \mathbf{R}$. Similar to population principal components, the j^{th} sample principal component of the standardized data is of the form $\hat{y}_j = \hat{\mathbf{e}}_j' \mathbf{z}$, $j = 1, 2, ..., p$, where $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq ... \geq \hat{\lambda}_p \geq 0$ are the eigenvalues of $\mathbf{R}$ with corresponding normalized eigenvectors $\hat{\mathbf{e}}_j$ (Johnson and Wichern, 2007). Analogous to population principal components, the total sample variance of the standardized data is $\sum_{j=1}^{p} \hat{\lambda}_j = p$, and the sample correlation coefficient between $\hat{y}_j$ and z_l is $r_{\hat{y}_j, z_l} = \hat{e}_{jl} \sqrt{\hat{\lambda}_j}$ for $j, l = 1, 2, ..., p$. Based on this result, it is clear that $\frac{\hat{\lambda}_j}{p}$ is the relative amount of the total sample variance accounted for by the j^{th} sample principal component for $j = 1, 2, ..., p$ (Johnson and Wichern, 2007).

2.1.5 Important Principal Components to Retain

Different statistical techniques have been designed to determine which principal components should be retained and which should be excluded from the data set, and the decision of which technique to use is not set in stone.

The most common way of doing this is to compare the sizes of the eigenvalues and keep the principal components corresponding to the largest eigenvalues (Anderson, 1963). Another useful way of determining which principal components to keep is to construct a scree plot. This is a two dimensional scatterplot with the eigenvalues $\hat{\lambda}_j$ plotted on the y-axis and the j values plotted on the x-axis, meaning that the eigenvalues are plotted from largest to smallest. To make it easier to see the relative sizes of the eigenvalues, the points in this plot are connected by straight lines (Johnson and Wichern, 2007). One of the points appears to form an elbow, meaning that the points to the left of this elbow point greatly differ in relative size and the points to the right of this elbow point do not differ very much in relative size. The principal components corresponding to the eigenvalues at the elbow point and to the left of the elbow point are retained for making statistical inferences. Sometimes, when doing principal component analysis, the smallest eigenvalue(s) may be approximately equal to 0 due to rounding errors. This indicates that at least one of the variables in the data matrix is redundant because it is a linear combination of other variables. These redundant variables should be removed from the data matrix to avoid having issues with data interpretation (Johnson and Wichern, 2007).

2.1.6 How to Interpret Sample Principal Components

There are many different ways to interpret the sample principal components. If the distribution of each row of $\mathbf{X}$ is approximately $N_p(\mathbf{b}, \mathbf{\Sigma})$ where $\mathbf{b} = [\mu_1, \mu_2, ..., \mu_p]'$ and there are n independent observations on $\mathbf{X}$, then for $j = 1, 2, ..., p$, the sample principal components of the form $\hat{y}_j = \hat{\mathbf{e}}_j'(\mathbf{x} - \overline{\mathbf{x}})$ can represent the population principal components of the form $Y_j = \mathbf{e}_j'(\mathbf{X} - \mathbf{b})$

where $\mathbf{b} = [\mu_1, \mu_2, ..., \mu_p]'$. The distribution of the population principal components is thus $N_p(\mathbf{0}, \mathbf{\Lambda})$ where $\mathbf{\Lambda}$ is a diagonal matrix with diagonal entries $\lambda_1, \lambda_2, ..., \lambda_p$ and each λ_j value, $j = 1, 2, ..., p$ is the j^{th} largest eigenvalue of $\mathbf{\Sigma}$ with corresponding normalized eigenvector $\mathbf{e}_j$ (Johnson and Wichern, 2007). The population mean vector $\mathbf{b}$ and population covariance matrix $\mathbf{\Sigma}$ can be estimated by the sample mean vector $\overline{\mathbf{x}}$ and sample covariance matrix $\mathbf{S}$ respectively. Assuming that $\mathbf{S}$ is a positive definite matrix, one can construct a contour plot of all the data vectors $\mathbf{x}_i$ for $i = 1, 2, ..., n$ that satisfy the constant density curve $(\mathbf{x}-\overline{\mathbf{x}})'\mathbf{S}^{-1}(\mathbf{x}-\overline{\mathbf{x}}) = c^2$ where $\mathbf{x}$ represents all of the $\mathbf{x}_i$ vectors and c is an arbitrary constant (Anderson, 1963; Johnson and Wichern, 2007). This curve is an estimate of the constant density curve $(\mathbf{x}-\mathbf{b})'\mathbf{\Sigma}^{-1}(\mathbf{x}-\mathbf{b}) = c^2$, and the principal components can easily be seen by the axes of this elipsoid. Even though the normality assumption can be very useful for interpreting principal components, it is not needed for calculating the sample principal components from an $n \times p$ data matrix $\mathbf{X}$ (Anderson, 1963; Johnson and Wichern, 2007).

Regardless of whether the data are normally distributed and form an elliptical p-dimensional plot, the n data points can still be plotted in a p-dimensional plot with the eigenvectors of $\mathbf{S}$ as the positions of the new axes. A hyperellipsoid centered at the coordinate $\overline{\mathbf{x}}$ with axis lengths $c_j\sqrt{\hat{\lambda}_j}$ can be fitted to the data for $j = 1, 2, ..., p$ where c_j is an arbitrary constant and $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq ... \geq \hat{\lambda}_p \geq 0$ represent the eigenvalues of $\mathbf{S}$ (Johnson and Wichern, 2007). It can be shown that the j^{th} sample principal component expressed as $\hat{y}_j = \hat{\mathbf{e}}_j'(\mathbf{x} - \overline{\mathbf{x}})$ is found on top of the j^{th} hyperellipsoid axis for $j = 1, 2, ..., p$. Ultimately, the sample principal components yield a transformed data set that, when plotted in p-dimensions, has each j^{th} axis running through the coordinate $\overline{\mathbf{x}}$ in the direction of the maximum variance $\hat{\lambda}_j$ for $j = 1, 2, ..., p$ (Johnson

and Wichern, 2007). When $\hat{\lambda}_k > \hat{\lambda}_l$, then the ellipse formed by the k^{th} and l^{th} axes is non-circular where the diameter of the ellipse is larger along the k^{th} axis. When $\hat{\lambda}_k = \hat{\lambda}_l$, then the ellipse formed by the k^{th} and l^{th} axes is circular, and both axes are perpendicular to each other in any direction. This means that the k^{th} and l^{th} sample principal components are also perpendicular to each other in any direction (Hotelling, 1933; Johnson and Wichern, 2007). In practice, the eigenvalues are never exactly the same. When all of the eigenvalues of $\mathbf{S}$ are approximately equal to each other, the variation in the data is approximately the same for all of the p sample principal components, meaning that eliminating any of these principal components would be inappropriate (Johnson and Wichern, 2007).

2.1.7 The Asymptotic Behaviour of Sample Principal Components

Johnson and Wichern (2007) explain the asymptotic behaviour of sample principal components for large sample sizes. Lets suppose that $\mathbf{\Lambda}$ is a $p \times p$ diagonal matrix with respective diagonal entries λ_1, λ_2, ..., λ_p, which are the eigenvalues of $\mathbf{\Sigma}$ in decreasing order with corresponding eigenvectors $\mathbf{e}_1$, $\mathbf{e}_2$,...,$\mathbf{e}_p$. It is also important to note that $\hat{\lambda}_1$, $\hat{\lambda}_2$, ..., $\hat{\lambda}_p$ are the eigenvalues of $\mathbf{S}$ with corresponding eigenvectors $\hat{\mathbf{e}}_1$, $\hat{\mathbf{e}}_2$, ..., $\hat{\mathbf{e}}_p$ (Johnson and Wichern, 2007). One can show that $\sqrt{n}(\hat{\mathbf{w}}-\mathbf{w})$ has an $N_p(\mathbf{0}, 2\mathbf{\Lambda}^2)$ asymptotic distribution where $\mathbf{w}$ and $\hat{\mathbf{w}}$ are both vectors containing the true eigenvalues and the eigenvalue estimates respectively. It is also important to mention that $\sqrt{n}(\hat{\mathbf{e}}_j - \mathbf{e}_j)$ has an $N_p(\mathbf{0}, \mathbf{A}_j)$ asymptotic distribution where $\mathbf{A}_j = \lambda_j \sum_{t \in C} \frac{\lambda_t}{(\lambda_t - \lambda_j)^2} \mathbf{e}_t \mathbf{e}_t'$ and C is the set $[1, 2, ..., j-1, j+1, j+2, ..., p]$ (Johnson and Wichern, 2007). Using these results, the asymptotic $100(1-\alpha)\%$ confidence interval of λ_j is derived to be $\left[\frac{\hat{\lambda}_j}{1+z_{\alpha/2}\sqrt{2/n}}, \frac{\hat{\lambda}_j}{1-z_{\alpha/2}\sqrt{2/n}}\right]$ where $P(Z > z_{\alpha/2}) = \alpha/2$ and Z is a random

variable from the standard normal distribution. Lets suppose that $\hat{\mathbf{A}}_j$ is calculated the exact same way as the matrix $\mathbf{A}_j$ is calculated except all of the $\hat{\mathbf{e}}_j$ vectors are in place of the corresponding $\mathbf{e}_j$ vectors and all of the $\hat{\lambda}_j$ values are in place of the corresponding λ_j values. Then the square root of the s^{th} diagonal element of $\frac{1}{n}\hat{\mathbf{A}}_j$ is approximately the standard error of $\hat{e}_{js}$, which is the s^{th} element of the j^{th} eigenvector (Johnson and Wichern, 2007).

2.2 The Jackknife

The jackknife is a statistical technique used to reduce the bias of a particular estimator that provides a biased estimate of the parameter that it is estimating. In general, a random sample $Y_1, Y_2, ..., Y_n$ can be divided into u subsamples, each consisting of v observations with $n = uv$ (Miller, 1964, 1974^a, 1974^b). Lets define $\hat{\theta}$ and $\hat{\theta}_{-j}$ as two estimators of θ computed in the exact same way except $\hat{\theta}$ uses all n sample values and, for $j = 1, 2, ..., u$, $\hat{\theta}_{-j}$ omits the j^{th} subsample values from all n sample values. Then $\tilde{\theta} = u\hat{\theta} + (1-u)\frac{1}{u}\sum_{j=1}^{u}\hat{\theta}_{-j}$ is the jackknife estimator of θ, which estimates θ with less bias than $\hat{\theta}$ does (Miller, 1974^b). In particular, Miller (1964) explains how using the jackknife removes the $\frac{\alpha_1}{n}$ term from $E(\hat{\theta})$ of the form $E(\hat{\theta}) = \theta + \frac{\alpha_1}{n} + \frac{\alpha_2}{n^2} + O(n^2)$, making the expectation closer to θ thus reducing the bias. Quenouille (1956) also explains this latter point in detail.

2.2.1 The Most Common Form of the Jackknife

In practice, using $u = n$ and $v = 1$ is usually the optimal way of deriving the jackknife estimate. This treats each of the n observations as its own subsample with one observation, and is the most common form of the jackknife used for reducing bias (Miller, 1974^b). Efron and Stein (1981) explained how

to estimate the variance of a particular estimator using the jackknife for this $u = n$ and $v = 1$ scenario. Let $\mathbf{X} = (X_1, X_2, ...X_n)$ be a random sample of size n and let $\hat{\theta}(\mathbf{X})$ be the estimator for some parameter θ. Then the jackknife estimate of the variance of $\hat{\theta}(\mathbf{X})$ is $\hat{Var}\hat{\theta}(\mathbf{X}) = \frac{n-1}{n}\sum_{j=1}^{n}[\hat{\theta}_{-j} - \hat{\theta}_{(.)}]^2$. In this formula, $\hat{\theta}_{-j}$ and $\hat{\theta}(\mathbf{X})$ are both calculated in the exact same way except the j^{th} observation is omitted from $\mathbf{X}$ when calculating $\hat{\theta}_{-j}$, and $\hat{\theta}_{(.)} = \frac{1}{n}\sum_{j=1}^{n}\hat{\theta}_{-j}$ (Efron and Stein, 1981). Efron and Stein (1981) also showed that the bias of $\hat{\theta}(\mathbf{X})$ is estimated by $\hat{Bias}\hat{\theta}(\mathbf{X}) = (n-1)(\hat{\theta}_{(.)} - \hat{\theta}(\mathbf{X}))$ when using the jackknife.

2.2.2 A Specific Example of How to Use the Jackknife

For example, one can use $\hat{\mu^2} = \overline{x}^2$ to estimate the true value of μ^2 where $\mathbf{x} = [x_1, x_2, ..., x_n]$ is an *i.i.d.* sample of size n from a finite population or a probability density distribution. It is known that $E(\overline{X}^2) = \mu^2 + \frac{\sigma^2}{n}$ for all distributions, meaning that $\overline{x}^2$ is an asymptotically unbiased estimator for μ^2 when n is large, but is always biased especially when n is small. In particular, $E(\overline{X}^2) > \mu^2$ means that $\overline{x}^2$ overestimates μ^2 especially when n is small. In order to reduce the bias of this biased estimate, one can use the jackknife to estimate μ^2. In general, it has been shown that using the jackknife methods eliminates the bias terms of the form $\frac{\alpha_1}{n}$. In this example, there is only one bias term of the form $\frac{\alpha_1}{n}$, which is the $\frac{\sigma^2}{n}$ term from $E(\overline{X}^2)$ that happens to represent the entire bias. Using the jackknife expression discussed above, the jackknife estimate is

$$\begin{aligned}\tilde{\mu^2} &= u\hat{\mu^2} + (1-u)\frac{1}{u}\sum_{j=1}^{u}\hat{\mu^2}_{-j} \\ &= n\overline{X^2} + (1-n)\frac{1}{n}\sum_{j=1}^{n}\overline{X^2}_{-j}\end{aligned}$$

in this case. The expected value of this expression is

$$\begin{aligned}
E(\tilde{\mu^2}) &= E\left[n\overline{X^2} + (1-n)\frac{1}{n}\sum_{j=1}^{n}\overline{X^2}_{-j}\right] \\
&= nE\left[\overline{X^2}\right] + (1-n)\frac{1}{n}\sum_{j=1}^{n}E\left[\overline{X^2}_{-j}\right] \\
&= n\left[\mu^2 + \frac{\sigma^2}{n}\right] + \frac{1-n}{n}\sum_{j=1}^{n}\left[\mu^2 + \frac{\sigma^2}{n-1}\right] \\
&= n\mu^2 + \sigma^2 + \frac{1-n}{n}\left[n\mu^2 + \frac{n\sigma^2}{n-1}\right] \\
&= n\mu^2 + \sigma^2 + (1-n)\mu^2 - \sigma^2 \\
&= n\mu^2 + \sigma^2 + \mu^2 - n\mu^2 - \sigma^2 \\
&= \mu^2
\end{aligned}$$

which means that $\tilde{\mu^2}$ is always an unbiased estimate of μ^2. Therefore, the bias term in $E(\overline{X}^2) = \mu^2 + \frac{\sigma^2}{n}$ is completely removed, and this verifies that using the jackknife methods in this case completely removes the bias in the estimate of μ^2.

2.2.3 A Concrete Illustration of This Example

To illustrate a specific case of this example, let $\mathbf{x} = [x_1, x_2, ..., x_n]$ be an *i.i.d.* sample of size $n = 20$ sampled from a normal distribution with $\mu = 20$ and $\sigma = 1$. Using the codes presented in the Appendix section, the biased estimate and jackknife estimate of μ^2 were calculated to be $\hat{\mu^2} = \overline{x}^2 = 406.3838$ and $\tilde{\mu^2} = 406.3538$ respectively. The bias and variance estimates of the jackknife estimate were calculated using the respective formulas defined above, and they are 0.03000 and 48.7416 respectively. Since the samples were obtained from a normal distribution with $\mu = 20$ and $\sigma = 1$, the true value of μ^2 is known

to be $20^2 = 400$. Since the jackknife estimate is closer to this value than the biased estimate is, this suggests that using the jackknife methods reduces bias in the estimate.

It is important to note that these calculations appear to be based on a sample of size $n = 20$ that coincidentally represents the normal distribution with $\mu = 20$ and $\sigma = 1$ very accurately. This is not always the case, because having a small sample size means having a larger variance, and it is more probable of having a sample with a sample mean that is much larger or smaller than $\mu = 20$. This is due to sampling error, which is more prominent in smaller sample sizes. It is also important to note that the jackknife estimate is always smaller than the original biased estimate in this particular example, because $\overline{x}^2$ overestimates μ^2 and the Jackknife method reduces the bias by decreasing the value of $\overline{x}^2$ to obtain the unbiased jackknife estimate $\tilde{\mu^2}$. This means that the biased estimates calculated in any samples skewed left are decreased to obtain the jackknife estimate. As a result, the jackknife estimate is further away from $\mu^2 = 400$ than the original biased estimate is. This would be problematic, because the true value of the parameter that one would like to estimate is unknown in practice, and there is no way of telling how good the jackknife estimate really is. Nevertheless, the codes in the Appendix section for this particular example were run numerous times, and it was found that the jackknife estimate of μ^2 is closer to $\mu^2 = 400$ than the biased estimate is in most of the cases.

2.3 The Bootstrap

Despite the fact that the bias and variance of a particular random variable based on a random sample can be estimated using the jackknife, the

bootstrap methods usually produce more reliable estimations of the bias and variance than the jackknife does. One can bootstrap a random sample of n observations collected from a population with a common but unknown distribution F where $\mathbf{X} = [X_1, X_2, ..., X_n]$ is the vector of random variables representing the randomly sampled observations and $\mathbf{x} = [x_1, x_2, ..., x_n]$ represents the observed random sample values (Efron, 1979). The objective of bootstrapping is to derive an expression to estimate the sampling distribution of some random variable $R(\mathbf{X}, F)$ by using the observed $\mathbf{x}$ values. Normally, one of two expressions of $R(\mathbf{X}, F)$ is used to do the jackknife, which can be used to approximate the results obtained from using the bootstrap. One of these expressions is $R(\mathbf{X}, F) = t(\mathbf{X}) - \theta(F)$ where $\theta(F)$ is a particular parameter describing the function of F and $t(\mathbf{X})$ is an expression derived to estimate the true value of $\theta(F)$. The other expression is $R(\mathbf{X}, F) = \frac{t(\mathbf{X}) - \hat{Bias}(t) - \theta(F)}{\sqrt{\hat{Var}(t)}}$. Both $\hat{Bias}(t)$ and $\hat{Var}(t)$ represent the estimates of the bias and variance of $t(\mathbf{X})$, which are calculated using the n vectors of the $\mathbf{x}$ vector with the i^{th} element deleted, $i = 1, 2, ..., n$ (Efron, 1979).

When one random sample of $\mathbf{x}$ values is drawn from a population with distribution F, the sample distribution of the $\mathbf{x}$ values, denoted $\hat{F}$, has a probability distribution function of $\frac{1}{n}$ at each of the $\mathbf{x}$ values and 0 elsewhere. To do the bootstrap, n values are randomly sampled with replacement from the n $\mathbf{x}$ values to generate a bootstrap sample with random variables $\mathbf{X}^* = [X_1^*, X_2^*, ..., X_n^*]$ and observed values $\mathbf{x}^* = [x_1^*, x_2^*, ..., x_n^*]$ (Efron, 1981; Efron, 1982). The bootstrap forms its own distribution, and the bootstrapped random variable $R(\mathbf{X}^*, \hat{F})$ is derived to estimate the distribution of some random variable $R(\mathbf{X}, F)$. Since the distributions of $R(\mathbf{X}^*, \hat{F})$ and $R(\mathbf{X}, F)$ are equal when $F = \hat{F}$, this means that the distribution of $R(\mathbf{X}^*, \hat{F})$ better ap-

proximates the distribution of $R(\mathbf{X}, F)$ when the $\mathbf{x}$ values accurately represent the population from where they were sampled. The actual function chosen for $R(\mathbf{X}, F)$ also affects the effectiveness of its estimator $R(\mathbf{X}^*, \hat{F})$ (Efron, 1979).

2.3.1 How to Determine the Bootstrap Distribution of Estimators

In practice, figuring out the bootstrap distribution is usually a hard and tedious process, and three methods have been formulated to derive this distribution. One way is to calculate the exact distribution based on theoretical results. One can also take N separate random samples from $\mathbf{x}$ where each sample is denoted as $\mathbf{x}_j^*$, $j = 1, 2, ..., N$, and then calculate all of the corresponding $R(\mathbf{x}_j^*, \hat{F})$ values to form a distribution that approximates the bootstrap distribution (Efron, 1979). Efron (1981) mentions that the estimates of the mean and standard deviation of the bootstrap distribution are simply the sample mean and sample standard deviation of the $R(\mathbf{x}_j^*, \hat{F})$ values. Analogous to jackknifing, the third method is to estimate the distribution of $R(\mathbf{X}^*, \hat{F})$ by estimating its mean and variance based on the Taylor series (Efron, 1979).

To explain this third method, lets suppose that the bootstrap sample $\mathbf{x}^*$ of size n is obtained with replacement from the observed values $\mathbf{x}$. For $i = 1, 2, ..., n$, let $W_i^* = \frac{1}{n} N_i^*$ where the random variable N_i^* represents the number of occurrences of x_i in the bootstrap sample (Efron, 1979; Efron and Stein, 1981). It has been established that $\mathbf{W}^* = [W_1^*, W_2^*, ..., W_n^*]$ is multinomially distributed with mean vector $\frac{1}{n}\mathbf{s}$ and covariance matrix $\frac{1}{n^3}(n\mathbf{I} - \mathbf{J})$ where $\mathbf{s} = [1, 1, ..., 1]$ is the vector with n elements of 1, $\mathbf{I}$ is the $n \times n$ identity matrix and $\mathbf{J}$ is the $n \times n$ matrix with all n^2 elements of 1 (Efron, 1979; Efron and Stein, 1981). The sampling distribution of the random variable $R(\mathbf{X}, F)$ can be estimated by the distribution of $R(\mathbf{X}^*, \hat{F})$. Using the Taylor series, $R(\mathbf{X}^*, \hat{F}) =$

$R(\mathbf{W}^*)$ can be expanded to

$$R\left[\mathbf{W}^*\right] \approx R\left[\frac{1}{n}\mathbf{s}\right] + \left(\mathbf{W}^* - \frac{1}{n}\mathbf{s}\right)\mathbf{A} + 0.5\left(\mathbf{W}^* - \frac{1}{n}\mathbf{s}\right)\mathbf{B}\left(\mathbf{W}^* - \frac{1}{n}\mathbf{s}\right)' \quad (1)$$

where $\mathbf{A}$ is the $n \times 1$ vector with i^{th} element $a_i = \frac{\partial R(\mathbf{W}^*)}{\partial W_i^*} \mid_{\mathbf{W}^*=\frac{1}{n}\mathbf{s}}$ for $i = 1, 2, ..., n$ and $\mathbf{B}$ is the $n \times n$ matrix with ij^{th} element $b_{ij} = \frac{\partial^2 R(\mathbf{W}^*)}{\partial W_i^* \partial W_j^*} \mid_{\mathbf{W}^*=\frac{1}{n}\mathbf{s}}$ for $i, j = 1, 2, ..., n$ (Efron, 1979). Using these results, Efron (1979) illustrated how the bias and variance of $R(\mathbf{W}^*)$ are approximately $\frac{1}{2n^2}\sum_{i=1}^{n} b_{ii}$ and $\frac{1}{n^2}\sum_{i=1}^{n} a_i^2$ respectively, and how these expressions are very similar to the corresponding expressions for the jackknife.

The bootstrapping techniques can also be applied to two independent random samples $\mathbf{x} = [x_1, x_2, ..., x_n]$ and $\mathbf{y} = [y_1, y_2, ..., y_n]$ collected from two distinct populations with respective distributions F and G. Similar to the one-sample case, the objective is to derive an estimate of the distribution of some random variable $R(\mathbf{X}, \mathbf{Y}, F, G)$ using the observed $\mathbf{x}$ and $\mathbf{y}$ values with respective distributions $\hat{F}$ and $\hat{G}$ (Efron, 1979). A bootstrap sample of size n, denoted $\mathbf{x}^* = [x_1^*, x_2^*, ..., x_n^*]$, can be drawn with replacement from the observed $\mathbf{x}$ values, and a different bootstrap sample independent of the previous sample of size n, denoted $\mathbf{y}^* = [y_1^*, y_2^*, ..., y_n^*]$, can be drawn with replacement from the observed $\mathbf{y}$ values. One can use these bootstrap samples to determine the distribution of $R(\mathbf{X}^*, \mathbf{Y}^*, \hat{F}, \hat{G})$ to estimate the distribution of $R(\mathbf{X}, \mathbf{Y}, F, G)$, and the distribution of $R(\mathbf{X}^*, \mathbf{Y}^*, \hat{F}, \hat{G})$ is determined using one of the three ways discussed in the one-sample case (Efron, 1979).

2.3.2 An Extenstion to the Example Discussed in the Jackknife Section Using the Bootstrapping Methods

The example discussed in the Jackknife section was applied to the bootstrap by using $\overline{x}^2$ to estimate the true value of μ^2 where $\mathbf{x} = [x_1, x_2, ..., x_n]$ is an *i.i.d.* sample of size n from a finite population or a probability density distribution. Again, $\mathbf{x} = [x_1, x_2, ..., x_n]$ is an *i.i.d.* sample of size $n = 20$ sampled from a normal distribution with $\mu = 20$ and $\sigma = 1$. Using the codes presented in the Appendix section, the bootstrap mean, standard deviation and deviation (the absolute difference between the true value and boostrap mean) were calculated to be 406.3488, 6.7920 and 0.03503 respectively. As discussed in the example in the Jackknife section, the true value of μ^2 is known to be $20^2 = 400$. The bootstrap mean is also very close to this value, and it is closer to μ^2 than the Jackknife estimate $\tilde{\mu^2}$ is. However, since the bias is of order $\frac{1}{n}$, one would expect that the Jackknife estimate would be closer to μ^2 since the Jackknife removes all of the bias terms of order $\frac{1}{n}$, meaning that the jackknife technique completely removes the bias unlike the bootstrap technique in this particular example. This suggests that the jackknife technique reduces the bias more effectively than the bootstrap technique does, which would not be true when the bias contains terms that are higher order than $\frac{1}{n}$. The bootstrap standard deviation also appears to be slightly smaller than the jackknife estimate of the standard deviation, which may also suggest that the bootstrap methods are more optimal to use.

As mentioned in the Jackknife section, the calculations for the bootstrap appear to be based on a sample of size $n = 20$ which happens to coincidentally represent the normal distribution with $\mu = 20$ and $\sigma = 1$ very accurately.

This does not always happen in practice, because small samples have larger variances, and having a sample with a sample mean that is much larger or smaller than $\mu = 20$ is more probable. Again, the codes in the Appendix section for this particular example were run numerous times, and it was found that the bootstrap estimate of μ^2 is closer to $\mu^2 = 400$ than the biased estimate is in most of the cases. Another reason why the bootstrap technique is prefered over the jackknife technique is that the bootstrap method does not always decrease the biased estimate of μ^2 like the jackknife method does. This is most likely because the bootstrap technique estimates the sampling distribution. Thus, this phenomenon also supports the claim that the bootstrap estimate is more reliable than the jackknife estimate is.

2.4 Information About the Data

Edge *et al.* (2010) wished to investigate the water quality of the Bras d'Henri and Forchette River in Quebec (Bras), the Oldman River in Alberta (Old), the South Nation River in Ontario (South), and the Sumas River in British Columbia (Sumas). The exact geographical locations of these watersheds can be seen in Figure 1 and Figure 2. They selected these watersheds, because they are geographically close to agricultural farmlands that are mostly responsible for fecal waste entering Canada's rivers, and it makes sense to investigate the most extreme cases (Edge *et al.*, 2010). The Bras, Old, South and Sumas watersheds contain six, eight, five and four Agricultural (Ag) sites respectively, which are directly affected by the waste produced by the agricultural farmlands that enters the water. Each watershed also contains four Reference (Ref) sites that are upstream from the Ag sites and not affected by agricultural farmland waste (Edge *et al.*, 2010).

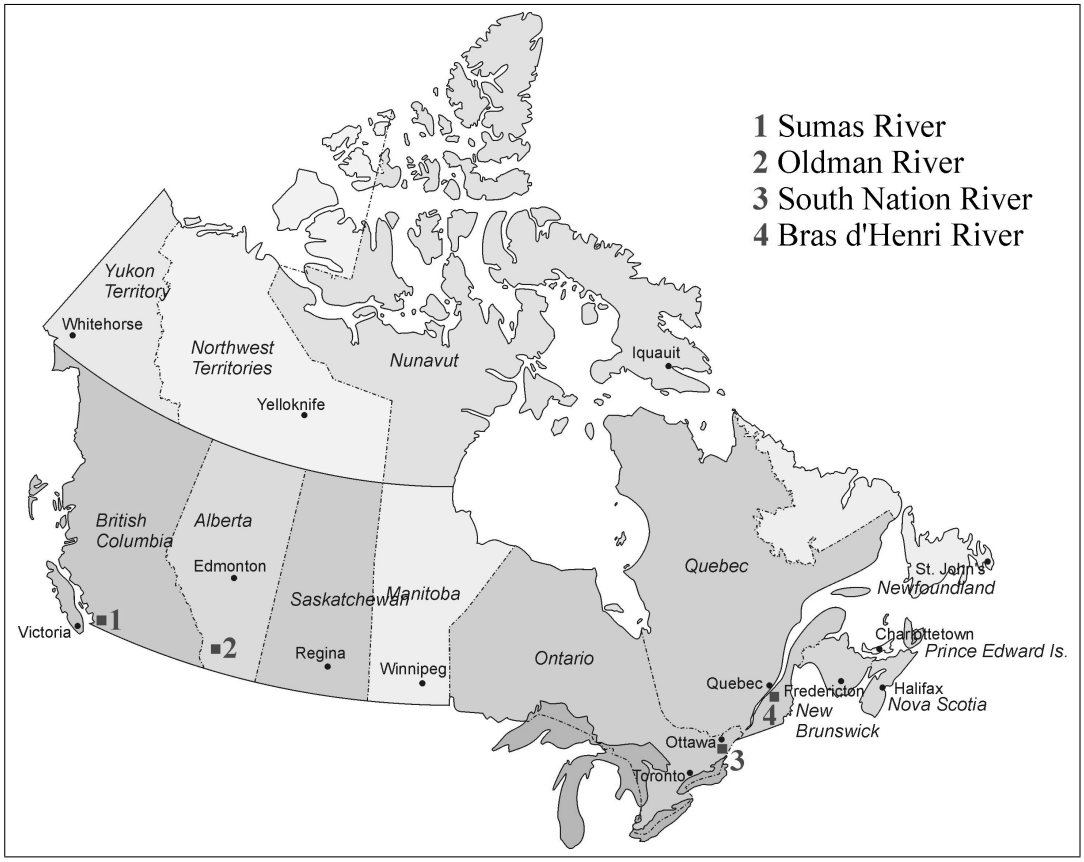
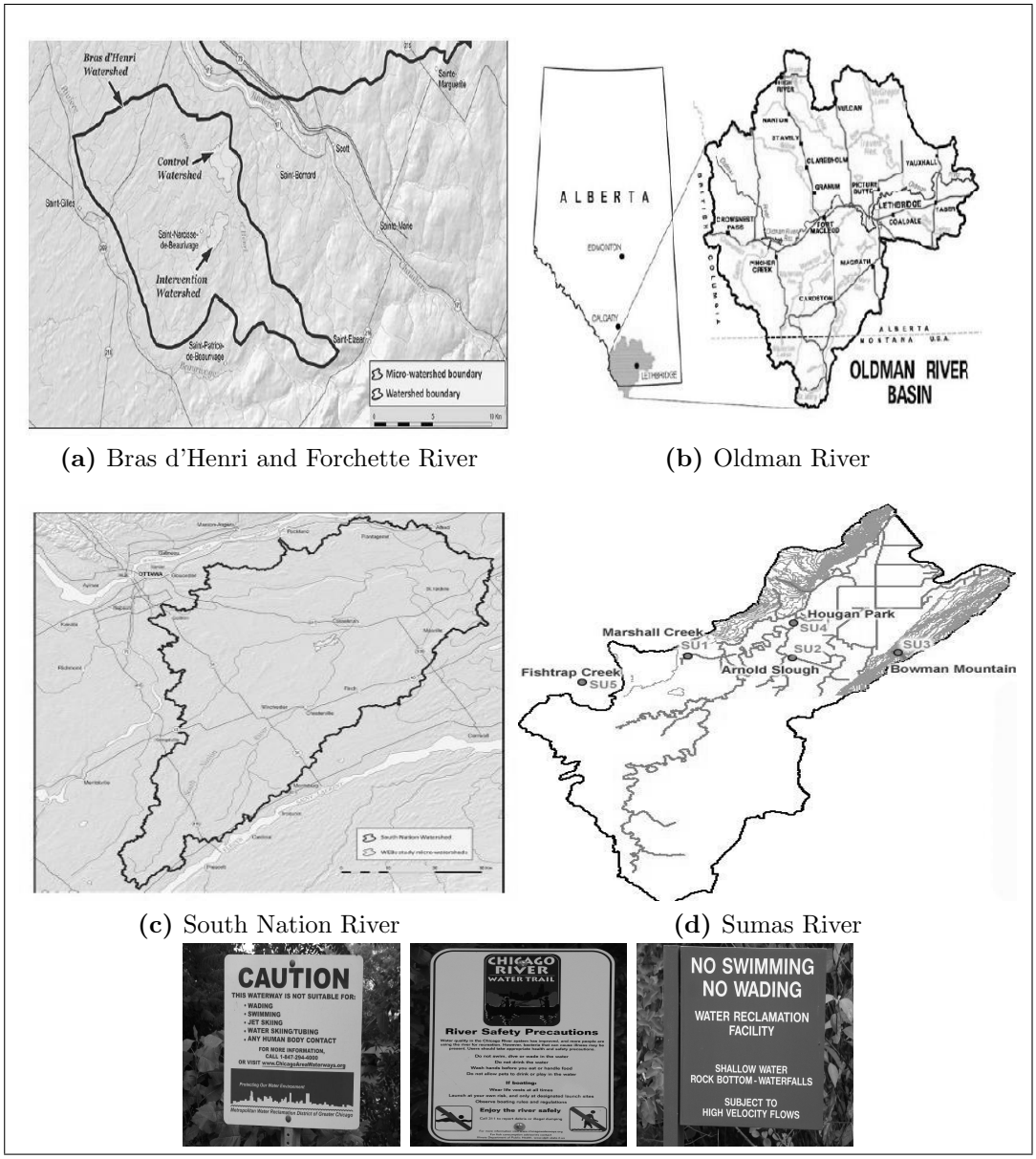


Figure 1: A map of Canada taken from Edge *et al.* (2010) showing the exact geographical locations of the four watersheds

There is potentially a major distinction in the water quality of these two types of watershed sites. The water in the Ref sites is not affected by the agricultural sites since it is not contaminated by any agricultural waste that may enter the streams (Edge *et al.*, 2010). In contrast, the water in the Ag sites may be affected by runoff water from agricultural farmlands. These two distinctions were used to compare the water quality before and after the water may become affected by agricultural waste to determine if agricultural waste significantly affects water quality in the watersheds. This is the main reason why the watershed sites are classified as Ag and Ref sites (Edge *et al.*, 2010).

Multiple different times from 2005 to 2007, Edge *et al.* (2010) collected water samples from each site when the water surface was not covered with ice.

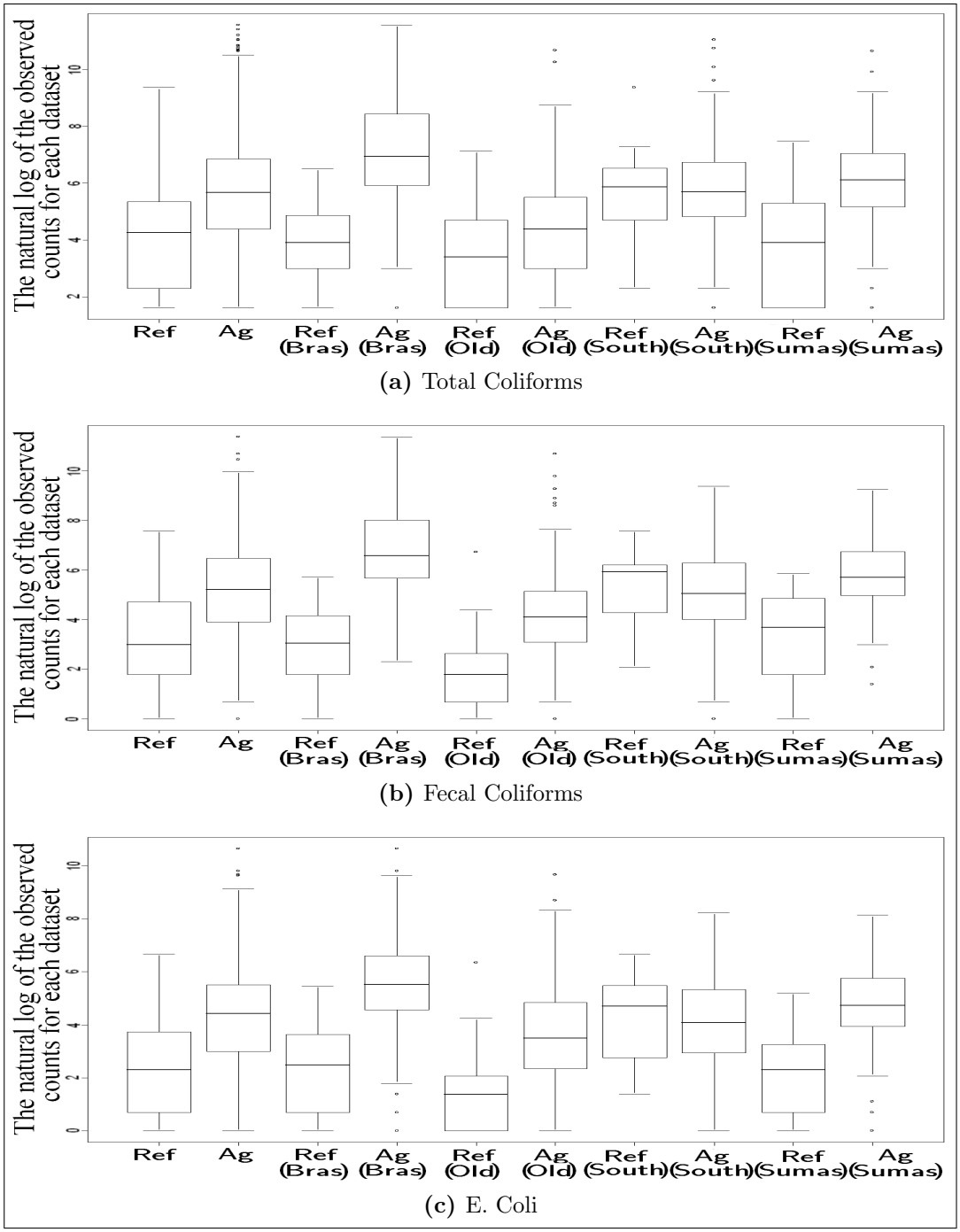


(a) Bras d'Henri and Forchette River **(b)** Oldman River

(c) South Nation River **(d)** Sumas River

Figure 2: A close-up geographical view of each watershed and a few of the warning signs found near the watersheds, which were all taken from Edge *et al.* (2010)

They used a sterilized bottle with a four liter capacity made from polypropylene to collect each water sample (Edge *et al.*, 2010). The samples, which were about 20 liters each, were stored in sealed plastic bags on top of ice to keep them frozen while being delivered to the places that tested the water quality. When testing the water samples, Edge *et al.* (2010) observed different factors



(a) Total Coliforms

(b) Fecal Coliforms

(c) E. Coli

Figure 3: The box plots of the Total Coliforms, Fecal Coliforms and E. coli natural log counts, each consisting of individual plots corresponding to the 10 datasets

pertaining to water quality (e.g. bacteria counts, water temperature, dissolved oxygen concentration, mineral concentrations and pH).

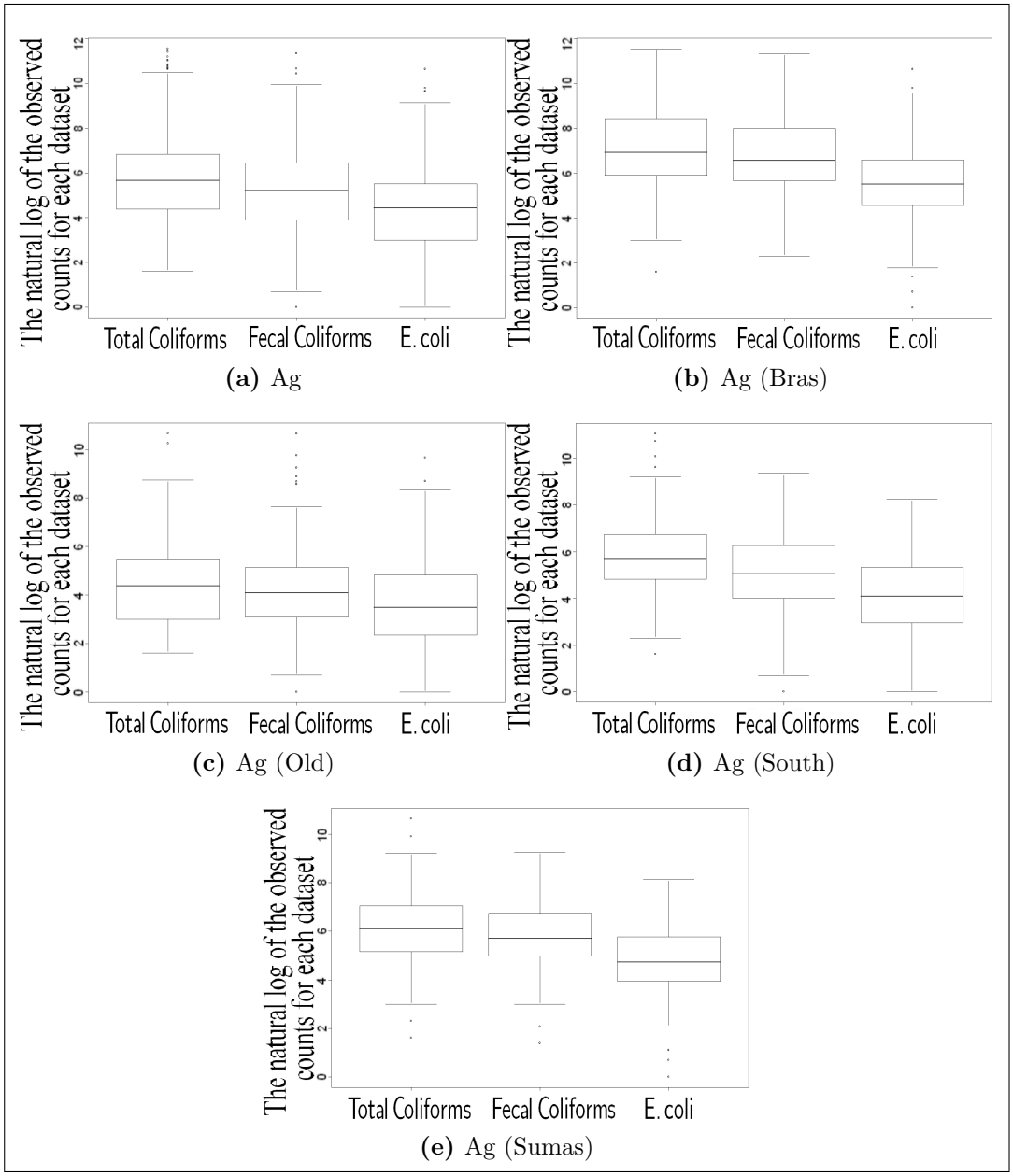
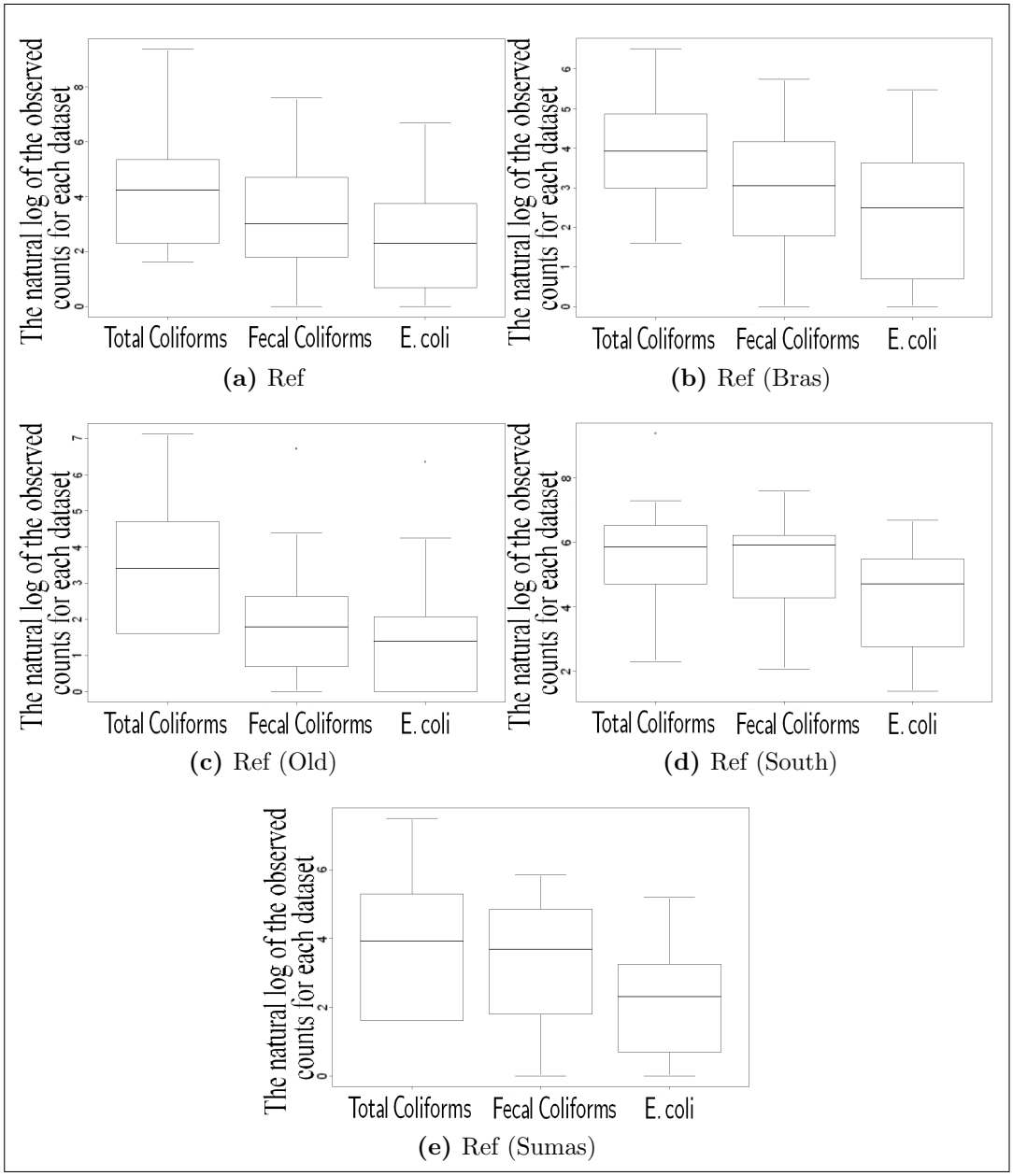


Figure 4: The box plots of the five datasets pertaining to the sites affected by agriculture, each consisting of individual plots corresponding to the Total Coliforms, Fecal Coliforms and E. coli natural log counts

The data used in this paper was borrowed from Edge *et al.* (2010). It is divided into Ag and Ref sites, each consisting of the four different watersheds. For both the Ag and Ref sites, there are five data sets; four of these data sets correspond to each of the four watersheds and the fifth data set corresponds to



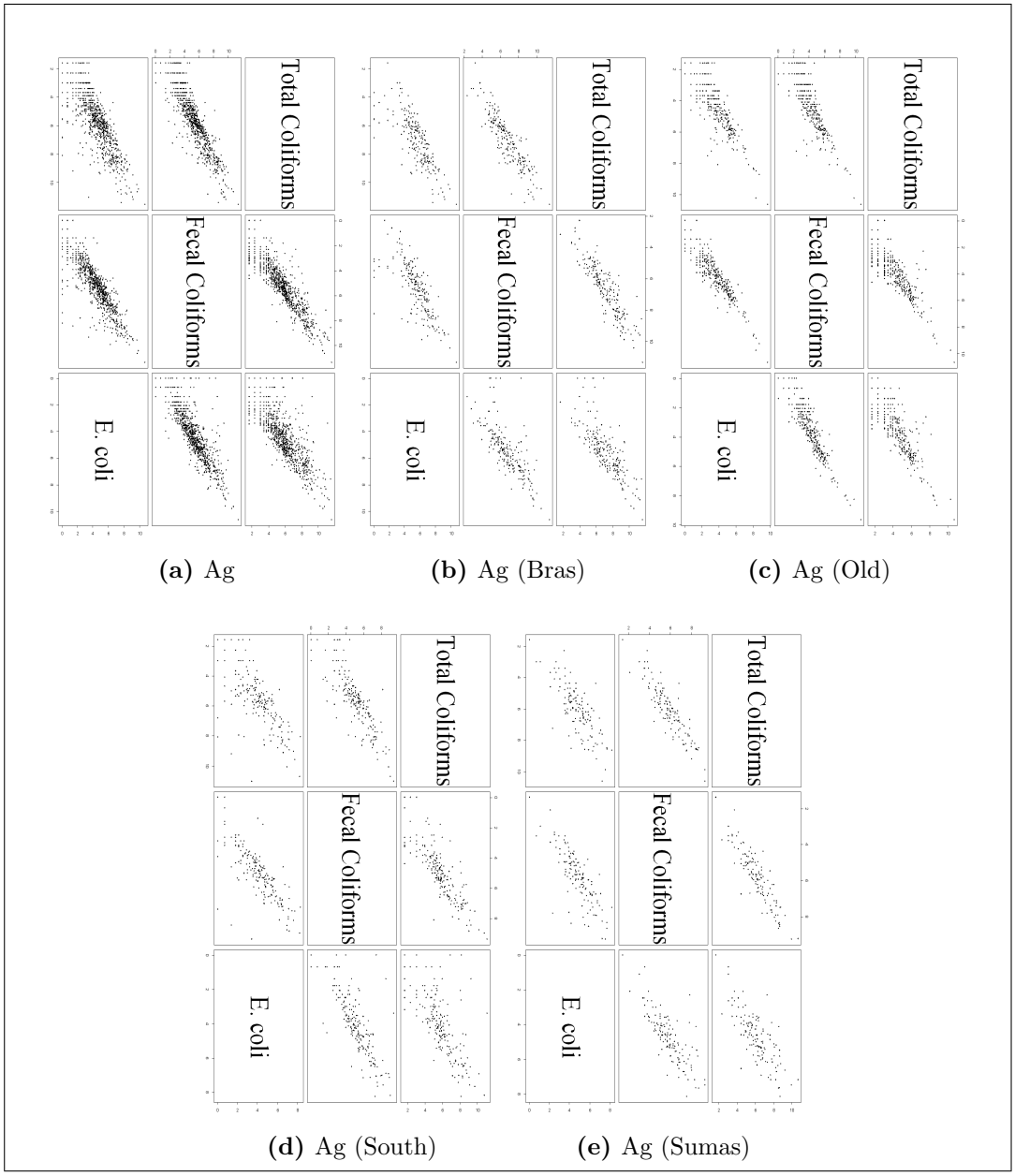
(a) Ref **(b)** Ref (Bras)

(c) Ref (Old) **(d)** Ref (South)

(e) Ref (Sumas)

Figure 5: The box plots of the five datasets pertaining to the sites not affected by agriculture, each consisting of individual plots corresponding to the Total Coliforms, Fecal Coliforms and E. coli natural log counts

all of the observations from all four watersheds pooled into one huge data set. Ultimately, there are a total of ten distinct data sets used throughout this paper for data analysis. Each data set contains four columns of values observed from the applicable watershed sites; the first column contains the values representing



(a) Ag **(b)** Ag (Bras) **(c)** Ag (Old)

(d) Ag (South) **(e)** Ag (Sumas)

Figure 6: The scatter matrices of the five datasets pertaining to the sites affected by agriculture, each consisting of individual plots corresponding to the Total Coliforms, Fecal Coliforms and E. coli natural log counts

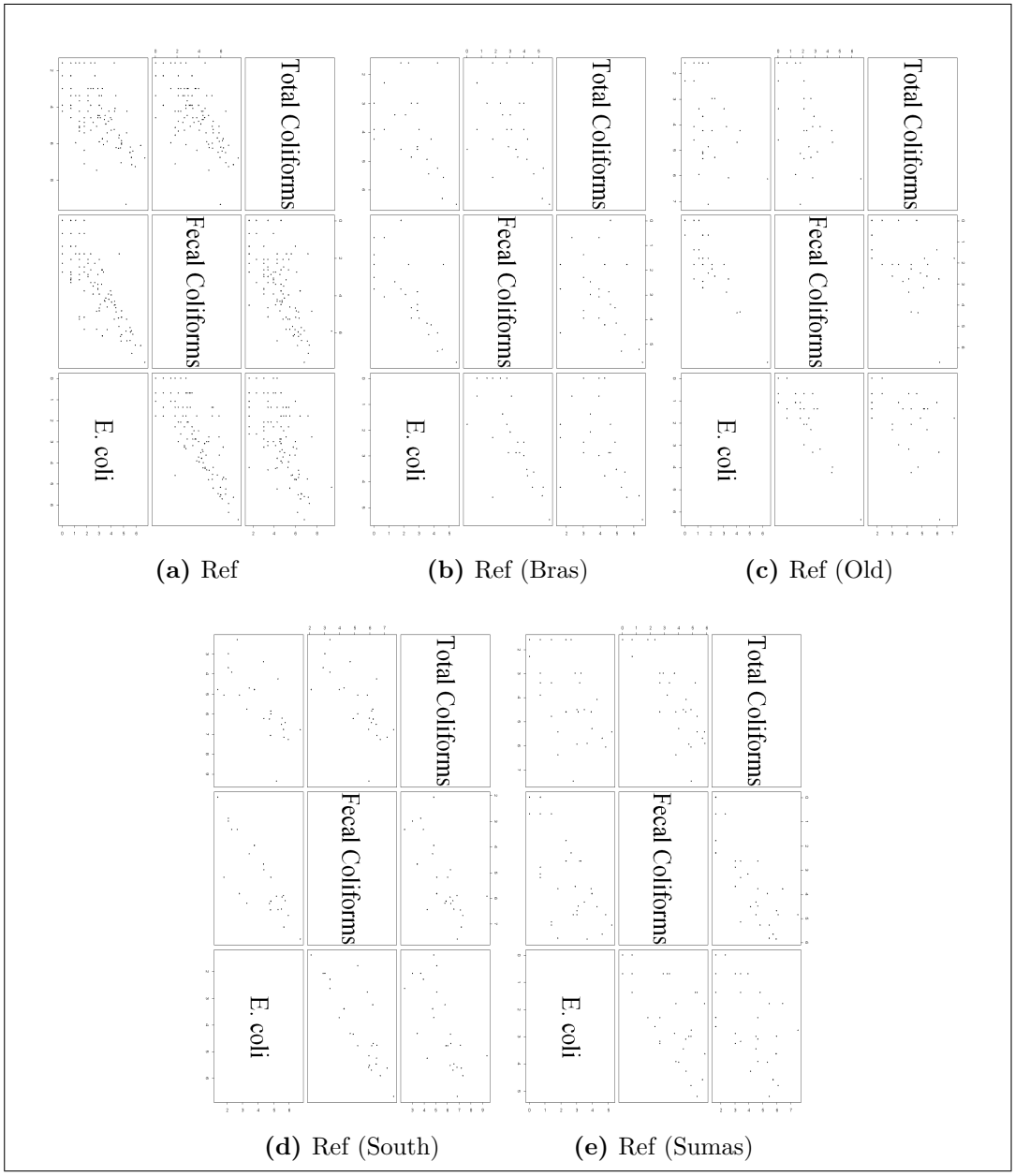
when the observations were made and the other three columns contain the natural logarithm of the Total coliform, Fecal coliform and E. coli counts. In each data set, the values in the first column are referred to as Julian days, which are chosen so that the earliest observation has a value of zero and each of

the remaining values is the number of days after the first observation is made. The data sets, denoted Ag, Ag (Bras), Ag (Old), Ag (South), Ag (Sumas), Ref, Ref (Bras), Ref (Old), Ref (South) and Ref (Sumas), contain 825, 206, 304, 187, 128, 121, 26, 37, 25 and 33 rows respectively.

It is important to note that some of the counts in the raw data set had a value of zero or were missing. Leaving these observations in the data set would have been problematic, because it would be very difficult or impossible to do data analysis when the columns are not equal length and the natural logarithm of zero is undefined. To avoid these issues, the rows containing these observations were excluded from the data set and the row counts mentioned earlier. This is reasonable, because the sample sizes of each of the ten data sets are large, and excluding a few observations would not really affect the sample sizes. An alternative solution that was considered was to replace the counts that are missing or zero with estimated values using missing data mechanisms. This approach was not used, because the calculations are very tedious and it is not worth going through all the work since the first solution discussed works fine with large sample sizes.

3 Preliminary Data Analysis

All of the organized data were inputted into the program R. Several different statistical techniques were executed in R either by using the libraries in R or implementing codes without the use of libraries. Before doing any statistical calculations, box plots and scatter matrices were created in R for visualizing trends in the data. In order to compare the counts of each pathogen amongst the ten different watershed sites, three sets of boxplots were created

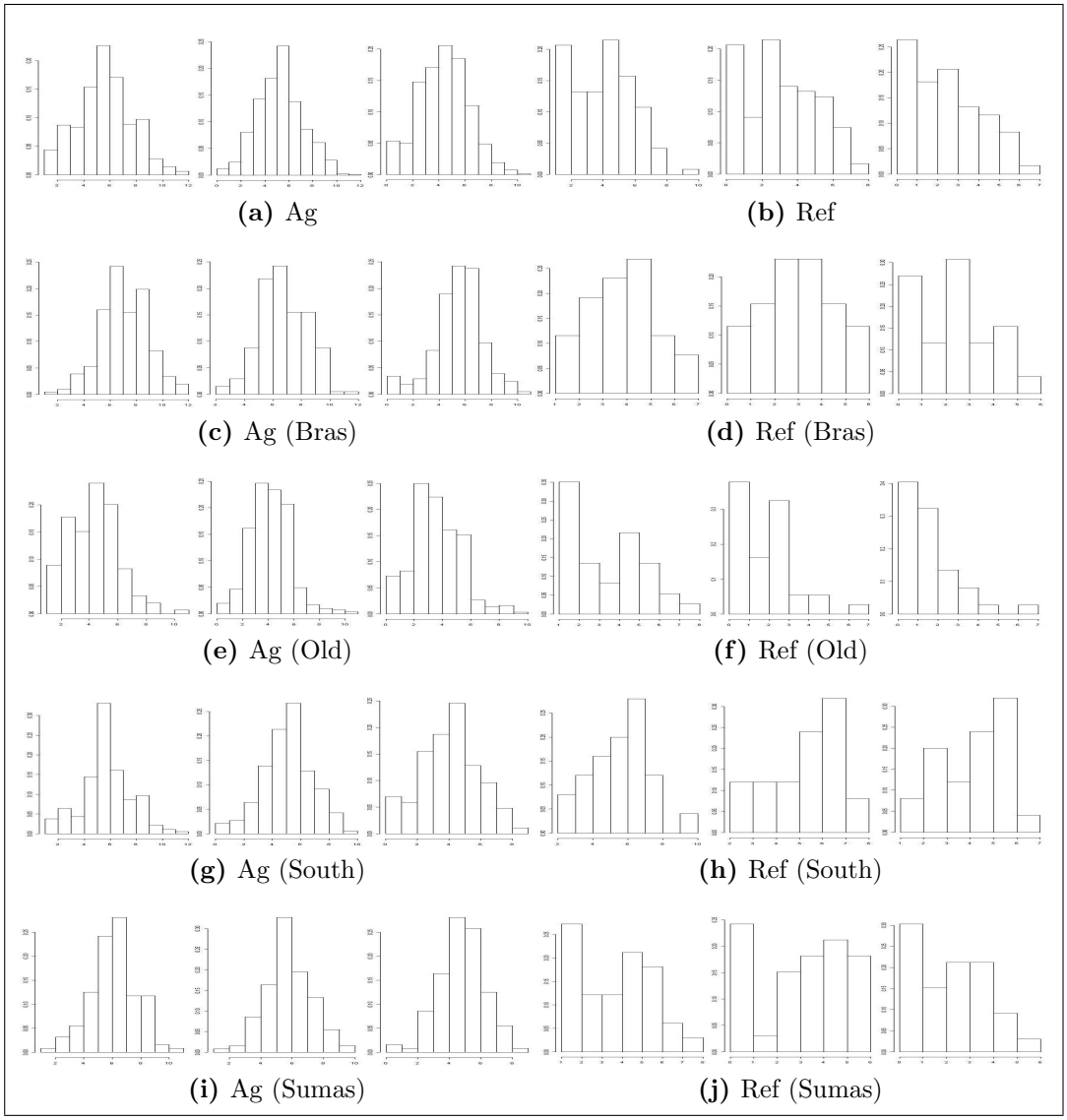


(a) Ref **(b)** Ref (Bras) **(c)** Ref (Old)

(d) Ref (South) **(e)** Ref (Sumas)

Figure 7: The scatter matrices of the five datasets pertaining to the sites not affected by agriculture, each consisting of individual plots corresponding to the Total Coliforms, Fecal Coliforms and E. coli natural log counts

in R corresonding to the Total Coliforms, Fecal Coliforms and E. coli natural log counts. Each of these plots consists of ten boxplots made from the ten data sets, and are displayed in Figure 3 with the ten watershed site names along the horizontal axes. It is seen in Figure 3 that the boxplots for the



(a) Ag **(b)** Ref

(c) Ag (Bras) **(d)** Ref (Bras)

(e) Ag (Old) **(f)** Ref (Old)

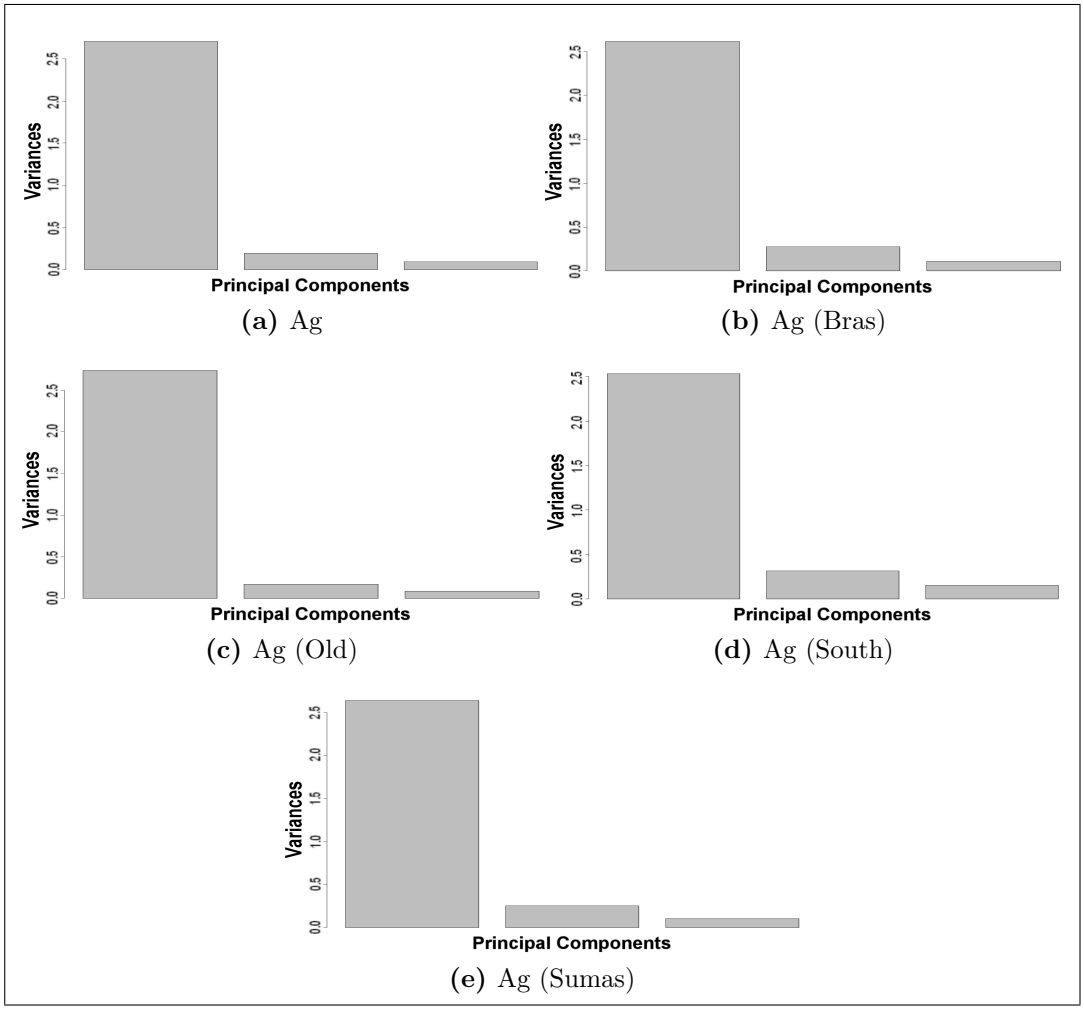
(g) Ag (South) **(h)** Ref (South)

(i) Ag (Sumas) **(j)** Ref (Sumas)

Figure 8: The histograms of the ten datasets (for each dataset, the plot on the left, in the middle and on the right correspond to the Total Coliforms, Fecal Coliforms and E. coli natural log counts respectively)

Ref sites are much lower than those for the Ag sites except the boxplots are roughly the same for the South watershed for each of the three pathogen types. This is a strong indication that all of the agricultural farmlands adjacent to the watersheds except for those adjacent to the South watershed significantly contribute to watershed contamination.

The Mann-Whitney U test for 2 independent samples was done for the



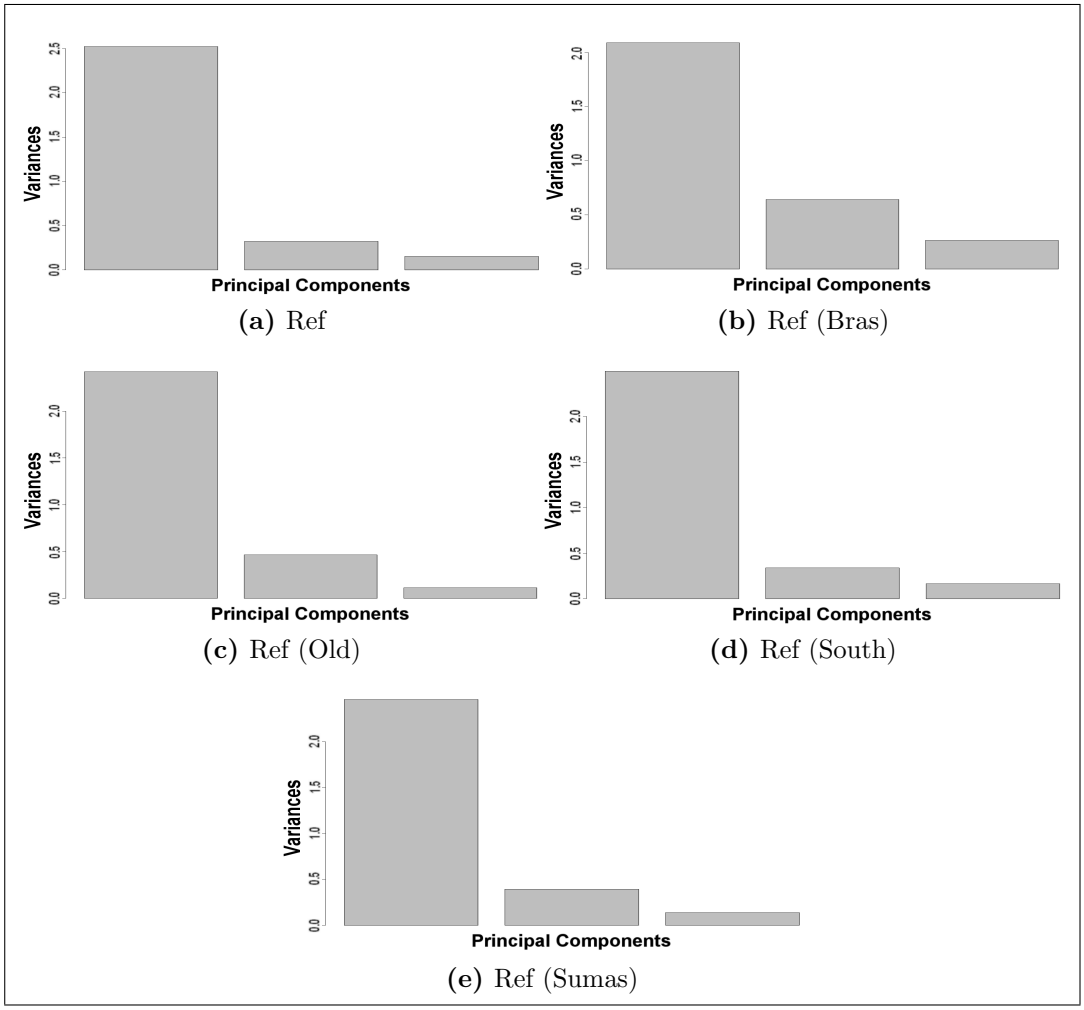
(a) Ag **(b)** Ag (Bras)

(c) Ag (Old) **(d)** Ag (South)

(e) Ag (Sumas)

Figure 9: The screeplots of the five datasets pertaining to the sites affected by agriculture, which display the variance of each principal component

E. coli natural log counts to test if there is a significant difference between the means of the E. coli natural log counts for the Ag and Ref data. There were five Mann-Whitney U tests done in total. They corresponded to the Ag and Ref datasets, the Ag (Bras) and Ref (Bras) datasets, the Ag (Old) and Ref (Old) datasets, the Ag (South) and Ref (South) datasets, and the Ag (Sumas) and Ref (Sumas) datasets. The respective W test statistics are 76498.5, 4841.5, 9441.5, 2203.5 and 3738.5 with respective p-values of $< 2.2 \times 10^{-16}$, 1.975×10^{-11}, 1.531×10^{-11}, 0.643 and 9.752×10^{-12}. Based on the p-values, it is



(a) Ref **(b)** Ref (Bras) **(c)** Ref (Old) **(d)** Ref (South) **(e)** Ref (Sumas)

Figure 10: The screeplots of the five datasets pertaining to the sites not affected by agriculture, which display the variance of each principal component

evident that the difference in mean values is significant for all the watersheds except the South watershed, which supports the observations made from the boxplots in Figure 3.

For each of the ten different data sets, one plot containing three box plots corresponding to the Total Coliform, Fecal Coliform and E. coli natural log counts was constructed in R with the pathogen names along the horizontal axis (see Figure 4 and Figure 5). For each data set, it is evident that the natural log concentration of pathogens is highest for Total Coliforms, second

highest for Fecal Coliforms and lowest for E. Coli. However, the difference in these three boxplots is very small, which indicates that there is slightly less E. Coli than Fecal Coliforms and that there are not that many other pathogens since the Total Coliform values are slightly larger than the other two values. One scatter matrix was also constructed in R for each of the ten data sets, each consisting of the scatter plots of all possible pairwise combinations of the Total Coliform, Fecal Coliform and E. coli natural log counts (see Figure 6 and Figure 7). All of the scatterplots for each of the ten different data sets appear to be weakly linearly correlated and positively correlated. This suggests that the Total Coliforms, Fecal Coliforms and E. Coli natural log counts are somewhat pairwise positively linearly correlated. This presumably means that a high presence of one pathogen type implies a high presence of the other two pathogen types and that a low presence of one pathogen type implies a low presence of the other two pathogen types. However, this noticable correlation appears not to be very strong, meaning that there may be a slight chance that the correlation is not completely linear.

Table 1: The variances of the principal components of **R** for each of the ten datasets

Dataset	Variances		
	sd_1^2	sd_2^2	sd_3^2
Ag	2.7057	0.1966	0.0978
Ag (Bras)	2.6140	0.2762	0.1099
Ag (Old)	2.7394	0.1725	0.0882
Ag (South)	2.5303	0.3152	0.1545
Ag (Sumas)	2.6403	0.2547	0.1050
Ref	2.5217	0.3255	0.1530
Ref (Bras)	2.0898	0.6450	0.2653
Ref (Old)	2.4205	0.4636	0.1159
Ref (South)	2.4948	0.3407	0.1645
Ref (Sumas)	2.4621	0.3978	0.1401

For each of the ten data sets, the three principal components and their

Table 2: The coefficients of the first principal component of **R** for each of the ten datasets

Dataset	First Principal Components		
	Total Coliforms	Fecal Coliforms	E.Coli
Ag	-0.5782	-0.5865	-0.5672
Ag (Bras)	-0.5900	-0.5847	-0.5568
Ag (Old)	-0.5679	-0.5830	-0.5810
Ag (South)	-0.5754	-0.5953	-0.5609
Ag (Sumas)	0.5838	0.5886	0.5592
Ref	-0.5557	-0.5936	-0.5821
Ref (Bras)	-0.5001	-0.6006	-0.6239
Ref (Old)	0.5235	0.60312	0.6018
Ref (South)	0.5537	0.5922	0.5854
Ref (Sumas)	0.5881	0.6007	0.5415

Table 3: The coefficients of the second principal component of **R** for each of the ten datasets

Dataset	Second Principal Components		
	Total Coliforms	Fecal Coliforms	E.Coli
Ag	0.5534	0.2288	-0.8009
Ag (Bras)	0.3332	0.4518	-0.8276
Ag (Old)	0.8213	-0.3543	-0.4472
Ag (South)	0.6056	0.1509	-0.7814
Ag (Sumas)	-0.4535	-0.3349	0.8259
Ref	0.8186	-0.2684	-0.5078
Ref (Bras)	0.8565	-0.4494	-0.2539
Ref (Old)	-0.8519	0.3626	0.3778
Ref (South)	0.8285	-0.3218	-0.4582
Ref (Sumas)	-0.4691	-0.2920	0.8335

respective standard deviations were calculated in R using the principal component analysis algorithm built into R. It is important to note that the standardized version of the Ag data with mean vector $[0, 0, 0]$ and 3×3 covariance matrix $\mathbf{R}$ with unit variances was used in the principal component analysis calculations. Using the standardized data instead of the non-standardized data yields the exact same principal components and corresponding standard deviations as using the correlation matrix instead of the covariance matrix. For each of the ten data sets, the principal components were calculated in the form of three vectors with three elements, each vector representing the linear com-

bination of the Total Coliform, Fecal Coliform and E. coli natural log counts. The standard deviation of each principal component and the nine principal component coefficients corresponding to the three principal components were computed (see Table 1 for the three standard deviations and Tables 2, 3 & 4 for the nine coefficient values). A scree plot was also produced in R to graphically display the variance of each principal component for each data set (see Figure 9 and Figure 10). In Table 1, Figure 9 and Figure 10, it is very obvious that the first eigenvalue is way larger than the other two eigenvalues for each of the ten data sets. For the Ag dataset, the first PC, the first two PCs and all three PCs account for 90.2%, 96.7% and 100% of the total variation in the data respectively. Since the data variation is mostly explained by just the first PC, this means that the other two PCs are not necessary. This trend is applicable to the other datasets. Therefore, for each dataset, only the first principal component is needed since it accounts for most of the variation in the data.

Table 4: The coefficients of the third principal component of **R** for each of the ten datasets

Dataset	Third Principal Components		
	Total Coliforms	Fecal Coliforms	E.Coli
Ag	-0.5995	0.7770	-0.1923
Ag (Bras)	-0.7355	0.6737	0.07178
Ag (Old)	0.05492	-0.7311	0.6800
Ag (South)	0.5497	-0.7892	0.2737
Ag (Sumas)	-0.6734	0.7358	-0.07133
Ref	0.1452	-0.7587	0.6351
Ref (Bras)	-0.1279	-0.6614	0.7391
Ref (Old)	-0.009643	0.7104	-0.7037
Ref (South)	-0.08298	0.7387	-0.6689
Ref (Sumas)	0.6588	-0.7442	0.1101

4 Precision and Stability of Principal Components

Table 5: The sample characteristics of the non-parametric bootstrap samples of the three principal components and their associated eigenvector components corresponding to the Ag dataset (true value is value calculated from original data, deviation is absolute difference between mean and true value, both the standard deviation and deviation values are in $\times 10^{-5}$, sd_j is the j^{th} PC standard deviation, and e_{ji} is the i^{th} element of the j^{th} PC)

Parameter	True Value	Mean	Standard Deviation	Deviation
sd_1	1.6449	1.6446	766.51	28.955
sd_2	0.4434	0.4442	2287.9	82.278
sd_3	0.3127	0.3118	1360.6	87.061
e_{11}	0.5782	0.5782	203.03	4.9238
e_{21}	0.5534	0.5574	5264.4	401.41
e_{31}	0.5995	0.5914	4974.7	809.67
e_{12}	0.5865	0.5866	157.38	6.6767
e_{22}	0.2288	0.2200	6713.5	882.35
e_{32}	-0.7770	-0.7763	1797.0	60.587
e_{13}	0.5672	0.5671	266.63	3.0936
e_{23}	-0.8009	-0.7957	2091.1	513.56
e_{33}	0.1923	0.1996	7007.4	737.53

Using the program R, the sampling distributions of the 12 principal component analysis parameters corresponding to the Ag data set were estimated using the non-parametric bootstrap technique. This was done using the corresponding codes in the Appendix, and the values obtained by running these codes are in Table 5. First, in each of the $N = 510$ iterations, a bootstrap sample of size $n = 825$ was created within a loop by randomly sampling from the rows of the Ag data set with replacement. This was done so that each bootstrap sample consisted of the same number of rows as in the original Ag data set. For each bootstrap sample that was created in each iteration, the three principal components and corresponding standard deviations were calculated. These resulting 12 bootstrap parameter values (three standard deviation and nine principal component coefficient values) were stored into 12

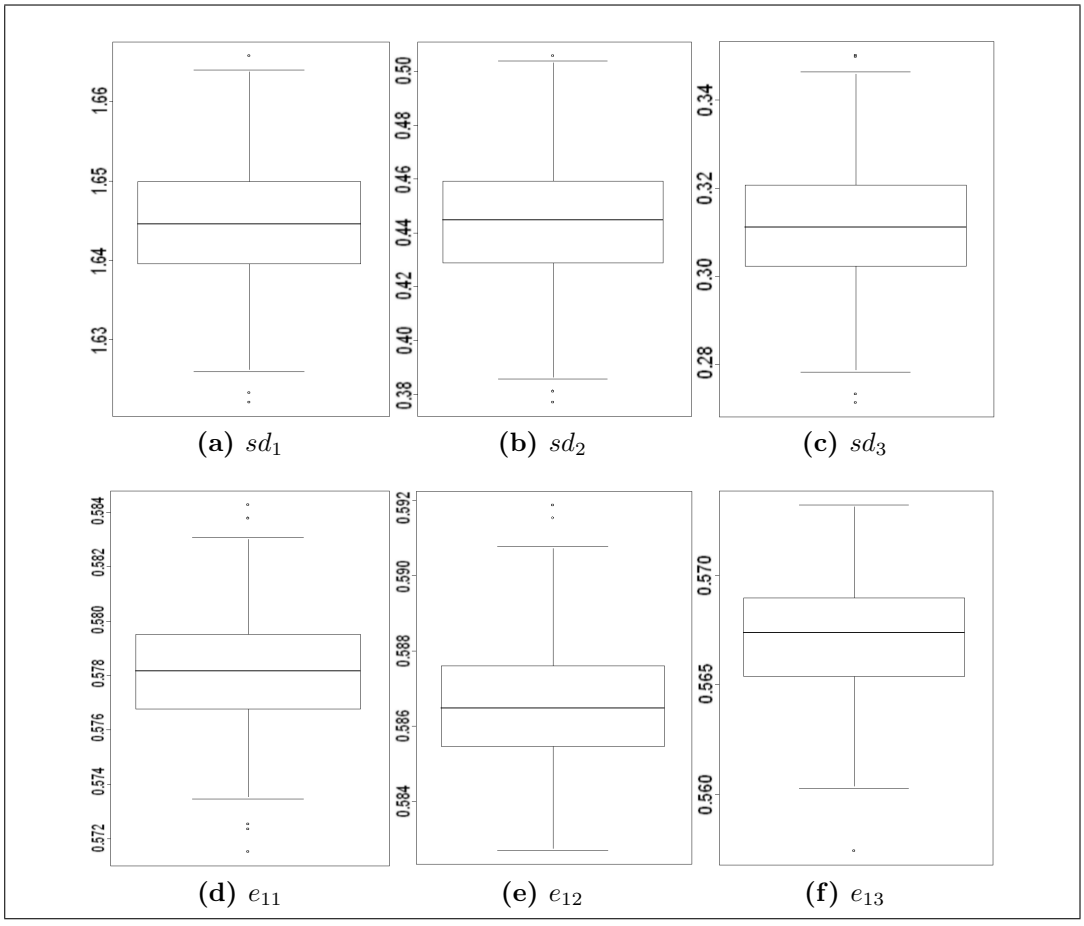
separate vectors, and this was done by combining the 12 vectors of values from the previous iterations with the 12 newly calculated values from the current iteration. After all of these $N = 510$ iterations were executed, the sample mean, sample standard deviation and deviation (absolute difference between the sample mean and the true value calculated from the original data set) were computed for each of the 12 vectors. The values in Table 5 were computed from this sample using the R codes in the Appendix section, and they are used to estimate the sampling distributions of the principal component analysis parameters.

Table 6: The 95% confidence interval of the non-parametric bootstrap samples of the three principal components and their associated eigenvector components corresponding to the Ag dataset (the one-sample t intervals were calculated using the pivot of the t-statistic random variable, the bootstrap intervals were calculated using the ordered values of the bootstrap samples, the asymptotic intervals were calculated based on the asymptotic distributions of the eigenvalue of $\mathbf{R}$, sd_j is the j^{th} PC standard deviation, and e_{ji} is the i^{th} element of the j^{th} PC)

Parameter	One-sample t	Bootstrap	Asymptotic
sd_1	1.6439, 1.6453	1.6285, 1.6596	1.5001, 1.8206
sd_2	0.4422, 0.4462	0.3970, 0.4883	0.4043, 0.4907
sd_3	0.3107, 0.3130	0.2867, 0.3393	0.2852, 0.3461
e_{11}	0.5780, 0.5784	0.5746, 0.5819	N.A.
e_{21}	0.5528, 0.5620	0.4642, 0.6635	N.A.
e_{31}	0.5871, 0.5957	0.4771, 0.6657	N.A.
e_{12}	0.5864, 0.5867	0.5834, 0.5896	N.A.
e_{22}	0.2142, 0.2258	0.07366, 0.3262	N.A.
e_{32}	-0.7779, -0.7748	-0.8073, -0.7413	N.A.
e_{13}	0.5669, 0.5674	0.5613, 0.5721	N.A.
e_{23}	-0.7975, -0.7939	-0.8222, -0.7446	N.A.
e_{33}	0.1935, 0.2057	0.08308, 0.3458	N.A.

In order to verify the precision of these results, the codes used to produce them were run numerous times and, for each time they were run, the values in the output were very close to those in Table 5. The results in Table 5 are very close to those produced from the subsequent times the codes were run. This confirms the stability of the bootstrap distributions of the 12 parameters, because the codes produce virtually the same values every time they are run.

Also, as seen in the first two columns of Table 5, the bootstrap mean values for the 12 parameters are very close to their respective true values. This means that the deviations, the absolute difference between these two values found in the last column of Table 5, are very close to zero. Table 5 also contains the standard deviations of the bootstrap values, which are also very close to zero. This strongly suggests that the bootstrap distributions of the 12 parameters are very stable, because the 12 parameter estimates appear to efficiently estimate the 12 respective true parameter values with very little bias.



(a) sd_1 **(b)** sd_2 **(c)** sd_3

(d) e_{11} **(e)** e_{12} **(f)** e_{13}

Figure 11: The boxplots of the non-parametric bootstrap samples generated using the Ag dataset for the six parameters for the standard deviations of the principal components and the elements of the first principal component (sd_j is the j^{th} PC standard deviation and e_{ji} is the i^{th} element of the j^{th} PC)

In addition, two different types of 95% confidence intervals were con-

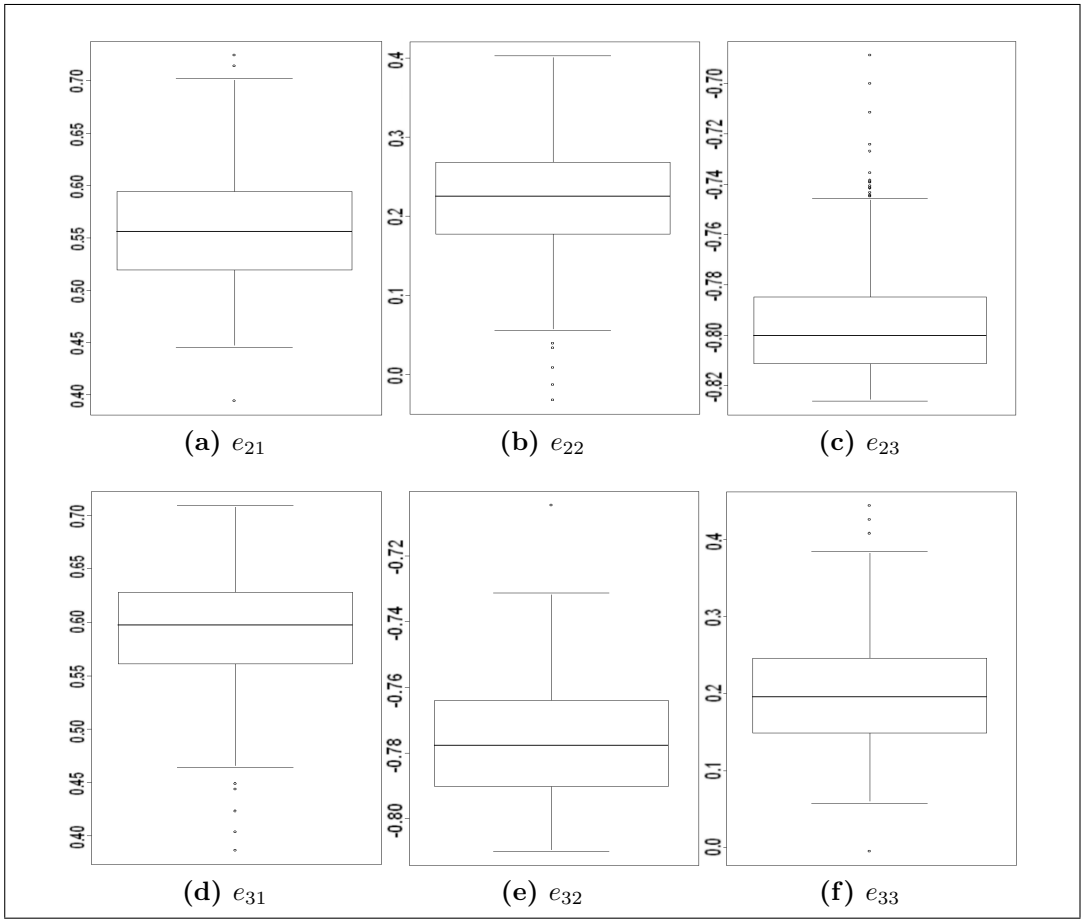


Figure 12: The boxplots of the non-parametric bootstrap samples generated using the Ag dataset for the six parameters for the elements of the second and third principal components (e_{ji} is the i^{th} element of the j^{th} PC)

structed in R for each of the 12 bootstrap parameters (see Table 6). One of these intervals was parametric, because it was based on the assumption that the bootstrap data is normally distributed. The derivation of this confidence interval makes use of the pivot of the one-sample t-statistic, and both confidence bounds are derived in the case where both tails are of size $\frac{\alpha}{2} = 0.025$. The other confidence interval was non-parametric, because it was not assumed that the bootstrap data is normally distributed. For this non-parametric confidence interval, the bootstrap sample was ordered from smallest to largest and two of the ordered observations were chosen such that $\frac{\alpha}{2} = 0.025$ of the or-

dered observations are less than or equal to the lower bound and $\frac{\alpha}{2} = 0.025$ of the ordered observations are greater than or equal to the upper bound. When looking at the bounds of both confidence intervals found in Table 6, it is seen that both bounds of both intervals are somewhat close to the true values and mean values in Table 5, which also suggests that the 12 bootstrap distributions are very stable with a reasonably small variance.

Table 7: The mean and standard deviation of the three principal components and their associated eigenvector components using different sample sizes (each sample was created using the parametric bootstrap technique which was sampled from three *i.i.d.* standard normal distributions, sd_j is the j^{th} PC standard deviation, and e_{ji} is the i^{th} element of the j^{th} PC)

Parameter	Mean			Standard Deviation		
	$n = 50$	$n = 100$	$n = 500$	$n = 50$	$n = 100$	$n = 500$
sd_1	1.1042	1.0750	1.0343	0.04454	0.03270	0.01485
sd_2	0.9962	0.9982	1.0000	0.02835	0.01989	0.008918
sd_3	0.8847	0.9193	0.9642	0.05285	0.03642	0.01599
e_{11}	0.5463	0.5246	0.5505	0.1915	0.1982	0.1895
e_{21}	0.4658	0.5093	0.4705	0.3360	0.3334	0.3284
e_{31}	0.5429	0.5226	0.5448	0.2013	0.2050	0.1879
e_{12}	-0.009879	0.01050	-0.003367	0.5753	0.5864	0.5786
e_{22}	0.01981	-0.007536	0.0002627	0.5818	0.5574	0.5728
e_{32}	0.008437	0.009556	0.005376	0.5753	0.5883	0.5814
e_{13}	0.02083	-0.02135	0.02494	0.5780	0.5846	0.5712
e_{23}	-0.03611	0.01064	0.0001778	0.5751	0.5651	0.5861
e_{33}	0.007492	0.01177	-0.01601	0.5783	0.5824	0.5748

In addition to the parametric bootstrap technique with three independent normal samples, another parametric bootstrap technique was performed using a different trivariate standard normal sample. The codes for both parametric bootstrap techniques are exactly the same except for the part that generates the samples. The mean vector and covariance matrix are defined in the portion of the codes that randomly generates the normal samples for the parametric bootstrap technique with the three dependent normal samples. The only thing that makes this parametric bootstrap technique unique from the other parametric bootstrap technique is that the covariance matrix $\mathbf{\Sigma}$ is

Table 8: The 95% confidence interval of the three principal components and their associated eigenvector components using different sample sizes (each sample was created using the parametric bootstrap technique which was sampled from three *i.i.d.* standard normal distributions, sd_j is the j^{th} PC standard deviation, and e_{ji} is the i^{th} element of the j^{th} PC)

Parameter	$n = 50$	$n = 100$	$n = 500$
sd_1	1.0297, 1.1974	1.0214, 1.1497	1.0099, 1.0676
sd_2	0.9368, 1.0540	0.9521, 1.0360	0.9799, 1.0178
sd_3	0.7622, 0.9685	0.8410, 0.9789	0.9311, 0.9905
e_{11}	0.05832, 0.7493	0.04760, 0.7459	0.06006, 0.7536
e_{21}	0.004728, 0.9756	0.008125, 0.9849	0.004155, 0.9746
e_{31}	0.05403, 0.7619	0.03839, 0.7566	0.07676, 0.7444
e_{12}	-0.7343, 0.7267	-0.7377, 0.7371	-0.7336, 0.7329
e_{22}	-0.9576, 0.9675	-0.9417, 0.9513	-0.9559, 0.9415
e_{32}	-0.7319, 0.7384	-0.7300, 0.7304	-0.7446, 0.7337
e_{13}	-0.7319, 0.7293	-0.7297, 0.7297	-0.7324, 0.7397
e_{23}	-0.9613, 0.9600	-0.9587, 0.9499	-0.9417, 0.9605
e_{33}	-0.7439, 0.7363	-0.7318, 0.7461	-0.7452, 0.7393

defined to be the correlation matrix of the Ag data set instead of the 3×3 identity matrix. This part of the codes requires the use of the R package mvtnorm to randomly generate the samples from a multivariate normal distribution. In order to generate these samples, the square root matrix of $\mathbf{\Sigma}$ must be calculated, which can be done in a few ways. In R, the default method of doing this calculation is eigenvalue decomposition, which was done in the simulations since the method was not specified in the R codes in the Appendix section (Genz *et al.*, 2010). Similar to the bootstrap technique for the three independent standard normal samples, the characteristics of the 12 bootstrap parameters were calculated. The results are found in Table 9 and Table 10, which are completely different from the corresponding values in Table 7 and Table 8.

The codes in the Appendix section used to calculate the confidence intervals in Table 6 were run numerous times just like the codes used to calculate the values in Table 5. Likewise, each time the codes for the conficence intervals

Table 9: The mean and standard deviation of the three principal components and their associated eigenvector components using different sample sizes (each sample was created using the parametric bootstrap technique which was sampled from the trivariate standard normal distribution using the correlation matrix of the Ag dataset for $\mathbf{\Sigma}$, sd_j is the j^{th} PC standard deviation, and e_{ji} is the i^{th} element of the j^{th} PC)

Parameter	Mean			Standard Deviation		
	$n = 50$	$n = 100$	$n = 500$	$n = 50$	$n = 100$	$n = 500$
sd_1	1.6442	1.6442	1.6447	0.02130	0.01465	0.006378
sd_2	0.4455	0.4452	0.4442	0.05844	0.04040	0.01809
sd_3	0.3041	0.3094	0.3117	0.04144	0.02958	0.01305
e_{11}	0.5783	0.5782	0.5782	0.004743	0.003257	0.001378
e_{21}	0.5394	0.5490	0.5527	0.1288	0.09117	0.03778
e_{31}	0.5861	0.5909	0.5980	0.1203	0.08299	0.03372
e_{12}	0.5866	0.5866	0.5866	0.005213	0.003618	0.001582
e_{22}	0.2151	0.2241	0.2279	0.1832	0.1154	0.04760
e_{32}	-0.7492	-0.7688	-0.7756	0.1214	0.03515	0.01354
e_{13}	0.5668	0.5671	0.5671	0.005952	0.004065	0.001735
e_{23}	-0.7721	-0.7909	-0.7992	0.1287	0.03470	0.01283
e_{33}	0.1765	0.1923	0.1924	0.1866	0.1213	0.05036

Table 10: The 95% confidence interval of the three principal components and their associated eigenvector components using different sample sizes (each sample was created using the parametric bootstrap technique which was sampled from the trivariate standard normal distribution using the correlation matrix of the Ag dataset for $\mathbf{\Sigma}$, sd_j is the j^{th} PC standard deviation, and e_{ji} is the i^{th} element of the j^{th} PC)

Parameter	$n = 50$	$n = 100$	$n = 500$
sd_1	1.5965, 1.6785	1.6110, 1.6687	1.6324, 1.6564
sd_2	0.3368, 0.5649	0.3732, 0.5326	0.4068, 0.4784
sd_3	0.2292, 0.3929	0.2541, 0.3695	0.2862, 0.3385
e_{11}	0.5687, 0.5875	0.5718, 0.5844	0.5756, 0.5808
e_{21}	0.2590, 0.7687	0.3499, 0.7146	0.4777, 0.6217
e_{31}	0.2889, 0.7663	0.4016, 0.7334	0.5316, 0.6591
e_{12}	0.5779, 0.5978	0.5800, 0.5945	0.5837, 0.5899
e_{22}	-0.1901, 0.5088	-0.01377, 0.4504	0.1379, 0.3163
e_{32}	-0.8108, -0.5897	-0.8098, -0.6769	-0.7978, -0.7460
e_{13}	0.5533, 0.5766	0.5579, 0.5739	0.5635, 0.5704
e_{23}	-0.8288, -0.5902	-0.8258, -0.6994	-0.8187, -0.7712
e_{33}	-0.2036, 0.5016	-0.06058, 0.4284	0.09580, 0.2852

were run, both confidence intervals were very close to both of the confidence intervals in Table 6 for each of the 12 parameters. This also confirms the stability of the bootstrap distributions of the 12 parameters, because the confidence interval codes produce virtually the same confidence intervals every time they are run. In addition, it is also seen in Table 6 that the confidence

intervals based on the normal assumption are much narrower than the non-parametric confidence intervals. Also, a boxplot for each of the 12 bootstrap parameters was created in R (see Figures 11 & 12). All of the boxplots appear to be roughly symmetric but appear to peak a lot in the middle. Both of these observations suggest that the bootstrap data may not be normally distributed.

Using the program R, a parametric bootstrap technique similar to the non-parametric bootstrap technique described above was performed on three independent standard normal samples of size n. This means that this artificial data set is actually a trivariate normal sample with mean vector $[0, 0, 0]$ and the 3×3 identity matrix as the covariance matrix. The characteristics of the 12 bootstrap parameters were calculated using the exact same type of codes used for the non-parametric bootstrapping technique except the calculations of the principal component analysis were based on this artificial data set. This parametric bootstrapping technique was run three separate times for $n = 50$, $n = 100$ and $n = 500$, and the results obtained from these calculations are presented in Table 7 and Table 8. As seen in Table 7, the standard deviations of the three principal components are very close to one. This is expected, because the three samples used to determine the 12 bootstrap values for each element of the bootstrap samples are all sampled from a standard normal distribution and are all independent of each other. The results in Table 5 and Table 6 are different from the results in Table 7 and Table 8, suggesting that the 12 bootstrap samples based on the Ag data set are not normally distributed or are not all independent of each other.

5 Discussion

The main objective of the statistical techniques used in the Preliminary Data Analysis section is to investigate if the waste produced in agricultural farmlands has an effect on the water quality of Canada's four major watersheds. For each of the three pathogen types in Figure 3, the boxplots for the Ref sites are much lower than those for the Ag sites with the exception in the South watershed in which case the boxplots are roughly the same. This indicates that the amount of waste from the agricultural farmlands near the South watershed does not hugely impact the water quality in the South watershed. However, this strongly suggests that the amount of waste produced at the agricultural sites near the Bras, Old and Sumas watersheds dramatically increases the bacteria counts. These inferences are based on the fact that the Ref sites are upstream from the Ag sites. This means that any contamination contributed by agricultural lands would have no affect on the Ref sites and that the pathogen counts measured in the Ref sites are a basal measure of the amount of pathogens in a control setting without any agricultural impacts. Therefore, ground contaminants that enter the watersheds appear to increase the bacteria counts in the watersheds thus decreasing the water quality, and this trend appears mainly in the Bras, Old and Sumas watersheds.

In all of the watersheds in Figures 4 & 5, it was observed that Fecal Coliforms are most present out of all the bacteria types, but not that much more abundant. This may be an indication that fecal waste is the most prominent type of agricultural waste contributing to watershed contamination. However, since all of the bacteria types appear to be similar in concentration, this presumably means that there are other sources of contamination even though they

are most likely not as prominent as fecal contamination. In Figures 6 & 7, it was observed that all of the scatterplots appear to be positively and linearly correlated, suggesting that the Total Coliforms, Fecal Coliforms and E. Coli natural log counts are pairwise positively linearly correlated. This is very desirable, because the linear regression model is the easiest model to fit to any type of data and there are many well known statistical tests for analyzing data modelled in this way. However, since these positive linear correlations are weak, it may not be ideal to fit the linear regression model. Therefore, other nonlinear models should be considered, because the data may be strongly correlated in a nonlinear way, which would mean not strongly linearly correlated.

As mentioned in the Preliminary Data Analysis section, the three principal components and their respective standard deviations were calculated for the standardized version of each of the ten data sets. The variation of each of the three variables has no effect on the variation of the principal components, because using the standardized data instead of the non-standardized data and using the correlation matrix instead of the covariance matrix both yield the exact same results as discussed in the Methodology section. Instead, the variation in the principal components is completely due to the nature of the data itself. The main objective of doing principal component analysis is to derive linear combination(s) of the original variables in such a way that reduces the amount of data used in the actual data analysis. As discussed in the Methodology section, using the transformed data simplifies data analysis, and the scatterplots of the data can be visualized in fewer dimensions.

For each of the ten data sets, it was observed that the first eigenvalue is much larger than the other two eigenvalues in Table 1 and Figures 9 & 10. This means that the first principal component accounts for most of the variation in

the data for each data set since the eigenvalues are the standard deviations of the principal components. Even though this has not been done in this paper, this means that a one-dimensional column of data can represent the original three columns of data for each of the ten data sets. This would be done by transforming the three original variables into a linear combination of these variables, and this linear combination would be constructed in such a way that the variation in the data is maximized. Ultimately, the variable that is defined to represent this linear combination would be the first principal component containing most of the variation in the data. It is important to note that the remaining two principal components would represent very little variation in the data and would not be needed for data analysis.

In the Precision and Stability of Principal Components section, the bootstrap distributions of the 12 parameters were estimated. It was found that the deviation and standard deviation values were very small for all 12 parameters. This means that the bootstrap distributions are very stable with small variance, meaning that the bootstrap estimators are very efficient. As mentioned in the Methodology section, using the bootstrap technique reduces the sampling bias of the bootstrap estimates, which is apparent in this case since the bias is very small. Therefore, it is evident that using the bootstrap technique significantly reduces the bias and results in stable sampling distributions. It was also observed that the bounds of both confidence intervals in Table 6 are somewhat close to the true values and mean values in Table 5, further suggesting that all of the 12 bootstrap distributions are very stable. Also, the codes for this non-parametric bootstrapping technique were run numerous subsequent times. It was observed that the values for each subsequent simulation are very close to those in Table 5 and Table 6, further suggesting

that the bootstrap distributions are very stable.

In addition, there is evidence that the bootstrap data may depart from normality. In Table 6, it was noted that the confidence intervals based on the normal assumption are much narrower than the non-parametric confidence intervals. This is a very strong indication that the 12 bootstrap distributions are not normally distributed, because both sets of 95% confidence intervals are different in size and magnitude, and one would expect both sets of confidence intervals to be aproximately the same if the bootstrap samples were normally distributed. Furthermore, it was observed that all of the boxplots for the 12 bootstrap parameters in Figures 11 & 12 look roughly symmetric but appear to peak a lot at the mean. This also suggests that the bootstrap data is not normally distributed, because the normal distribution does not have such a high kurtosis since it does not peak so abruptly at its mean. For each eigenvalue in Table 6, the asymptotic confidence interval is wider than the corresponding other two intervals, further suggesting that the bootstrap data may not be normally distributed.

The other two bootstrapping techniques discussed in the Precision and Stability of Principal Components section are both parametric assuming normality. Because the first parametric bootstap method assumes that the bootstrapping distribution for the three variables is $i.i.d.$ $N(0,1)$, one would expect that the results from this bootstrapping technique would be roughly the same as the results of the non-parametric bootstrap technique if the three columns of the Ag data are approximately $i.i.d.$ $N(0,1)$. Instead, it was observed that the results in Table 5 and Table 6 are not the same as the results in Table 7 and Table 8. This suggests that the 12 bootstrap samples based on the Ag data set are not normally distributed or all three variables in the Ag data set

are not all independent of each other. This is ambiguous, because it is unclear as to why the results in both sets of tables are different. It could be because the Ag data are normally distributed or the three variables are independent of each other. However, it has already been established that the Ag data are not normally distributed, so both of these reasons must be true.

To further look into this ambiguity, the second parametric bootstrap technique based on three standard normal samples with the correlation matrix computed from the Ag data set was run. If the three columns of the Ag data are all approximately normally distributed, one would expect that the results of this parametric bootstrap technique would be roughly the same as the results of the non-parametric bootstrap technique. Instead, it was observed that the results in Table 9 and Table 10 are completely different from those in Table 7 and Table 8. This suggests that the three variables in the Ag dataset are not all independent. Furthermore, this may be the foundation of a possible hypothesis test of independence.

6 Conclusions

From the Preliminary Data Analysis section, it was found that contaminants entering the Bras, Old and Sumas watersheds seem to increase the bacteria counts in the watersheds and decrease the water quality, but this does not appear to happen as much in the South watershed. It was also found that there is some positive correlation between each type of bacteria type, indicating that higher amounts of one pathogen type generally means higher amounts of other pathogen types. These findings suggest that water contamination is a major issue in today's society that urgently needs to be addressed and re-

solved. This means that industries and people in general need to eliminate or at least reduce the amount of waste and pollution that they create, especially the agricultural farmlands near the Bras, Old and Sumas watersheds which appear to negatively impact the water quality of these three watersheds. Protecting Canada's watersheds is very important, because clean drinking water is a huge necessity and fisheries depend on having clean water for fishing. Thus, it is very clear that government agencies urgently need to make stronger laws to protect the water quality of Canada's watersheds and other water resources.

The calculations for the principal component analysis were done in R using the standardized version of the data sets. It is usually better to standardize the data, because the variation in the principal components is completely unaffected by the variation in each of the three variables. This means that one should use the standardized version of data to do the calculations for principal component analysis to ensure that all of the variables in the data have unit variance and mean of zero. For each of the ten data sets, it was found that the first principal component accounts for most of the variation in the data, meaning that all three variables can be represented as a linear combination of these variables constructed in such a way that the variation in the data is maximized. In this case, this would be useful for analyzing the ten data sets represented by only one variable if one would choose to further analyze the water quality data further. This is because there are many one sample tests that can be used to analyze the data represented by the first principal component and one-dimensional confidence intervals concerning these data can be easily calculated. On the other hand, it is usually more difficult to construct three-dimensional confidence regions and to do hypothesis tests with the data represented by the original three variables.

In the Precision and Stability of Principal Components section, it was found that all of the 12 non-parametric bootstrap distributions are very stable, because the bias and variation of the bootstrap sample are both very small for each of the 12 principal component analysis parameters. This is an important characteristic, because it means that the estimates of the principal component analysis parameters are very reliable since they are very efficient with negligable bias. This characteristic was also found for the two parametric bootstrap techniques, which means that the corresponding bootstrap distributions are all stable since the estimates of the principal component analysis parameters are also very efficient with negligable bias. It was also found that the results from all three bootstrap techniques are all completely different from each other, which is a very strong indication that the three variables in the Ag data set are not *i.i.d.* $N(0,1)$. This phenomenon can very well be the basis of a statistical hypothesis test of independence between two or more variables.

7 Future Work

In this paper, the three bootstrap techniques were only applied to the Ag dataset. One could also apply these bootstrap techniques to the other nine data sets to further investigate the properties of the corresponding bootstrap distributions. In addition, the data used in this paper were actually a small subset of a larger dataset. Only some of the columns of the original dataset were used; the columns that were omitted include other bacteria counts as well as different abiotic factors such as temperature and pH. One could generalize the analyses done in this paper by including these other datasets to further investigate the external effects on freshwater contamination in the four water-

sheds. It was previously mentioned that the correlations amongst the three variables for each dataset may be nonlinear. One could further investigate the correlations in the scattermatrices to determine if the correlations are truely linear. If there is significant evidence that they are nonlinear, then one could attempt fitting a different regression model to the data that would minimize the residuals.

In this paper, the estimates of the principal component analysis parameters were estimated using the bootstrap methods. In addition, one could use the jackknife technique to calculate the jackknife estimates of these parameters, and then calculate the jackknife estimates of the bias and variance of these jackknife estimates. In the Methodology section, there was an example illustrating how to calculate both the jackknife and bootstrap estimates, and then comparing these two estimates with the original biased estimate. In the case of the principal component analysis parameters, one could compare the 12 jackknife estimates with the corresponding bootstrap estimates. The purpose of doing these comparisons would be to verify that using the bootstrap methods reduces the bias of the principal component analysis estimates, and reduces the bias more effectively than using the jackknife methods does.

The two parametric bootstrap techniques done in this paper were based on standard normal samples of size $n = 50$, $n = 100$ and $n = 500$. One could try running these two parametric bootstrap techniques for several different larger values of n to determine how different values of n affect the bootstrap results. Moreover, one could investigate the asymptotic behaviour of the bootstrap estimates as n approaches infinity. Similarly, one could also explore how several different larger values of N affect the bootstrap results, thus determining the asymptotic behaviour of the bootstrap estimates as the bootstrap

sample size N approaches infinity. This work could also be done to determine how large N and n should be when running the bootstrap to ensure that the bootstrap results are roughly asymptotic.

It was briefly mentioned in the Discussion and Conclusions sections that one can use the parametric and non-parametric bootstrap techniques to investigate the normality and independence of two or more variables. There are several different hypothesis tests of independence and normality, and these techniques may potentially be the foundation of creating more of these hypothesis tests. In particular, one could express the values calculated for all three bootstrap methods as random variables and then derive a test statistic based on these random variables. The distribution of this test statistic can then be used to derive a critical region of maximum power to formulate an appropriate hypothesis test. Thus, one could possibly derive a statistical hypothesis of independence and normality to test for normality of certain variables and for independence amongst these variables.

8 References

Anderson, T. W. (1963). Asymptotic Theory for Principal Component Analysis. *The Annals of Mathematical Statistics* **34,** 122-148.

Edge, T. A., El-Shaarawi, A., Gannon, V., Jokinen, C., Kent, R., Khan, I. U. H., Koning, W., Lapen, D., Miller, J., Neumann, N., Phillips, R., Robertson, W., Schreier, H., Scott, A., Shtepani, I., Topp, E., Wilkes, G., and van Bochove, E. (2010). Development of an *Escherichia coli* Environmental Benchmark for Waterborne Pathogens in Agricultural Watersheds in Canada. *Internal Report, Canada Centre for Inlet Waters.* Burlington, Ontario, Canada.

Efron, B. (1979). Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics* **7,** 1-26.

Efron, B. (1981). Nonparametric estimates of standard error: The jackknife, the bootstrap and other methods. *Biometrika* **68,** 589-599.

Efron, B. (1982). *The Jackknife, the Bootstrap and Other Resampling Plans.* **38**. Society for Industrial and Applied Mathematics, Philadelphia.

Efron, B., and Stein, C. (1981). The Jackknife Estimate of Variance. *The Annals of Statistics* **9,** 586-596.

Genz, A., Bretz, F., Miwa, T., Mi, X., Leisch, F., Scheipl, F., and Hothorn, T. (2010). Multivariate Normal and t Distributions. *Package mvtnorm* 1-13.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology* **24,** 417-441.

Hotelling, H. (1936). Simplified calculation of principal components. *Psychometrika* **1,** 27-35.

Johnson, R. A., and Wichern, D. W. (2007). *Applied Multivariate Statistical Analysis, Sixth Edition.* Pearson Prentice Hall, New Jersey.

Jolliffe, I. T. (2002). *Principal Component Analysis, Second Edition.* Springer-Verlag, New York.

Miller, R. G. (1964). A Trustworthy Jackknife. *The Annals of Mathematical Statistics* **35,** 1594-1605.

Miller, R. G. (1974^a). An Unbalanced Jackknife. *The Annals of Statistics* **2,** 880-891.

Miller, R. G. (1974^b). The jackknife - a review. *Biometrika* **61,** 1-15.

Quenouille, M. H. (1956). Notes on Bias in Estimation. *Biometrika* **43,** 353-360.

R Development Core Team (2008). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL `http://www.R-project.org`.

9 Appendix (all R codes used to get results)

9.1 The Jackknife and Bootstrap examples in Methodology section

```
n=20
u=n
M=10000
minus_terms=NULL
```

```
boot_est=NULL
x=rnorm(n, mean=20, sd=1)
muSq_hat=(mean(x) )^2                    #true value of estimate
for(j in 1:n){                           #jackknife loop
minus_terms=cbind(minus_terms, (mean(x[-c(j)]))^2)}
muSq_tilde=u * muSq_hat + (1 - u) * 1/u * sum(minus_terms)
muSq_dot=1/n*sum(minus_terms)
variance=(n-1)/n * sum((minus_terms - muSq_dot)^2)
bias=(n-1) * (muSq_dot - muSq_hat)
for(j in 1:M){                           #bootstrap loop
x_=x[sample(1:n,replace=T)] #Samples n values w/ replacement
boot_est=cbind(boot_est, (mean(x_))^2)}
muSq_hat         #this is the biased estimate
#here are the calculations for the jackknife:
muSq_tilde   #jackknife estimate of mu squared
bias               #jackknife estimate of bias
variance         #jackknife estimate of variance
#here are the calculations for the bootstrap:
mean(boot_est)     #bootstrap mean
abs(mean(boot_est)-muSq_hat) #the bootstrap deviation
sqrt(1/(M-1)*sum((boot_est-mean(boot_est))^2)) #the bootstrap stdev
```

9.2 The preliminary data analysis section

```
setwd("E:/Thesis/Data") #file directory
PCA_f=function(data){ #fct for PCA
PCA=prcomp(data,center=TRUE,scale=TRUE)
print(PCA)
screeplot(PCA)}  #end of fct
#Ag dataset:
data=na.omit(read.csv(file="Ag.csv")[1:3])
plot(data)
boxplot(data)
PCA_f(data)
#Ag(Bras) dataset:
data=na.omit(read.csv(file="Ag(Bras).csv")[1:3])
plot(data)
boxplot(data)
PCA_f(data)
#Ag(Old) dataset:
data=na.omit(read.csv(file="Ag(Old).csv")[1:3])
plot(data)
boxplot(data)
PCA_f(data)
#Ag(South) dataset:
```

```
data=na.omit(read.csv(file="Ag(South).csv")[1:3])
plot(data)
boxplot(data)
PCA_f(data)
#Ag(Sumas) dataset:
data=na.omit(read.csv(file="Ag(Sumas).csv")[1:3])
plot(data)
boxplot(data)
PCA_f(data)
#Ref dataset:
data=na.omit(read.csv(file="Ref.csv")[1:3])
plot(data)
boxplot(data)
PCA_f(data)
#Ref(Bras) dataset:
data=na.omit(read.csv(file="Ref(Bras).csv")[1:3])
plot(data)
boxplot(data)
PCA_f(data)
#Ref(Old) dataset:
data=na.omit(read.csv(file="Ref(Old).csv")[1:3])
plot(data)
boxplot(data)
PCA_f(data)
#Ref(South) dataset:
data=na.omit(read.csv(file="Ref(South).csv")[1:3])
plot(data)
boxplot(data)
PCA_f(data)
#Ref(Sumas) dataset:
data=na.omit(read.csv(file="Ref(Sumas).csv")[1:3])
plot(data)
boxplot(data)
PCA_f(data)
fct=function(data){  #fct for 3 boxplots
logdata=log(data)
boxplot(logdata)}  #end of fct
fct(data=read.csv(file="TotalColiforms.csv"))
fct(data=read.csv(file="FecalColiforms.csv"))
fct(data=read.csv(file="EColi.csv"))
library(lattice)
par(mfrow=c(1,3))
Variable_f=function(XData, MainTitle, XTitle){ #fct for each variable
hist(XData,data=X,freq=FALSE,type="density",xlab=XTitle,main=
MainTitle,ylab="")
```

```
}   #end of fct
Data_f=function(datatype){   #fct for each dataset
Variable_f(X$log.Total.coliforms,"","T")
Variable_f(X$log.Fecal.coliforms,datatype,"F")
Variable_f(X$log.E..coli,"","E")
}   #end of fct
X=na.omit(read.csv(file="Ag.csv")[1:3])
Data_f("Ag")
X=na.omit(read.csv(file="Ref.csv")[1:3])
Data_f("Ref")
X=na.omit(read.csv(file="Ag(Bras).csv")[1:3])
Data_f("Ag (Bras)")
X=na.omit(read.csv(file="Ref(Bras).csv")[1:3])
Data_f("Ref (Bras)")
X=na.omit(read.csv(file="Ag(Old).csv")[1:3])
Data_f("Ag (Old)")
X=na.omit(read.csv(file="Ref(Old).csv")[1:3])
Data_f("Ref (Old)")
X=na.omit(read.csv(file="Ag(South).csv")[1:3])
Data_f("Ag (South)")
X=na.omit(read.csv(file="Ref(South).csv")[1:3])
Data_f("Ref (South)")
X=na.omit(read.csv(file="Ag(Sumas).csv")[1:3])
Data_f("Ag (Sumas)")
X=na.omit(read.csv(file="Ref(Sumas).csv")[1:3])
Data_f("Ref (Sumas)")
x=as.matrix(na.omit(read.csv(file="Ag.csv")[3]))
y=as.matrix(na.omit(read.csv(file="Ref.csv")[3]))
wilcox.test(x,y) # 2-sample (Mann-Whitney U) test
x=as.matrix(na.omit(read.csv(file="Ag(Bras).csv")[3]))
y=as.matrix(na.omit(read.csv(file="Ref(Bras).csv")[3]))
wilcox.test(x,y) # 2-sample (Mann-Whitney U) test
x=as.matrix(na.omit(read.csv(file="Ag(Old).csv")[3]))
y=as.matrix(na.omit(read.csv(file="Ref(Old).csv")[3]))
wilcox.test(x,y) # 2-sample (Mann-Whitney U) test
x=as.matrix(na.omit(read.csv(file="Ag(South).csv")[3]))
y=as.matrix(na.omit(read.csv(file="Ref(South).csv")[3]))
wilcox.test(x,y) # 2-sample (Mann-Whitney U) test
x=as.matrix(na.omit(read.csv(file="Ag(Sumas).csv")[3]))
y=as.matrix(na.omit(read.csv(file="Ref(Sumas).csv")[3]))
wilcox.test(x,y) # 2-sample (Mann-Whitney U) test
```

9.3 The precision and stability of principal components section

9.3.1 The non-parametric bootstrap

```
setwd("E:/Thesis/Data") #file directory
n=825              #size of Ag data set
M=510              #size of bootstrap sample
Data=na.omit(read.csv(file="Ag.csv")[1:3])
PCA=prcomp(Data,center=TRUE,scale=TRUE) #PCA
#3 sample standard deviations:
sd1_t=PCA$sd[1]
sd2_t=PCA$sd[2]
sd3_t=PCA$sd[3]
#9 sample rotation values:
if (PCA$r[1,1]<0) {r11_t=-PCA$r[1,1]} else {r11_t=PCA$r[1,1]}
if (PCA$r[1,2]<0) {r12_t=-PCA$r[1,2]} else {r12_t=PCA$r[1,2]}
if (PCA$r[1,3]<0) {r13_t=-PCA$r[1,3]} else {r13_t=PCA$r[1,3]}
if (PCA$r[1,1]<0) {r21_t=-PCA$r[2,1]} else {r21_t=PCA$r[2,1]}
if (PCA$r[1,2]<0) {r22_t=-PCA$r[2,2]} else {r22_t=PCA$r[2,2]}
if (PCA$r[1,3]<0) {r23_t=-PCA$r[2,3]} else {r23_t=PCA$r[2,3]}
if (PCA$r[1,1]<0) {r31_t=-PCA$r[3,1]} else {r31_t=PCA$r[3,1]}
if (PCA$r[1,2]<0) {r32_t=-PCA$r[3,2]} else {r32_t=PCA$r[3,2]}
if (PCA$r[1,3]<0) {r33_t=-PCA$r[3,3]} else {r33_t=PCA$r[3,3]}
sd1_=NULL
sd2_=NULL
sd3_=NULL
r11_=NULL
r12_=NULL
r13_=NULL
r21_=NULL
r22_=NULL
r23_=NULL
r31_=NULL
r32_=NULL
r33_=NULL
for(i in 1:M){    #beginning of bootstrap loop
Data_=Data[sample(1:n,replace=T),] #Samples n rows w/ replacement
PCA=prcomp(Data_,center=TRUE,scale=TRUE)   #PCA
#storing 3 bootstrap standard deviation values:
sd1_=cbind(sd1_, PCA$sd[1])
sd2_=cbind(sd2_, PCA$sd[2])
sd3_=cbind(sd3_, PCA$sd[3])
#9 bootstrap rotation values:
if (PCA$r[1,1]<0) {r11=-PCA$r[1,1]} else {r11=PCA$r[1,1]}
```

```
if (PCA$r[1,2]<0) {r12=-PCA$r[1,2]} else {r12=PCA$r[1,2]}
if (PCA$r[1,3]<0) {r13=-PCA$r[1,3]} else {r13=PCA$r[1,3]}
if (PCA$r[1,1]<0) {r21=-PCA$r[2,1]} else {r21=PCA$r[2,1]}
if (PCA$r[1,2]<0) {r22=-PCA$r[2,2]} else {r22=PCA$r[2,2]}
if (PCA$r[1,3]<0) {r23=-PCA$r[2,3]} else {r23=PCA$r[2,3]}
if (PCA$r[1,1]<0) {r31=-PCA$r[3,1]} else {r31=PCA$r[3,1]}
if (PCA$r[1,2]<0) {r32=-PCA$r[3,2]} else {r32=PCA$r[3,2]}
if (PCA$r[1,3]<0) {r33=-PCA$r[3,3]} else {r33=PCA$r[3,3]}
#storing 9 above values:
r11_=cbind(r11_, r11)
r12_=cbind(r12_, r12)
r13_=cbind(r13_, r13)
r21_=cbind(r21_, r21)
r22_=cbind(r22_, r22)
r23_=cbind(r23_, r23)
r31_=cbind(r31_, r31)
r32_=cbind(r32_, r32)
r33_=cbind(r33_, r33)}       #end of bootstrap loop
fct=function(x_,x_t){  #fct for bootstrap results
mu=mean(x_) #bootstrap mean value
dev=abs(mu-x_t) #bootstrap deviation value
stdev=sqrt(sum((x_-mu)^2)/(M-1)) #bootstrap stdev value
a=12        #quartile for lower bound of CI
b=498       #quartile for upper bound of CI
bootCI=cbind(sort(x_)[a],sort(x_)[b]) #95% bootstrap CI
print(list(mu=mu,dev=dev,stdev=stdev,bootCI=bootCI,t.test(x_)))
boxplot(c(x_))} #end of fct
fct(sd1_,sd1_t)
fct(sd2_,sd2_t)
fct(sd3_,sd3_t)
fct(r11_,r11_t)
fct(r12_,r12_t)
fct(r13_,r13_t)
fct(r21_,r21_t)
fct(r22_,r22_t)
fct(r23_,r23_t)
fct(r31_,r31_t)
fct(r32_,r32_t)
fct(r33_,r33_t)
```

9.3.2 The parametric bootstrap using $\Sigma = I_3$

```
n=500              #size of standard normal samples
M=1000             #size of bootstrap sample
sd1_=NULL
```

```
sd2_=NULL
sd3_=NULL
r11_=NULL
r12_=NULL
r13_=NULL
r21_=NULL
r22_=NULL
r23_=NULL
r31_=NULL
r32_=NULL
r33_=NULL
for(i in 1:M){               #beginning of bootstrap loop
x1=rnorm(n, mean=0, sd=1)    #first normal sample
x2=rnorm(n, mean=0, sd=1)    #second normal sample
x3=rnorm(n, mean=0, sd=1)    #third normal sample
PCA=prcomp(cbind(x1,x2,x3),center=TRUE,scale=TRUE) #PCA
#storing 3 standard deviation values:
sd1_=cbind(sd1_, PCA$sd[1])
sd2_=cbind(sd2_, PCA$sd[2])
sd3_=cbind(sd3_, PCA$sd[3])
#9 bootstrap rotation values:
if (PCA$r[1,1]<0) {r11=-PCA$r[1,1]} else {r11=PCA$r[1,1]}
if (PCA$r[1,2]<0) {r12=-PCA$r[1,2]} else {r12=PCA$r[1,2]}
if (PCA$r[1,3]<0) {r13=-PCA$r[1,3]} else {r13=PCA$r[1,3]}
if (PCA$r[1,1]<0) {r21=-PCA$r[2,1]} else {r21=PCA$r[2,1]}
if (PCA$r[1,2]<0) {r22=-PCA$r[2,2]} else {r22=PCA$r[2,2]}
if (PCA$r[1,3]<0) {r23=-PCA$r[2,3]} else {r23=PCA$r[2,3]}
if (PCA$r[1,1]<0) {r31=-PCA$r[3,1]} else {r31=PCA$r[3,1]}
if (PCA$r[1,2]<0) {r32=-PCA$r[3,2]} else {r32=PCA$r[3,2]}
if (PCA$r[1,3]<0) {r33=-PCA$r[3,3]} else {r33=PCA$r[3,3]}
#storing 9 above values:
r11_=cbind(r11_, r11)
r12_=cbind(r12_, r12)
r13_=cbind(r13_, r13)
r21_=cbind(r21_, r21)
r22_=cbind(r22_, r22)
r23_=cbind(r23_, r23)
r31_=cbind(r31_, r31)
r32_=cbind(r32_, r32)
r33_=cbind(r33_, r33)}       #end of bootstrap loop
boot_info=function(x_){  #fct for bootstrap results
mu=mean(x_) #bootstrap mean value
stdev=sqrt(sum((x_-mu)^2)/(M-1)) #bootstrap stdev value
a=25         #quartile for lower bound of CI
b=976       #quartile for upper bound of CI
```

```
bootCI=cbind(sort(x_)[a],sort(x_)[b]) #95% bootstrap CI
print(list(mu=mu,stdev=stdev,bootCI=bootCI))} #end of fct
boot_info(sd1_)
boot_info(sd2_)
boot_info(sd3_)
boot_info(r11_)
boot_info(r12_)
boot_info(r13_)
boot_info(r21_)
boot_info(r22_)
boot_info(r23_)
boot_info(r31_)
boot_info(r32_)
boot_info(r33_)
```

9.3.3 The parametric bootstrap using $\Sigma = \Re$

```
library(mvtnorm)
setwd("E:/Thesis/Data") #file directory
Data=na.omit(read.csv(file="Ag.csv")[1:3])
n=500              #size of standard normal samples
M=1000             #size of bootstrap sample
sd1_=NULL
sd2_=NULL
sd3_=NULL
r11_=NULL
r12_=NULL
r13_=NULL
r21_=NULL
r22_=NULL
r23_=NULL
r31_=NULL
r32_=NULL
r33_=NULL
for(i in 1:M){               #beginning of bootstrap loop
X=rmvnorm(n,mean=c(0,0,0),sigma=cor(Data))
PCA=prcomp(X,center=TRUE,scale=TRUE) #PCA
#storing 3 standard deviation values:
sd1_=cbind(sd1_, PCA$sd[1])
sd2_=cbind(sd2_, PCA$sd[2])
sd3_=cbind(sd3_, PCA$sd[3])
#9 bootstrap rotation values:
if (PCA$r[1,1]<0) {r11=-PCA$r[1,1]} else {r11=PCA$r[1,1]}
if (PCA$r[1,2]<0) {r12=-PCA$r[1,2]} else {r12=PCA$r[1,2]}
if (PCA$r[1,3]<0) {r13=-PCA$r[1,3]} else {r13=PCA$r[1,3]}
```

```
if (PCA$r[1,1]<0) {r21=-PCA$r[2,1]} else {r21=PCA$r[2,1]}
if (PCA$r[1,2]<0) {r22=-PCA$r[2,2]} else {r22=PCA$r[2,2]}
if (PCA$r[1,3]<0) {r23=-PCA$r[2,3]} else {r23=PCA$r[2,3]}
if (PCA$r[1,1]<0) {r31=-PCA$r[3,1]} else {r31=PCA$r[3,1]}
if (PCA$r[1,2]<0) {r32=-PCA$r[3,2]} else {r32=PCA$r[3,2]}
if (PCA$r[1,3]<0) {r33=-PCA$r[3,3]} else {r33=PCA$r[3,3]}
#storing 9 above values:
r11_=cbind(r11_, r11)
r12_=cbind(r12_, r12)
r13_=cbind(r13_, r13)
r21_=cbind(r21_, r21)
r22_=cbind(r22_, r22)
r23_=cbind(r23_, r23)
r31_=cbind(r31_, r31)
r32_=cbind(r32_, r32)
r33_=cbind(r33_, r33)}       #end of bootstrap loop
boot_info=function(x_){  #fct for bootstrap results
mu=mean(x_) #bootstrap mean value
stdev=sqrt(sum((x_-mu)^2)/(M-1)) #bootstrap stdev value
a=25         #quartile for lower bound of CI
b=976       #quartile for upper bound of CI
bootCI=cbind(sort(x_)[a],sort(x_)[b]) #95% bootstrap CI
print(list(mu=mu,stdev=stdev,bootCI=bootCI))} #end of fct
boot_info(sd1_)
boot_info(sd2_)
boot_info(sd3_)
boot_info(r11_)
boot_info(r12_)
boot_info(r13_)
boot_info(r21_)
boot_info(r22_)
boot_info(r23_)
boot_info(r31_)
boot_info(r32_)
boot_info(r33_)
```

9.3.4 The asymptotic confidence intervals for non-parametric bootstrap

```
z=qnorm(0.975,mean=0,sd=1)    #right alpha/2 quantile
lambda=c(1.6448828,0.4433664,0.3127090) #eigenvalues for Ag dataset
lambda_f = function(n){   #fct for CI calculations
for(i in 1:3){
print(lambda[i]*c(1/(1+z*sqrt(2/n)),1/(1-z*sqrt(2/n))))}} #end of fct
lambda_f(825)
```