Using Mass Balance, Factor Analysis, and Multiple Imputation to Assess Health Effects of Water Quality

BY

CHIPING NIEH
B.S., National Taiwan University, 1999
M.S., The University of North Carolina at Chapel Hill, 2002

THESIS

Submitted as partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Public Health Sciences
in the Graduate College of the
University of Illinois at Chicago, 2012

Chicago, Illinois

Defense Committee:

Peter Scheff, Chair
Samuel Dorevitch
Rachael Jones
Li Liu
Richard Whitman, U.S. Geological Survey

ACKNOWLEDGMENT

I am heartily thankful to my advisor, Peter Scheff, and all my committee members whose encouragement, supervision, and support from the preliminary to the concluding level enabled me to develop an understanding of the subject. I also would like to make a reference to the Metropolitan Water Reclamation District of Greater Chicago for funding this project and kindly providing relevant data. I thank Mr. Ramon Lopez for his patience in sharing an office, and partnership through the ups and downs of graduate school. Lastly, I owe my deepest gratitude to my family and my partner Mr. Greg Nicpon for their support over these four years.

TABLE OF CONTENTS

TABLE OF CONTENTS (Continued)

CHAPTER **PAGE**

TABLE OF CONTENTS (Continued)

CHAPTER **PAGE**

LIST OF TABLES

LIST OF TABLES (Continued)

TABLE **PAGE**

LIST OF TABLES (Continued)

TABLE **PAGE**

LIST OF TABLES (Continued)

LIST OF TABLES (Continued)

TABLE **PAGE**

LIST OF TABLES (Continued)

LIST OF FIGURES

LIST OF FIGURES (Continued)

LIST OF ABBREVIATIONS

AGI	Acute gastrointestinal illness
AL	Alsip
AMI	Arithmetic mean imputation
ANN	Artificial neural network
BA	Beaubien Woods
BR	Bridge Street
CAWS	Chicago Area Waterway Systems
CFA	Confirmatory factor analysis
CHEERS	Chicago Health, Environmental Exposure, and Recreation Study
CMB	Chemical nass balance
CO	Canal Origins
CP	Clark Park
CSO	Combined sewage overflow
CSSC	Chicago Sanitary and Shipping Canal
DA	Date augmentation

LIST OF ABBREVIATIONS (Continued)

DML	direct maximum likelihood
DNA	Deoxyribonucleic
DO	Dissolved oxygen
EFA	Exploratory factor analysis
EM	Expectation maximization
FIB	Fecal indicator bacteria
IPCB	Illinois Pollution Control Board
LA	Lincoln Avenue
LAW	Lawrence Fisheries
MAR	Missing at random
MCAR	Missing completely at random
MCMC	Markov chain Monte Carlo
MI	Multiple imputation
MNAR	Missing not at random
MS	Main Stem
MSE	Mean square error

LIST OF ABBREVIATIONS (Continued)

MWRDGC	Metropolitan Water Reclamation District of Greater Chicago
NAM	North Avenue
ND	Non-detect
NPS	Non-point source
PAHs	Polycyclic aromatic hydrocarbons
PCA	Principal component analysis
PM	Particulate matters
PMF	Positive matrix factorization
PS	Point source
PT	Ping Tom
QC	Quality control
QPCR	Quantitative polymerase chain reaction
RM	Riverdale Marina
RP	River Park
SCRAM	Support Center for Regulatory Air Models
SK	Skokie Rowing Center

LIST OF ABBREVIATIONS (Continued)

SMC	Squared multiple correlation coefficient
SRI	Stochastic regression imputation
SVM	Support vector machine
U.S. EPA	U.S. Environmental Protection Agency
VIF	Variance inflation factor
VOCs	Volatile organic contaminants
WE	Wester Avenue
WLS	Weighted least square
WO	Worth
WS	Willow Springs
WRP	Water reclamation plant.

SUMMARY

This dissertation evaluates the performance of multiple imputation method in filling in missing microbial data, and utilizes the chemical mass balance model and the exploratory factor analysis for the identification of sources of fecal contamination.

The multiple imputation method was applied on surface water measurements collected on the Chicago River from 2007 to 2009. The method was used to fill in missing values and the original dataset was compared to the imputed dataset. Descriptive statistics show that the imputed dataset has a similar distribution as the original dataset. In order to further evaluate the performance of the imputation, a portion of the original dataset was deleted, and the missing values were filled in using multiple imputation method. Results show that the imputed dataset can provide inferential parameter estimates, and that multiple imputation can fill in missing microbial data without distorting the distribution of the original dataset.

The chemical mass balance model and exploratory factor analysis were then utilized to identify sources of fecal contamination in the river system. Sources identified included physicochemical and densities of other microbes. Results from both methods suggested that microbial sources may vary at different locations on the river system. Identified sources were then used in predicting of the risk of Acute Gastrointestinal (AGI) illness among water users. However, no association between pollutant sources and health risk was identified.

CHAPTER 1

INTRODUCTION

Recreational water quality continues to be an important issue for the public's health. Poorly monitored water systems prevent timely warning of high levels of fecal contamination to recreators, and therefore, pose a threat to the health of water users. Health risks associated with recreating on water systems with high levels of fecal contamination include: gastrointestinal illnesses (GI illness), such as vomiting and diarrhea; respiratory symptoms including cough and sore throat; eye and ear infection; and skin rash(1). Shuval(2) estimated that coastal waters polluted by wastewater generate approximately 120 million episodes of GI illness and 50 million acute respiratory diseases globally annually, costing $12 billion per year.

It is essential to understand water quality pollutant sources in order to develop new policies and management strategies to protect water quality and human health. Currently, fecal indicator bacteria (FIB) are used to indicate the water quality. Fecal indicator bacteria are bacteria, such as *E. coli*, enterococci, male-specific coliphages, and somatic coliphages, which don't frequently cause human illness, but are present in high density in fecal material. The guidelines were established by the U.S. Environmental Protection Agency (U.S. EPA) using epidemiological studies assessing the risk of GI illness among swimmers exposed to fecal contaminated waters(3), where level of fecal contamination was indicated by the density of FIB. However, FIB inputs to water are diverse, including sewage spills, urban runoff, pet/livestock waste, waste from

wildlife(4; 5), and environmental (soil, sediment, or water) reservoirs(6). This means that the presence of FIB does not exclusively indicate the presence of human waste.

Domestic discharge and urban runoff have been identified as two major sources of water pollution in the US(7), upgrades in sanitation infrastructure has reduced the release of domestic waste water into the environment. As a result, urban runoff is now considered to be the predominant source of water pollution. Urban runoff itself is a diverse source, and contains pollutants from storm waters, households, and even untreated raw human sewage with high levels of toxic pollutants, and infectious bacteria and pathogens(8; 9).

Bacteria, such as *E. coli* and enterococci, as well as coliphage viruses have been used to identify fecal contamination in water(10). Specifically, the 1986 U.S. EPA standards for recreational water quality use *E. coli* and enterococci as single sample concentration and 30 day geometric means to define acceptable water quality. Although these indicators are generally not the etiology of recreational waterborne illness, their density in recreational waters have been found to be indicators of health risk(11; 12; 13). Wade et al.(14) conducted a meta-analysis to evaluate the risks of GI illness among swimmers in fresh and marine water with point source (PS) pollution. The authors reviewed 27 studies and concluded the use of enterococci and *E. coli* in marine and fresh water respectively as the indicators of GI illness. In addition to the association between microbial indicators and health risk, sampling and analytical methods are another consideration because the analyses of certain bacterial levels in water require less time and money.

Water quality standard levels developed by U.S. EPA are based on relationship between health risks and water quality observed in epidemiological studies. This type of studies rely on multivariable models which do not perform well when there is a large portion of missing values. However, majority of these studies often run into missing data problem due to personnel or laboratory factors, or participant dropoff. Gap filling techniques using imputation have been tested in psychological and environmental fields. However, no one has applied imputation method on microbial water quality data sets. Since the presence of missing values is an issue in the water quality field, it is important to evaluate if imputation can also fill in unbiased values for microbial missing data.

Bacteria concentration criteria used to monitor water quality have been set to protect swimmers. However, there are people conducting other recreational activities. Based on the amount of contact with water, water recreation can be classified as "full contact" and "limited contact". Full contact water recreation includes activities such as swimming, diving, or jet skiing, while limited contact water recreation includes canoeing, kayaking, or fishing. Individuals performing activities that fall into full contact recreation face higher risk of developing waterborne diseases due to higher amount of water exposure(15; 16; 17; 18). However, these articles found associations in water bodies with PS of sewage treatment plants. The associations between water bodies with only non-point source (NPS) pollution and health risks remain unknown.

Health risks associated with limited contact water activities have not been well studied or understood, even though, over 17.8 million Americans participated in activities such as kayaking, canoeing, and rafting in 2008(19). In addition, the number of boats registered every year has

been increasing from 8 million in 1980 to more than 12 million in 2009(20). With a growing percentage of US population who regularly conduct limited contact water activities, it is important to understand the relationship between limited contact water recreation and waterborne diseases.

The Chicago Area Waterway Systems (CAWS) was the target waterway for this study. To protect Lake Michigan as a drinking water source, in 1900 the Metropolitan Water Reclamation District of Greater Chicago (MWRDGC) redirected the Chicago River away from Lake Michigan by building a number of canals and locks. Today, this canal system is referred to as the CAWS(21). The CAWS consists of the two main branches of the Chicago River (North Branch and South Branch), as well as the North Shore Channel, the Cal-Sag Channel, and the Chicago Sanitary and Ship Canal(22) (Figure 1). According to MWRDGC, the waterways receive over 300 million gallons per day of filtered but not disinfected wastewater that has been treated using aerobic digestion. The discharge has bacterial counts between 700 and 340,000 fecal coliforms per 100 mL. It is estimated that wastewater effluent accounts for 70% of the total flow in the CAWS, and during dry weather, it accounts for over 90% of the flow. The system was never intended to be used for recreation. However, people now consider the CAWS a source of recreational opportunities. The Illinois Pollution Control Board (IPCB) designating most CAWS segments for limited contact recreational use. The number of people using the CAWS has been increasing over the years(23). Overall, there are over 150 access points along the system and in the summer 2008, two canoe liveries each reported over 20,000 recreators. With

the increasing number of water recreators, it is important to understand whether water quality in the CAWS adversely impacts the health of recreators.

This study is part of the Chicago Health, Environmental Exposure, and Recreation Study (CHEERS), funded by MWRDGC. CHEERS is a cohort study designed to find out what the health risks are of using the Chicago River system for recreation. The study focused on the health risks of limited contact water activities such as canoeing, fishing, kayaking, rowing, and motor boating on the CAWS. During the study period, water samples were collected in CAWS to analyze the concentrations of FIB and pathogens. Water recreators in CAWS were recruited as participants for the study. Runners, walkers, and bicyclists were also recruited as the control group. The goal of CHEERS is to use logistic regression analysis to identify the association between water quality and health risks among water users. In the CHEERS, we encountered similar problem of a non-neglectable amount of missing values. In addition, the association between FIB concentrations and health risks remained unexplained.

This study explored the use of three analytical methods to improve the utility of microbial water quality data in predicting health risk among water users. The three methods were multiple imputation (MI), chemical mass balance (CMB) model, and exploratory factor analysis (EFA). The linkage between microbial water quality data and health risk has been limited due to two major issues. First, missing values of microbial data resulted in a limited amount of data that could be linked to health outcomes. This issue was addressed by evaluating if MI method can be applied to microbial data and provide unbiased results. Secondly, the inconsistent performance of indicator bacteria in indicating the presence and concentration of pathogens

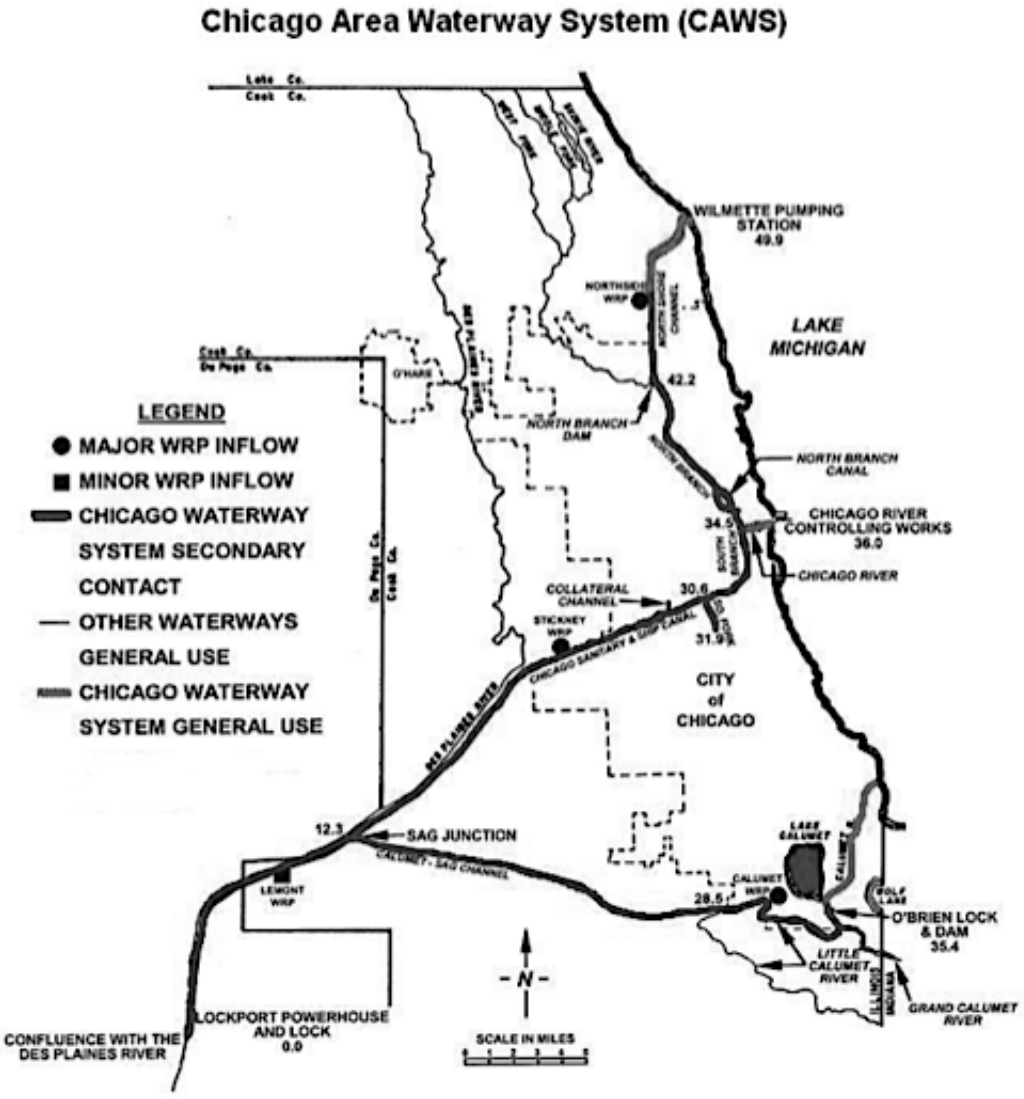


Figure 1. Map of Chicago Area Water Systems

also made it difficult to address water quality and health risk effectively. To address this issues, two different approaches, source apportionment and factor analysis, were utilized for identification of pollutant sources. The goal was to examine if the two methods can signal sources that are posing human health risk.

Based on the two major issues of linking microbial water quality data to health risk, the research hypotheses were developed accordingly. The hypotheses are discussed in the following section.

1.1 Research Hypotheses

The questions this study aimed to answer by exploring the three analytical techniques are:

1.1.1 Hypothesis A

Can multiple imputation be used to generate unbiased values for missing microbial data while maintaining overall variance? The underlying hypotheses are:

1.1.1.1 Assumption 1

The multiple imputation method can fill in missing values for microbial data while preserving the structure of the original dataset and maintaining the variability as it is in the observed data.

1.1.1.2 Assumption 2

The imputed water quality dataset can provide inferential statistic estimates as the original dataset using a logistic regression model to regress water user health outcomes.

1.1.2 Hypothesis B

How well can a receptor model on urban waterbodies be applied to identify sources of contamination? The underlying hypotheses are:

1.1.2.1 Assumption 1

Microbial number balance approach can reach the goal of source apportionment using a chemical mass balance model.

1.1.2.2 Assumption 2

Source profiles developed in the study using observed water quality data can be inferential of pollutant sources in waterbodies.

1.1.2.3 Assumption 3

A logistic regression model using sources as covariates can predict acute gastrointestinal illness (AGI) rates among water users: AGI = f(source1, source2, ...)

1.1.3 Hypothesis C

How well can a statistical approach, exploratory factor analysis (EFA), identify microbial pollution sources? The underlying hypotheses are:

1.1.3.1 Assumption 1

Exploratory factor analysis can provide inference of factors that have impacts on water quality at different pollution levels.

1.1.3.2 Assumption 2

A logistic regression model using factors as covariates can predict AGI rates among water users and reduce multicollinearity problems: AGI = f(factor1, factor2, ...)

The approaches used to test the hypotheses are discussed in each corresponding chapter.

CHAPTER 2

DATA COLLECTION

2.1 Water Quality Measurements

2.1.1 Sampling Strategy

Data used in the analysis portion of the study was taken from water sample and survey data that were collected along the CAWS during three summers between 2007 and 2009. Water samples were collected at the site of recreation or entry onto a body of water while participant recruitment into the survey portion of the study was ongoing. Water samples collected every two hours during participant recruitment were analyzed for Indicator microbes (*E. coli*, enterococci, male-specific coliphages and somatic coliphages). As a result, water quality was measured within two hours of recreation of each study participant. The pathogenic organisms (*Giardia, Cryptosporidium*) are generally at low levels and seldom detected, therefore the samples were only collected every six hours during participant recruitment.

2.1.2 Sampling Methods

Samples were collected following the U.S. EPA protocols. Grab sampling was used for indicators and large-volume sampling was used for pathogens.

Grab samples were collected using a telescopic pole with a bottle attached at front facing upstream to collect water. During the field seasons in 2007 and 2008, grab samples were

TABLE I

METHODS USED TO MEASURE INDICATOR AND PATHOGEN ORGANISMS

Indicators and Pathogens	Method
E. coli	U.S. EPA Method 1603
Enterococci	U.S. EPA MEthod 1600
Coliphages (male-specific, somatic)	U.S. EPA Method 1602
Giardia, Cryptosporidium	CFC (U.S. EPA Method 1623)

collected in individual containers for each indicator method. In 2009, one 2-liter grab sample was collected and then distributed to individual containers for each indicator method.

Large-volume samples were collected by pumping 10-liter of water into containers and filtered in the University of Illinois at Chicago laboratory before being shipped to the laboratory for analysis.

All samples were collected by trained water sampling specialists and samples were placed into coolers with ice and transported to various laboratories for analysis. U.S. EPA methods utilized in measuring water quality for this study are listed in Table I. While collecting water quality samples, samplers also recorded water chemistry data, such as dissolved oxygen (DO), pH, conductivity, and turbidity. The instruments used to measure water chemistry data are listed in Table II.

TABLE II

INSTRUMENTS USED FOR WATER CHEMISTRY MEASUREMENTS

Water Property	Instrument
pH DO	Accumet AP84 Portable Waterproof pH/DO Meter
Conductivity	Oakton CON 6 Hand-Held Conductivity/TDS Meter
Turbidity	HF Scientific MicroTPW Field Portable Turbidimeters, model 20000

2.1.3 Sampling Locations

The sampling sites for the North Branch System and Cal-Sag Channel are presented in Figure 2 and Figure 3 respectively. The sampling locations in North Branch System included Bridge Street (BR), Skokie Rowing Center (SK), Lincoln Avenue (LA), River Park (RP), Clark Park (CP), and North Avenue (NAM). Bridge Street and Skokie Rowing Center are located 4.2 and 0.7 km upstream of the North Side water reclamation plant (WRP). Lincoln Avenue, River Park, Clark Park, and North Avenue were located 3.2, 5.8, 9.1, and 14.6 km downstream of the WRP. Skokie Rowing Center was geographically upstream of the WRP, but considered a downstream location due to the dispersion of effluent from the North Side WRP in the vicinity of this location. Therefore, in the data analysis portion, SK was considered as a downstream site. The sampling locations in the Cal-Sag Channel included Beaubien Woods (BA), Riverdale Marina (RM), Alsip (AL), and Worth (WO). Beaubien Woods was located 1.3 km upstream of the Calumet water reclamation plant. Riverdale Marina, Alsip, and Worth were located 4.8, 14.6, and 18.8 km downstream of the Calumet WRP, respectively.

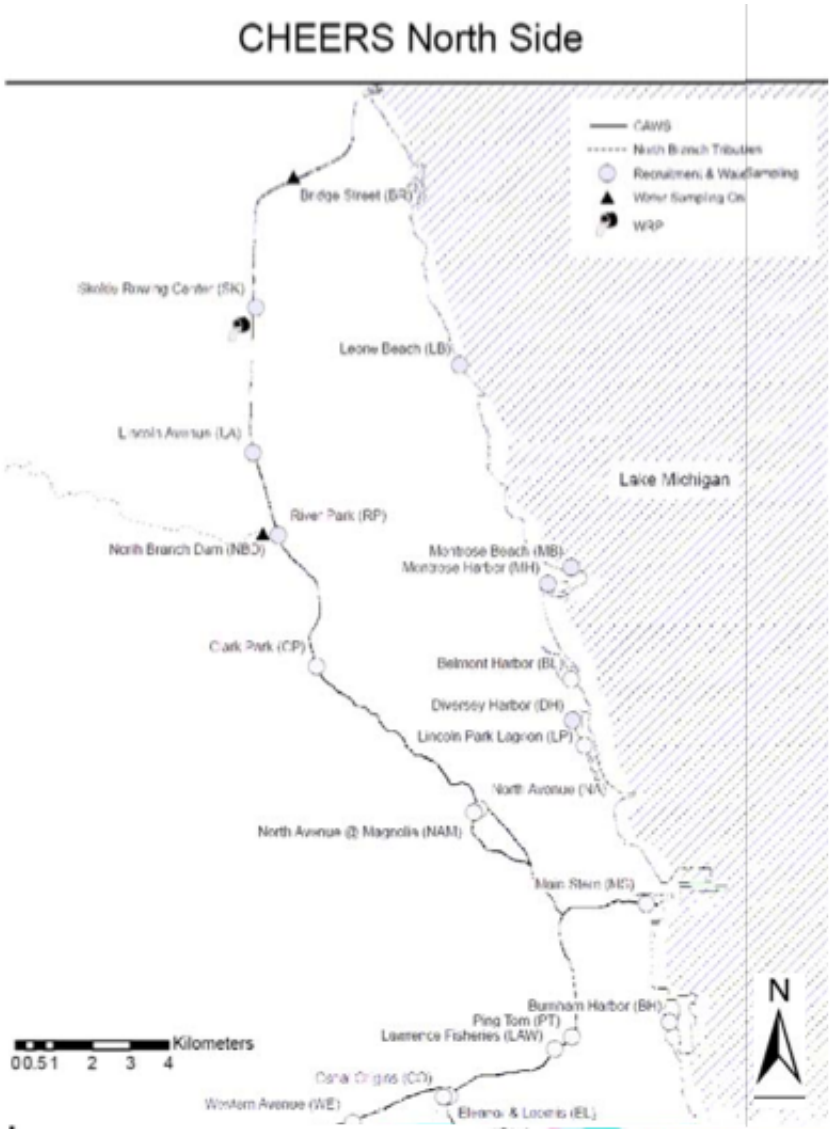


Figure 2. CAWS North Branch System sampling sites

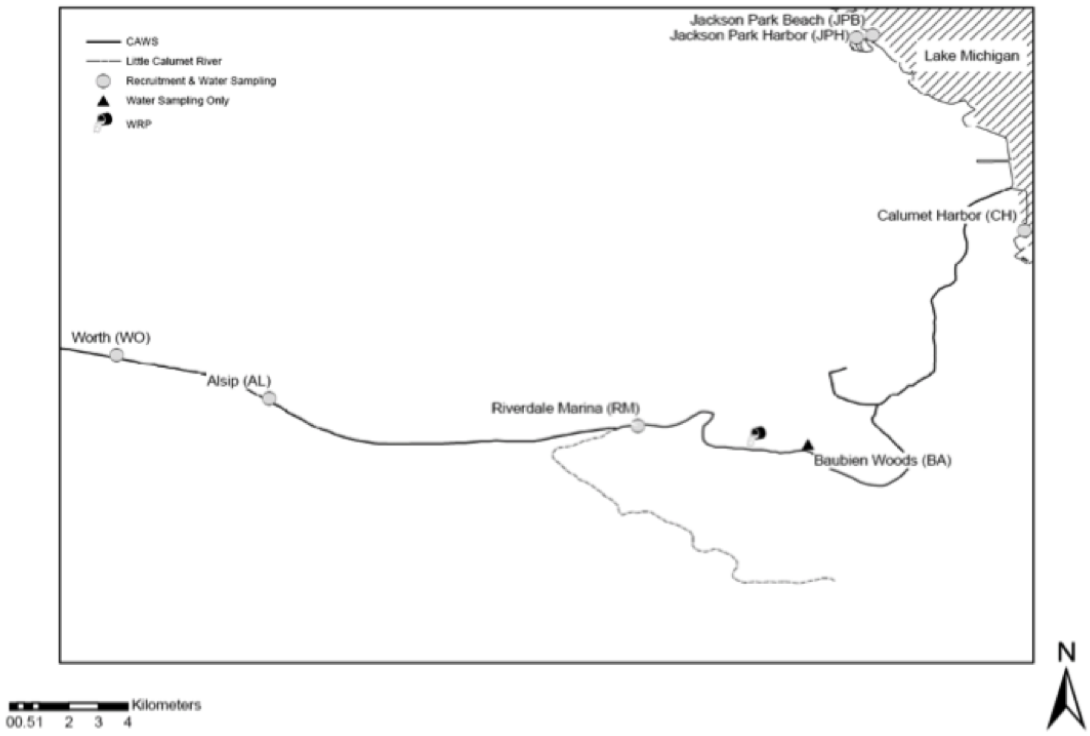


Figure 3. CAWS Cal-Sag Channel sampling sites

The South Branch sampling locations included Ping Tom (PT), Lawrence Fisheries (LAW), Canal Origins (CO), and Wester Avenue (WE). All of the South Branch sampling sites are located downstream of the North Side WRP, but due to the long distance from the plant, they are separate from the North Brach group. The rest of sampling sites in CAWS are Willow Springs (WS) and Main Stem (MS). Willow Springs, downstream of the Stickney WRP, is located on the Chicago Sanitary and Shipping Canal (CSSC). Main Stem is located downstream of the Chicago Locks and Controlling Works on Lake Michigan.

2.1.4 Data Quality

During the study, external quality control (QC) was performed using blank samples, split samples, and spiked samples. Split samples were used to evaluate precision; spike samples were used to assess accuracy.

Autoclaved deionized water was used as the blank media. Results of blank samples are showed in Table III. For *E. coli*, male-specific coliphages, and somatic coliphages, only two samples of each microbe had non-zero counts. There were four enterococci blank samples had non-zero counts.

Splits samples were collected from the water using one large container and separated into two sample bottles. Spearman correlation coefficients in split samples are showed in Table IV. The split analysis shows high level of precision among all microbes except *Cryptosporidium* ($\rho = 0.57$). No specific pattern of disagreement between splits of *Cryptosporidium* was observed.

TABLE III

FIELD BLANK RESULTS

	Indicator Bacteria (Method Blanks)		Coliphages (Method Blanks)	
Quantile	*E. coli* (n=249)	Enterococci (n=249)	Male specific (n=242)	Somatic (n=242)
100% Max	4000	547	13576	50
99%	220	200	52	20
95%	0	13	0	0
90%	0	3	0	0
75% Q3	0	0	0	0
50% Median	0	0	0	0
25% Q1	0	0	0	0
10%	0	0	0	0
5%	0	0	0	0
1%	0	0	0	0
0%	0	0	0	0

TABLE IV

SPEARMAN CORRELATION COEFFICIENTS BETWEEN SPLIT SAMPLES

	Indicator Bacteria		Coliphages		Protozoan Pathogens	
	E. coli	Enterococci	Male specific	Somatic	*Giardia*	*Crypto*
No. of pairs	183	184	184	184	14	14
Correlation coefficients	0.90	0.88	0.90	0.92	0.96	0.57
P-value	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	0.0332

TABLE V

RECOVERIES FOR SPIKED SAMPLES

	Indicator Bacteria		Coliphages		Protozoan Pathogens	
	E. coli	Enterococci	Male specific	Somatic	*Giardia*	*Crypto*
Count	118	130	115	115	34	34
Average	110%	68%	94%	61%	22%	21%
EPA criteria	17-117%	63-110%	Detect to 120%	48-291%	15-118%	13-111%

Spiking involved the subdivision of a water sample into two samples. A known concentration of microbes was added into the first sample and the second sample was not manipulated. Recovery was then calculated by dividing the microbe concentration detected in the spiked sample, by the sum of the expected concentration added to the spiked sample and the microbe concentration detected in the non-spiked sample. The calculated recoveries, which all fell within U.S. EPA criteria, are showed in Table V.

During the period between 9/2008-5/2009 poor recovery results of *E. coli* and enterococci were identified. Although the average recovery fell within the range recommended by the U.S. EPA (17%-117% for *E. coli*, and 63%-110% for enterococci), a significant number of recoveries landed outside the recommended range. Among all the samples collected 2007 - 2009, 7% of both *E. coli* and enterococci samples, had undetectable levels of microbes. This result is surprising since over 50% of these samples were collected downstream of the WRP. The WRP continuously releases these microbes, and microbial concentrations are consistently in the range of 1,000 CFU/100mL. Therefore, these 7% samples are interpreted to indicate laboratory error.

A method was developed to determine questionable samples. The method utilized involved taking a running average of *E. coli* and enterococci recoveries for three consecutive sampling days. If the three-day running average of recoveries fell outside the U.S. EPA recommended range, all samples for that particular microbe on the middle day of the three-day period were excluded from the dataset. After eliminating questionable data points, the final dataset had 302 (27%) and 425 (38%) missing values of *E. coli* and enterococci respectively. It limited the number of available data points in modeling water quality. In addition, a large portion of health data could not match to water quality data due to the missingness, impacting the statistical power to detect associations between water quality and health risk among water users.

2.2 Meteorology Data

Precipitation data was obtained from the Illinois State Water Survey(24) from multiple monitoring stations. Each location was matched to the nearest precipitation gauges. Rain gauge data was recorded hourly everyday with the amount (inch) of precipitation. Before linking the precipitation data to the water quality data, the information was converted to the magnitude and duration of last rain, and also the duration between last rain and sampling time.

Data of combined sewage overflows (CSOs) was obtained from quarterly reports provided by the MWRDGC to the Illinois EPA. Combined sewage overflows data includes the magnitude and duration of last CSO event, and also the duration between the last event and the sampling time. Each CSO event in the North Branch System and Cal-Sag Channel was defined by release from any outfall with either branch. Two events were distinguished by 1 hour without CSO.

2.3 Health Data

Water users health history and water activity were collected at the sampling sites. Using interview surveys conducted, participants were then followed for 21 days through 3 phone interviews to investigate the development of AGI and other illnesses, such as respiratory symptoms and skin rash.

The analysis presented here uses health outcomes data collected via phone interviews in the 21 days subsequent to recreation. However, a large portion of data used is based on information obtained on the day of recreation. Specifically, demography and participants' health history were obtained prior to recreation. Immediately after recreation, information about the type of water activities, how wet did the participant get during the course of activity, and the water sport concern one had were obtained.

For water sport concern, participants were asked to rank the health risk of conducting water activities on Chicago River on a scale of 0 to 10.

After recreation, participants were also asked to report water exposure by body part (face/head, torso, hands/arms, and feet/legs), on a scale of 0 to 5 from "not wet at all" to "submerge". Each body part was weighted differently assuming the risk of getting sick varies by where the participant has water exposure. A weighted average of these response was then used as the wet score of each participant.

A summary of participants responses is showed in Table VI.

TABLE VI

SURVEY DATA RESULTS

Variable	Mean or Percentage
Age	35
Age 10 and lower	7.9%
Age 65 and over	2.89%
Gender: female	49%
Gender: male	51%
Race: black	7.69%
Race: hispanic	4.87%
Race: other	8.22%
Race: white	79.22%
Motor boating	16.7%
Canoeing	22.3%
Fishing	10.7%
Kayaking	34.2%
Rowing and other limited contact	16.1%
Pre-existing GI illness	4.11%
Previously exposed to GI illness	3.04%
Wet scores	3.83
Water sport concern	4.70
GI illness	4.42%

CHAPTER 3

MULTIPLE IMPUTATION OF MISSING MICROBIAL WATER QUALITY DATA

3.1 Literature Review

Missing data is a frequently encountered problem in environmental health research due to the ubiquity of long term field sampling campaigns. Long term field sampling campaigns are vulnerable to missing data as personnel, weather and equipment issues disrupting sampling. Another cause of missing values is censored data, which are values under minimum detection limit reported as non-detect (ND). The presence of missing data can effect the ability to perform modeling and statistical analysis, and it can affect the observation of an annual or seasonal patterns. In general, if the proportion of missing values is small, listwise deletion, e.g., the omission of the entire observation when any of the variables are missing, is a reasonable approach. However, as the proportion of missing data increases, deletion can introduce bias and inaccuracies in further analyses, especially if the pattern of missing data is not completely random. Listwise deletion also decreases the sample size, which may reduce the ability of the study to detect a true association. Consequently, proper handling of missing data is important in order to achieve the specified research goals.

3.1.1 Missing Mechanisms

Rubin(25) categorized missing patterns into three types, missing at random (MAR), missing completely at random (MCAR), and missing not at random (MNAR) based on the different mechanisms that caused missing data. The mechanisms describe the relationship between missing data and other measured variables in a dataset. Missing at random indicates that missingness of the variable is not related to its value, but to the value of some other variable. In other words, the probability that variable, Y, is missing depends on another observed variable, X, but not on variable Y itself. Missing completely at random describes the pattern that the missing values are not related to any other observed data. This means the distribution of the variable with missing values and the relationships between variables are preserved in the dataset, and the missing data is a random subset of the original dataset (26). Missing not at random means the probability of missing data being associated with the values that are missing.

If the data are MCAR or MAR, a gap filling technique, such as mean replacement or multiple imputation, can be applied because the missing data is a random subset of the original data. However, if the data are NMAR, the missing mechanism cannot be simply ignored. There are two methods used for NMAR data to model missing mechanism, selection models and pattern mixture (27; 28). In a selection model, one simultaneously models the variable with missing values and the probability of the values being missing. In a pattern mixture approach, one performs multiple imputation with different assumptions about the missing mechanism. This

allows exploration of how sensitive the results are to the assumption about missing mechanism.

There is no specific method to determine the missing mechanism in a dataset. Instead, the data is explored to determine what assumptions about the missing mechanism are plausible and appropriate. It is more important to determine if the data are MNAR than if the data are MCAR or MAR because techniques of the treatment of MNAR data are unique from these for MCAR or MAR.

The two major causes of missing values are (1) lack of sample collection, and (2) censored data. In both cases, the missing patterns are seldom MCAR. In the first case, the missing values could be systematically related to a certain location, instrument, or personnel. In the later case, the missing values are clustered at low values. These missing mechanisms can, however, be classified as MAR as long as variables causing the missingness are included in the data set.

Collins et al.(29) tested many realistic cases and concluded that an incorrect assumption of MAR would only cause slight effect on parameter estimates and standard errors. When a dataset needs a treatment of missing data and clearly the missing mechanism is not MNAR, a gap filling method should be considered to fill in missing values.

3.1.2 Traditional Gap-Filling Methods

Arithmetic mean imputation (AMI) is one of many single imputation methods. It replaces missing values with the arithmetic mean of the non-missing values. The advantage of AMI is

the generation of unbiased parameter estimates if the data is MCAR because the missing values are being replaced by the mean which lands on the regression line. However, the limitation of AMI is that it changes the distribution of the original data by narrowing the variance (30). Similarly, in median replacement, missing values are replaced by the median. This method shares the same limitation as AMI method.

In contrast with AMI or median replacement, regression imputation regresses the variable with missing values using other variables in the dataset and improves some of the problems in reducing the variance AMI and median replacement methods encounter (31). However, since all the imputed values fall on the regression line, the variance is still not as large as the original dataset. Stochastic regression imputation (SRI) additionally addresses variance distortion by adding a randomly sampled residual to each imputed value. SRI is thus considered to be one of the best traditional missing-data techniques.

3.1.3 Modern Gap-Filling Methods

Improved computer technology has allowed researchers to apply computational methods to these problems. Modern missing-data techniques include direct maximum likelihood (DML) and multiple imputation (MI), both of which use a likelihood approach to address missing values and maintain the distribution of the original data set (28).

DML does not physically impute values to fill in missing data. Instead, it estimates model parameters and standard deviations directly using observed data. The DML procedure uses a particular set of parameters to estimate the likelihood that a missing value will occur. It runs

a series of iterations by replacing different values for the unknown parameters and converges to a single set of parameters with the highest probability of matching the observed data. The algorithm can be explained by the following equation (30):

$$\log L_i = K_i - \frac{1}{2}\log|\Sigma_i| - \frac{1}{2}(x_i - \mu_i)'\Sigma_i^{-1}(x_i - \mu_i) \tag{3.1}$$

Where, x_i is a vector of observed data of all available variables in observation i, μ_i is the vector of mean estimates of the available variables in i, Σ_i is the estimated covariance matrix, and K_i is a constant. Based on the missing pattern, the size and contents in the equation change accordingly for each observation i. For each observation i, the log likelihood is calculated, and the parameter estimates are determined by maximizing the log likelihood. Since DML does not replace the missing values, one cannot confirm that the fitted coefficients reflect an imputed data set, which has the same distribution as the set of measured values.

Multiple imputation is a simulation-based method that creates m data sets with replaced missing values. Multiple imputation enables the variation introduced by imputation to be compared across the m imputed data sets (27).

There are two major imputation algorithms, [1] propensity score with the approximate Bayesian bootstrapping technique and [2] regression model with data augmentation (DA) algorithm. In the propensity score method, a logistic regression is used in which the dependent variable is wether or not an observation is missing. The estimated logistic regression model is then used to calculate the probability of the observation being missing. According to the probability of the

observation being missing, a propensity score is generated for each missing value in each variable. The observations are then broken into groups based on the propensity scores and an approximate Bayesian bootstrap imputation is applied. The DA algorithm generates a regression equation from measured data set. The generated regression equation along with random noise are then used to fill in missing values. In the next step, it simulates a sequence of random draws of each parameter from the posterior distribution and then concludes another imputed data set. Allison (32) compared the two imputation algorithms, the propensity score and the DA algorithm, and concluded that when MAR was violated, the propensity score method generated more biased parameter estimates. In addition, propensity score method should be only applied to data sets with monotone missing patterns. A data set has a monotone missing pattern if an $\mathrm{i}^{th} observation has a missing value in variable \mathrm{X}_2$ then the $\mathrm{i}^{th} observation of variable \mathrm{X}_j$, $j > 2$, are missing as well. If a data set has a non-monotone missing pattern, DA approach should be utilized (33). In the current study, the missing data of indicator microbes does not follow a monotone pattern. Consequently, the DA method was preferred over the propensity score approach.

The DA algorithm was proposed by Schafer (34) to include cyclic repletion of an imputation step (I-step) and a posterior step (P-step). Initially, the DA uses the Expectation Maximization (EM) algorithm and ML estimation to generate a covariance matrix. During the I-step, missing values are filled in with imputed numbers based on a multivariate regression equation. To avoid the imputed values falling on the regression surface, random noise is added to each imputed value by randomly selecting a residual from a normal distribution. In the end of the I-step,

a new covariance matrix and mean vector are generated using the complete imputed dataset. In the P-step, random noise is added to the regression equation previously used to generate the imputed data, this prevents MI from generating the same sets of missing values. During P-step, new elements of a covariance matrix and mean vector are randomly selected from a posterior distribution based on the imputed data in I-step. The newly constructed estimates of the covariance matrix and mean vector are used to predict a new set of imputed values for the next, I-step. The iteration between I-step and P-step is repeated a large number of times until the regression model converges, at which point the "final" set of imputed data is generated. The iteration is repeated until m sets of imputation data is generated (usually m=5 - 10).

In addition to Schafer's (34) DA algorithms, Richman et al. (35) proposed machine learning algorithms for use with MI methods. Support vector machines (SVM) and artificial neural networks (ANN) were proposed to fill in missing values following four steps of imputation:

1. Separate the original dataset into two datasets, one with no missing values (set 1), and the second with missing values (set 2).

2. Begin the first iteration for each variable that has missing data in set 2 using the regression equation constructed from set 1.

3. Begin the second iteration by merging set 1 with the imputed set and constructing the regression equation using the merged set. Fill in missing values in set 2 by using the newly constructed regression equation.

4. Repeat step 3.

The SVM was developed by Vapnik (36). With a set of training rules $\{(x_i, y_i)\}_{i=1}^{l}$ of l observations, the target is to construct a function for the dependent variable y, using the following equation:

$$y = f(\vec{x}) = \vec{w} \cdot \vec{x} + b \tag{3.2}$$

Where, $\vec{w}$ represents the weight vector and b is bias. The formula of SVMs is explained in detail by Vapnik(36). Artificial neural network is a network constructed with computational nodes of one input layer, one or more hidden layers, and an output layer (37). Figure 4 shows a design of an ANN with four inputs, two hidden layers, and an output layer. The first hidden layer contains eight variables and the second layer contains four variables. The model theory and construction of ANNs is explained by Haykin (37).

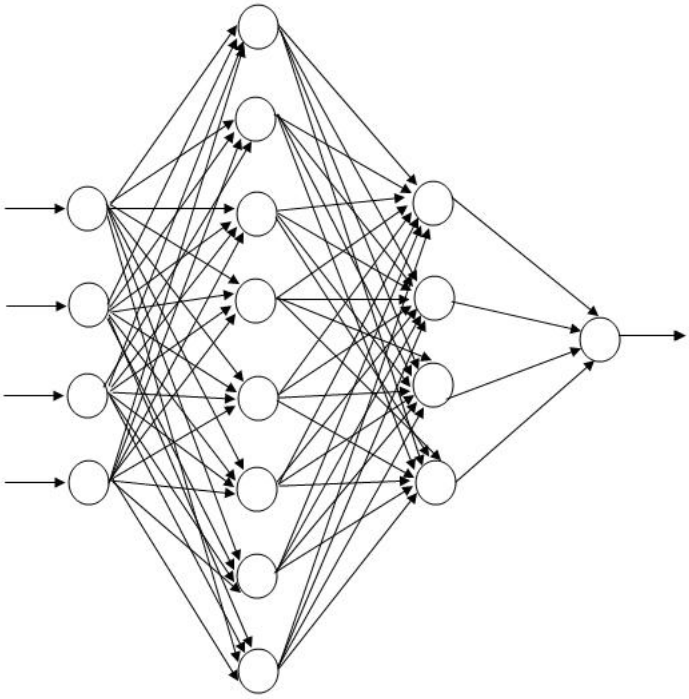


Figure 4. An artificial neural network design

3.1.4 Comparison Between Traditional and Modern Gap-Filling Methods

Multiple studies have compared MI to traditional methods including listwise deletion, mean imputation, and median replacement using artificial data sets (30; 38; 31). All studies concluded that MI constructs the closest variance to the original dataset and yields a better estimation of the mean in comparison with traditional gap-filling methods.

Richman et al. (35) compared results of using MI with machine learning algorithms (SVM and ANN) to data deletion, mean replacement, simple linear regression, and multiple regression. The methods were tested against a data set with 5%, 10%, and 20% of missing values. Overall, SVM has the least errors in the mean and variance of the estimates when compare to traditional missing-data methods. Machine learning algorithms yield the best parameter estimates in the data set with the least amount of missing data. As the percentage of missing data increases, the ANN algorithm performance, measured by the errors in the parameter estimates, decreases and may reach performance levels similar to those obtained by traditional mean substitution methods.

MI has been evaluated for treatment of left-censored data. Because non-detect data clusters at the low range of values which imply a MNAR mechanism, it is questionable that MI is adequate for gap-filling. However, Chen et al. (39) applied MI in a simulation study to a pesticide data set to handle non-detect data and concluded that imputed and complete data sets can yield consistent parameter estimates.

The goal of filling in missing values in this study is to provide inferential statistic estimates from an incomplete dataset by preserving the structure of the original dataset and also ensuring that uncertainty remains as it is in the data. All the studies found that modern missing-data techniques can better achieve the goal than tradition methods. Therefore, modern imputation techniques were applied to replace missing microbial indicator data in this study.

3.1.5 Appropriate Variables for Gap-Filling Methods

A challenge in the use of the MI method is the identification of variables included in the imputation model. Collins et al. (29) addressed this question by comparing parameter estimates obtained using various amount of variables in imputation model and found that with the more number of variables in the model, the imputation results improved.

A study conducted by Zhou et al. (38) also used MI on an dataset with artificially created missing values and tested the parameter estimates obtained using imputed data sets. The authors imputed four sets of data using different numbers of variables in the imputation model. All the models were found to introduce bias into the dataset. However the model with the most variables introduced the least amount of bias.

3.1.6 Treatments of Missing Data in Microbial Water Quality

In water quality research, data sets often contain missing values. Whitman et al. (40) found that one of five years of *E. coli* data was missing, which limited observation of *E. coli* patterns in that year. Whitman et al. (40) decided to omit the data from that year and only reported years with available data. While Nevers et al. (41) chose to fill in missing values using the average of the three previous and three subsequent values. The authors from both of the studies did not report the percent of missing values in the data and the reasons that led to the treatment of missing values. In the context of censored data, both Whitman et al. (40), and Boehm et al. (42) in identifying pollutant sources encountered either values under or over detection limit. In

both studies, the authors determined that the percent of censored data did not have significant effect on statistical inferences. Therefore, omitted the censored data from any analyses.

Based on my review of the methodology, under certain circumstances researchers should consider applying the MI method to solve major missing data problems. These circumstances are:

1. when the proportion of missing values reaches a level that is not ignorable;
2. when the assumption of MAR is plausible which includes cases that the factors causing MNAR are also in the imputation model as variables;
3. and, the percent of missing values makes designed statistical analysis less reliable or impossible;

Imputation technique should be limited to independent variables. Even though Schafer et al. (43) stated that missing values on independent variables and missing values on dependent variables are not fundamentally different, in order to avoid the concern of "making up data," gap filling techniques should be only utilized to fill in missing values of independent variables, so in such matters all the observed values of the dependent variable can be used in modeling.

3.2 Methods

Data collected for the CHEER study encountered missing data problems. The samples were collected and analyzed. However, due to questionable laboratory performance, *E. coli* and enterococci concentrations measured on selected dates were excluded, and considered as "missing". Out of the 1,123 water samples collected and analyzed, 302 sample (27%) results for *E. coli* and

425 (38%) for enterococci were excluded due to quality control issues. The large proportion of missing data limited the number of water users which could be matched to water quality measurements, thereby reducing the sample size for the analysis of associations between health outcomes and water quality.

A MI method, Schafer's (34) DA algorithm was utilized to replace missing values of the two independent variables, *E. coli* and enterococci, in this study. Missing patterns were examined first to ensure the plausibility of MAR assumption. The goal was to identify a gap filling method that can be applied to microbial water quality data and provide unbiased statistic inference if missing values are not ignorable. Multiple imputation method is easy-to-use and there are a number of freeware of commercial software packages supporting MI method, it is selected over machine learning approach.

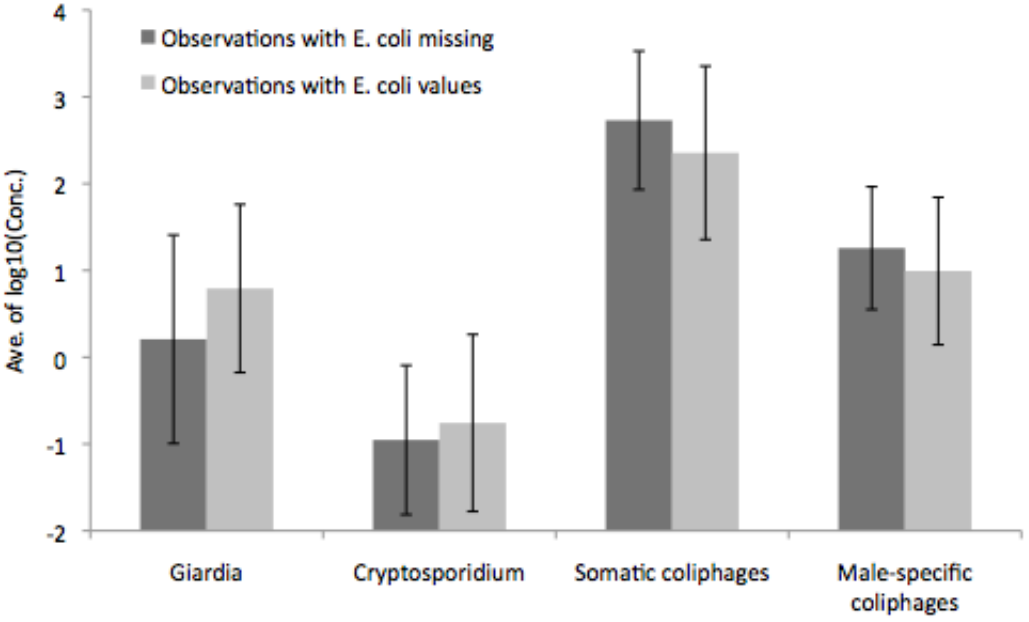
First, the imputed data set was compared to the original data set using descriptive statistics. Next, a subset dataset was created excluding all the observations with a missing value of either *E. coli* or enterococci. This complete data set was used to fit a logistic regression model to predict GI illness. The obtained coefficient estimates were considered as the "true parameter estimates". I then calculated new probability of GI illness data using the true parameter values and converted it to a binary outcome variable. In the next step, a portion of *E. coli* and enterococci values was randomly deleted to create an artificial data set. Multiple imputation was applied to fill in missing values in the artificial data set. The imputed dataset was used to fit the logistic regression model to predict newly calculated GI illness outcomes. The resulting parameter estimates were then compared to the true parameter estimates.

In this chapter, "original data set" indicates the data set of 1,123 observations collected in the summer from 2007 to 2009. In addition, "complete data set" represents the subset of data, containing 573 observations, from the original dataset with no missing values of *E. coli* or enterococci. Another term, "artificial data set", represents two types of data: one with partial *E. coli* values deleted and one with partial enterococci values deleted. Missing values of artificial data sets were then imputed using MI method.

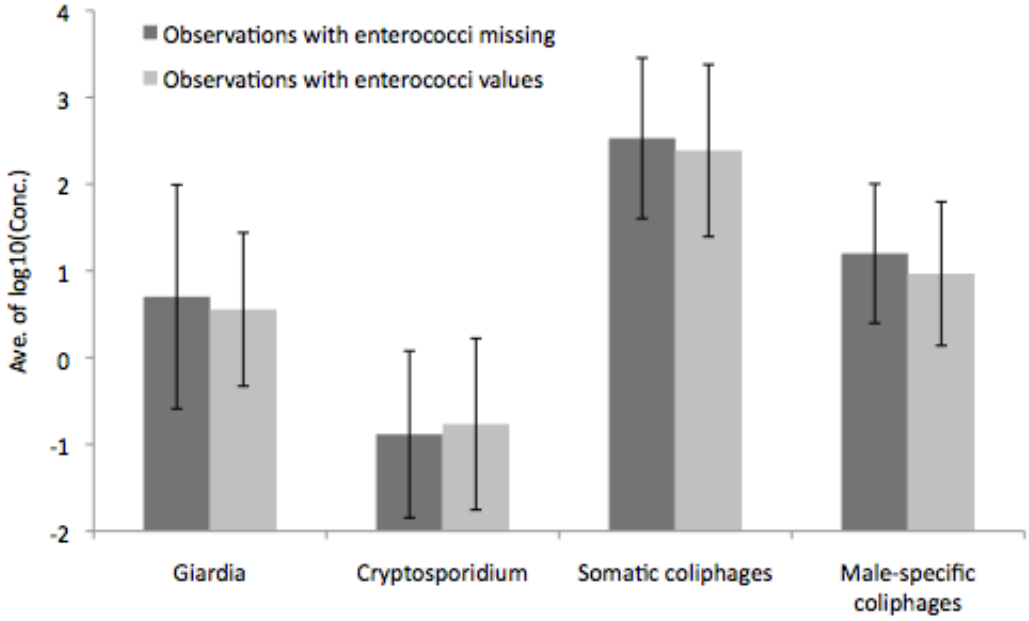
3.2.1 Orignal Data Set

Before the application of multiple imputation, the observations with or without *E. coli* values were grouped separately in order to examine the missing mechanism. For each group, the mean and standard deviation of each variable were calculated. If the means are not different between missing value group and non-missing value group, the assumption of MAR is plausible. Same procedure was repeated for enterococci data. Figure 5 is an example of the means and standard deviations of measured microbes compared between two groups, and it shows no significant difference between groups.

S-PLUS® software was used for MI method, implementing Markov chain Monte Carlo (MCMC) imputation mechanism. The function used to impute missing values was S+MISSINGDATA, available by loading library named "missing". Yuan (33) showed marginal relative efficiency increased significantly when the sets of imputation changed from 3 to 5 and remained constant when the number increased from 5 to 10. Therefore, five sets of imputation were generated in this study. One thousand iterations within each imputation in a single chain were conducted.



(a) With/Without *E. coli* values



(b) With/Without enterococci values

Figure 5. Comparison of mean and standard deviation between observations with missing values and observations without missing values

The first 100 iterations of each imputation were discarded to ensure the independence between each single imputation.

E. coli and enterococci data was log10-transformed to meet the normality assumption criteria. Based on Zhou's (38) finding that parameter estimate bias is minimized by the inclusion of numerous variables, I included all possible variables in the imputation. Specifically, variables included for both *E. coli* and enterococci models are: location, concentration of *Giardia*, *Cryptosporidium*, somatic coliphages, and male-specific coliphages, sampling hour, pH, dissolved oxygen, conductivity, turbidity, water temperature, time since last rain, time since last combined sewer overflow (CSO) event, duration of last rain/CSO, and magnitude of last rain/CSO. The variable location is categorical, while all others are continuous. In addition, for the *E. coli* imputation, enterococci was included, and visa-versa.

3.2.2 Artificial Dataset

The complete dataset with 573 observations was used to create an artificial data set with missing values. Both complete data set and artificial data set were used to fit the logistic regression model and the resulting parameter estimates were compared to each other.

3.2.2.1 Missing Pattern

In order to create a MAR pattern, the complete dataset was broken down into three subgroups based on the somatic coliphage concentrations. Log10-transformed somatic coliphage concentration was used as the criterion to decide the percent of data deletion. The concentrations ranged from 0 to 5 and were classified into three subgroups, low concentration (0 to <1.5),

TABLE VII

PERCENT AND NUMBER OF SAMPLES DELETED FROM THE ARTIFICIAL DATA SET

Number of samples in complete data set		Dataset1		Dataset2	
Concentration levels	Percent (Number)	*E. coli*	Enterococci	*E. coli*	Enterococci
LOW	30% (171)	15% (86)	15% (86)	5% (29)	9% (52)
MED	43% (246)	10% (57)	10% (57)	11% (63)	15% (86)
HI	27% (156)	10% (57)	10% (57)	8% (46)	12% (69)

medium concentration (1.5 to <3), and high concentration (3 to ≤5). In the original data set, there are 5%, 11%, and 8% of missing *E. coli* data in low, medium, and high concentration groups. Nine percent, 15%, and 12% of enterococci data are also missing in the three groups.

In order to test the effect of different missing patterns, two sets of missing data were generated. Data set 1 was created by randomly deleting 15%, 10%, and 10% of *E. coli* and enterococci data from the low, medium, and high concentration groups. Data set 2 was created following the missing patterns of the original data set. The number of total samples in each concentration level and the number of *E. coli* and enterococci in data set 1 and 2 are listed in Table VII.

For each set of missing data, MI was utilized to fill in five sets of imputed values. Each set of imputed data was used to fit a logistic regression model. Variables used in the *E. coli* and enterococci logistic regression models are listed in Table VIII and Table IX respectively. Covariates included in the logistic regression models were those having biological plausibility,

TABLE VIII

VARIABLES USED IN THE *E. COLI* LOGISTIC REGRESSION MODEL TO PREDICT GI ILLNESS

Dependent	GI illness
Independent	*E. coli* concentration
	Age 1-10
	Age 65over
	Gender
	Hispanic
	White
	Other race
	Pre-existing GI
	Previous exposed to GI
	Wet score
	Duration since last rain
	Boating
	Canoeing
	Kayaking
	Rowing
	Water sport concern
	CSO within 24 hours

and those identified as potential confounders. Backwards model selection was used to evaluate if the identified covariates should be included in the model, applying an $\alpha = 0.05$ significance criterion.

TABLE IX

VARIABLES USED IN THE ENTEROCOCCI LOGISTIC REGRESSION MODEL TO PREDICT GI ILLNESS

Dependent	GI illness
Independent	Enterococci concentration
	Age 1-10
	Age 65over
	Gender
	Hispanic
	White
	Other race
	Pre-existing GI
	Previous exposed to GI
	Wet score
	Duration since last rain
	Boating
	Canoeing
	Kayaking
	Rowing
	Water sport concern
	CSO within 24 hours
	Interaction term between enterococci conc. and wet score

The resulting parameter estimates and standard deviations were used to calculate the final parameter value and standard deviation following the method presented by Rubin (27).

Suppose that $\hat{Q}$ is the parameter estimate in the logistic model for the j^{th} imputed data set, where $j(j = 1, 2, ..., m)$, and U_j be the standard deviation of $\hat{Q}_j$. Then, the best estimate of $\hat{Q}$, the true parameter, is the average of the $\bar{Q}$.

$$\bar{Q} = \frac{1}{m}\sum_{j=1}^{m}\hat{Q}_{j.} \tag{3.3}$$

To calculate the overall variance, one has to calculate the within-imputation variance (Equation 3.4) and between-imputation variance (Equation 3.5):

$$\bar{U} = \frac{1}{m}\sum_{j=1}^{m}U_{j.} \tag{3.4}$$

$$B = \frac{1}{m-1}\sum_{j=1}^{m}(\hat{Q} - \bar{Q})^2 \tag{3.5}$$

The total variance, T, is:

$$T = \bar{U} + (1 + \frac{1}{m})B \tag{3.6}$$

The calculated parameter estimates and standard deviations were then compared between two missing pattern data sets for the examination of the impact caused by different missing patterns.

3.2.2.2 Using Artificial Dataset To Predict Health Outcomes

The complete *E. coli* data set was used to fit the logistic regression model against which the parameter estimate based on the imputed data are compared. The resulting parameter estimates were considered as true values. The fitted GI illness probabilities were then calculated based on the parameter estimates. Since logistic regression model is used for data with binary outcome variable, a threshold of 0.013 was used to convert the calculated probabilities back to binary GI illness outcomes. The value, 0.013, was the probability of GI illness contributed by conducting limited contact water recreation on the CAWS. Setting a threshold to convert the probabilities to binary outcomes unfortunately introduces uncertainty to the parameter estimates and they can no longer be considered as "true parameter estimates." However, using the threshold developed from this specific data set was the best approach to keep the uncertainty minimum. The imputed data sets were then used to regress the fitted GI illness outcomes. The resulting parameter estimates were compared to the ones obtained from using complete data set. This procedure was also repeated for the enterococci model.

In order to estimate the bias introduced by threshold 0.013, the data sets were used to regress fitted logit using a multivariate linear regression analysis. The fitted logit were calculated following Equation 3.7, where P_i is the calculated probability, β_S are the parameter estimates, and $X_{i,j}$ are the variables in the model. This procedure avoided using threshold, 0.013, to convert fitted probability to binary health outcome variable, and therefore eliminated the bias

introduced during the conversion. The resulting parameter values using the artificial data set were then compared to the ones obtained using the complete data set.

$$\log \frac{P_i}{1 - P_i} = \beta_0 + \beta_1 X_{1,i} + ... + \beta_k X_{k,j} \tag{3.7}$$

3.3 Results

Results of applying the MI method to original data set and artificial data set in addition to results of using imputed data to regress GI illness outcomes are discussed in the following sections. *E. coli* and enterococci models are evaluated separately.

3.3.1 Multiple Imputation Using Original Data Set

Descriptive statistics of both original and imputed data sets for *E. coli* and enterococci were compared in Table X and Table XI respectively. The imputed data set presents a similar distribution as the original data set. A boxplot of the two indicators in the original and imputed data sets shows the similarity in the distributions of the two sets, Figure 6. This analysis showed that MI provided a good simulation for gap filling and could be applied on water quality data sets.

TABLE X

COMPARISONS OF *E. COLI* VALUES IN ORIGINAL DATASET AND IMPUTED DATASET

Dataset	N	Means	Std. Dev	Q1	Q3	Minimum	Maximum
Original dataset	821	2.67	1.06	2.04	3.44	-1.00	5.23
Imputed dataset	1123	2.68	0.99	2.07	3.40	-1.00	5.23

TABLE XI

COMPARISONS OF ENTEROCOCCI VALUES IN ORIGINAL DATASET AND IMPUTED DATASET

Dataset	N	Means	Std. Dev	Q1	Q3	Minimum	Maximum
Original dataset	698	2.22	0.79	1.76	2.75	-1.00	4.50
Imputed dataset	1123	2.17	0.81	1.65	2.73	-1.00	4.50

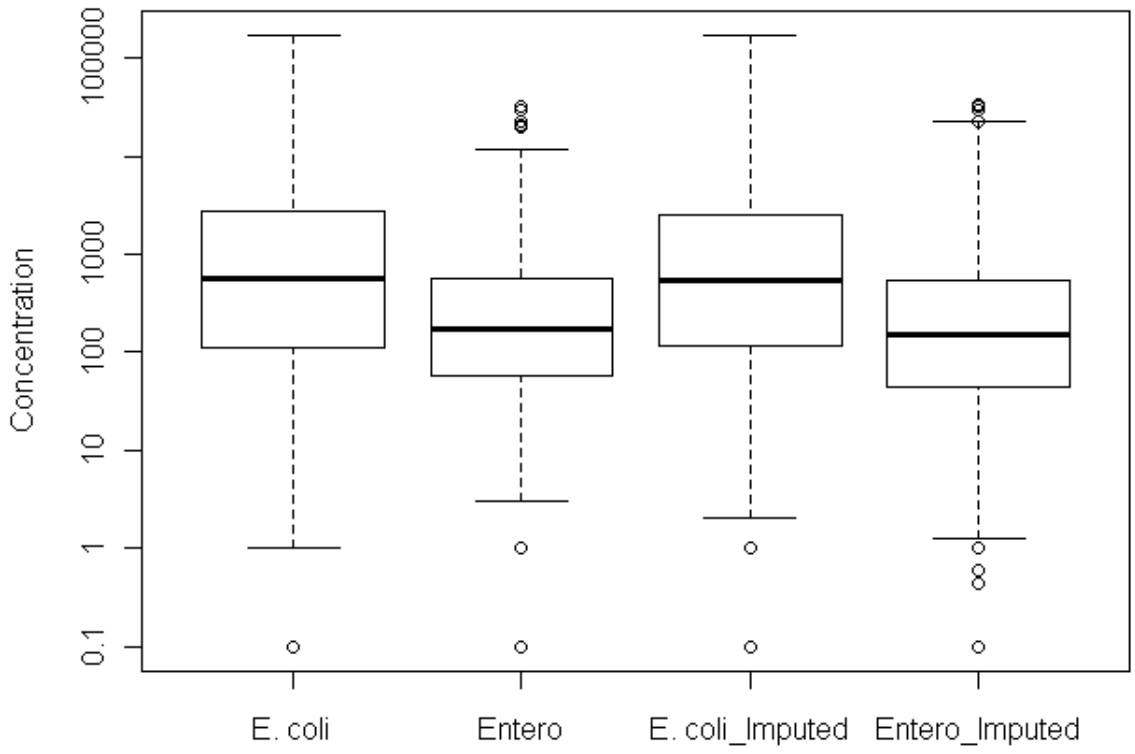


Figure 6. Comparison of distribution

TABLE XII

MEANS AND STANDARD ERRORS WITH DATA DELETION (DD), AVERAGE REPLACEMENT (AR), MEDIAN REPLACEMENT (MR), AND MULTIPLE IMPUTATION (MI) METHODS

E. coli	DD	AR	MR	MI
No.	821	1123	1123	1123
Mean	2.671	2.671	2.692	2.682
Standard Errors	1.057	0.903	0.903	0.987
Enterococci	DD	AR	MR	MI
No.	698	1123	1123	1123
Mean	2.224	2.224	2.228	2.175
Standard Errors	0.795	0.626	0.626	0.807

The means and standard errors using traditional gap filling techniques and multiple imputation were also calculated for comparison, Table XII. The results agree with the conclusions of the limitations of traditional methods, which average replacement and median replacement methods tend to narrow down the shape of the distribution and the data deletion method has limited number of samples.

TABLE XIII

COMPARISONS OF *E. COLI* VALUES IN ORIGINAL COMPLETE DATASET AND IMPUTED ARTIFICIAL DATASET

Dataset	N	Means	Std. Dev	Q1	Q3	Minimum	Maximum
Complete dataset	573	2.72	0.98	2.13	3.43	-1.00	4.78
Data set 1	573	2.71	0.93	2.15	3.40	-1.00	4.78
Data set 2	573	2.69	0.95	2.11	3.39	-1.00	4.37

TABLE XIV

COMPARISONS OF ENTEROCOCCI VALUES IN ORIGINAL COMPLETE DATASET AND IMPUTED ARTIFICIAL DATASET

Dataset	N	Means	Std. Dev	Q1	Q3	Minimum	Maximum
Complete dataset	573	2.18	0.80	1.71	2.72	-1.00	4.46
Data set 1	573	2.16	0.75	1.68	2.72	-1.00	4.35
Data set 2	573	2.16	0.78	1.67	2.72	-1.00	4.46

3.3.2 Multiple Imputation Using Artificial Data Set

Descriptive statistics of *E. coli* and enterococci for the complete data set, and data sets 1 and 2 (different artificial missing patterns) are listed in Table XIII and Table XIV. The parameter estimates and standard deviations obtained from the two artificial data sets are listed in Table XV and Table XVI. The two data sets did not yield major different imputation results or parameter estimates. Accordingly, data set 2, the one created using the original missing pattern, was selected for further analysis of MI performance.

TABLE XV

ESTIMATED COEFFICIENT PARAMETER OF *E. COLI* AND ITS ASSOCIATED STANDARD DEVIATION

Dataset	*E. coli* Coefficient Parameter	Estimated Standard Deviation
Data set 1	0.2187	0.5631
Data set 2	0.1723	0.4919

TABLE XVI

ESTIMATED COEFFICIENT PARAMETER OF ENTEROCOCCI AND ITS ASSOCIATED STANDARD DEVIATION

Dataset	Enterococci Coefficient Parameter	Estimated Standard Deviation
Data set 1	-0.1281	0.6471
Data set 2	-0.0804	0.5938

3.3.3 Multiple Imputation in Logistic Regression Models

Descriptive statistics between the original complete data set and imputed artificial data sets for *E. coli* and enterococci are listed in Table XVII and Table XVIII respectively. The imputed dataset and the complete dataset show the same shape of distribution with relatively identical means.

Table XIX shows the modeling results using the *E. coli* complete dataset. P-value, 0.5950, indicates that *E. coli* is not a significant parameter in the model to predict GI illness status. Similar results were also observed in enterococci model (Table XX).

TABLE XVII

COMPARISONS OF *E. COLI* IN COMPLETE DATASET AND IMPUTED ARTIFICIAL DATASET

Dataset	N	Means	Std. Dev	Q1	Q3	Minimum	Maximum
Complete set	878	2.89	0.83	2.32	3.57	0.778	4.51
Imputed artificial	878	2.88	0.87	2.31	3.15	0.778	4.62

TABLE XVIII

COMPARISONS OF ENTEROCOCCI IN COMPLETE DATASET AND IMPUTED ARTIFICIAL DATASET

Dataset	N	Means	Std. Dev	Q1	Q3	Minimum	Maximum
Complete set	878	2.13	0.78	1.85	2.72	-1.00	4.46
Imputed artificial	878	2.13	0.75	1.80	2.74	-1.00	4.46

The parameter estimates of *E. coli* and enterococci obtained using artificial data set 2 are listed in Table XXI and Table XXII, along with the parameter values from the complete data set. The percent difference of parameter values, β, were calculated following Equation 3.8.

$$Bias(\%) = \frac{\beta_{ArtificialDataWithImputedValues} - \beta_{CompleteData}}{\beta_{CompleteData}} \tag{3.8}$$

where, $\beta_{ArtificialDataWithImputedValues}$ and $\beta_{CompleteData}$ are the parameter estimates of *E. coli* from the artificial and complete data sets respectively.

TABLE XIX

PARAMETER ESTIMATES OF *E. COLI* MODEL USING COMPLETE DATASET

Response	GI illness: 35 No GI illness: 843		
Parameter	Estimate	Standard Error	p-value
Intercept	-3.86	1.16	0.0008
E. coli	0.14	0.27	0.5950
Age 10 and under	-0.70	0.80	0.3861
Gender	0.16	0.35	0.6562
Age 65 and over	-12.64	619.0	0.9837
$\text{Race}_{Hispanic}$	1.04	0.84	0.2172
Race_{White}	0.18	0.73	0.8046
Race_{Other}	0.81	0.80	0.3098
Pre-exist GI	0.46	0.78	0.5537
Previously exposed to GI	-12.90	612.6	0.9832
Wet score	0.09	0.07	0.2023
Duration of last rain	-0.03	0.04	0.5336
Boat	-1.46	0.72	0.0420
Canoe	-1.25	0.73	0.0872
Kayak	-1.49	0.72	0.0388
Row	-1.20	0.78	0.1260
Water sport concern	0.13	0.07	0.0493
Previous CSO	-12.62	1234.2	0.9918

TABLE XX

PARAMETER ESTIMATES OF ENTEROCOCCI MODEL USING THE COMPLETE DATASET

Response	GI illness: 35 No GI illness: 843		
Parameter	Estimate	Standard Error	p-value
Intercept	-3.24	1.08	0.0027
Enterococci	-0.12	0.33	0.7273
Age 10 and under	-0.69	0.80	0.3877
Gender	0.13	0.36	0.7180
Age 65 and over	-12.60	628.3	0.9840
$\text{Race}_{Hispanic}$	1.07	0.84	0.2015
Race_{White}	0.17	0.73	0.8157
Race_{Other}	0.81	0.80	0.3103
Pre-exist GI	0.37	0.79	0.6410
Previously exposed to GI	-12.88	622.6	0.9835
Wet score	-0.13	0.18	0.4611
Duration of last rain	-0.03	0.04	0.5415
Boat	-1.27	0.74	0.0867
Canoe	-1.17	0.74	0.1123
Kayak	-1.41	0.73	0.0540
Row	-1.06	0.82	0.1998
Water sport concern	0.13	0.07	0.0661
Previous CSO	-12.81	1229.0	0.9917
Enterococci*Wet score	0.10	0.07	0.1703

TABLE XXI

ESTIMATED COEFFICIENT PARAMETER OF *E. COLI* AND ITS ASSOCIATED STANDARD DEVIATION USING ARTIFICIAL DATA SET AND COMPLETE DATA SET TO REGRESS FITTED OUTCOMES

Dataset	Coefficient Parameter	Estimated Standard Deviation	Percent Difference
Complete Data Set	6.98	3.04	
Artificial Data Set	-0.29	0.53	-104%

TABLE XXII

ESTIMATED COEFFICIENT PARAMETER OF ENTEROCOCCI AND ITS ASSOCIATED STANDARD DEVIATION USING ARTIFICIAL DATA SET AND COMPLETE DATA SET TO REGRESS FITTED OUTCOMES

Dataset	Coefficient Parameter	Estimated Standard Deviation	Percent Difference
Complete Data Set	5.70	3.06	
Artificial Data Set	-0.26	0.56	-104%

The results show the estimated parameters using imputed data sets are extremely different than the ones obtained from the complete data sets with -104% bias. This finding was consistent in both *E. coli* model and enterococci model.

The results of using artificial data sets and complete data sets to regress fitted logit using a multivariate linear regression model are showed in Table XXIII and Table XXIV. By avoiding setting a threshold, the bias dropped to 2% and -33% for *E. coli* and enterococci data. The bias introduced by the MI method dramatically improved in *E. coli* model but not in enterococci model. There is no standard deviation for the complete dataset because the fitted logits were

TABLE XXIII

ESTIMATED COEFFICIENT PARAMETER OF *E. COLI* AND ITS ASSOCIATED STANDARD DEVIATION USING ARTIFICIAL DATA SET AND COMPLETE DATA SET TO REGRESS CALCULATED LOGITS

Dataset	Coefficient Parameter	Estimated Standard Deviation	Percent Difference
Complete Data Set	0.14	N/A	
Artificial Data Set	0.14	0.29	2%

TABLE XXIV

ESTIMATED COEFFICIENT PARAMETER OF ENTEROCOCCI AND ITS ASSOCIATED STANDARD DEVIATION USING ARTIFICIAL DATA SET AND COMPLETE DATA SET TO REGRESS CALCULATED LOGITS

Dataset	Coefficient Parameter	Estimated Standard Deviation	Percent Difference
Complete Data Set	-0.12	N/A	
Artificial Data Set	-0.08	0.34	-33%

calculated using the complete dataset, and therefore, all the fitted points fell on the regression line. The parameter estimate is the true parameter with no standard deviation.

3.4 Conclusions

3.4.1 Summary of Findings

In the original data set, there were a total of 1,123 samples, out of which 302 (27%) and 425 (38%) QC sample parameters for *E. coli* and enterococci, respectively, were not satisfied. Multiple imputation was utilized to fill in the missing values for subsequent analysis. Testing

confirmed the MI method was successful in filling in these gaps with values which generated the same distribution as the original data set.

For the artificial data set, MI was applied to two different artificial missing patterns created under the assumption of MAR. Results indicate that the MI method can be used to fill in missing values and still produces the same distribution as the original dataset. In other words, the imputed data set has a mean and a standard deviation that are close to the original data set and this approach solves the distortion problem of the data distribution caused by the use of traditional imputation methods.

3.4.2 Implication of the Multiple Imputation

In comparison to other gap filling methods, MI is a technique to fill in missing values with the consideration of the uncertainty of missing data. It does not invent any more information than traditional gap filling methods, such as mean replacement or single imputation methods, which consider no uncertainty between imputations and view imputed values as the same as the observed values.

Case deletion is the most simple and straightforward method to deal with missing values. When the proportion of missing values is small, less than 5%, case deletion could be an acceptable approach. However, if the proportion of missing values is larger than 5%, for the purpose of data modeling, case deletion method would generate invalid inferences unless the data is MCAR.

Therefore, prior to any data analysis, the amount and pattern of missing data in microbial indicator data sets should be examined first. If the purpose of a study is to use microbial data for modeling and the proportion of the missing values is not ignorable, then the proper approach of handling missing values should be determined accordingly. Among all the gap filling methods, MI should be considered for its consistent performance in imputing missing values for data sets of indicator microbes. However, if the goal of a study is to simply examine the patterns of microbial counts due to climate change or other effects, one should only use available data instead of imputed data to observe patterns.

In the analysis of artificial data set, we observed 2% and -33% of the bias being introduced by the MI method to *E. coli* and enterococci missing data. The percentage of missing values of *E. coli* and enteroccoci, 24% and 36% respectively, can be the cause of the different results of bias. Further study should evaluate MI with different percentages of microbial missing values to demonstrate the maximum percentage of missing values for which MI can generate parameter estimates with bias under certain level.

3.4.3 Strengths

Multiple imputation method has previously not been applied to fill in missing microbial data. In this study, we have a large sample size that allows us to create an artificial data set with a MAR pattern of *E. coli* and enterococci missing values to compare the imputed values to the observed values. This approach shows that even with the missing of 24% of *E. coli*, MI can provide parameter estimates of microbial data that are close to the ones generated using the

complete data set. The standard deviations of the parameter estimates are also close to the ones from the complete data set.

3.4.4 Limitations

Since there is no standard method to test if a data is MCAR, MAR, or MNAR, it is the major limitation of MI method. Even though Collins et al. (29) concluded that in majority of the cases the violation of MAR assumption would only result in minor impact on inferences, it is essential to examine the missing mechanism to ensure that there is no evidence of MNAR. In order to investigate any violation of MAR assumption, a sensitivity analysis of results to departures can be conducted. Various realistic assumptions of variables using in imputation model can be applied to examine how sensitive the parameter estimates are. One can also apply MI to a microbial data set with known MNAR mechanism to evaluate the MI performance if MAR assumption is violated.

In addition, while major software programs provide MI method function, many of them employ regression or propensity scores methods as default, both of which require a monotone missing pattern. However, for an arbitrary missing pattern, such as the data set in this study, MCMC method should be applied instead. All the software programs provide a function of checking the missing pattern. Researchers need to have an understanding of the assumptions of the MI method in order to conduct the analysis properly.

CHAPTER 4

SOURCE IDENTIFICATION ANALYSIS USING RECEPTOR MODELING

4.1 Literature Review

4.1.1 The Advantage of Source Identification

Indicator bacteria have been used as water quality criteria since 1986. However, bacterial indicators come from various sources, such as animal waste, treated human waste, or untreated raw human waste (4; 5; 6). Not all the sources are equally dangerous to water recreators. In addition, the associations between water users health and water bodies with only non-point source (NPS) pollution, land runoff, precipitation, or atmospheric deposition, remain uncertain.

The need of identifying sources of contamination has stimulated the development of microbial source tracking, a method that identifies the original source of fecal waste. There are two types of microbial source tracking, library-dependent and library-independent. Both types face disadvantages in application. For each study area, the library-dependent method requires a great amount of resources to collect, identify, and store all the potential organisms that might be present. The library-independent method uses genome sequences (RNA and DNA) to identify genomic markers of organisms which are unique to different sources, such as human or

animal feces. However, the source markers do not exist in large quantities in the environment and QPCR detects all cellular DNA regardless of viability. The two methods were compared by Santo Domingo et al. (44) and the authors concluded that library-dependent method tends to result in larger number of false positives and false negatives then the library-independent method.

Identifying fecal pollutant sources is essential for water quality management. Two approaches, other than microbial source tracking, were used to identify sources of contamination in this study. One was utilizing receptor modeling (current chapter) to identify pollutant sources, and another approach (Chapter 5) was performing exploratory factor analysis (EFA) to indicate factors that affect bacteria levels in the system.

4.1.2 Receptor Modeling

Various receptor models have been applied in air quality management during the last three decades (45; 46; 47). These types of models identify different pollutant sources using particle size, chemical composition, and concentration patterns. The proportion of contribution of each source to the receptor is calculated using mathematical or statistical approaches. The modeling process does not require information such as pollution emission rates or environmental fate of chemical transformation. Therefore, it can be applied with limited emission information. There are two fundamental assumptions in receptor models (48),

1. The mass of pollutants at the receptor is a linear combination of the mass released from the sources;

2. The mass and pollutant composition remains constant from emission to collection, and there is no interaction between pollutants.

Receptor modeling works by comparing the profiles of pollution sources with pollution measured at locations of interest to infer the relative contribution of each source. These type of models consistently provide close predictions of sources in air quality data (49; 50; 51). The U.S. EPA has developed three main receptor models, Chemical Mass Balance (CMB) model, UNMIX models, and the Positive Matrix Factorization (PMF) model, using in air quality management (52). Among all, CMB model was the only one that has been applied to water pollution data.

The CMB model has been used for air quality monitoring (53; 54), and the performance of the model has been adequate that the U.S. EPA has suggested to use it as a regulatory monitoring technique. Many approved State Implementation Plans have also used CMB as the analytical method (55).

Only a few attempts have applied the CMB model to water pollution data, which limited the literature review that follows. Li et al. (56) used the CMB model to identify the main sources of polycyclic aromatic hydrocarbons (PAHs) measured in the sediment of Lake Calumet in Chicago. The authors used source profiles selected from literature and the model can almost explained 100% of the total PAHs (R^2=0.993). The identified main contributors were coke ovens used in steel making and traffic. This study demonstrated the applicability of the CMB model to identify various sources in an urban river system.

Saada (57) also applied the CMB model to a water related dataset, analyzing volatile organic contaminants (VOCs) in ground water, Battle Creek, Michigan. The CMB model was found to be a valuable method for source apportionment in a water setting as well.

4.1.3 Receptor Modeling in CAWS

The goal of this study was to examine the ability of the CMB receptor model to identify the relative contributions of four sources (background, rain, combined sewer overflows, water reclamation plants) to microbial density in surface water. Chemical mass balance model was utilized in this study because it has been previously applied to water data sets in providing inference of source contributions and the results were satisfactory. The model can be explained using Equation 4.1:

$$C_i = \sum_{j=1}^{J} M_{ij} * S_j + e_i \tag{4.1}$$

Where, C_i is the total concentration of a certain elemental contaminant i (ex. *E. coli* or enterococci), and S_j is the jth sources included in the model. M_{ij} is the fraction of contribution from each source j to the elemental component i. Only if i is larger than j, a unique solution for each M_{ij} can then be calculated using weighted least square (WLS) method, which calculates the smallest sum of squares of the differences between measured mass and modeled mass from Equation 4.1.

The CAWS are designated as limited contact use waters and the water quality in CAWS varies in a wide range of bacterial levels between locations. It is crucial to identify the sources of contamination in the system in order to develop appropriate solutions to maintain the water quality and protect public health. In this study, instead of using the CMB to idenfity chemical sources, we focused on microbial concentrations in water samples, using a number balance approach.

4.2 Methods

According to the users manual (58), the following five steps should be implemented in CMB modeling:

1. Identify the types of source contributing to the system;
2. determine the proper species to be included in the model calculations;
3. use observed data or information in the literature to develop source profiles of the fraction of each species presenting in each source type;
4. evaluate the uncertainty present in both concentration data and source profiles;
5. use source profiles, concentration data, and uncertainty information to calculate the chemical mass balance equations.

The five steps were followed in this study and then the model results were evaluated by comparing the modeled concentrations to the measured concentration. In addition, the predicted

sources were used to fit a logistic regression model to predict health outcomes and the results were compared to the ones obtained using water quality parameter as predictors.

The EPA-CMB8.2 software was used in this analysis. The software can be downloaded from the U.S. EPA's Support Center For Regulatory Air Models (SCRAM) website (http://www.epa.gov/scram001). The Users Manual (58) provided a detailed explanation of the model. The minimum requirements for running EPA-CMB8.2 software include:

- IBM® PC compatible desktop, portable, or laptop computer with 386 processor and 16MB RAM.
- Hard disk drive with storage of 4MB.
- Windows® 9x or higher operating system.

The recommended hardware configuration is:

- IBM® compatible Intel Pentium® with 64MB RAM and 100MB storage.
- Super VGA video graphics adapter and monitor.
- Graphics capable printer
- Windows® XP or NT 4.0 operating system.

In this study, four contributing sources were first identified, water reclamation plant, river background, combined sewer overflow (CSO), and rainfall. The reclamation plant was chosen because it discharges high amount of indicator microbes. River background was chosen to explain environmental and human contributions to water quality through routs other than

the WRP, CSO, or precipitation, such as the effluent from the lake or sources from upriver. Combined sewer overflow events were chosen because they are a source of raw sewage and bacteria following storm events. Edge et al. (59) demonstrated that in urban waters, the effluent from CSOs is particularly important since the levels of indicator microbes are much higher than in wet-weather flows or stormwater. Consequently, CSO events need to be addressed in monitoring recreational water quality. The last source, rainfall, was selected because studies have shown a significant association between rainfall and indicator bacteria concentrations and pathogen detection (60; 61; 62).

The CMB model has not been used to balance the number of microbes in water, and there is a lack of information in the literature that can be used as source profiles. Instead, the four indicators (*E. coli*, enterococci, male-specific/somatic coliphages) and two pathogens (*Giardia, Cryptosporidium*) collected in the CHEERS were used to develop source profiles.

The CAWS North Branch System and CAWS Cal-Sag Channel were analyzed separately. The upstream sampling sites in the North Branch System and Cal-Sag Channel were Bridge Street (BR) and Baubien Woods (BA). The downstream sampling sites right after the North Side WRP and Calumet WRP were Skokie Rowing Center (SK) and Riverdale Marina (RM) respectively.

Criteria used for defining each source profile are listed in Table XXV. Samples collected at two upstream locations, BR and BA, and two downstream locations, SK and RM, were used for developing profiles. The number of microbes going through each sampling location per second

TABLE XXV

CRITERIA USED FOR SOURCE PROFILE DEVELOPMENT

Source	Background	Rain	CSO	Plant
Criteria	Last CSO $>$ 144 hrs Last rain $>$ 144 hrs	Last CSO $>$ 144 hrs Last rain $\leq$ 24 hrs Accumulation $>$ 0.5 in.	Last CSO $\leq$ 48hrs	Last CSO $>$ 144 hrs Last rain $>$ 144 hrs

were calculated by multiplying the concentration (counts/100mL) by the river flow (100mL/sec). Mass measured at two upstream sites, BR and BA, were used to develop background, rain, and CSO source profiles. To do so, samples that satisfied the listed criteria were included and the average mass of each species was calculated. In order to develop the plant profile, samples collected upstream and downstream were matched by sampling date and hour. The mass for each species was then calculated by subtracting the counts at upstream location from the downstream location. It was considered as the plant contribution, and the samples that satisfied the criteria were used to calculate the average mass of each species. Each source profile was then defined by the fraction of every species in the specific source.

The number of samples for each microbe used in developing source profiles are showed in Table XXVI. In order to create clean source profiles, only samples collected under extreme weather conditions were used. This resulted in part of the source profiles being determined by a small number of samples.

TABLE XXVI

NUMBER OF SAMPLES PER MICROBE USED IN SOURCE PROFILE DEVELOPMENT

North Branch System	*Giardia*	*Crypto*	*E. coli*	somcoli	malcoli	enterococci
Background	5	5	5	5	5	5
Rainfall	4	4	10	10	10	10
CSO	30	30	46	41	41	46
WRP	4	4	5	5	5	5
Cal-Sag Channel	*Giardia*	*Crypto*	*E. coli*	somcoli	malcoli	enterococci
Background	6	6	10	10	10	10
Rainfall	4	4	5	4	4	5
CSO	2	2	3	3	3	3
WRP	5	5	9	9	9	9

Once source profiles were developed, the uncertainty for each microbe was determined by the measured concentration in the system and the detection limit of the analysis method. The difference between detection limits for each microbe was within a certain range (less than ten fold), but the measured concentration varied significantly. Therefore, uncertainties of 10% and 20% were assigned to indicators (*E. coli*, enterococci, male-specific coliphages, and somatic coliphages) and pathogens (*Giardia* and *Cryptosporidiums*) respectively. In general, the CMB model requires a species that represents the composition of the majority of the chemicals present such as particulate matters (PM). However, in our study, a similar species was not present. Therefore, a species named "total" was created with the sum of the concentrations of all six species. A 20% uncertainty was assigned to the total species.

Once the source profiles and uncertainties were determined, EPA CMB8.2 was used for source proportion analyses. Since the North Branch System and the Cal-Sag Channel had their own unique source profiles, the two systems were analyzed separately. Concentration data of all six species from different locations and various weather conditions were used as inputs to the model. For upstream locations, only background, rain, and CSO sources were included in the analyses and for downstream sites all four sources were applied. Modeling results were then compared to the rain and CSO patterns to examine if the CMB model can accurately identify the rain and CSO sources.

The agreement between measured microbial concentrations and the predicted concentrations were plotted against each other, using a perfect fit line ($y = x$) as an agreement standard. Bland-Altman plot was also applied to examine the association between the agreement and the concentrations of microbes.

The data with estimated source contributions was then matched with the recreators GI illness status. The *E. coli* logistic regression model used in chapter 3 was applied using sources as independent variables instead of the observed parameters. The results were examined for any advantages of using pollution sources as predictors of recreators GI illness.

4.3 Results

The EPA CMB8.2 model was fitted using source profiles developed individually for the two systems, North Branch System and Cal-Sag Channel. The percent mass explained by the

model using pollutant sources were calculated for the examination of model performance. The association between pollutant sources and GI illness was evaluated.

4.3.1 Source Profiles

Source profiles of the six microbes developed for the North Branch System along with the percentage of each microbe in each source are showed in Figure 7 and Table XXVII. Source profiles for the Cal-Sag Channel along with the percentage of each microbe in each source are showed in Figure 8 and Table XXVIII. The profile patterns from different sources are more distinguishable in the Cal-Sag Channel then the North Branch System. In addition, Cal-Sag Channel and the North Branch System share the same plant profile.

In the North Branch System, over 90% of the plant source is contributed by *E. coli* (63.7%) and somatic coliphages (27.4%). *E. coli*, 94.7%, is the dominant species in the rain source. Over 90% of CSO source is attributed by *E. coli* (52.1%), enterococci (22.3%), and somatic coliphages (22.1%). *E. coli* (54.1%) and enterococci (36.8%) attribute over 90% of background source.

In the Cal-Sag Channel, over 90% of the plant source is contributed by *E. coli* (57.6%) and somatic coliphages (36.6%). Over 90% of the rain source is attributed by *E. coli* (70.2%) and enterococci (25.3%). CSO source is dominated by somatic coliphages (49.9%), male-specific coliphages (29.3%), and *E. coli* (19.1%). Over 90% of the background source is attributed by *E. coli* (71.8%), enterococi (15.25%), and somatic coliphages (10.2%).

TABLE XXVII

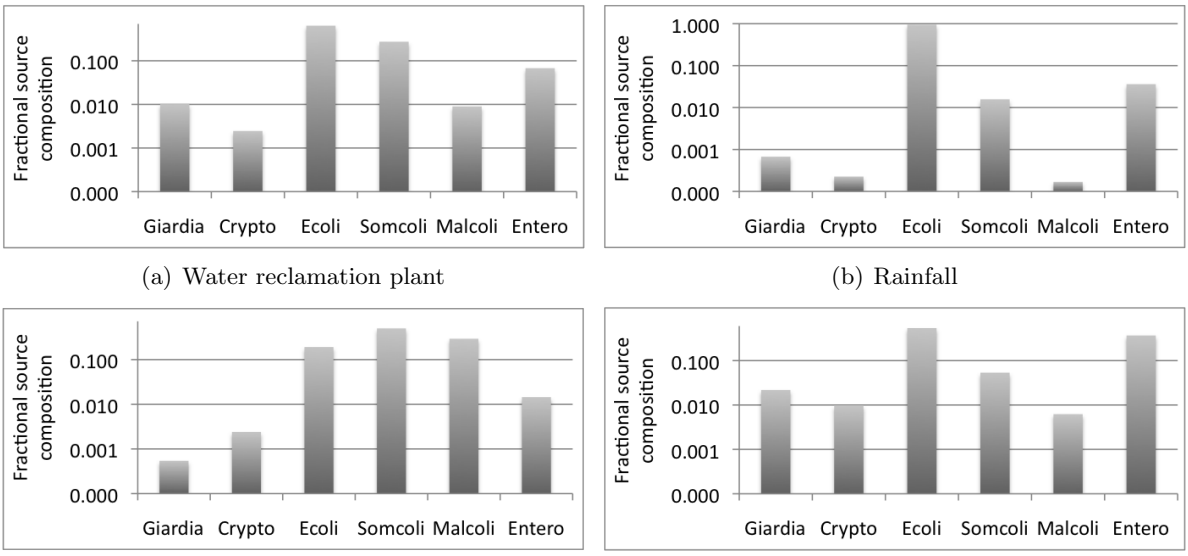
PERCENTAGE OF MICROBES IN PLANT, RAIN, CSO, AND BACKGROUND SOURCE PROFILES IN THE NORTH BRANCH SYSTEM. VALUES IN THE PARENTHESES INDICATE STANDARD DEVIATIONS.

Source	*Giardia*	*Crypto*	*E. coli*	somcoli	malcoli	enterococci
Plant	1.02 (0.48)	0.25 (0.20)	63.73 (37.42)	27.35 (20.86)	0.90 (0.56)	6.76 (8.30)
Rain	0.07 (0.02)	0.02 (0.04)	94.70 (235.31)	1.58 (3.89)	0.02 (0.01)	3.62 (4.40)
CSO	0.16 (0.32)	0.29 (1.02)	52.08 (194.1)	22.08 (54.01)	3.06 (9.47)	22.34 (71.58)
BKGD	2.18 (2.61)	0.97 (1.53)	54.06 (62.21)	5.35 (6.74)	0.62 (0.86)	36.82 (43.06)

TABLE XXVIII

PERCENTAGE OF MICROBES IN PLANT, RAIN, CSO, AND BACKGROUND SOURCE PROFILES IN CAL-SAG CHANNEL

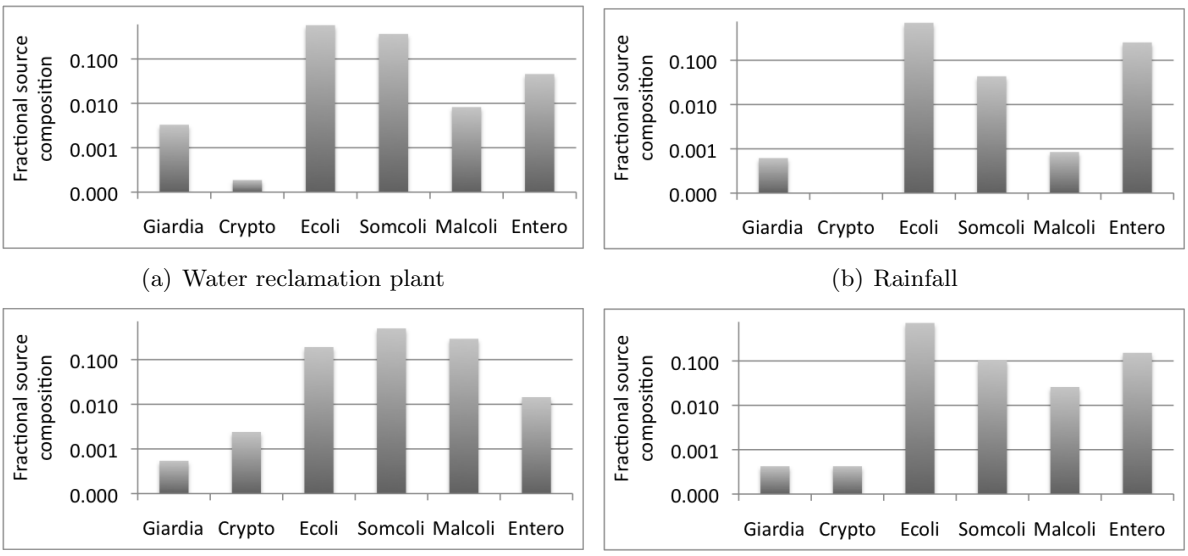
Source	*Giardia*	*Crypto*	*E. coli*	somcoli	malcoli	enteroccoci
Plant	0.33 (0.34)	0.02 (0.03)	57.63 (41.93)	36.61 (25.01)	0.83 (1.01)	4.58 (5.65)
Rain	0.06 (0.07)	0.00 (0.00)	70.17 (94.90)	4.35 (5.62)	0.09 (0.11)	25.33 (30.39)
CSO	0.05 (0.07)	0.24 (0.08)	19.11 (25.93)	49.90 (15.82)	29.26 (40.28)	1.45 (1.92)
BKGD	0.04 (0.05)	0.04 (0.05)	71.84 (84.03)	10.24 (16.43)	2.59 (7.27)	15.25 (18.01)



(a) Water reclamation plant (b) Rainfall

(c) CSO (d) Background

Figure 7. Source profiles of the North Branch System



(a) Water reclamation plant (b) Rainfall

(c) CSO (d) Background

Figure 8. Source profiles of the Cal-Sag Channel

4.3.2 North Branch System

Table XXIX shows CMB model setting for each location in the North Branch System. The overall results of the analyses are showed at the bottom of Table XXIX, note that large sample sizes at all locations except River Park. R-squared values between the predicted and measured total concentrations (sum of six microbes) were all above 0.60. At Clark Park, Lincoln Avenue, and North Avenue, the model presented stronger fits, with r-squared values of 0.87, 0.84, and 0.80 respectively. Percent mass (%Mass) indicates the percentage of observed mass being explained by the model and the results show the CMB model explains the downstream locations more precisely than the upstream location except River Park, which might be due to an insufficient sample size.

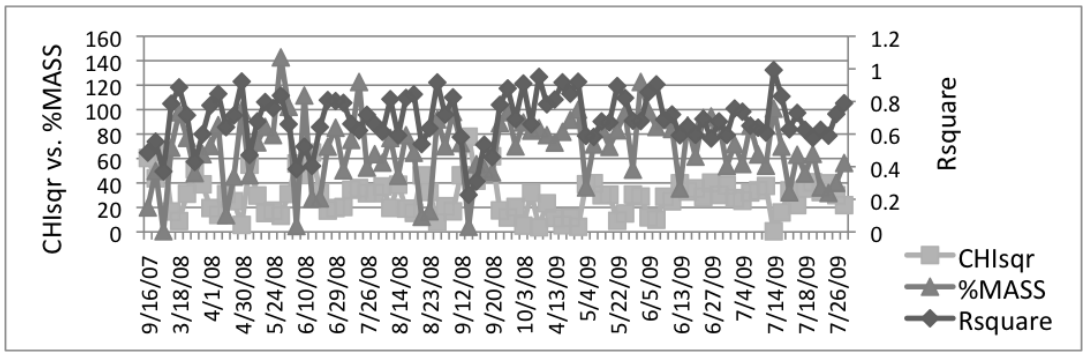
At each location, the three measurements of model performance (r-squared, chi-squared, and %Mass) are plotted against each sample, in Figure 9 and Figure 10. In general, the model performance is considered consistent and can closely predict concentrations. However, the three indicators of model performance varies widely across samples at upstream location, Bridge Street. It can be the cause of fewer sources in the model creating a greater proportion of unexplained information. Downstream locations, on the other hand, have relatively consistent results.

Model results were further examined using Bland-Altman analysis. Firstly, the paired results of measured and calculated concentrations for each sample were plotted with the y=x line, to visually present the agreement between the pairs, Figure 11. The closer the data points are to

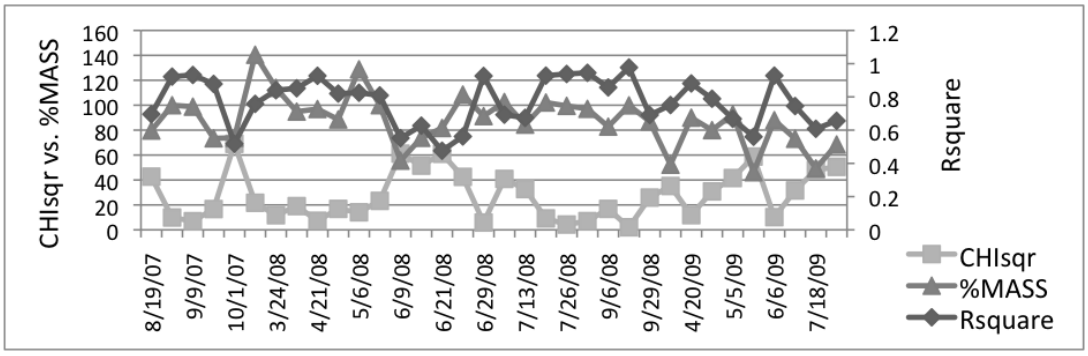
TABLE XXIX

SUMMARY OF CMB MODEL SETTINGS AND RESULTS IN THE NORTH BRANCH SYSTEM. %MASS REPRESENTS THE PERCENTAGE OF TOTAL MEASURED CONCENTRATIONS BEING EXPLAINED BY THE MODEL.

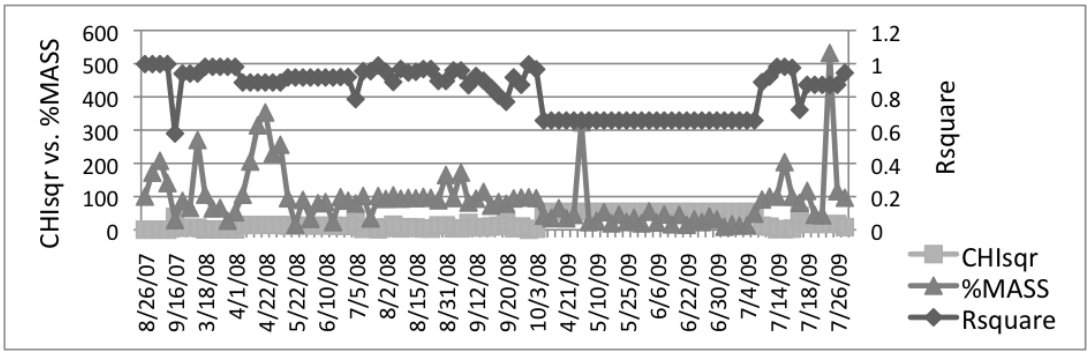
OPERATION #	1	2	3	4	5	6
Location	BR	CP	LA	NAM	RP	SK
SPECIES INCLUDED						
E. coli	x	x	x	x	x	x
Entercocci	x	x	x	x	x	x
Male-specific coliphages	x	x	x	x	x	x
Somatic coliphages	x	x	x	x	x	x
Giardia	x	x	x	x	x	x
Cryptosporidium	x	x	x	x	x	x
SOURCES INCLUDED						
Plant		x	x	x	x	x
Rain	x	x	x	x	x	x
CSO	x	x	x	x	x	x
Background	x	x	x	x	x	x
RESULTS						
n	90	26	94	30	8	34
R square	0.69	0.87	0.84	0.80	0.66	0.77
Chi square	29.07	19.11	23.68	24.36	45.52	27.60
%Mass	67.88	111.87	92.89	83.17	68.47	88.11



(a) Bridge Street

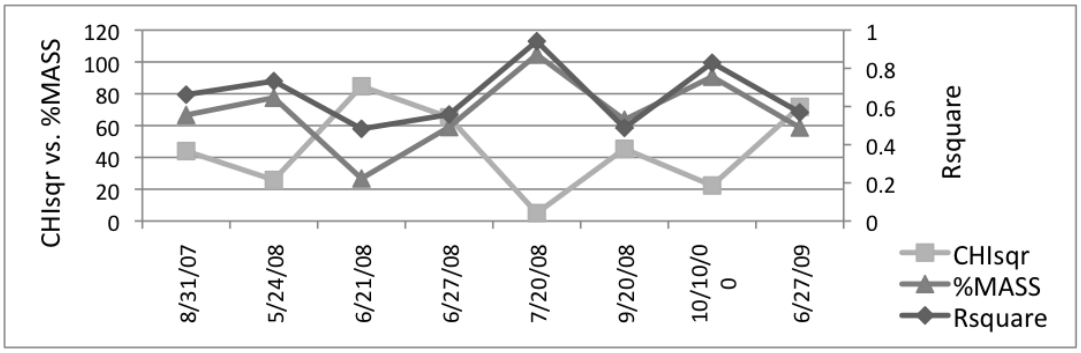


(b) Skokie Rowing Center

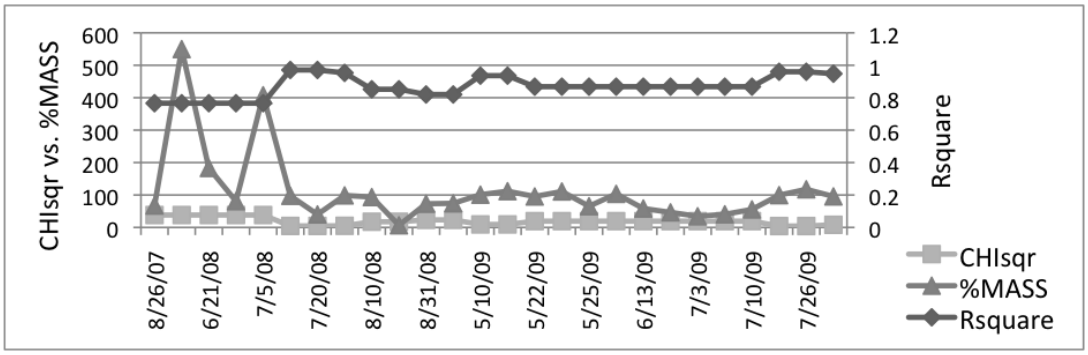


(c) Lincoln Avenue

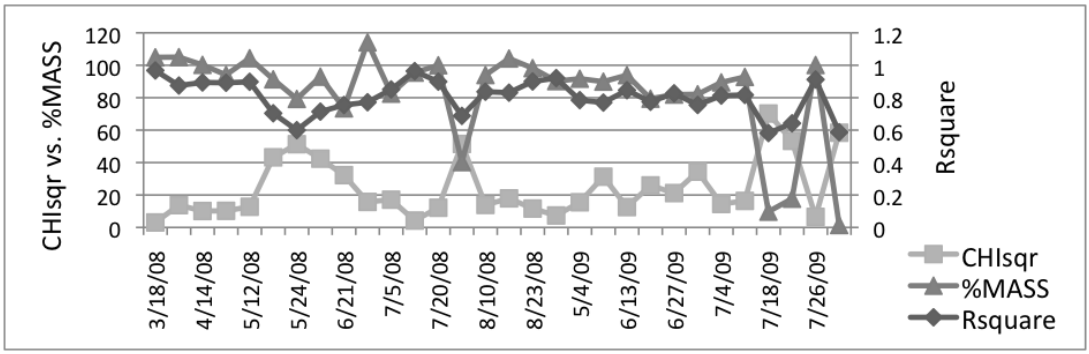
Figure 9. CMB model performance by sample in the North Branch System



(a) River Park



(b) Clark Park



(c) North Avenue

Figure 10. CMB model performance by sample in the North Branch System (continued)

the line, the higher agreement between the measured and the calculated concentration. The plot shows that approximately half of the points fall roughly along the straight line, however a portion of the data points are under-estimated. Secondly, the difference between the measured and calculated concentrations was plotted against their average in order to identify trends in agreement related to microbial concentrations, Figure 12. The solid line represents the average of the difference and the dashed lines are the upper and lower bounds for the two standard deviations. The plot shows that more dispersions occurred at higher concentration levels than at lower levels. The spearman correlation coefficient between the measured and the calculated concentrations is 0.8852 with a p-value less than 0.0001.

The total measured concentrations at each location are listed in Table XXX along with the calculated contributions from all four sources. Majority of the microbial counts are attributed from the plant and rain sources. Background source contributes less than 2% in each location.

Table XXXI shows results of sampling days with extremely large magnitude of CSOs (values above the 90 percentile), indicating as CSO_{HiDays}, and low magnitude of CSOs (values below the 10 percentile), $\mathrm{CSO}_{LowDays}$, separately. Model predicted CSO contributions were calculated for these two groups, along with the percentage of total measured microbes explained by CSO source. This approach helps us to examine if CMB accurately predicts CSO events. Same analysis was applied to sampling days with extreme high (Rain_{HiDays}) or low ($\mathrm{Rain}_{LowDays}$) amount of precipitations (Table XXXII). Location RP only has one data point for each extreme condition, and therefore, is not listed.

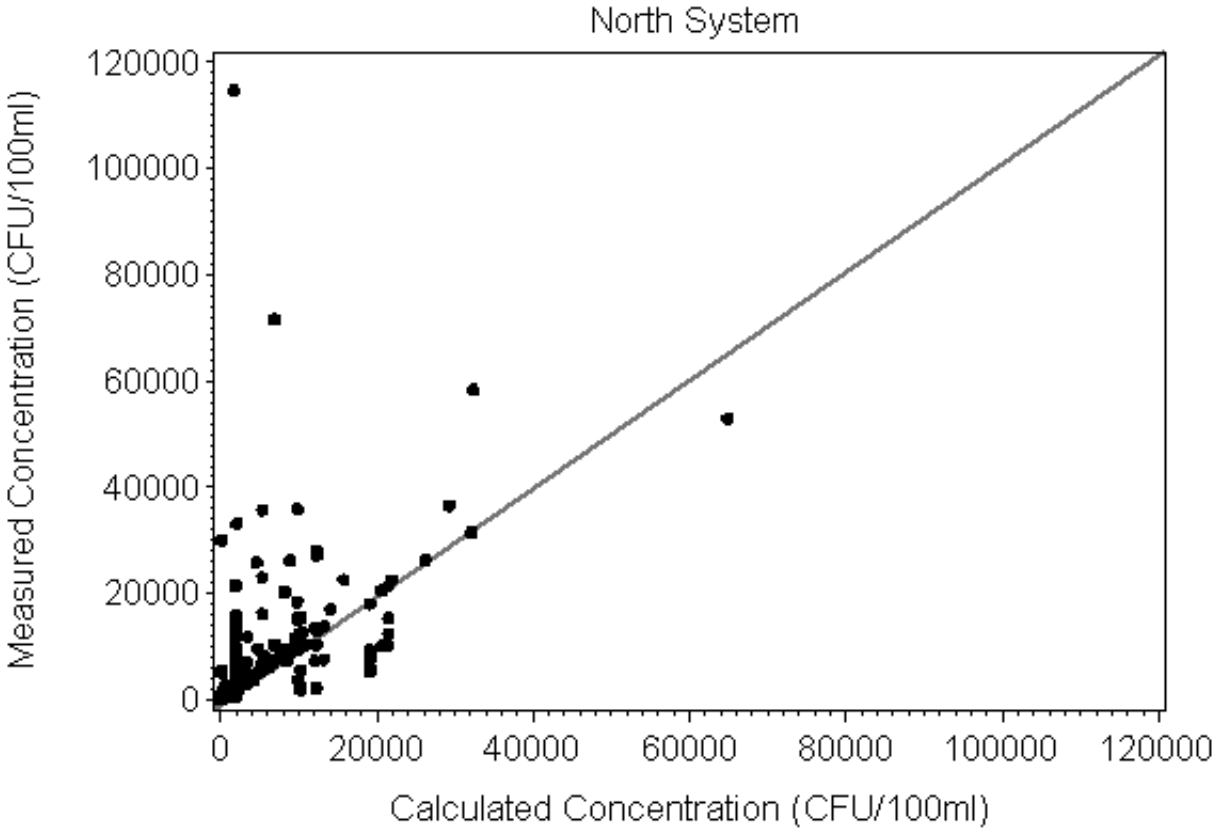


Figure 11. Scatter plot of measured and calculated concentration using CMB model in the North Branch System

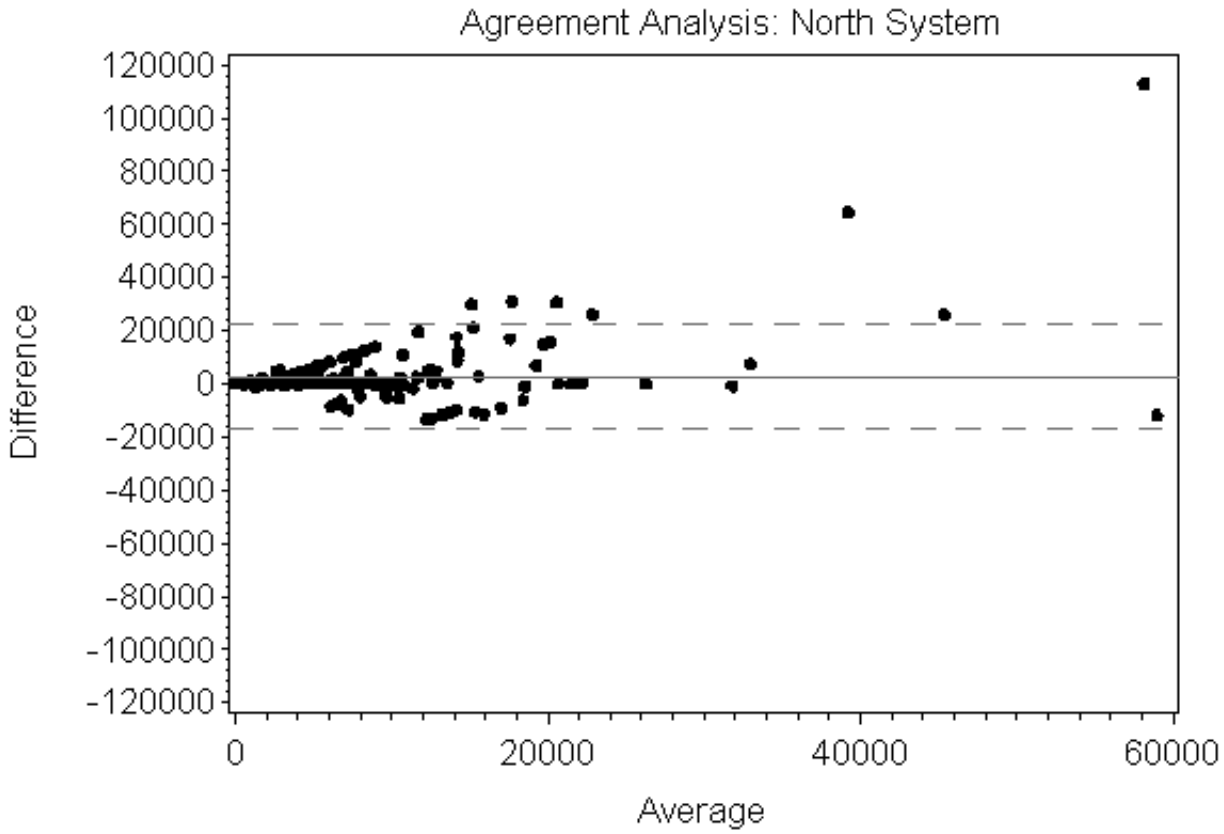


Figure 12. Bland-Altman plot of measured and calculated concentration using CMB model in the North Branch System

TABLE XXX

THE AVERAGE OF TOTAL MEASURED MICROBIAL CONCENTRATIONS (COUNTS/100ML) AND CALCULATED CONTRIBUTIONS FROM PLANT, RAIN, CSO, AND BACKGROUND SOURCES BY LOCATION IN NORTH BRANCH SYSTEM. STANDARD DEVIATIONS ARE SHOWED IN THE PARENTHESES.

	Total Conc.	Plant	Percentage	Rain	Percentage
BR	1960.1 (6817.7)	NA	NA	897.3 (6295.1)	45.8%
CP	7121.5 (6857.5)	1723.1 (1707.5)	24.2%	936.9 (1424.6)	13.2%
LA	10276.7 (6901.6)	4082.5 (3989.8)	39.7%	1263.2 (2389.0)	12.3%
NAM	10708.5 (23990.4)	2008.4 (2147.6)	18.8%	1517.1 (3361.5)	14.2%
RP	4125.3 (3982.0)	1603.1 (2345.5)	38.9%	772.6 (1601.4)	18.8%
SK	5823.6 (12585.0)	1469.3 (3014.3)	25.2%	2784.4 (7515.3)	47.8%
	Total Conc.	CSO	Percentage	Background	Percentage
BR	1960.1 (6817.7)	425.0 (2375.4)	21.7%	1.7 (278.5)	0.1%
CP	7121.5 (6857.5)	324.1 (808.0)	4.6%	20.5 (201.3)	0.3%
LA	10276.7 (6901.6)	919.7 (2097.8)	8.9%	133.0 (474.9)	1.3%
NAM	10708.5 (23990.4)	1239.6 (1756.6)	11.6%	148.8 (457.0)	1.4%
RP	4125.3 (3982.0)	2540.4 (3766.6)	61.6%	32.4 (87.6)	0.8%
SK	5823.6 (12585.0)	1299.2 (2269.6)	22.3%	68.1 (153.3)	1.2%

The results (Table XXXI and Table XXXII) show that the model can capture the real CSO events, except in SK. The model predicts higher microbial concentration on CSO_{HiDays} at location SK, but the percentage of total measured concentrations attributed by CSO source is lower in comparison to $CSO_{LowDays}$ data. The large standard deviations indicate that the performance of the model is inconsistent. The model can also capture precipitation impacts in terms of predicted concentrations, except at Lincoln Avenue (LA) where the model predicted higher contribution of rain source on low precipitation days than high precipitation days. The large standard deviations of the rain source contributions also show the inconsistency of the model performance. The percentage of total measured concentrations being explained by the rain source is not as stable as the CSO source performance given that at BR and SK, it has lower percent on high precipitation days then the low precipitation days. The inconsistent results can indicate that the model does not provide a good prediction of the rain impact. It could also be the reason that in the system rainfalls bring up microbial concentrations from all other sources as well, and therefore, the percentage does not increase as much as expected.

TABLE XXXI

MEAN PREDICTED CSO CONTRIBUTIONS (COUNTS/100ML), STANDARD DEVIATIONS (INDICATED IN THE PARENTHESES), AND PERCENTAGE OF TOTAL MEASURED MICROBES ON SAMPLING DAYS WITH EXTREME CSO EVENTS IN NORTH BRANCH SYSTEM. CSO HIDAYS ARE DAYS WITH MAGNITUDE OF CSOS ABOVE THE 90 PERCENTILE. CSO LOWDAYS ARE DAYS WITH MAGNITUDE OF CSOS BELOW THE 10 PERCENTILE.

Location	CSO_{HiDays}	Percentage	$CSO_{LowDays}$	Percentage
Overall	845.6 (1341.8)	30.11%	361.8 (1321.3)	13.40%
BR	507.2 (930.3)	66.97%	35.4 (55.7)	29.16%
CP	751.0 (1413.3)	8.64%	15.3 (14.6)	0.63%
LA	1246.6 (1887.6)	51.36%	701.1 (2063.3)	11.47%
NAM	910.0 (895.8)	83.62%	532.3 (416.0)	37.00%
SK	94.5 (185.7)	0.87%	61.5 (26.5)	16.87%

TABLE XXXII

MEAN PREDICTED RAIN CONTRIBUTIONS (COUNTS/100ML), STANDARD DEVIATIONS (INDICATED IN THE PARENTHESES), AND PERCENTAGE OF TOTAL MEASURED MICROBES ON SAMPLING DAYS WITH EXTREME PRECIPITATIONS IN NORTH BRANCH SYSTEM. RAIN HIDAYS ARE DAYS WITH MAGNITUDE OF RAIN ABOVE THE 90 PERCENTILE. RAIN LOWDAYS ARE DAYS WITH MAGNITUDE OF RAIN BELOW THE 10 PERCENTILE.

Location	$Rain_{HiDays}$	Percentage	$Rain_{LowDays}$	Percentage
Overall	2310.6 (6006.9)	21.52%	316.2 (420.1)	20.75%
BR	619.9 (1072.8)	29.84%	115.0 (169.1)	38.20%
CP	1255.0 (1475.6)	18.11%	448.8 (429.0)	13.61%
LA	465.1 (713.4)	8.53%	559.6 (431.9)	5.17%
NAM	7774.3 (8086.8)	37.72%	0 (0)	0.00%
SK	8000.0 (14259.4)	41.08%	334.7 (632.6)	66.13%

A close look at source proportions for each sample collected at each location coupled with the patterns of magnitude of last rain and CSO events are showed in Figure 13, Figure 14, Figure 15, Figure 16, Figure 17, and Figure 18 for BR, SK, LA, RP, CP, and NAM respectively. At downstream locations, matching patterns of rain and CSO events with the predicted sources are easily identified. By contrast, the rain and CSO patterns do not particularly follow the predicted sources at upstream site.

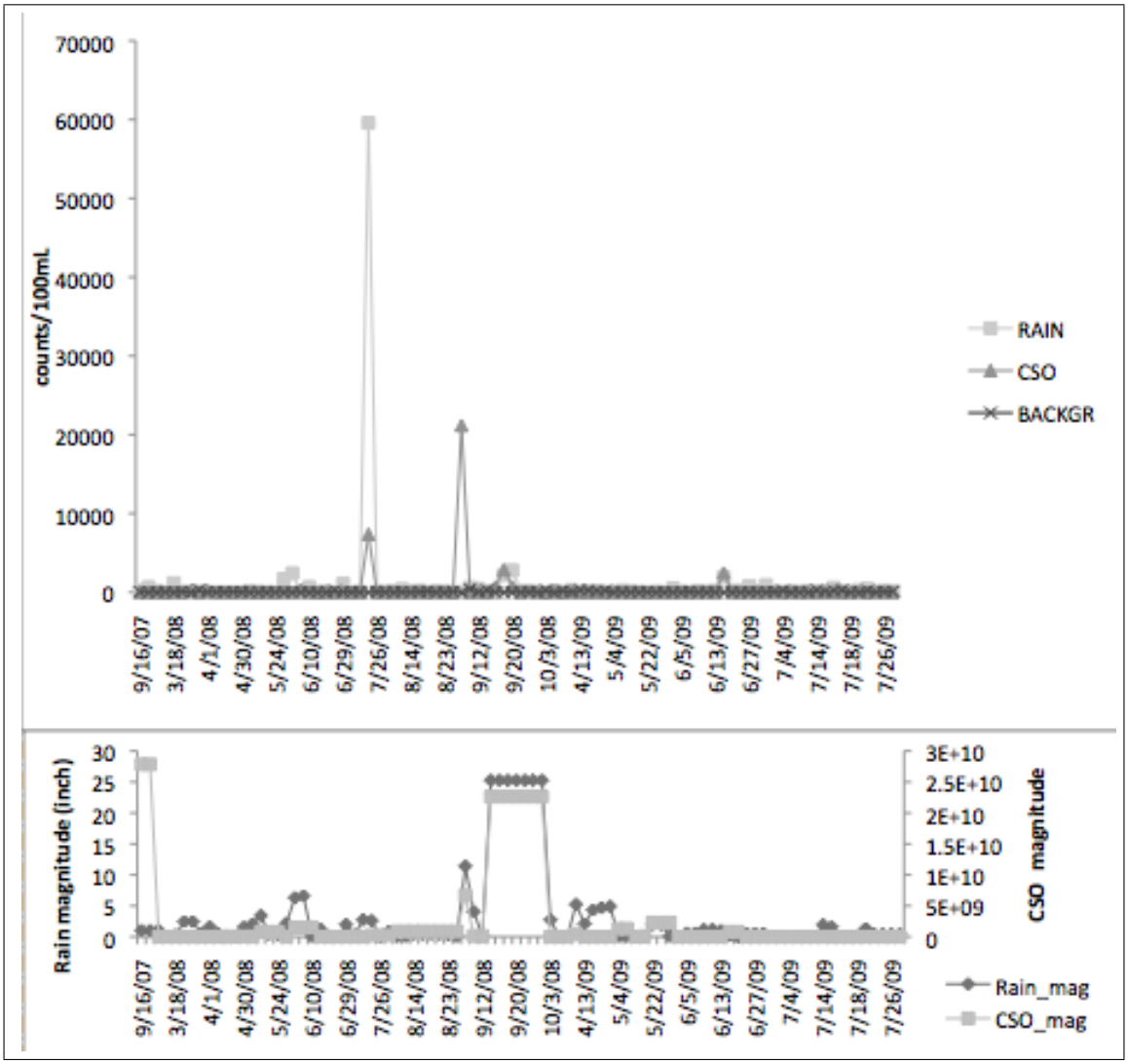


Figure 13. Bridge Street source proportion

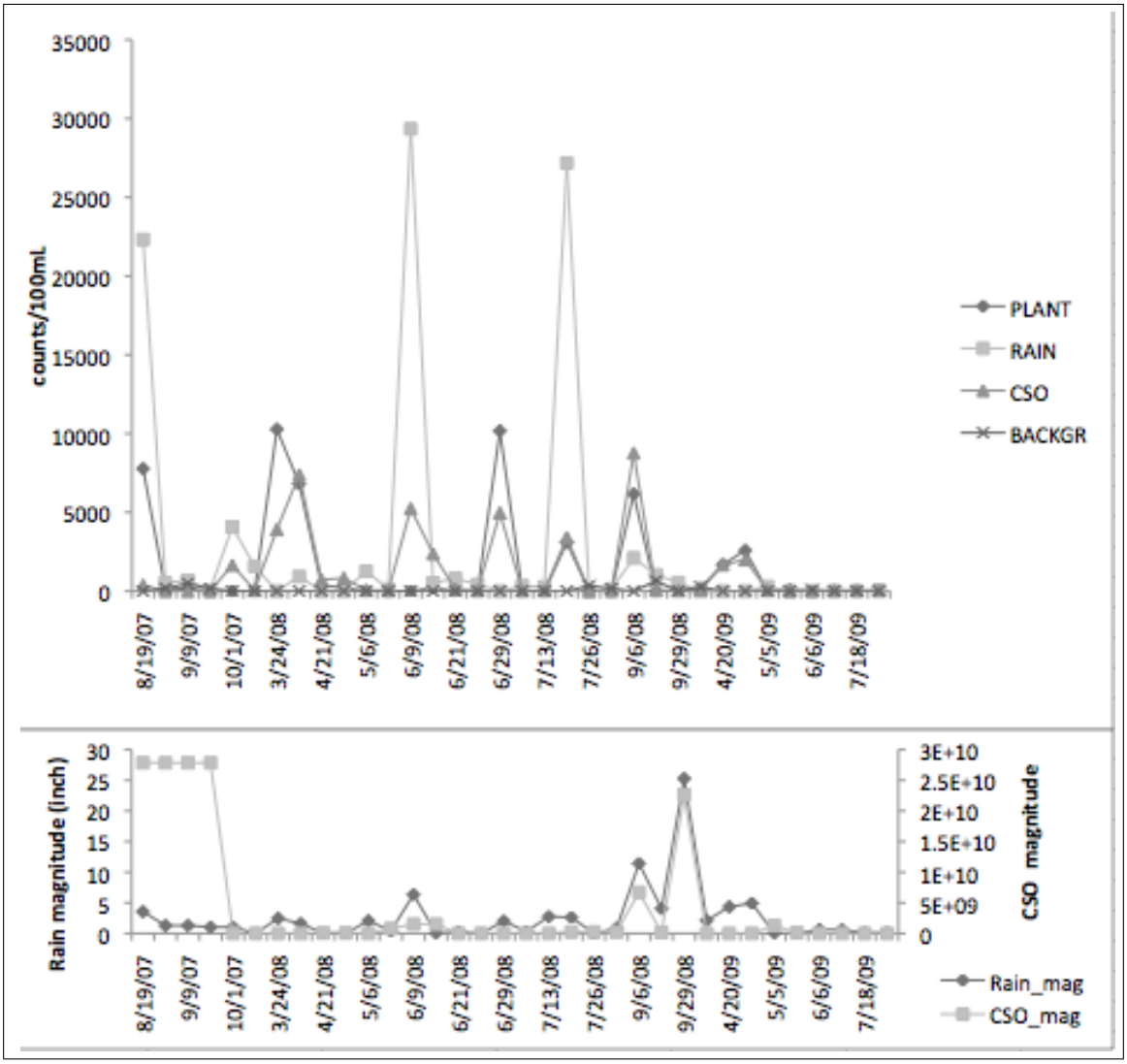


Figure 14. Skokie Rowing Center source proportion

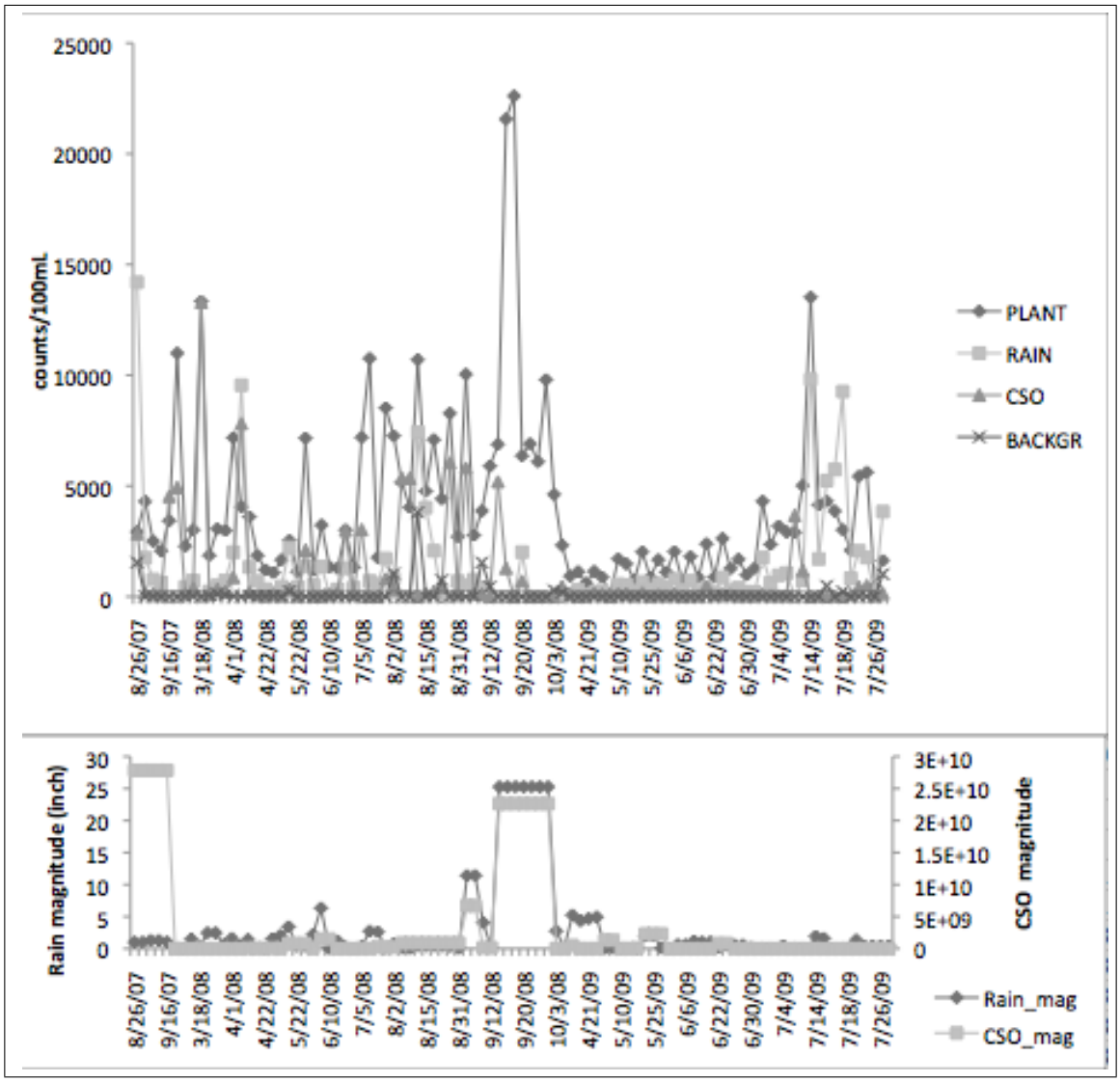


Figure 15. Lincoln Avenue source proportion

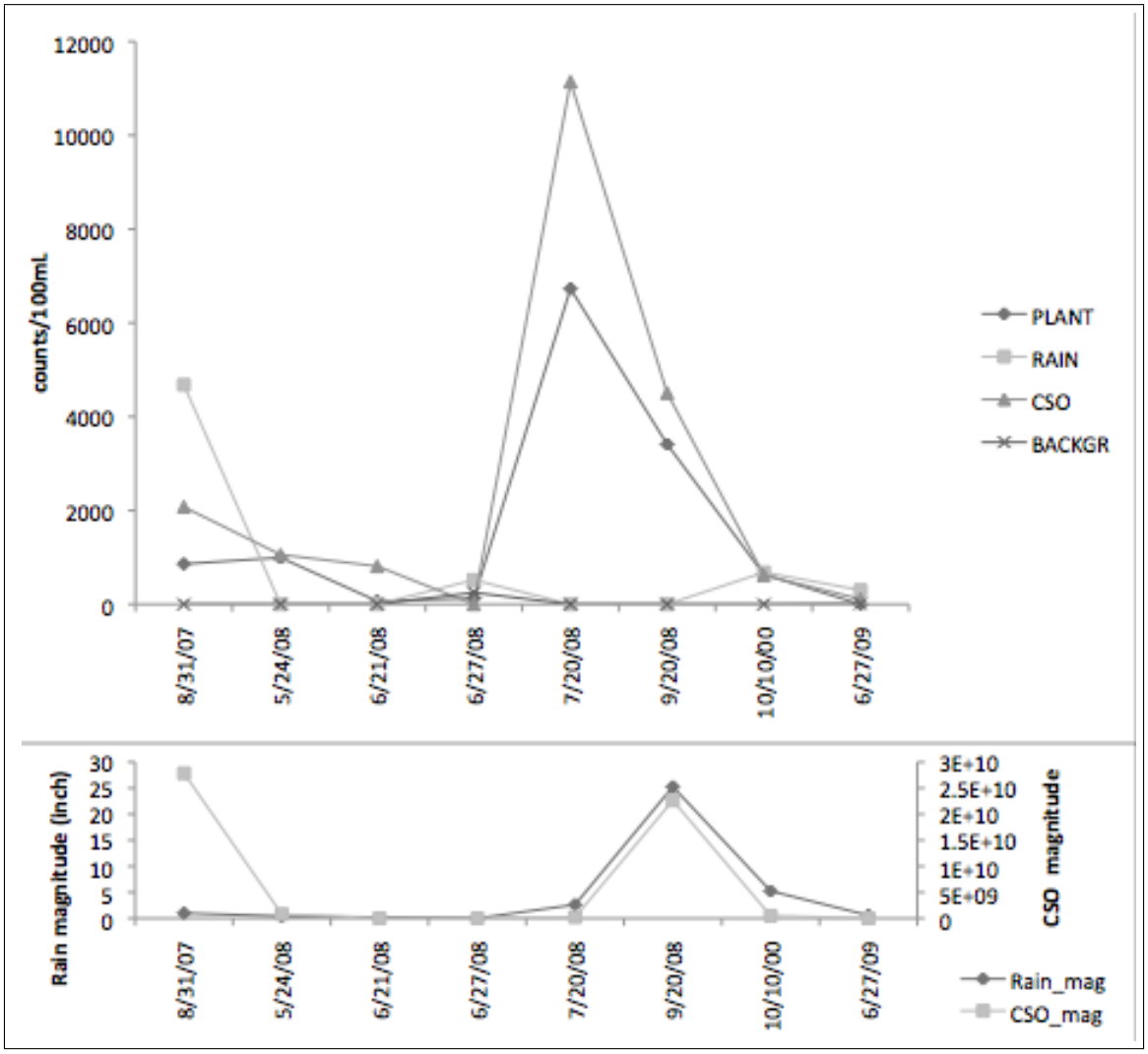


Figure 16. River Park source proportion

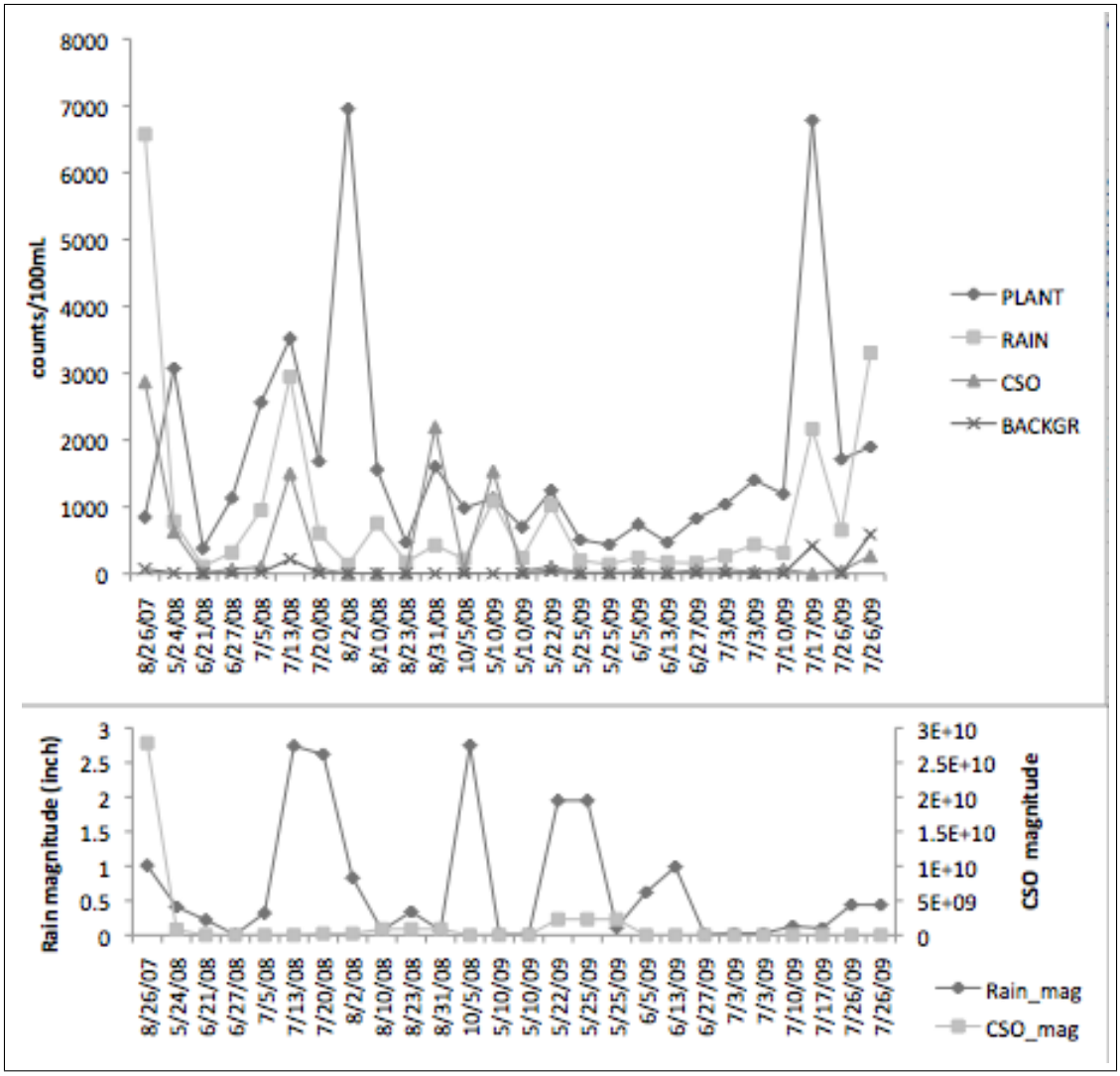


Figure 17. Clark Park source proportion

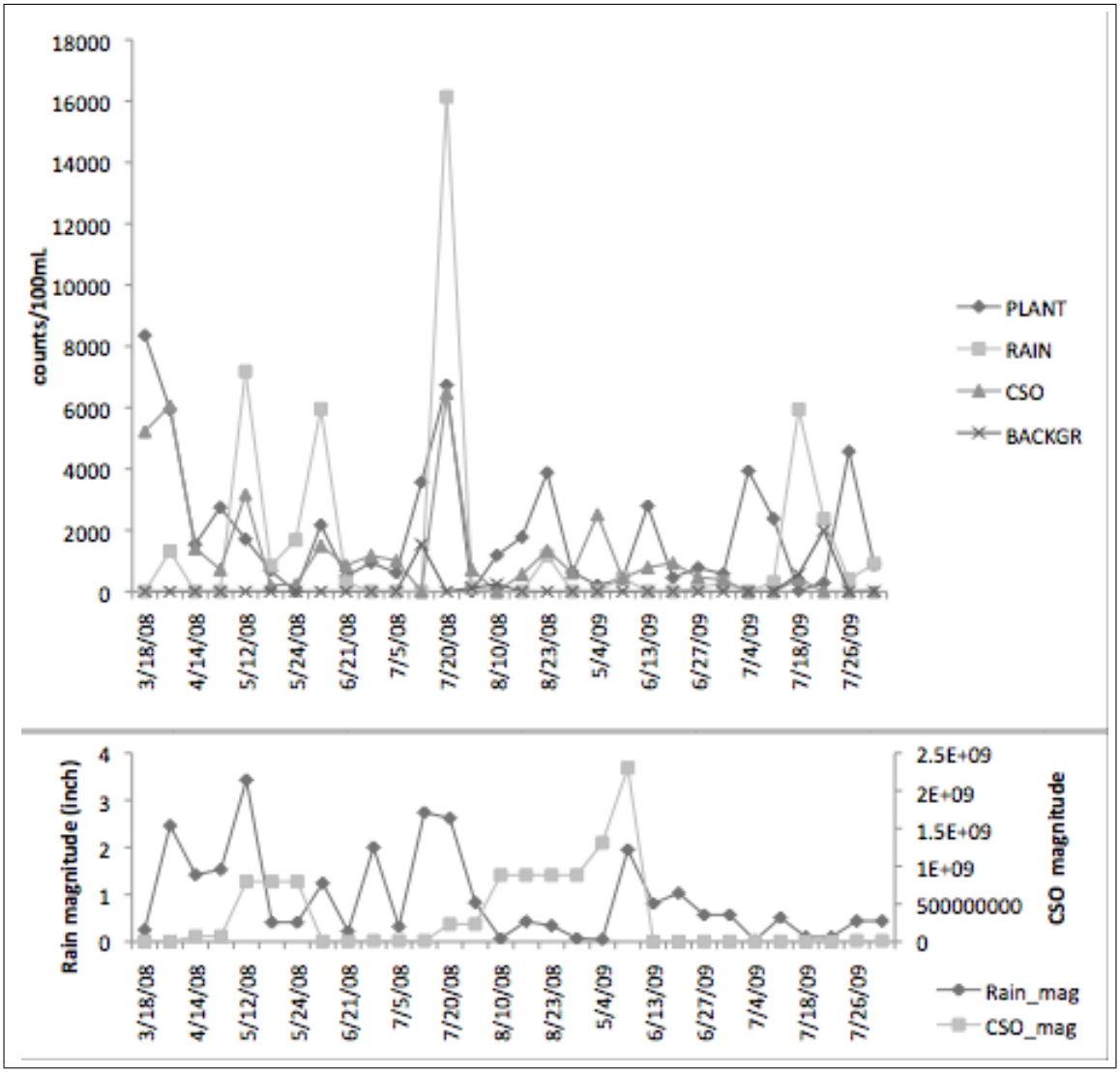


Figure 18. North Avenue source proportion

4.3.3 Cal-Sag Channel

Table XXXIII shows four sets of operations conducted using the CMB model in Cal-Sag Channel. Sample sizes of the four locations are relatively equal, and r-squared values between measured and predicted total concentrations are all above 0.7, except upstream location Beaubien Woods which has a r-squared value equal to 0.64. Percent mass indicates that the model explained the concentrations well at all three downstream locations. Conversely, the percent mass at upstream location explained by the model is comparatively low which was consistent with the CMB model results from the North Branch System. The measurements of model performance, r-squared, chi-squared, and %mass, per sample at each location are showed in Figure 19.

Comparisons of the two systems, Figure 19 for the Cal-Sag Channel versus Figure 9 and Figure 10 for the North Branch System, the model performance measurements are more inconsistent in the Cal-Sag Channel.

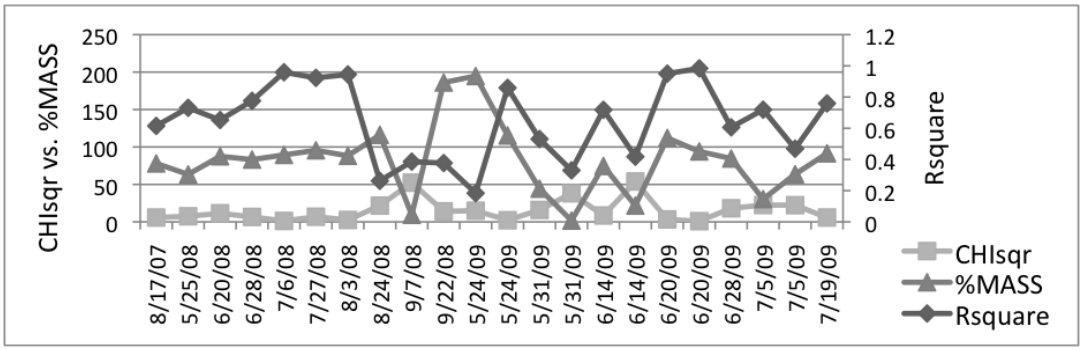
Figure 20 shows the majority of the points fall closely along the straight line, and only a small proportion of the data points are under-estimated. Figure 21 shows that all points fall within the two standard deviations, but more dispersions appeared at higher concentration levels. The spearman correlation coefficient (between the measured and the calculated concentrations) is 0.8932 with a p-value less than 0.0001.

The total measured concentrations at each location are listed in Table XXXIV along with the calculated contributions from all four sources. in contrast to the North Branch System,

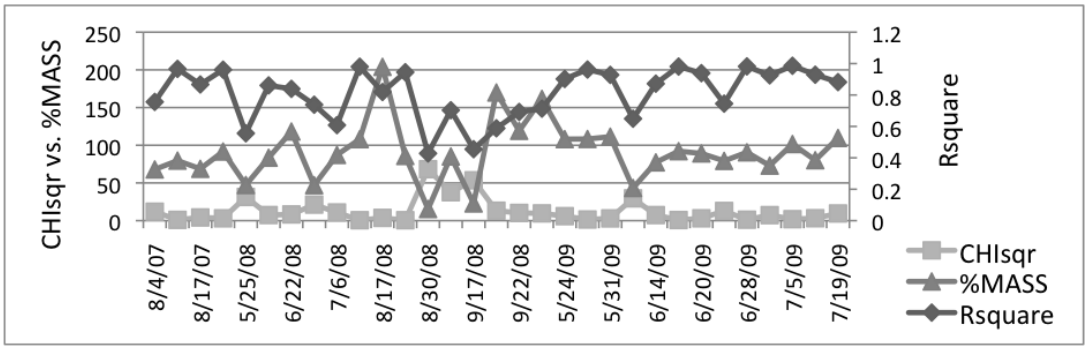
TABLE XXXIII

SUMMARY OF CMB MODEL SETTINGS AND RESULTS IN THE CAL-SAG CHANNEL. %MASS REPRESENTS THE PERCENTAGE OF TOTAL MEASURED CONCENTRATIONS BEING EXPLAINED BY THE MODEL.

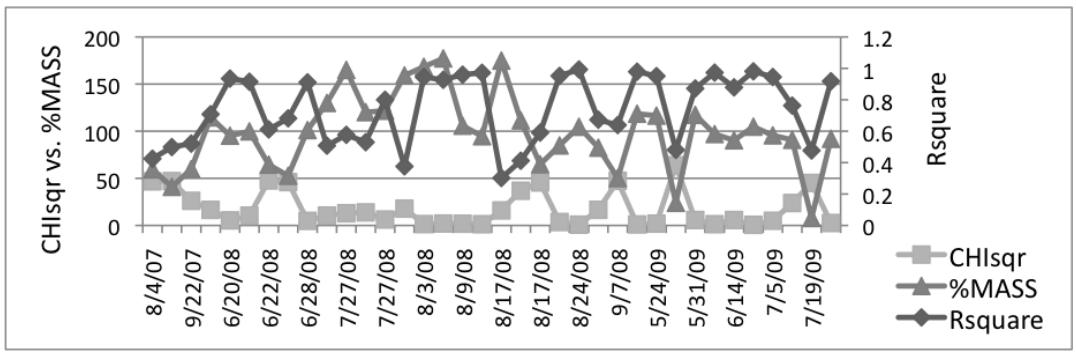
OPERATION #	1	2	3	4
Location	AL	RM	BA	WO
SPECIES INCLUDED				
E. coli	x	x	x	x
Entercocci	x	x	x	x
Male-specific coliphages	x	x	x	x
Somatic coliphages	x	x	x	x
Giardia	x	x	x	x
Cryptosporidium	x	x	x	x
SOURCES INCLUDED				
Plant	x	x		x
Rain	x	x	x	x
CSO	x	x	x	x
Background	x	x	x	x
RESULTS				
n	36	31	22	21
R square	0.74	0.81	0.64	0.78
Chi square	17.86	12.39	15.31	11.86
%Mass	98.85	91.19	82.99	103.87



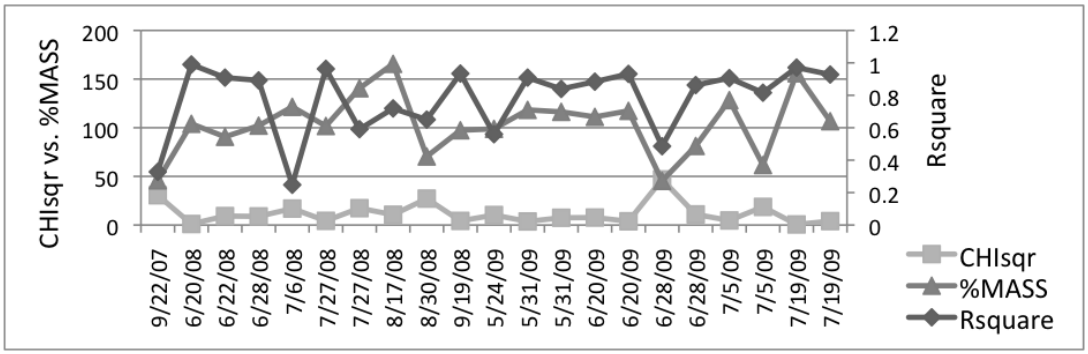
(a) Beaubien Woods



(b) Riverdale Marina



(c) Alsip



(d) Worth

Figure 19. CMB model performance by sample in Cal-Sag Channel

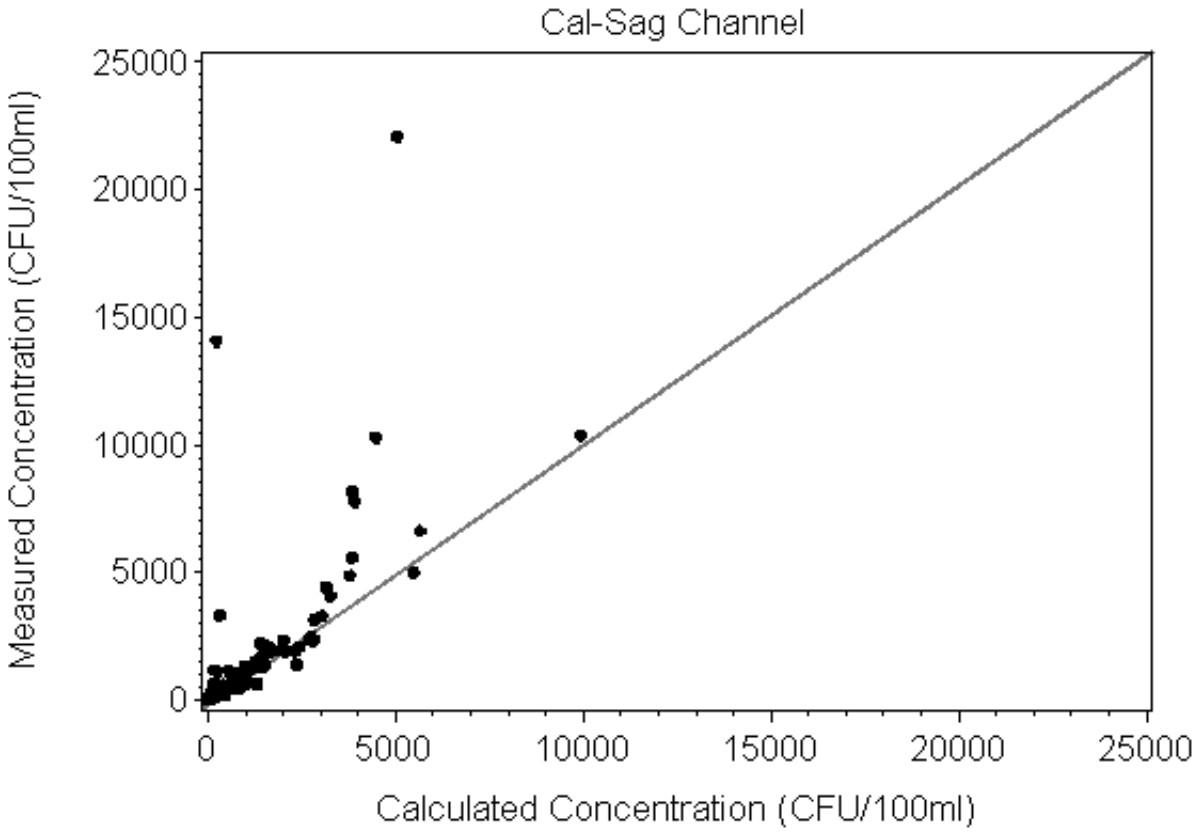


Figure 20. Scatter plot of measured and calculated concentration using CMB model in the Cal-Sag Channel

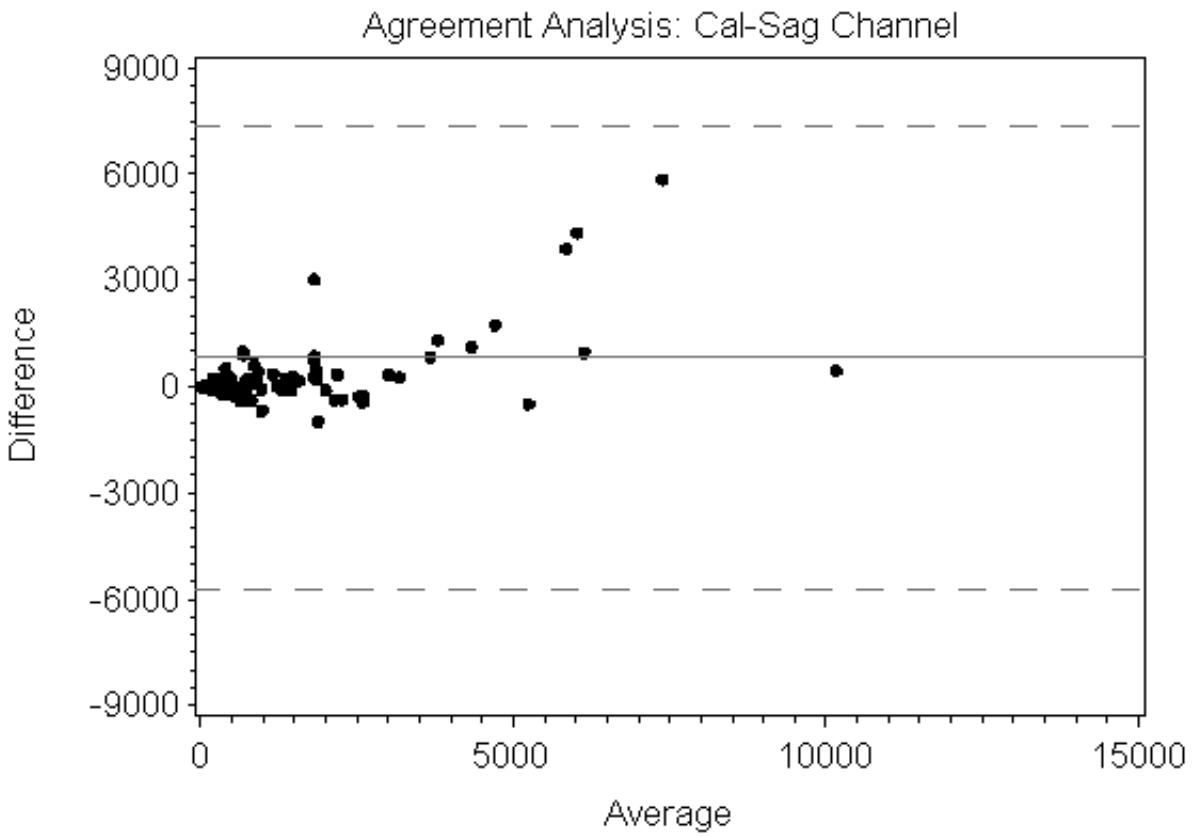


Figure 21. Bland-Altman plot of measured and calculated concentration using CMB model in the Cal-Sag Channel

TABLE XXXIV

THE AVERAGE OF TOTAL MEASURED MICROBIAL CONCENTRATIONS (COUNTS/100ML) AND CALCULATED CONTRIBUTIONS FROM PLANT, RAIN, CSO, AND BACKGROUND SOURCES BY LOCATION IN CAL-SAG CHANNEL. STANDARD DEVIATIONS ARE SHOWED IN THE PARENTHESES.

	Total Conc.	Plant	Percentage	Rain	Percentage
BA	1113.0 (2994.0)	NA	NA	122.4 (288.9)	11.0%
AL	1853.2 (4460.1)	451.1 (602.4)	24.3%	42.3 (2543.7)	2.3%
RM	3252.7 (4161.6)	1617.3 (1281.0)	49.7%	1656.3 (1878.5)	50.9%
WO	815.6 (1080.6)	420.2 (496.1)	49.5%	557.3 (1923.4)	65.6%
	Total Conc.	CSO	Percentage	Background	Percentage
BA	1113.0 (2994.0)	16.2 (68.0)	1.5%	509.0 (1075.8)	45.7%
AL	1853.2 (4460.1)	9.2 (171.2)	0.5%	1100.4 (1639.1)	58.5%
RM	3252.7 (4161.6)	140.5 (163.4)	4.3%	1847.6 (6506.7)	56.8%
WO	815.6 (1080.6)	81.3 (146.4)	10.0%	100.7 (1515.3)	11.9%

the background source is one of the major contributors to the microbial concentrations in the Cal-Sag Channel.

Model predicted CSO and rain contributions were calculated for these sampling days with extreme CSO or rain events, Table XXXV and Table XXXVI respectively. The results show that the model can capture the real CSO events by predicting higher microbial concentrations attributed by CSO source, except for location Riverdale Marina (RM). The large standard deviations indicate that the performance of the model is inconsistent. The results of percentage, however, show inconsistency in comparison between CSO_{HiDays} and $\mathrm{CSO}_{LowDays}$ data. The model can also capture precipitation impacts except at Riverdale Marina (RM) and Baubien Woods (BA) where the model predicted higher contribution of rain source on low precipitation

TABLE XXXV

MEAN PREDICTED CSO CONTRIBUTIONS (COUNTS/100ML), STANDARD DEVIATIONS (INDICATED IN PARENTHESES), AND PERCENTAGE OF TOTAL MEASURED MICROBES ON SAMPLING DAYS WITH EXTREME CSO EVENTS IN CAL-SAG CHANNEL. CSO HIDAYS ARE DAYS WITH MAGNITUDE OF CSO ABOVE THE 90 PERCENTILE. CSO LOWDAYS ARE DAYS WITH MAGNITUDE OF CSO BELOW THE 10 PERCENTILE.

Location	CSO_{HiDays}	Percentage	$CSO_{LowDays}$	Percentage
Overall	137.7 (210.6)	6.64%	78.2 (132.0)	10.21%
BA	112.3 (180.0)	8.17%	3.7 (6.4)	3.86%
RM	127.8 (225.5)	0.58%	174.0 (234.4)	5.16%
WO	267.8 (278.9)	17.73%	42.8 (42.8)	6.65%
AL	36.8 (73.5)	0.47%	64.0 (72.2)	17.54%

days than high precipitation days. More inconsistency is observed in terms of the predicted percentage of microbial concentrations attributed by rain source. Overall, the model performance is not as stable as it is in the North Branch System.

A close look at source proportion at each location along with the patterns of magnitude of last rain and CSO events are showed in Figure 22, Figure 23, Figure 24, and Figure 25 for BA, RM, AL, and WO respectively. In contrast to the North Branch System, background source in the Cal-Sag Channel is the main contributor out of all sources. This finding indicates that there could be other sources affecting water quality from the upriver, and therefore, cause unstable pattern of background source. At downstream sites, predicted plant source patterns are also closely related to the patterns of rain and CSO events. In addition, at two downstream sites, Worth and Alsip, well matched patterns of rain source and rain events are identified. At the

TABLE XXXVI

MEAN PREDICTED RAIN CONTRIBUTIONS (COUNTS/100ML), STANDARD DEVIATIONS (INDICATED IN PARENTHESES), AND PERCENTAGE OF TOTAL MEASURED MICROBES ON SAMPLING DAYS WITH EXTREME PRECIPITATIONS IN CAL-SAG CHANNEL. RAIN HIDAYS ARE DAYS WITH MAGNITUDE OF RAIN ABOVE THE 90 PERCENTILE. RAIN LOWDAYS ARE DAYS WITH MAGNITUDE OF RAIN BELOW THE 10 PERCENTILE.

Location	Rain_{HiDays}	Percentage	$\text{Rain}_{LowDays}$	Percentage
Overall	769.7 (1336.6)	81.05%	367.6 (673.8)	179.22%
BA	0 (0)	0.00%	38.8 (45.4)	21.64%
RM	1309.5 (2619.0)	5.92%	1034.4 (958.7)	504.04%
WO	1024.8 (499.6)	216.46%	105.8 (211.5)	18.14%
AL	595.6 (727.2)	81.46%	205.7 (503.8)	120.98%

two locations, the effect of plant source is decreasing, and therefore, the rain and CSO sources become more significant. Source proportion results per sample at different locations also reveal more noise in the Cal-Sag Channel than in the North Branch System.

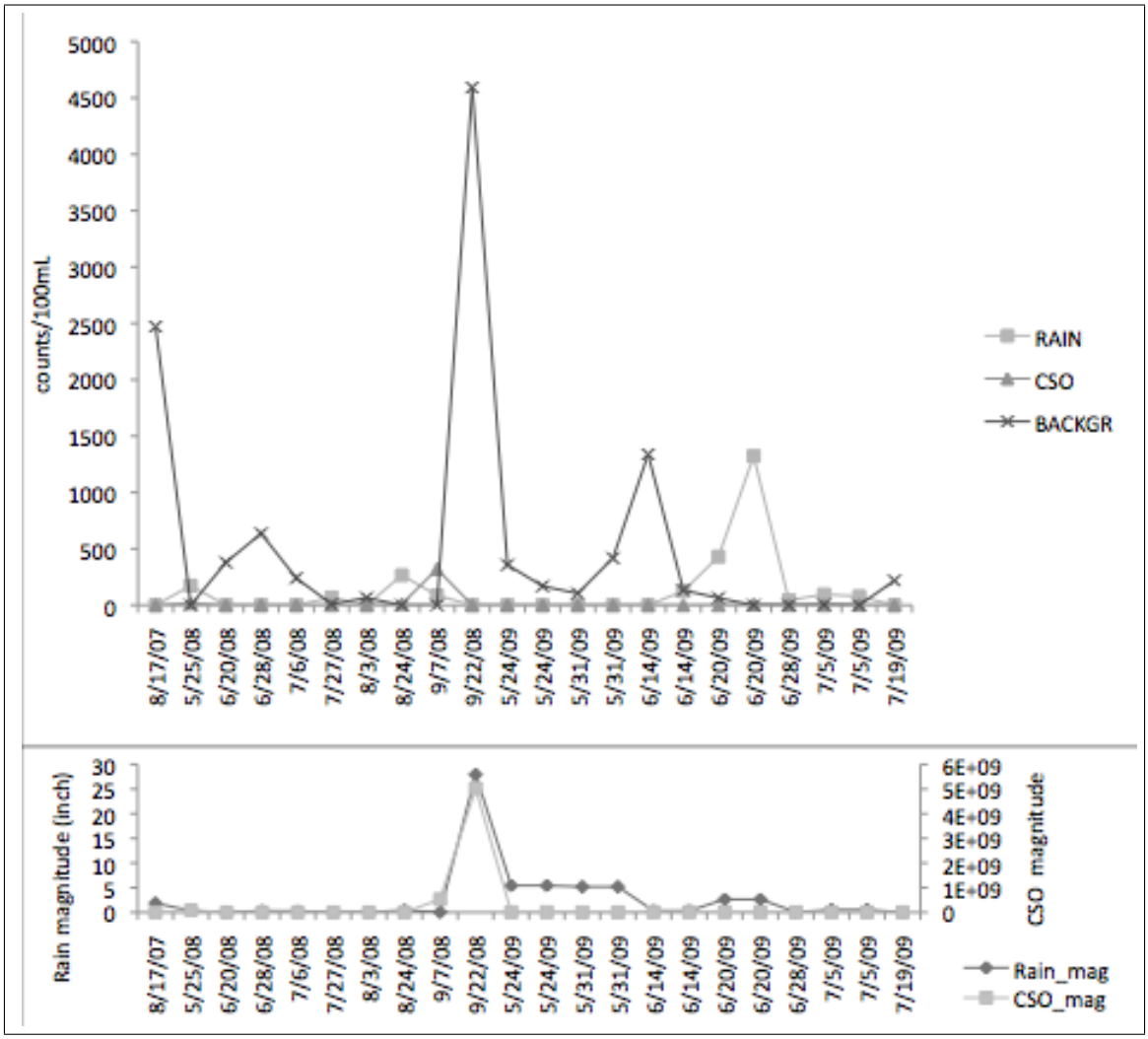


Figure 22. Beaubien Woods source proportion

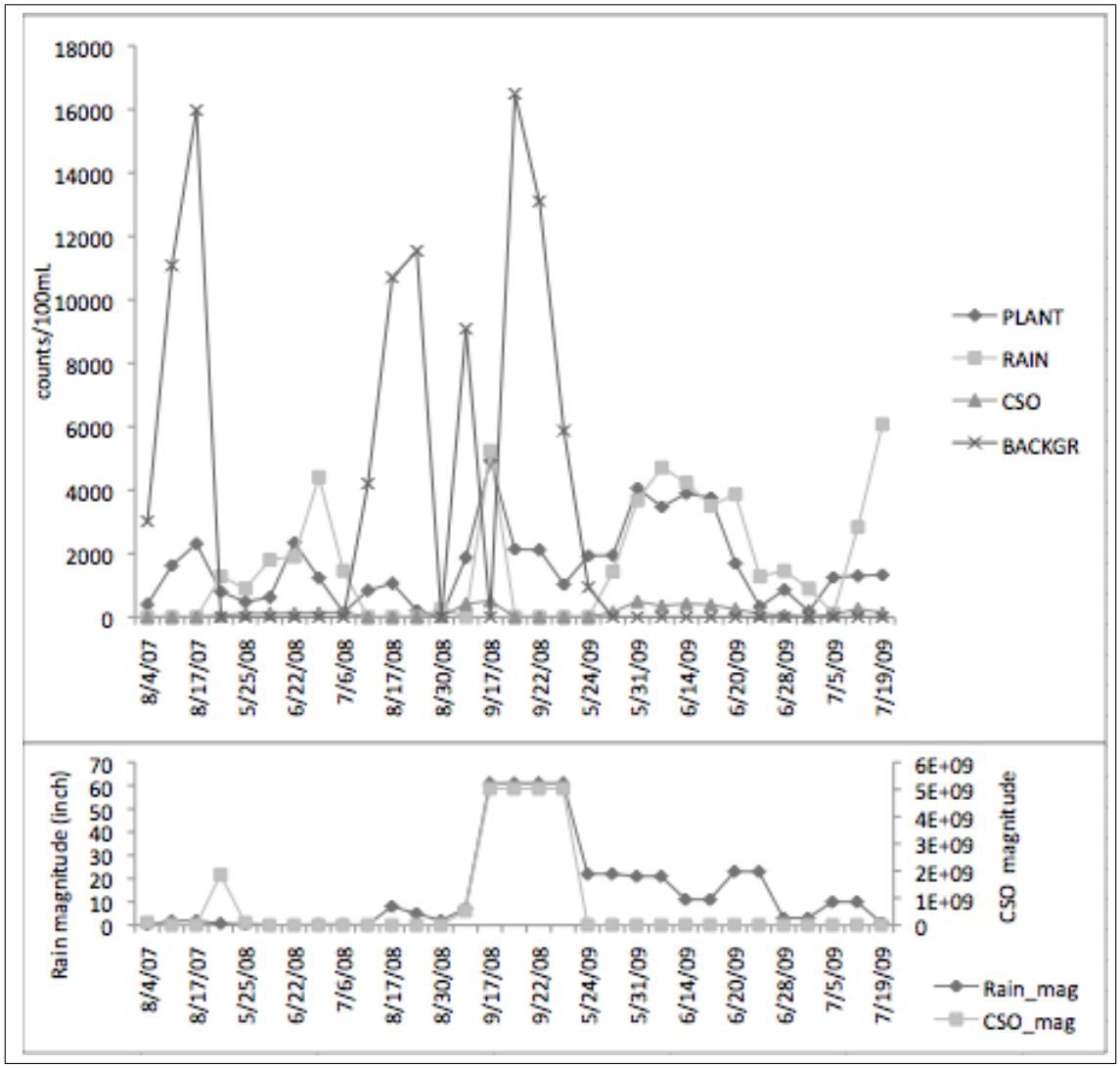


Figure 23. Riverdale Marina source proportion

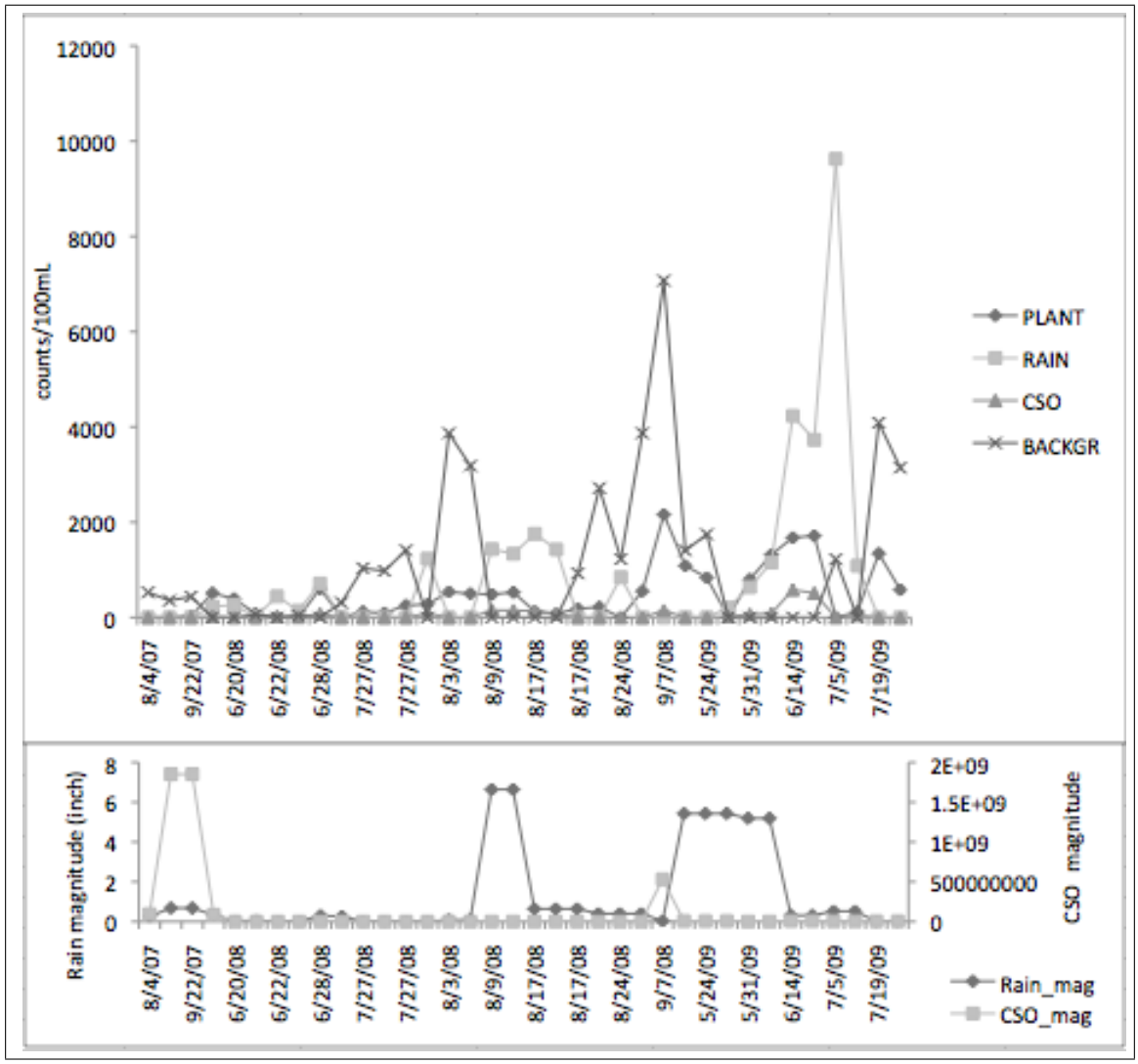


Figure 24. Alsip source proportion

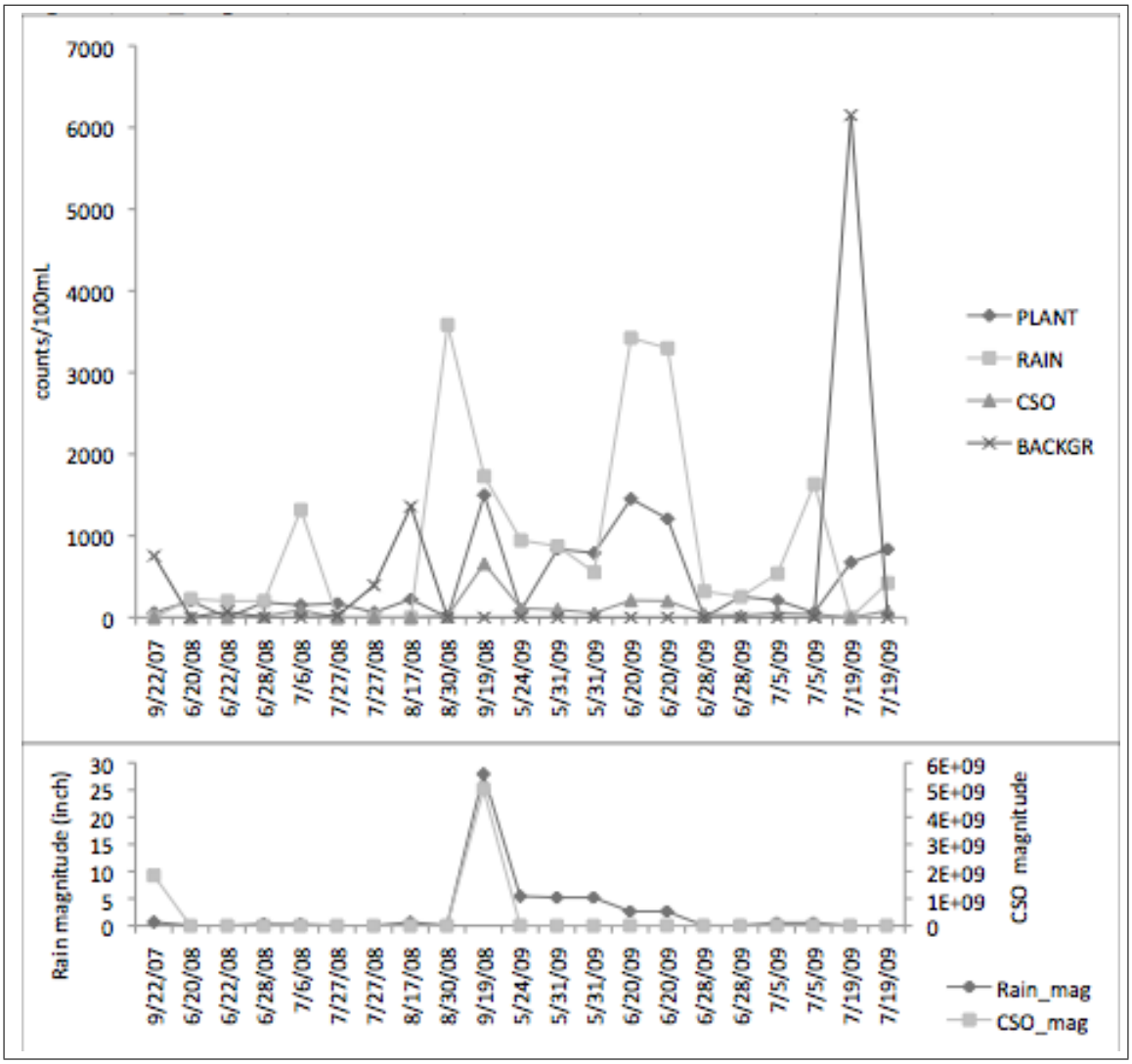


Figure 25. Worth source proportion

4.3.4 Using Pollutant Sources as Predictors of GI Illness

Estimated source contributions were used to fit the *E. coli* logistic regression model and evaluate the effectiveness of using pollutant sources to predict water users health outcomes. Two *E. coli* logistic regression models were tested. The first was the original model using both water and subject parameters as predictors, the second model only included water parameters as independent variables. The reason of utilizing the second model approach was to avoid the influence of subject parameters in the model. As described previously, *E. coli* is not a significant parameter in the model to predict health outcomes. Therefore, with the subject parameters in the model, it could be more difficult for us to observe any advantages of using sources over water quality parameters as predictors.

4.3.4.1 Model with Subject Variables

The *E. coli* logistic regression model in previous section was used for testing the power of pollutant sources on predicting health outcomes. The original model and the model with water parameters replaced by sources were compared to investigate any improvements of using sources as predictors. The North Branch System and the Cal-Sag Channel were evaluated separately.

4.3.4.1.1 North Branch System

The modeling results of using observed water variables or sources are listed in Table XXXVII and Table XXXVIII, respectively. Statistically speaking, both models are not significant and the results are relatively similar. Only the"age 10 or younger" parameter is marginal significant

in predicting GI illness. Variance inflation factor (VIF) (63) is a measurement of multicolinearity. Any variable with a VIF larger than 10 indicates that it has a linear relationship with another variable in the model. In both models, water activities, canoeing and kayaking, have multicollinearity problem. Overall, no improvement is observed using pollutant sources over water quality parameters. The expected observation of plant or CSO sources, which are considered as human fecal contamination, as strong predictors of GI illness is not found.

TABLE XXXVII

E. COLI LOGISTIC REGRESSION MODEL RESULTS IN THE NORTH BRANCH SYSTEM

Parameter	β	SE (β)	p	VIF
Intercept	-15.6266	196.1	0.9365	0
E. coli	1.2164	1.0521	0.2476	4.0545
Age0-10	2.0238	1.0952	0.0646	1.0861
Gender	0.6329	0.7579	0.4037	1.0926
Age 65 and over	-10.4638	517.1	0.9839	1.1063
$\text{Race}_{Hispanic}$	11.9870	196.0	0.9512	1.6731
Race_{White}	9.8334	196.0	0.9600	3.6840
Race_{Other}	-0.5931	301.8	0.9984	3.3062
Pre-exist GI	-10.1366	362.0	0.9777	1.0888
Previously exposed to GI	-9.8601	649.7	0.9879	1.0431
Wet score	0.1026	0.1885	0.5861	1.4339
Duration of last rain	0.0340	0.1102	0.7573	2.0945
Boat	-12.8950	348.1	0.9704	2.3467
Canoe	-3.3301	2.3113	0.1496	12.3077
Kayak	-3.9179	2.6772	0.1433	10.3161
Row	-1.7477	1.8198	0.3369	8.9807
Water sport concern	0.1238	0.1675	0.4597	1.1486
Previous CSO	-0.5419	1.6748	0.7463	1.8557
Test			p	
Overall model evaluation				
Likelihood ratio test			0.6784	
Score test			0.4992	
Wald test			0.9379	

TABLE XXXVIII

E. COLI LOGISTIC REGRESSION MODEL RESULTS USING SOURCES AS PREDICTORS IN THE NORTH BRANCH SYSTEM

Parameter	β	SE (β)	p	VIF
Intercept	-12.6136	193.7	0.9481	0
Plant source	0.3108	0.3375	0.3570	1.5857
Rain source	0.2981	0.8436	0.7238	6.7081
CSO source	-0.2921	0.3587	0.4153	1.1958
Background source	0.0075	0.3458	0.9827	1.3144
Age0-10	2.2435	1.0792	0.0376	1.0611
Gender	0.4879	0.7727	0.5277	1.1143
Age 65 and over	-10.6380	607.8	0.9860	1.0798
$\text{Race}_{Hispanic}$	11.5451	193.7	0.9525	1.6664
Race_{White}	9.7833	193.6	0.9597	3.6275
Race_{Other}	-1.0763	315.4	0.9973	3.3108
Pre-exist GI	-10.1247	404.4	0.9800	1.0922
Previously exposed to GI	-10.1373	692.2	0.9883	1.0424
Wet score	0.0889	0.1736	0.6087	1.3289
Boat	-12.4618	420.5	0.9764	2.2313
Canoe	-2.8753	3.5466	0.4175	16.3570
Kayak	-3.7637	4.1251	0.3616	14.4067
Row	-1.9884	1.9572	0.3097	9.4425
Water sport concern	0.1098	0.1662	0.5088	1.1436
Test			p	
Overall model evaluation				
Likelihood ratio test			0.6931	
Score test			0.5096	
Wald test			0.9513	

4.3.4.1.2 Cal-Sag Channel

Modeling results of using original observed water parameters and pollutant sources as predictors are listed in Table XXXIX and Table XL, respectively. During the analysis of the original model, four variables were excluded due to linearity with other variables in the model. These four variables were three water activity types (canoeing, kayaking and rowing) and the hours since previous CSO event. In the source model, three water activity types (canoeing, kayaking, and rowing) were also excluded from the analysis due to the same reason. None of the variables were significant as predictors of health outcomes in both models. In addition, the improvement of predictive power of using the source model over the water parameter model could not be identified.

TABLE XXXIX

E. COLI LOGISTIC REGRESSION MODEL RESULTS IN THE CAL-SAG CHANNEL

Parameter	β	SE (β)	p	VIF
Intercept	-22.5304	512.2	0.9649	0
E. coli	0.5504	2.5711	0.8305	1.7596
Age0-10	-9.6288	256.6	0.9701	1.4048
Gender	-0.3531	1.3200	0.7891	1.0878
Age 65 and over	-9.7291	267.0	0.9709	1.3858
$\text{Race}_{Hispanic}$	-0.0511	489.6	0.9999	4.4048
Race_{White}	9.0876	401.3	0.9819	4.9427
Race_{Other}	-3.1857	860.1	0.9970	1.9285
Pre-exist GI	-9.1844	401.3	0.9817	1.3704
Previously exposed to GI	-8.4210	484.1	0.9861	1.0906
Wet score	0.1853	0.1786	0.2996	1.2607
Duration of last rain	-0.0349	0.1568	0.8239	1.4389
Boat	9.7868	318.2	0.9755	1.1869
Canoe	0	.	.	.
Kayak	0	.	.	.
Row	0	.	.	.
Water sport concern	-0.1470	0.2026	0.4681	1.2878
Previous CSO	0	.	.	.
Test			p	
Overall model evaluation				
Likelihood ratio test			0.9788	
Score test			0.9839	
Wald test			0.9994	

TABLE XL

E. COLI LOGISTIC REGRESSION MODEL RESULTS USING SOURCES AND SUBJECT INFORMATION AS PREDICTORS IN THE CAL-SAG CHANNEL

Parameter	β	SE (β)	p	VIF
Intercept	-16.6025	304.6	0.9565	0
Plant source	-0.6242	0.6203	0.3143	2.1368
Rain source	-0.2908	0.9915	0.7693	5.8721
CSO source	0.7758	1.3933	0.5776	8.5447
Background source	0.0041	0.9867	0.9967	8.6349
Age0-10	-7.6863	142.4	0.9570	1.4323
Gender	-0.0560	1.5218	0.9706	1.0960
Age 65 and over	-7.7125	139.5	0.9559	1.3854
$\text{Race}_{Hispanic}$	-0.2101	276.8	0.9994	4.0525
Race_{White}	7.2028	254.3	0.9774	4.5909
Race_{Other}	-3.5744	526.8	0.9946	1.8616
Pre-exist GI	-6.9554	254.3	0.9782	1.4134
Previously exposed to GI	-8.4850	261.8	0.9741	1.2297
Wet score	0.1223	0.1950	0.5304	1.3223
Boat	8.3243	167.7	0.9604	1.2913
Canoe	0	.	.	.
Kayak	0	.	.	.
Row	0	.	.	.
Water sport concern	-0.3170	0.2658	0.2330	1.5983
Test			p	
Overall model evaluation				
Likelihood ratio test			0.9432	
Score test			0.9163	
Wald test			0.9979	

4.3.4.2 Models with Only Water Parameters

Logistic regression analyses using only water parameters were performed separately in the North Branch System and the Cal-Sag Channel. In the water parameter model, the six microbes, (*E. coli*, enterococci, somatic coliphages, and male-specific coliphages, *Giardia* and *Cryptosporidium*), which were used for development of source profiles, were utilized as independent variables.

4.3.4.2.1 North Branch System

The results of the two models are listed in Table XLI and Table XLII. Overall, the model which uses the six microbes as predictors is significant. Among the six microbes, *Cryptosporidium* is the strongest predictor in the model (p-value = 0.0191). *Giardia* is marginally significant (p-value = 0.0575). The four indicators are not significant as predictors of health outcomes. In addition, the analysis of VIF indicates that the model has multicollinearity problems given that male-specific coliphages and somatic coliphages both have VIFs around 10.

The model of using four sources as predictors is not significant in predicting health outcomes. However, this model does not encounter any multicollinearity problems.

TABLE XLI

E. COLI LOGISTIC REGRESSION RESULTS USING INDICATORS AND PATHOGENS AS PREDICTORS IN THE NORTH BRANCH SYSTEM

Parameter	β	SE (β)	p	VIF
Intercept	-8.1868	2.2389	0.0003	0
Cryptosporidium	-1.2364	0.5275	0.0191	1.6466
Giardia	2.3107	1.2163	0.0575	1.6374
Somatic coliphages	0.8389	1.5966	0.5993	9.1760
Male-specific coliphages	-1.5039	1.4035	0.2839	10.1973
E. coli	0.5930	0.8405	0.4805	3.7577
Enterococci	-0.5865	0.8518	0.4912	1.8106
Test			p	
Overall model evaluation				
Likelihood ratio test			0.0042	
Score test			0.0042	
Wald test			0.0204	

TABLE XLII

E. COLI LOGISTIC REGRESSION MODEL RESULTS USING ONLY SOURCES AS PREDICTOR IN THE NORTH BRANCH SYSTEM

Parameter	β	SE (β)	p	VIF
Intercept	-3.5680	0.8661	<0.0001	0
Plant source	0.3102	0.2667	0.2447	1.1335
Rain source	-0.0259	0.2010	0.8975	1.0388
CSO source	-0.1062	0.2545	0.6763	1.0313
Background source	-0.0696	0.2672	0.7944	1.1761
Test			p	
Overall model evaluation				
Likelihood ratio test			0.7587	
Score test			0.7874	
Wald test			0.8099	

TABLE XLIII

E. COLI LOGISTIC REGRESSION MODEL RESULTS USING INDICATORS AND PATHOGENS AS PREDICTORS IN THE CAL-SAG CHANNEL

Parameter	β	SE (β)	p	VIF
Intercept	-3.2482	447.4	0.9942	0
Cryptosporidium	7.9377	77.7807	0.9187	1.9698
Giardia	-1.4824	112.3	0.9895	1.8088
Somatic coliphages	7.7053	195.6	0.9686	4.5129
Male-specific coliphages	-2.8647	135.2	0.9831	5.0766
E. coli	-15.9473	166.7	0.9238	6.7967
Enterococci	11.9979	130.5	0.9267	8.2656
Test			p	
Overall model evaluation				
Likelihood ratio test			0.5865	
Score test			0.6169	
Wald test			1.0000	

4.3.4.2.2 Cal-Sag Channel

The results of the two models applied to the Cal-Sag Channel are listed in Table XLIII and Table XLIV. Both models are not significant and none of the regressors is a strong predictor. This finding differs from the result of the North Branch System. In addition, the model in the Cal-Sag Channel using sources as predictors does not provide too much improvement in multicollinearity problem, which however, was not significant in both models (VIFs < 10).

TABLE XLIV

E. COLI LOGISTIC REGRESSION MODEL RESULTS USING ONLY SOURCES AS PREDICTORS IN THE CAL-SAG CHANNEL

Parameter	β	SE (β)	p	VIF
Intercept	-1.9783	1.7193	0.2499	0
Plant source	-0.4607	0.4469	0.3026	1.5122
Rain source	-0.3968	0.8921	0.6565	4.7502
CSO source	0.1605	1.5460	0.9173	7.5021
Background source	-0.3674	0.9553	0.7005	7.5903
Test			p	
Overall model evaluation				
Likelihood ratio test			0.6891	
Score test			0.6140	
Wald test			0.6982	

4.4 Conclusions

4.4.1 Summary of Findings

This study shows that the CMB model can well explain observed fecal indicator bacteria concentrations using microbial number balance approach and provide inference of source distributions in a urban river environment.

In our study area, generally the percentage of enterococci presenting in each source is higher in the North Branch System than in the Cal-Sag Channel. The plant source profiles for the two systems are almost identical, which indicates the combinations of microbes in the discharge from the two plants are similar. Except of plant source, other source profiles are different between the North Branch System and Cal-Sag Channel. In the North Branch System, *E. coli* and enterococci are two dominant species in the background source profile while in the Cal-Sag Channel the profile is dominated by *E. coli*. In the North Branch System, *E. coli* is dominant (over 90%) in the rain source profile and in the Cal-Sag Channel, the source profile is dominated by *E. coli* and enterococci. For the CSO source, in the North Branch System, *E. coli*, enterococci and somatic coliphages are dominant species in the source profile and in the Cal-Sag Channel the dominant species are somatic coliphages followed by male-specific coliphages and *E. coli* respectively.

Bland-Altman analysis shows good agreement between total measured microbial concentrations and the concentrations calculated by the CMB model. The Spearman correlation coefficients in the Cal-Sag Channel and in the North Branch System are 0.893 and 0.885 respectively.

However, for each sample the percent of mass being explained by the model is less consistent in the Cal-Sag Channel than in the North Branch System. The r-squared values between estimated and measured concentrations are above 0.8 in the North Branch System and above 0.7 in the Cal-Sag Channel. Overall, the model can well reproduce observed microbial concentrations using source profiles. A future study comparing microbial source tracking results in the system to CMB model results using conventional FIB data can further validate if suggested CSO and plant sources are human fecal contamination.

4.4.2 Implication of the CMB Model

As mentioned previously, for each sample, the model does not reproduce the measured concentrations in the Cal-Sag Channel as well as in the North Branch System. The difference in topography and uses of the two systems is a potential cause of the differential predictive ability. Table XXIX and Table XXXIII show that the percentage of mass in the North Branch System and Cal-Sag Channel being explained by CMB are very different. The Cal-Sag Channel is a deeper and wider channel (around 200 feet wide by 26 feet deep) and therefore, it has heavy commercial boat traffic. The channel of North Branch System varies in width from 90 to 200 feet with the depth of 21 feet and the system is used primarily for canoeing and kayaking. According to data file kindly provided by the MWRDGC, the average flows from Lake Michigan into the Cal-Sag Channel is three times more than the North Branch System, 150 and 40 cubic feet per second (cfs) respectively. Therefore, the Cal-Sag Channel may have more complicated sources that were not considered in the model.

In the North Branch System, the dominant sources are rain, CSO, and plant. Except of plant source which is a unique source in our system, CSO events and urban runoff are two common sources of bacteria in urban waterbodies. The finding of urban runoff as one leading sources of FIB concentrations agrees with EPA National Water Quality Inventory report (64) which states that runoff from urban area is the major source of impairments to estuaries and lakes.

4.4.2.1 Approaches for Mitigation of Pollutant Sources

In May 2011, U.S. EPA ordered the City of Chicago to disinfect the discharge from the water reclamation plant and make the river swimmable. According to MWRDGC, the discharge, which has bacterial counts between 700 and 340,000 fecal coliforms per 100 mL, accounts for 70% and 90% of the total flow during wet and dry weather respectively. Once disinfection is implemented, it should reduce the water quality impact of plant source.

Animal sources of fecal indicator bacteria in urban runoff include pets or waste of urban wildlife such as geese, pigeons, and deer. One way to reduce the source is to educate pet owners to pick up pet waste from the street. Another way is to control the wildlife populations. In the Great Lakes area, more than 50% of *E. coli* in the lakes is contributed by gulls (65). Therefore, controlling wildlife population can well reduce the bacteria in urban runoff. Furthermore, wetlands can act as a sink for FIB which will later be diminished by sunlight (66). For new developments, having vegetation instead of pavement when possible can reduce the amount of urban runoff so less animal waste will flow into the water.

Chicago has a combined septic and stormwater system. A new design separating stormwater drainage and sewage systems is in theory a way to eliminate CSO source of bacteria into waterbodies. Separation devices of sudden large rainfall can also be installed in existing storm sewer system, however, the work can only be done during major upgrades (67).

4.4.3 Strengths

This is the only study applying the CMB model to evaluate sources of microbial contaminations, unlike other published studies, which use microbial source tracking to evaluate sources of FIB. The area of this study is a relatively uncomplicated system with a major point source of FIB from the water treatment plant. One can easily determine the sources in this type of systems and in this case CMB is an adequate approach to evaluate the effects of sources. For a specific water system, once the source profiles are developed, this approach can explain the contribution of fecal bacterial concentration from each source without the knowledge of emissions from the sources in a timely manner. This can help researchers to quickly identify impacting sources.

The large sample size with variability of weather conditions in this study allows us to develop a clean profile for each source using extreme weather conditions to evaluate rain and CSO effects. This can help the model easily to identify the difference between any two sources. In addition, multiple sampling sites (upstream and downstream from the WRPs) under different weather conditions also allow us to develop clean profile for plant source for evaluation of the effects.

The requirement of disinfection will be beneficial to evaluate the predictions of the CMB model. Before disinfection is taking place, CMB model can be utilized to predict the water quality with limited plant impact in the model. The results can then be compared to the observed concentrations after disinfection to validate the model performance.

4.4.4 Limitations

The design of this study was to identify the association between water quality and water users health. Therefore, water samples were only collected while participants were conducting water recreations. This resulted in the cancellation of water sampling events during heavy rainfall or storms and these events always relate to higher counts of FIB. It limits the patterns of FIB concentrations driven by rain or CSO events we could possibly observe.

Due to the cost of sampling and analytical method of pathogens, and unknown of which pathogens are the real cause of GI illness, we only collected one water sample a day analyzed for *Giardia* and *Cryptosporidium*. Therefore, only a small amount (less than 10) of pathogen data was included in each source profile. Since pathogens are the cause of illness, the small amount presenting in the source profile could be one of the reasons that we did not observe any association between the sources and the health outcomes. A way to solve this problem is to use a less restrict criteria to define weather conditions. This approach will allow more samples to be considered, however, it is a trade-off between sample size and clean source profiles. A further study should evaluate how the trade-off affects the model in predicting source contributions.

In this study, CMB model is utilized to partition sources in a system where potential sources of FIB are easily to determine. However, in order to model a more complex system with unknown sources, a strong knowledge of modeling site is crucial to identify potential emission sources. It requires personnel who are familiar with the system and a large number of data in order to develop source profiles properly. Furthermore, in a complex system, one could combine microbial source tracking to identify potential sources and CMB model to predict the source contributions. A harbor estuary in a highly urbanized area with inputs of multiple rivers could be considered as a complex system because of the multiple input directions of human sewage (from WRPs or CSOs), urban runoffs, leaking sewer systems, or upriver non-point sources.

The CMB model explains the concentrations received at the endpoint using source profiles to trace back the contributions from each source. It assumes the direction of pollutants coming straight for the emission point to the endpoint. Therefore, it can be applied to a river setting when the direction of pollutant movement is consistent with limited diffusion. However, it is not adequate to apply this method to model ocean or harbor systems where flow patterns are more complicated.

The association between pollution sources and health risks remains unexplained in this study. The rationale of using sources of FIB to predict health is that the sources can be classified as human sources and non-human sources and human sources of FIB, especially untreated human sewage, are more likely to have pathogens that could infect other humans. Therefore, I was

expecting to see the CSO and plant sources being strong predictors of GI illness. The lack of the association will be discussed in Chapter 6.

CHAPTER 5

FACTOR ANALYSIS OF THE EFFECTS OF WATER QUALITY ON HEALTH

5.1 Literature Review

Water quality measurements are generally highly correlated to one another. Hidden variables, such as seasons, weather, or rainfall, maybe responsible for the correlation. Factor analysis is a method to identify these hidden variables. It is a multivariate statistical technique that describes the variance of observed variables using a minimum number of latent variables called factors.

Loehlin (68) classified factor analysis into two types, exploratory factor analysis (EFA) and confirmatory factor analysis (CFA). In EFA, one seeks to identify the latent variables that can cause the interrelations of a set of observed variables. On the other hand, in CFA, one tests a hypothesized structure using available data to see how it relates to the observed variables. In this research, the goal was to indicate any latent variables that could account for the relationships between observed covariates in the dataset, therefore, the EFA method was used. Exploratory factor analysis is a method to identify the relationships among observed variables. The underlying assumption of the analysis is that the factors can be used to explain the correlation among the observed variables. If all factors are held constant, theoretically correlation

among observed variables would then be zero and each variable could be represented in a linear regression of factors, Equation 5.1.

$$x_i = \alpha_{i1} f_1 + \alpha_{i2} f_2 + \alpha_{i3} f_3 + ... + \alpha_{ik} f_k + \varepsilon_i \tag{5.1}$$

Where, x_i is the ith observed variable, f_1 to f_k are the k factors, and ε_i is the residual of x_i on the factors. The loadings of each factor on variable x_i is indicated by α_{i1} to α_{ik}.

An example of a single layer of paths between latent variables and factors is showed in Figure 26. F1 and F2 are two common factors shared by observed variables X1, X2, X3, and X4. The unique factors for each observed variable are represented by e1, e2, e3, and e4. This specific model assumes that no correlation between these unique factors exist, allowing the two factors to explain all of the correlation found between pairs of observed variables.

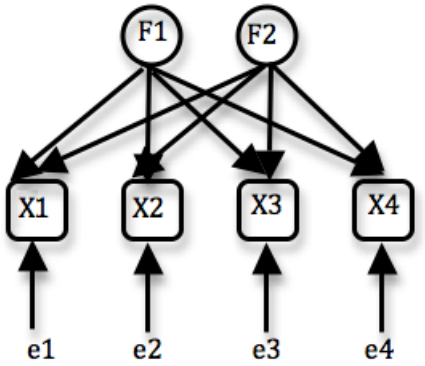


Figure 26. Example of a factor analysis model

Loehlin (68) suggested a series of steps of conducting the EFA. First, select a factor extraction method, then estimate communalities, determine the proper number of factors, establish the method of rotations, and finally interpret the factors and estimate factor scores. Details of each step are explained in the follow paragraphs.

5.1.1 Factor Extraction Methods

There are several different factor extraction methods including maximum-likelihood (ML) factor, principal component, principal factor, unweighted least-square factor, and alpha factor. The two most common factor extraction methods are ML factor and principal factor. The ML method requires variables to be normally distributed, principal factor method in contrast, because violation of the normality assumption causes distortion (69). The major limitation of the principal factor method is that it cannot compute condifence intervals and significant tests. Fabrigar et al. (70) compared the two methods in regards to the violation of the normal distribution assumption, and suggested only when transformations fail to create a normal distribution, should principal factor method be used.

5.1.2 Communalities

Once the factor extraction method is determined, the communalities can be calculated for the examination of the relationships between observed variables. The communality of a variable is the proportion of the variance that is shared with other variables. Estimation of communalities is needed for EFA since the analysis aims to recognize the latent variables that explain the common variance. The most straight forward way to estimate communalities is to use the

largest value of absolute correlation (Pearson or Spearman correlation based on the normality of variables) between the variable of interest and all other variables as the estimate. A more advanced method of estimating communalities utilizes the squared multiple correlation between the variable of interest and all other variables as the estimate.

5.1.3 Number of Factors

Determination of the number of factors is the most important and difficult step in EFA. Common methods for determining the factors include Kaiser-Guttman rule, scree plot test, percentage of variance, and parallel analysis. The Kaiser-Guttman rule extracts the number of factors that have an eigenvalue greater than one. The threshold is set to one because of the assumption that factors should have variances as large as the unity. Kaiser-Guttman rule is the only method used in Principal Component Analysis (PCA) component extraction. Principal Component Analysis is also a variable reduction technique that has been frequently mistaken as EFA. The major difference between the two techniques is that PCA is to discover components that can explain the maximal amount of variance among observed variables. On the other hand, EFA is a technique to identify the factors that account for common variance among observed variables (71). Since the goal of this study was to identify the latent variables that are influencing the observed variables, EFA was utilized.

Another common method is examining a scree plot which represents a relationship curve between eigenvalues and the number of extracted factors; an example of a scree plot is showed in Figure 28. Interpreting a scree plot involves locating the “elbow” in the plot, where the slope

of the line changes. For example, Figure 28 shows an obvious slope change at the fourth factor. In this case, the scree plot suggests four factors be extracted. However, some scree plots may have a more ambiguous shape and it could result in various interpretations.

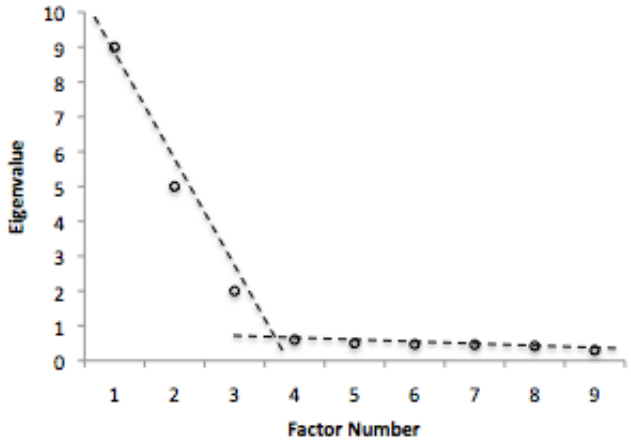


Figure 27. Example of scree plot

The percentage of variance is another criterion used for determining the number of factors. In this method, the researcher decides the preferred amount of variance being explained by the factors and then the number of factors that fit this criterion are extracted accordingly.

Another common method is the parallel analysis method, which is based on the concept that the eigenvalues of the reduced correlation matrix should be bigger than the expected eigenvalues

derived from repeated random datasets with the same sample size and number of variables. The number of factors extracted are determined at the point where the eigenvalue drops lower than the expected value.

Finally, the fifth common method is the interpretability. In EFA, the goal is to identify the hidden factors and their influence on observed variables, therefore, interpretable factors supply more information than non-interpretable factors. In other words, a four factor model with interpretability is more useful then a five factor model with an extra factor with no explanations. Most researchers tend to use one method to decide the number of factors and ignored the criterion of interpretability. In spite of this, interpretability is the most essential criterion and should always be included in an EFA study.

5.1.4 Methods of Rotation

Once factors are extracted, rotation is applied in order to obtain more interpretable factors. The factors are rotated simultaneously to reach as many zero loadings for each factor as possible that is to get as many coefficients equal to zero. This highlights which variables are important. Based on the correlation between factors, a set of data can be either orthogonal or oblique, that require different rotation methods. Data is considered orthogonal when all the extracted factors are orthogonal (not correlated). Otherwise, data is considered oblique. When data was considered orthogonal, the most common method is varimax method. For oblique data, promax is preferred (68).

5.1.5 Interpretation

Interpretation is the final step in factor analyses. Based on the factor loadings of each variable, each factor is given a meaningful name that relate to a potential latent variable in the system. Researchers can then use these factors to conduct further analyses.

Many water quality measures are correlated to each other and are affected by other variables, such as locations or seasons. A regular regression model would run into problems of multicollinearity. If the goal is simply predict Y, such as health outcomes, then multicollinearity is not a problem. The overall model predictions will still be accurate. However, if the goal is to understand how the various water parameters, such as microbial concentrations, pH, or water temperature, impacting Y, then multicollinearity would be a big problem. The individual P-value can be non-significant even though a variable is important, and therefore, mislead the interpretation. The use of EFA can avoid multicollinearity problem and pinpoint the real variables that relate to the variations in the water quality. Furthermore, the observed variables might not be significant in a regression analysis, mainly because the variance shared between the variables yields too much noise. Regular water quality measures such as pH, water temperature, or dissolved oxygen (DO) vary seasonally, and the impact of these measures on water quality change by season as well. Therefore, EFA is a tool to evaluate the temporal and spacial variations of water quality by grouping variables with similar variations together.

Exploratory Factor Analysis has been used on water quality data in many studies to identify different pollutant sources at sites with various pollution levels (72; 73; 74). In these studies, the extracted factors also provided a way to observe the seasonal variations in water quality.

Recently, Li and Zhang (75) used EFA to successfully identify the seasonal metal sources in the Upper Han River in China and then conducted health risk assessment accordingly. Since the metal sources changed seasonally, the health risk assessment combined with EFA could estimate the risk more accurately by taking seasonal effects into account. Li and Zhang showed that using the latent variables identified by EFA can help to observe the association between water quality and human health.

Applying EFA to this study can help us to identify the different sources of bacterial microbes in various levels of polluted areas. Despite the number of studies that have applied EFA to water quality data, none of them examined if factors extracted from a water quality dataset can be used to predict disease rates among water users. The goal of this study is to evaluate the performance of using factors in predicting health risks.

5.2 Methods

The following approaches were developed in order to test the research hypothesis that one can apply exploratory factor analysis to water quality data and use the factors to regress water users health outcomes.

i) Apply EFA on North Branch System and Cal-Sag Channel separately to identify the different factors impacting the water quality.

ii) Perform EFA based on different levels of pollution (upstream or downstream of the WRPs) to test if influencing factors vary considerably by pollution levels.

iii) Evaluate the performance of using factors as predictors of health outcomes.

The analyses were conducted in SAS®, version 9, using command PROC FACTOR. Exploratory factor analysis was applied to all the samples collected in the CAWS and to the two systems separately to examine any different factors influencing water quality presenting in the systems. Furthermore, for each system, upstream and downstream sites were also analyzed individually for identification of potential different factors at various pollution levels. In this study, measured values of water parameters varied in a wide range in regards to different locations; complete results are described in Appendix A. This could indicate the presence of certain temporal unobserved variables.

The variables used to extract factors were four indicators, dissolved oxygen (DO) concentration, pH, turbidity, conductivity, water temperature, accumulated solar radiation, last CSO event (hours since last CSO), magnitude of last CSO, duration of last CSO, intensity (gal/hour) of last CSO, last rain event (hours since last rain), magnitude of last rain, duration of last rain, and intensity (inch/hour) of last rain event. Transformations were used to convert the variables to normal distributions, however, even with the transformation, violations of normality, skewness > 2 or kurtosis > 7, were still present in some of the variables. Hence, principle factor method was used in this study for factor extraction.

After the extraction method had been selected, the communalities between observed variables were estimated. In this step, the squared multiple correlation coefficient (SMC) of one variable with all other variables, also known as r-squared, was used for communality estimation. The SMC of a variable multiplied by 100 represented the percent of variation of the variable being explained by all other variables. This was the portion which EFA attempted to explain by extracting latent variables.

Once communalities were calculated, the number of factors was determined. Initially, six factors were extracted based on four criteria: Kaiser-Guttman rule, percentage of variance preferred to be explained, scree plot test, and parallel analysis, Figure 28. However, while examining the variable loadings for each factor, not all the factors were interpretable. Factor five and six seemed to present the unexplained information remaining in the system but they were not inferential. Therefore, four factors were chosen for this study.

The correlation between the four extracted factors was examined. Ideally, in factor analysis, the goal is to extract non-correlated factors. However, in this study, the four selected factors were not orthogonal. In this case, one of the options was to apply EFA on the four factors in order to obtain a second tier of factors to resolve the correlation problem, but this would result in fewer factors with no interpretability, so this approach was not considered for the purposes of this study.

Before giving any interpretations to the four factors, factor rotations were performed. Promax rotation was selected over varimax method since the four factors were not orthogonal. The

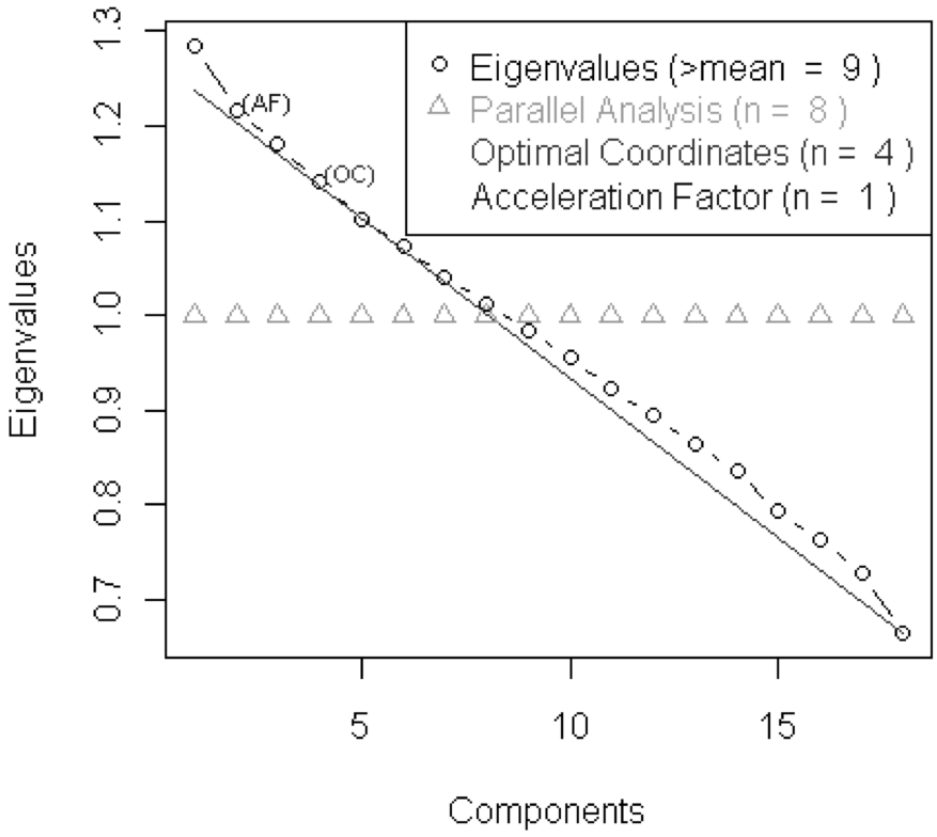


Figure 28. Scree plot and non-graphical solutions of scree test

rotated factors were then interpreted based on the combination of variables with high loading in each factor. Factor scores for CAWS, Cal-Sag Channel, North Branch System, upstream and downstream of the North Branch System were calculated and compared for the identification of different factor patterns from various sites.

The goal of applying the EFA method to this study was to identify the sources that are impacting the water quality and examined how sources can predict the health outcomes among water users. In order to accomplish the goal, the data with factor scores was matched with the health outcome data and then used to fit the *E. coli* logistic regression model previously used. The parameter estimates were compared with the ones obtained from using individual water quality parameters as predictors.

5.3 Results

The EFA method was applied on the CAWS, the North Branch System and the Cal-Sag Channel, and, in addition to the whole North Branch System, upstream and downstream sites in the North Branch System. Initially the EFA method was to be applied on the upstream and downstream sites in the Cal-Sag Channel, but due to an insufficient sample size it was impossible to conduct EFA.

5.3.1 CAWS Overall

Factor loadings in CAWS are showed in Table XLV. 47% variation was explained by the four factors. The standard criteria used to determine the significancy of a factor loading are: factor loadings from 0.3 to less than 0.6 are considered moderate, indicating the factor only has

moderate effects on the variables, and loadings of 0.6 and above are defined as strong, indicating strong factor effect. In Table XLV bold and underlined values indicated strong and moderate loadings respectively. Based on the factor loadings, factor 1 was named as the weather factor and factor 2 was identified as the indicator factor. The separation of an indicator factor from a weather factor was apparent. Factor 3 and 4, on the other hand, were not as manifest as factor 1 or 2. Factor 3 was identified as the unexplained influence of weather factor and factor 4 was considered water chemistry factor. The CAWS contains a mix of sampling sites with various pollution levels, upstream and downstream, and also unique location characteristics. Therefore, I hypothesized that it is complicated to observe meaningful factors without grouping samples by pollution levels or location characteristics. Figure 29 shows a two dimensional map of variable loadings of factor 1 and 2 of the CAWS. One can notice that the separation of variables started to occur while only two factors were considered. It indicates the presence of latent variables that are impacting the observed water parameters.

5.3.2 North Branch System Versus Cal-Sag Channel

Factor loadings in the North Branch System and Cal-Sag Channel are listed in Table XLVI. In the North Branch System, a strong indicator factor, factor 1, exists while by contrast, the same factor is not identified in the Cal-Sag Channel. In addition, the separation of an indicator factor from a weather factor is more clear in the North Branch System than in the Cal-Sag Channel. In the Cal-Sag Channel, indicators are distributed across different factors. Also, the factor loading for conductivity is related to the indicator factor in the North Branch System.

TABLE XLV

FACTOR LOADINGS AT CAWS LOCATIONS

	Variation explained by factors: 47% (n=559)			
	Factor 1	Factor 2	Factor 3	Factor 4
Somatic coliphages (PFU/100mL)	0.23	**0.87**	0.07	0.18
Male-specific coliphages (PFU/100mL)	0.32	**0.83**	0.21	0.13
E. coli (CFU/100mL)	0.03	**0.73**	0.08	0.06
Enterococci (CFU/100mL)	-0.26	**0.74**	0.01	-0.14
DO (mg/L)	-0.13	-0.31	-0.06	-0.43
pH	0.14	-0.28	-0.01	-0.21
Turbidity (NTU)	0.07	0.17	0.07	0.37
Conductivity (mmho/cm)	0.17	0.32	-0.35	0.12
Water temperature (°C)	-0.12	-0.03	-0.12	**0.62**
Solar radiation (W/m^2)	-0.14	-0.03	-0.20	0.51
Last CSO (hour)	-0.14	-0.21	-0.50	0.11
CSO magnitude (gallon)	**0.75**	0.13	0.19	0.02
CSO duration (hour)	0.58	0.08	0.09	-0.09
CSO intensity (gallon/hour)	**0.60**	0.17	0.26	0.01
Last rain (hour)	0.58	-0.18	0.06	0.06
Rain magnitude (inch)	**0.85**	0.05	**0.68**	0.12
Rain duration (hour)	**0.78**	0.10	0.57	-0.05
Rain intensity (inch/hour)	0.42	0.14	0.59	0.12

Bold and underlined values indicate strong and moderate loadings respectively

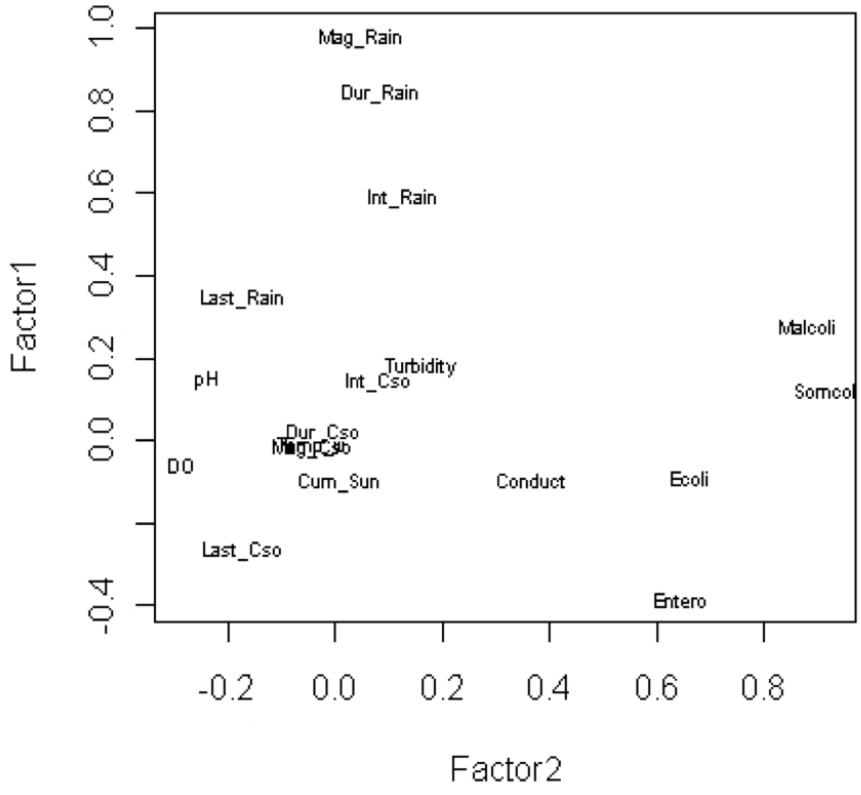


Figure 29. Factor loadings of factor 1 and factor 2 of CAWS

This finding is consistent with the overall CAWS and can indicate the relationship between indicators and conductivity. However, while a disparity between the two systems is noticeable, one has to keep in mind that the North Branch System is characterized by twice as many samples as the Cal-Sag Channel. The small sample size at the Cal-Sag Channel may impact identification of factors.

TABLE XLVI

FACTOR LOADINGS AT NORTH BRANCH SYSTEM AND CAL-SAG CHANNEL

	North Branch system Variation explained by factors: 51% (n=312)			
	Factor 1	Factor 2	Factor 3	Factor 4
Somatic coliphages (PFU/100mL)	**0.90**	0.07	0.19	0.16
Male-specific coliphages (PFU/100mL)	**0.86**	0.14	0.31	0.05
E. coli (CFU/100mL)	**0.77**	-0.08	0.15	0.24
Enterococci (CFU/100mL)	**0.72**	-0.36	0.04	0.18
DO (mg/L)	-0.28	-0.09	-0.14	-0.47
pH	-0.27	0.15	-0.02	-0.34
Turbidity (NTU)	0.10	-0.06	0.26	0.28
Conductivity (mmho/cm)	0.51	0.22	-0.12	-0.24
Water temperature (°C)	0.08	-0.09	0.08	**0.76**
Solar radiation (W/m^2)	-0.01	-0.11	-0.16	0.51
Last CSO (hour)	-0.06	-0.01	-0.34	-0.03
CSO magnitude (gallon)	0.09	**0.84**	0.40	-0.12
CSO duration (hour)	0.05	**0.64**	0.10	-0.12
CSO intensity (gallon/hour)	0.05	**0.64**	0.42	-0.09
Last rain (hour)	-0.17	**0.60**	0.08	-0.17
Rain magnitude (inch)	0.08	**0.73**	**0.85**	-0.29
Rain duration (hour)	0.13	**0.61**	**0.65**	-0.49
Rain intensity (inch/hour)	0.20	0.27	**0.65**	0.02
	Cal-Sag Channel Variation explained by factors: 64% (n=152)			
	Factor 1	Factor 2	Factor 3	Factor 4
Somatic coliphages (PFU/100mL)	0.18	**0.77**	0.27	0.22
Male-specific coliphages (PFU/100mL)	0.30	**0.85**	0.44	0.16
E. coli (CFU/100mL)	0.02	0.02	-0.05	**0.72**
Enterococci (CFU/100mL)	**-0.66**	0.17	-0.02	0.23
DO (mg/L)	0.01	-0.30	-0.01	**-0.63**
pH	0.28	-0.12	0.23	-0.31
Turbidity (NTU)	0.07	0.15	-0.22	0.33
Conductivity (mmho/cm)	-0.09	-0.39	-0.59	0.05
Water temperature (*ĉirc*C)	-0.29	-0.43	0.10	-0.38
Solar radiation (W/m^2)	-0.19	-0.09	-0.41	-0.02
Last CSO (hour)	-0.28	**-0.71**	-0.12	-0.45
CSO magnitude (gallon)	**0.82**	0.55	**0.80**	-0.01
CSO duration (hour)	**0.91**	0.56	**0.63**	-0.01
CSO intensity (gallon/hour)	0.45	**0.76**	**0.77**	-0.26
Last rain (hour)	**0.80**	0.14	-0.09	-0.10
Rain magnitude (inch)	**0.92**	0.55	**0.64**	0.15
Rain duration (hour)	**0.92**	**0.61**	0.54	0.23
Rain intensity (inch/hour)	**0.87**	0.41	0.30	0.26

Bold and underlined values indicate strong and moderate loadings respectively

5.3.3 North Branch System: Upstream Versus Downstream

Factor loadings for upstream and downstream of the North Branch System are showed in Table XLVII. It is noticeable that the strongest factor in the downstream system is the indicator factors. In the upstream location, conversely, the most significant factor is the weather factor. This implies that the observed indicator factors downstream could be the influence of the North Side WRP. In addition, the trend of conductivity correlated with the indicator factor is also identified in the downstream sites. Furthermore, separation of rain factor and CSO factor can be identified in the downstream sites as well. This is the first time this separation occurred in all analyses. Downstream sites have more samples with similar pollution levels and site characteristics, and it makes the identification of weather factors in the system more clear. This can be the reason that clustering of variables with high factor loadings are more apparent at downstream sites than at the upstream site.

TABLE XLVII

FACTOR LOADINGS AT CAWS NORTH BRANCH SYSTEM UPSTREAM AND DOWNSTREAM

	Upstream Variation explained by factors: 60% (n=63)			
	Factor 1	Factor 2	Factor 3	Factor 4
Somatic coliphages (PFU/100mL)	0.42	**0.77**	**0.72**	0.29
Male-specific coliphages (PFU/100mL)	0.51	**0.66**	**0.75**	0.27
E. coli (CFU/100mL)	-0.01	**0.72**	0.10	0.15
Enterococci (CFU/100mL)	-0.52	**0.65**	-0.11	0.18
DO (mg/L)	-0.25	-0.33	-0.32	**-0.67**
pH	0.01	-0.15	0.01	-0.54
Turbidity (NTU)	0.10	0.46	0.15	0.14
Conductivity (mmho/cm)	0.17	-0.11	**0.67**	-0.17
Water temperature (∘C)	0.07	0.28	-0.18	**0.74**
Solar radiation (W/m^2)	-0.09	-0.13	-0.17	0.44
Last CSO (hour)	-0.23	-0.45	0.02	-0.28
CSO magnitude (gallon)	**0.94**	0.29	**0.64**	0.18
CSO duration (hour)	**0.68**	-0.00	0.17	0.08
CSO intensity(gallon/hour)	**0.69**	0.42	0.54	0.23
Last rain (hour)	**0.73**	-0.16	0.04	-0.01
Rain magnitude (inch)	**0.92**	0.37	**0.71**	0.17
Rain duration (hour)	**0.72**	0.21	**0.76**	-0.01
Rain intensity (inch/hour)	**0.63**	0.45	0.40	0.14
	Downstream Variation explained by factors: 48% (n=249)			
	Factor 1	Factor 2	Factor 3	Factor 4
Somatic coliphages (PFU/100mL)	**0.84**	0.02	0.17	-0.03
Male-specific coliphages (PFU/100mL)	**0.80**	0.07	0.33	-0.13
E. coli (CFU/100mL)	**0.72**	-0.05	0.10	0.11
Enterococci (CFU/100mL)	**0.77**	-0.26	-0.01	0.03
DO (mg/L)	-0.09	-0.04	-0.01	-0.46
pH	-0.20	0.17	0.07	-0.26
Turbidity (NTU)	0.03	-0.09	0.19	0.38
Conductivity (mmho/cm)	0.42	0.23	-0.10	-0.42
Water temperature (∘C)	-0.01	-0.10	-0.01	**0.77**
Solar radiation (W/cm^2)	-0.14	-0.11	-0.17	0.45
Last CSO (hour)	-0.06	0.05	-0.31	-0.03
CSO magnitude (gallon)	0.01	**0.85**	0.37	-0.07
CSO duration (hour)	-0.01	**0.67**	0.12	-0.16
CSO intensity (gallon/hour)	-0.03	**0.63**	0.39	-0.07
Last rain (hour)	-0.28	0.58	0.10	-0.22
Rain magnitude (inch)	0.05	0.59	**0.92**	-0.19
Rain duration (hour)	0.12	0.50	**0.72**	-0.43
Rain intensity (inch/hour)	0.26	0.17	0.59	0.13

Bold and underlined values indicate strong and moderate loadings respectively

5.3.4 Using Factor Loadings As Predictors of Health Outcomes

Four factor model of the CAWS was used to fit the *E. coli* logistic regression model. The two variables, duration of last rain and hour since last CSO, were removed from the model since the rain and CSO variables were included in the four factors. However, the expected improvement was not observed. Possible reasons are: factors do not link to the health outcomes, the latent variables that cause the cluster of observed variables do not affect health outcomes, and some of the subject variables are significant in the model and affect any observation of improvements contributed by the factors. In order to test the second hypothesis, an approach of logistic regression modeling without subject variables was utilized.

5.3.4.1 Logistic Regression: Full Models with Subject Variables

The CAWS data was used to fit the logistic regression model using factor loadings and survey data. The results are listed in Table XLVIII.

The factor model was not significant. As mentioned previously, this might be because the subject variables in the model have a stronger effect predicting GI illness rates among water users. Since the goal of this part of study was to evaluate if factors can predict GI illness better than individual water parameters, the analysis was adjusted by using only water parameters or factors in the model to compare the results. The model with only water parameters contained only the eighteen variables used in extracting factors. The results are showed in the following sections.

TABLE XLVIII

LOGISTIC REGRESSION ANALYSIS USING FACTORS

Parameter	β	SE (β)	p	VIF
Intercept	-3.7476	0.9644	0.0001	0
Factor 1	-0.5558	0.5280	0.2925	2.2828
Factor 2	0.2870	0.3728	0.4414	3.1146
Factor 3	-0.6378	0.4027	0.1133	1.6103
Factor 4	-0.2911	0.3663	0.4269	1.7645
Age0-10	0.1374	0.6174	0.8239	1.1648
Gender	0.4741	0.3569	0.1841	1.0325
Age 65 and over	-12.5607	573.5	0.9825	1.0377
$\text{Race}_{Hispanic}$	0.7587	0.8031	0.3448	1.5270
Race_{White}	-0.3373	0.6467	0.6020	2.6441
Race_{Other}	0.3931	0.7609	0.6054	1.9920
Pre-exist GI	1.3983	0.6993	0.0455	1.0397
Previously exposed to GI	-12.8966	600.9	0.9829	1.0157
Wet score	0.2072	0.0591	0.0005	1.4662
Boat	-0.8611	0.6581	0.1907	2.9482
Canoe	-1.7350	0.772	0.0256	4.2993
Kayak	-2.0296	0.9043	0.0248	4.0406
Row	-1.4486	0.8442	0.0862	2.2032
Water sport concern	0.0448	0.0645	0.4868	1.0854
Test			p	
Overall model evaluation				
Likelihood ratio test			0.0240	
Score test			0.0241	
Wald test			0.0985	

5.3.4.2 Logistic Regression: Models With Only Water Parameters

The variables in the water parameter model are four indicators, (*E. coli*, enterococci, somatic coliphages, and male-specific coliphages), D.O., pH, water temperature, turbidity, conductivity, cumulative sun, hours since last CSO event and rain event, magnitude of last CSO and rain, duration of last CSO and rain, and interval of last CSO and rain. There are a total of 18 variables in the water parameter model and a total of four parameters in the factor model. The logistic regression models were applied to the complete CAWS dataset, Cal-Sag Channel, North Branch System, North Branch System downstream locations, and also upstream locations in the North Branch System. However, at the upstream site of North Branch System, the majority of the water parameter variables were linear combinations of the other variables in the model, which resulted in the model returning error message and the parameter estimates were not able to be obtained. This might be because all the data points were from one site (BR). The lack of variability of site characters might result in the linear relationship between the water parameter variables. The small sample size might also impact the ability to detect the true correlation. The results of the two different models at each location group are listed in the following sections.

5.3.4.2.1 CAWS

The results of the two logistic regression models are listed in Table XLIX and Table L. Again, the results show no improvement when using factors as predictors instead of water parameters individually. Overall, both models are not significant statistically. The model using water

parameters does not have strong multicollinearity problems given that all the VIFs are under 10. However, one can manifest the improvement of VIFs in the model using factors as predictors.

5.3.4.2.2 Cal-Sag Channel

The results of modeling Cal-Sag Channel are listed in Table LI and Table LII. Results indicate no improvement when using factors as variables instead of water parameters. Overall model evaluation indicates that neither models are statistically significant; the likelihood ratio test of p-value=0.03 in water parameter model is only marginally significant. The individual p-value for each factor in the factor model are all larger than 0.05, indicating that factors are not stronger predictors than water parameters. However, the factor model solves the multicollinearity problem the water parameter model encountered.

5.3.4.2.3 North Branch System

The results of modeling North Branch System are listed in Table LIII and Table LIV. Overall model evaluation results indicate that neither models are statistically significant. In addition, in the factor model, there is no multicollinearity problem that is observed in the water parameter model.

5.3.4.2.4 North Branch System - Downstream Sites

The results of modeling the North Branch System downstream sites are listed in Table LV and Table LVI respectively. There is no improvement noticed when using factors as variables instead

TABLE XLIX

LOGISTIC REGRESSION ANALYSIS USING ONLY WATER PARAMETERS: CAWS

Parameter	β	SE (β)	p	VIF
Intercept	-3.6549	4.9189	0.4575	0
Somatic coliphages	0.4006	0.4223	0.3427	7.0243
Male-specific coliphages	-0.6834	0.4355	0.1166	5.9349
D.O.	0.1533	0.1250	0.2201	1.4313
p.H.	0.1756	0.6602	0.7903	1.4097
Water Temperature	-0.1790	0.0757	0.0181	4.1218
Turbidity	0.0048	0.0295	0.8698	1.8487
Last CSO	0.0037	0.0013	0.0033	2.2850
Last rain	-0.0118	0.0047	0.0125	2.6049
CSO-Mag	-655E-13	5.2E-11	0.2077	8.2868
CSO-Duration	0.0060	0.0043	0.1599	4.8011
Rain-Intensity	0.0938	1.7085	0.9562	2.6913
CSO-intensity	3.37E-9	4.83E-9	0.4849	2.9684
Rain-Duration	-0.0639	0.0561	0.2546	3.8574
Rain-Mag	0.1197	0.2307	0.6039	6.2286
Conductivity	0.0007	0.0007	0.3278	1.8255
E. coli	0.4216	0.4335	0.3307	3.9047
Enterococci	-0.2572	0.5034	0.6094	3.7193
Sun-Cum	0.1878	0.0952	0.0484	2.3440
Test			p	
Overall model evaluation				
Likelihood ratio test			0.0881	
Score test			0.1037	
Wald test			0.1697	

TABLE L

LOGISTIC REGRESSION ANALYSIS USING ONLY FACTORS: CAWS

Parameter	β	SE (β)	p	VIF
Intercept	-3.1585	0.2449	<0.0001	0
Factor 1	-0.3017	0.4319	0.4848	2.0818
Factor 2	-0.1432	0.2385	0.5481	1.4813
Factor 3	-0.5852	0.3416	0.0867	1.4364
Factor 4	-0.0911	0.2625	0.7286	1.1083
Test			p	
Overall model evaluation				
Likelihood ratio test			0.2913	
Score test			0.3416	
Wald test			0.3317	

of water parameters individually. The results of the overall model evaluation, indicate that both models are not statistically significant, and consequently, factors are not stronger predictors than water parameters in predicting water users health outcomes. However, the factor model reduces the multicollinearity problem that occurred in the water parameter model.

TABLE LI

LOGISTIC REGRESSION ANALYSIS USING ONLY WATER PARAMETERS: CAL-SAG CHANNEL

Parameter	β	SE (β)	p	VIF
Intercept	-48.1751	56.1655	0.3910	0
Somatic coliphages	-2.5483	2.4892	0.3060	8.6040
Male-specific coliphages	-0.6860	2.4412	0.7787	7.5914
D.O.	-2.4773	1.4812	0.0944	6.5679
p.H.	4.2490	4.9360	0.3893	2.6327
Water Temperature	-0.1947	0.3775	0.6060	7.7722
Turbidity	-0.1067	0.1185	0.3678	2.2331
Last CSO	0.0460	0.0473	0.3312	7.7859
Last rain	-0.2085	0.2539	0.4115	22.7555
CSO-Mag	-7.47E-8	1.25E-7	0.5483	53.5393
CSO-Duration	0.9385	4.8614	0.8469	29.1803
Rain-Intensity	-5.9919	43.9390	0.8915	5.7478
CSO-intensity	1.00E-6	1.81E-6	0.5805	10.0624
Rain-Duration	0.9323	2.4249	0.7006	60.7826
Rain-Mag	0.2357	34.3882	0.9945	109.1240
Conductivity	0.0191	0.0127	0.1322	3.1859
E. coli	-2.9634	2.3915	0.2153	3.4544
Enterococci	3.6199	3.4536	0.2946	4.0588
Sun-Cum	1.0289	0.4941	0.0373	1.9584
Test			p	
Overall model evaluation				
Likelihood ratio test			0.0324	
Score test			0.1917	
Wald test			0.8215	

TABLE LII

LOGISTIC REGRESSION ANALYSIS USING ONLY FACTORS: CAL-SAG CHANNEL

Parameter	β	SE (β)	p	VIF
Intercept	-4.9877	1.6730	0.0029	0
Factor 1	-3.1945	2.5013	0.2016	1.4052
Factor 2	-1.3607	1.0032	0.1750	1.2570
Factor 3	-1.5167	2.4117	0.5294	1.4602
Factor 4	0.3375	0.9018	0.7082	1.1505
Test			p	
Overall model evaluation				
Likelihood ratio test			0.1462	
Score test			0.3746	
Wald test			0.4470	

5.4 Conclusions

5.4.1 Summary of Findings

In the comparison of the North Branch System to the Cal-Sag Channel, an indicator factor is only identified in the North Branch System. In the Cal-Sag Channel, the indicator microbes do not cluster together. Studies have found that the levels of indicator microbes were consistently higher in the North Branch System than in the Cal-Sag Channel (76; 77). It could be one of the reasons why a clear indicator factor is observed in the North Branch System.

The difference between the North Branch System and Cal-Sag Channel is also observed using CMB model for source apportionment. As mentioned previously, the heavy commercial traffic

TABLE LIII

LOGISTIC REGRESSION ANALYSIS USING ONLY WATER PARAMETERS: NORTH BRANCH SYSTEM

Parameter	β	SE (β)	p	VIF
Intercept	4.8419	13.0925	0.7115	0
Somatic coliphages	2.4510	1.1065	0.0268	8.8242
Male-specific coliphages	-0.8721	0.7030	0.2148	6.7993
D.O.	0.5250	0.3638	0.1490	2.7766
p.H.	-0.7192	1.7228	0.6763	1.7976
Water Temperature	-0.5492	0.2516	0.0290	8.1835
Turbidity	0.0740	0.0971	0.4461	3.7005
Last CSO	0.0058	0.0039	0.1355	4.9039
Last rain	-0.0275	0.0120	0.0213	3.4141
CSO-Mag	-184E-12	1.22E-10	0.1312	12.9076
CSO-Duration	0.0169	0.0083	0.0408	5.7385
Rain-Intensity	3.4552	3.4122	0.3113	3.5650
CSO-intensity	9.57E-9	9.32E-9	0.3044	3.7423
Rain-Duration	-0.0990	0.1016	0.3296	3.0721
Rain-Mag	0.0038	0.4730	0.9937	7.4537
Conductivity	-0.0011	0.0025	0.6599	3.2863
E. coli	0.2836	0.8843	0.7484	6.8341
Enterococci	0.2836	0.8843	0.7484	4.4760
Sun-Cum	0.4247	0.2225	0.0563	3.8208
Test			p	
Overall model evaluation				
Likelihood ratio test			0.1970	
Score test			0.3749	
Wald test			0.6352	

TABLE LIV

LOGISTIC REGRESSION ANALYSIS USING ONLY FACTORS: NORTH BRANCH SYSTEM

Parameter	β	SE (β)	p	VIF
Intercept	-3.1729	0.2632	<0.0001	0
Factor 1	-0.1268	0.3585	0.7237	1.9066
Factor 2	0.2473	0.3258	0.4479	2.1148
Factor 3	-0.0925	0.4592	0.8403	1.5462
Factor 4	-0.1330	0.4258	0.7548	1.5159
Test			p	
Overall model evaluation				
Likelihood ratio test			0.7782	
Score test			0.7774	
Wald test			0.7858	

and more complicated environment in the Cal-Sag Channel could also influences the separations between factors.

In the comparison between the North Branch upstream and downstream sites, separation of a rain factor and CSO factor is only observed in the downstream system. The downstream sites also show stronger separations between water treatment plant factor (indicator factor), rain factor, and CSO factor. The North Branch System is a narrow channel with limited traffic, majorly canoes or kayak. It does not have heavy commercial boat traffic as in the Cal-Sag Channel. This characteristic of the North Branch System can explain the reason why the potential sources impacting the water quality can be well explained by the plant, rain, and CSO factors.

TABLE LV

LOGISTIC REGRESSION ANALYSIS USING ONLY WATER PARAMETERS: NORTH BRANCH DOWNSTREAM SITES

Parameter	β	SE (β)	p	VIF
Intercept	4.2228	13.8357	0.7602	0
Somatic coliphages	2.5541	1.2196	0.0362	7.8942
Male-specific coliphages	-0.8857	0.7043	0.2085	5.7118
D.O.	0.4778	0.3854	0.2150	2.8347
p.H.	-0.6932	1.7987	0.6999	1.8267
Water Temperature	-0.5334	0.2571	0.0381	8.8915
Turbidity	0.0762	0.0951	0.4232	3.9877
Last CSO	0.0050	0.0046	0.2766	6.6496
Last rain	-0.0251	0.0142	0.0778	3.8427
CSO-Mag	-188E-12	1.26E-10	0.1351	12.9773
CSO-Duration	0.0176	0.0089	0.0476	5.9846
Rain-Intensity	2.9250	3.8550	0.4480	3.7844
CSO-intensity	9.83E-9	9.44E-9	0.2977	3.7768
Rain-Duration	-0.1173	0.1302	0.3677	3.3321
Rain-Mag	0.0325	0.4946	0.9476	7.4105
Conductivity	-0.0013	0.0026	0.6263	3.4363
E. coli	0.1995	0.9326	0.8306	6.0555
Enterococci	-1.3104	1.6977	0.4402	4.5753
Sun-Cum	0.4265	0.2235	0.0564	3.7941
Test			p	
Overall model evaluation				
Likelihood ratio test			0.2559	
Score test			0.4349	
Wald test			0.6717	

TABLE LVI

LOGISTIC REGRESSION ANALYSIS USING ONLY FACTORS: NORTH BRANCH DOWN STREAM SITES

Parameter	β	SE (β)	p	VIF
Intercept	-3.1933	0.2684	<0.0001	0
Factor 1	-0.1388	0.3084	0.6525	1.5343
Factor 2	0.1927	0.2568	0.4532	1.7300
Factor 3	-0.0310	0.4469	0.9447	1.3060
Factor 4	-0.1879	0.4103	0.6469	1.4207
Test			p	
Overall model evaluation				
Likelihood ratio test			0.7962	
Score test			0.8047	
Wald test			0.8115	

5.4.2 Implications of the Factor Analysis

In the downstream of North Branch System, plant factor is the strongest contributor of FIB. This finding can be validated once the disinfection of plant discharge is implemented. If the FA model can explaining the plant factor correctly, with the implementation of disinfection the domination of the plant factor is expected to be reduced.

In order to protect the health of water users, it is important to distinguish the human fecal contamination from animal waste in urban storm runoff. Several studies have demonstrate a strong relationship between FIB contributed from sewage outfalls and the rates of illness among water users (78; 15). However, our knowledge of the association between GI illness rates and FIB concentrations in water bodies without known source of domestic sewage is limited.

Fleisher et al. (79) addressed this issue in an epidemiological study and concluded that the risk of GI illness among exposed group is not significantly higher than the non-exposed group. In this study, we differentiated between point source (PS), such as the plant, and non-point source (NPS), such as rain, using factor analysis and examined how can PS and NPS predict GI illness rates. In the findings, the separation between CSO and rain factors in the downstream North Branch System suggests CSO input as human fecal contamination, and rain factor as non-point source urban runoff. A logistic regression model can then examine the association between NPS FIB concentrations and risks of illness by evaluating if the rain factor is a strong predictor.

5.4.3 Strengths

This is the first application of an exploratory factor analysis for source identification of microbial water quality data. This study shows that EFA is able to identify the latent variables using observed water quality variables. The factors identified may be related to particular sources affecting the CAWS. For example, indicator factor could be the plant source and rain factor could be the urban runoffs. The finding of strong indicator factor in the downstream locations agreed with the CMB model results (Chapter 4).

In comparison to the CMB model, EFA does not require any prior knowledge of the flow or source emission. Consequently, EFA can be applied more widely to systems with complicated flow patterns.

Majority of the physical phenomena of the environment are related. Correlation in these data was amplified by the "matching" of one measurement to many participants. This can cause multicollinearity situation and result in the parameter estimates changing inconsistently in respond to small changes in the data or the model. Furthermore, computer software packages will not be able to calculate parameter estimates with a high degree of multicollinearity. This study shows that EFA provides a way of capturing important information that is contained in latent variables.

5.4.4 Limitations

This study shows that the EFA can identify latent variables that are impacting the water quality. It is important to validate the finding before drawing any conclusion. One way to verify it is to compare the results to fecal Bacteroides spp. QPCR assay designed by Converse et al. (80). This method has been proved that it can detect human fecal contamination even when high concentration of animal contamination is in present, especially bird feces, which is the dominant animal contamination in the Great Lakes area. A well agreement between human fecal contamination from QPCR analysis and indicator factors can prove that the indicator factor is plant source in the system.

Exploratory factor analysis is a method requiring a large number of variables and observations. The preferred ratio of factors to variables is at least 1:4. Due to the constraint of sample size, the comparison between upstream and downstream in the Cal-Sag Channel could not be analyzed. *Giardia* and *Cryptosporidium* were excluded from the analysis because of the small number

of observations. However, *Giardia* and *Cryptosporidium* are the pathogens that cause illness among water users. Therefore, the exclusion of these two pathogens limits the potential link with health outcomes. Furthermore, for a system with more complex sources, more variables are needed to identify these sources. For example, in order to examine the source of boat traffic in the Cal-Sag Channel, information such as traffic counts, noise pollution, or even concentrations of liquid petroleum hydrocarbon for potential oil spills would be useful.

The association between FIB concentrations and rates of GI illness using factors as predictors remained unexplained. Possible reasons will be discussed in Chapter 6.

CHAPTER 6

DISCUSSION

6.1 Multiple Imputation On Microbial Water Quality Data

This is the first study to utilize multiple imputation to fill in missing microbial density values. In these data (n= 1,123), microbial density values were missing due laboratory quality issues: A total of 27% of E. coli and 38% of enterococci density values were missing.

Multiple imputation techniques are evaluated at two levels. On the first level, the method is evaluated with respect to how well the imputed data match the true values of the missing data. On the second level, the method is evaluated with respect to the bias introduced into coefficients of statistical models fitted with the imputed data relative to the coefficients fitted with the true data. To address both of these levels, I identified a subset of the dataset for which all observations were complete (n=573) , and introduced two patterns of missingness to these data, with an overall missingness of 24%. I found that the imputed values were similar to the true values (Table XIII and Table XIV). Coefficients in a logistic regression model of GI illness fitted with the computed data had bias less than 2% relative to those fitted with the true, complete data (Table XXIII). This strategy was taken because in a true application, the missing values are unknown. As a result, I evaluated the performance of multiple imputation

for the original data set (n = 1,123), and found that the probability distribution of the data after imputation was similar to that of the data before imputation.

Overall, this analysis suggests that multiple imputation may introduce some bias in the parameter estimates of statistical models. However, for these data as missingness of 24% was associated with bias of less than 2%. It may be that the bias introduced, however, is outweighed by the benefit of a larger sample size. This work suggests that multiple imputation may be a useful technique for the treatment of missing microbial water quality data. Other investigators considering the use of MI, should evaluate (1) the amount of missing values in the data set, (2) the missing mechanism, and (3) the analytical method that will be applied on the data set to determine if the sample size benefits of MI outweigh the bias introduced by MI, or the bias introduced by other methods.

6.2 Comparisons between EFA and CMB model

Objective 2 and 3 in this study explored the use of CMB model and EFA as alternatives for predicting health outcomes due to microbial problem. I hypothesized that sources of fecal contamination identified by CMB model can predict health outcomes, and the factors impacting the variance of microbial concentrations identified by EFA have an association with health outcomes.

The EFA is a method analyzing mathematical relationship of variance between sample to sample. Variables that are associated with potential factors have to be included in the EFA model in order for the extracted factors to make physical sense. Bzdusek et al. (81) applied EFA to

apportion the sources of PAHs and identified the impacting factors as the power plant, coke oven, gasoline engine, diesel engine, etc. The extracted factors are known for the emission of PAHs. Therefore, the source profiles can well explain the source attribution.

Variables which believed to be linked to potential factors in CAWS, such as WRP factor, rain factor, or CSO factor, were included in this study for EFA modeling. The EFA identified the clusters of indicator microbes, CSO events, and rain (Table XLV, Table XLVI, Table XLVII). However, a cluster of rain and indicator microbes, or CSO and indicator microbes was not observed. The lack of relationship indicates that the method needs other variables that can well associate with potential factors such as rain and CSO events.

The CMB model uses measured microbial concentrations to define sources in the system. It explains the biological association between microbial concentrations and pollutant sources in the system. This is the first study utilizing the CMB model to evaluate sources of microbial contaminations in a water system. The results show that CMB model perform well in the North Branch System for source attributions correctly predicting high and low CSO days (Table XXXI). Future studies are desired for the validation of CMB model results in the North Branch System. Such a study would determine the possibility of using CMB model to monitor water quality regularly in the CAWS. The proposed disinfection of wastewater "plant" inputs will be a natural experiment to evaluate the CMB model. Before disinfection is taking place, one can utilize CMB model to predict the water quality with limited plant impact in the model. The results can then be compared to the observed concentrations after disinfection under the same weather conditions. Microbial source tracking is another approach to validate the model

performance because genetic fingerprints could be used to differentiate the sources. Once the model is validated, it can be utilized to monitor the impacts of different sources in the system, and provide knowledge to direct mitigation strategies and new management.

The results in the Cal-Sag Channel have more impact from the background source than in the North Branch System, and the CSO events are not as well predicted in the Cal-Sag Channel than in the North Branch System. There are two possible explanations. First, the Cal-Sag Channel is a deep and wide channel (around 200 feet wide by 26 feet deep) with heavy commercial boat traffic. Second, the average flows from Lake Michigan into the Cal-Sag Channel is three times more than the North Branch System, according to data file provided by the MWRDGC. The Cal-Sag Channel might have more pollutant sources than what were modeled in this study. Further studies should explore the improvement of including source of boat traffic in explaining the total microbial concentrations in the Cal-Sag Channel. This approach requires the knowledge of traffic counts, noise pollution, or concentrations of liquid petroleum hydrocarbon for potential oil spills.

6.3 Pollutant Sources in CAWS and Potential Mitigations

The three major sources identified by the CMB model in the CAWS are plant source, rain source, and CSO source. The plant source of microbial concentrations would be reduced once the disinfection required by U.S. EPA is implemented. Approaches of the reduction of bacterial concentrations from urban runoff include educating pet owners to pick up pet waste from the street, controlling the wildlife populations, and the use of design of constructed wetlands. The

CSO source can be reduced through a new design separating stormwater drainage and sewage system. However, the work would require a lot of time and money.

6.4 Pollutant Sources as Health Predictors

The CMB model and the EFA were used to identify pollutant sources that are impacting FIB concentrations in the CAWS. The sources predicted by both approaches were then used as predictors to model rates of GI illness among water users. The association between sources and water users health was not identified in either approach in this study. Possible explanations are water exposure, dose-response relationship, data collection method, and study design.

The amount of water exposure during limited contact water activities could be a cause of the lack of association between FIB concentrations and GI illness rates. Previous studies that identified an association between FIB concentrations and health risks were focused on swimmers (11; 12; 13; 14). On average, limited contact water users ingest three times less water than swimmers during the activities (82) and therefore, the water exposure might be too low to observe any association.

Dose-response curves are in general S-shaped, which indicates that there is a threshold dose above which an effect manifests and also a maximum effect dose above which the effect stops increasing and remains constant. The CAWS receives wastewater directly from water treatment plants without disinfection and has higher concentration of pathegens than other recreational waterbodies in previous epidemiological studies. Awareness of the water quality could have caused some participants to be cautious while recreating and helped to significantly reduce

exposures, while other participants, who were not as cautious, might have high-dose exposure due to the high counts of microbes. These two types of exposures fall within the range of the two flat ends of the dose-response curve. Consequently, no dose-response relationship is observed.

Water samples were collected at times and locations where participants entered the water. However, participants could spend hours on the water, traveling long distances. Since we do not have the information of the location where water exposures happened, the measured FIB concentrations might not necessary relate to the water quality participants experienced while recreating. In addition, part of the survey data relied on self-reported information, such as water ingestion, onset time of the symptoms, and severity of symptoms which are susceptible to recall bias.

This study is a non-randomized cohort study with confounders. Examples of confounders in our study could be a group of CAWS users who exercise regularly and as a result are healthier than the general population; or a group of users who are aware of the water quality and minimize exposure. Even though efforts have been made to reduce the confounders during data collection and analysis, unknown confounders might be the cause of the unexplained association between water quality and health risks.

Conclusions made from this study regarding the association between water quality data and water users health outcomes are limited. Future studies should focus on estimating the water exposure for limited contact water activities, conducting water sampling during and near

where water exposures happen, distinguishing human and animal fecal contamination in urban stormwater, and evaluating rates of GI illness among water users on surface waters with and without point source pollution.

APPENDICES

Appendix A

DESCRIPTIVE STATISTICS OF WATER QUALITY PARAMETERS BY LOCATION

Appendix A (Continued)

TABLE LVII

WATER QUALITY PARAMETERS

Locations	AL				
Variables	Ave	SD	Median	Min	Max
$\log_{10}(E.\ coli)$	2.13	0.85	2.30	-1	4.40
$\log_{10}(Enterococci)$	1.91	0.64	1.85	0.30	3.76
$\log_{10}(Somcoli)$	2.29	0.54	2.34	1	3.83
$\log_{10}(Malcoli)$	0.93	0.65	0.85	0	2.72
DO (mg/L)	6.38	1.29	5.96	4.48	9.01
pH	6.83	0.34	6.83	5.74	7.52
Turbidity (NTU)	23.28	13.94	21.13	7.42	110
Conductivity (mmho/cm)	935.56	240.44	913.00	405	1434
Water temp (°C)	23.64	2.98	23.40	17.5	30.1
Solar radiation (W/m^2)	4.98	3.31	4.30	0.04	11.82
Last CSO (hour)	426.85	362.30	308.80	0	1365.8
CSO magnitude (gallon)	4E08	1E08	1E06	1E05	5E09
CSO duration (hour)	14.86	30.82	0.73	0.25	121.07
CSO interval (hour)	8E06	1E07	5E06	5E04	5E07
Last rain (hour)	84.98	81.95	58.00	0	282
Rain magnitude (inch)	2.52	6.44	0.31	0.01	27.94
Rain duration (hour)	11.06	13.82	7.00	1	61
Rain interval (hour)	0.10	0.12	0.05	0.005	0.46
Locations	BR				
Variables	Ave	SD	Median	Min	Max
$\log_{10}(E.\ coli)$	2.31	0.84	2.26	-1	4.78
$\log_{10}(Enterococci)$	2.11	0.67	2.09	0.16	4.35
$\log_{10}(Somcoli)$	1.49	0.79	1.00	0	3.97
$\log_{10}(Malcoli)$	0.36	0.80	0	-1	3.62
DO (mg/L)	8.34	3.23	7.94	0.96	19.09
pH	7.10	0.57	7.02	6.17	9.38
Turbidity (NTU)	14.76	11.48	12.54	4.53	85.34
Conductivity (mmho/cm)	476.69	379.42	304.00	184	1813
Water temp (°C)	19.16	4.59	20.40	3.3	28.6
Solar radiation (W/m^2)	3.20	2.78	2.32	0	10.4
Last CSO (hour)	158.74	173.22	92.43	0	980.3
CSO magnitude (gallon)	4E09	9E09	3E08	8E03	3E10
CSO duration (hour)	46.95	72.47	18.07	0.03	282.12
CSO interval (hour)	4E07	7E07	6E07	8E03	2E08
Last rain (hour)	53.94	64.47	35.00	0	353
Rain magnitude (inch)	3.30	6.82	0.59	0.01	25.27
Rain duration (hour)	14.12	16.20	7.00	1	67
Rain interval (hour)	0.15	0.18	0.08	0.003	0.97

Appendix A (Continued)

TABLE LVIII

WATER QUALITY PARAMETERS (CONTINUED-1)

Locations	BA				
Variables	Ave	SD	Median	Min	Max
$\log_{10}(E.\ coli)$	1.96	0.76	2.07	-1	4.15
$\log_{10}(Enterococci)$	1.48	0.62	1.46	0.09	3.45
$\log_{10}(Somcoli)$	1.40	0.64	1.00	0	3.29
$\log_{10}(Malcoli)$	0.38	0.76	0	0	3.37
DO (mg/L)	7.08	1.82	7.50	3.58	11.15
pH	7.22	0.47	7.15	6.02	8.09
Turbidity (NTU)	27.21	14.12	25.20	10.95	69.2
Conductivity (mmho/cm)	660.58	281.60	571.50	311	1096
Water temp (°C)	22.32	2.09	22.45	18.3	26.5
Solar radiation (W/m^2)	3.94	3.68	2.43	0.01	10.83
Last CSO (hour)	348.83	312.92	277.58	0	1365.8
CSO magnitude (gallon)	7E08	1E09	4E07	1E05	5E09
CSO duration (hour)	21.53	39.69	1.12	0.25	121.07
CSO interval (hour)	1E07	2E07	5E06	5E04	5E07
Last rain (hour)	80.65	79.44	55.00	0	277
Rain magnitude (inch)	4.19	8.91	0.33	0.01	27.94
Rain duration (hour)	14.49	18.36	7.00	1	61
Rain interval (hour)	0.13	0.16	0.06	0.004	0.46
Locations	SK				
Variables	Ave	SD	Median	Min	Max
$\log_{10}(E.\ coli)$	2.80	0.86	2.83	1.08	4.76
$\log_{10}(Enterococci)$	2.33	0.75	2.24	-1	4.50
$\log_{10}(Somcoli)$	1.86	1.02	1.48	0	3.81
$\log_{10}(Malcoli)$	0.65	0.90	0	0	2.45
DO (mg/L)	7.11	2.01	7.40	2.06	10.99
pH	6.92	0.57	6.88	5.44	8.3
Turbidity (NTU)	19.73	8.74	19.92	4.84	41.97
Conductivity (mmho/cm)	459.21	411.87	312.00	131.8	2470
Water temp (°C)	21.55	4.51	21.40	6.7	33.2
Solar radiation (W/m^2)	4.30	2.75	4.32	0	11.09
Last CSO (hour)	146.64	135.76	120.88	0	664.53
CSO magnitude (gallon)	5E09	1E10	1E08	8E03	3E10
CSO duration (hour)	46.04	68.48	28.02	0.03	209.4
CSO interval (hour)	5E07	8E07	6E07	8E03	2E08
Last rain (hour)	46.90	60.14	28.50	0	352
Rain magnitude (inch)	2.35	4.12	1.29	0.01	25.27
Rain duration (hour)	12.75	12.52	8.00	1	68
Rain interval (hour)	0.17	0.21	0.08	0.005	0.87

Appendix A (Continued)

TABLE LIX

WATER QUALITY PARAMETERS (CONTINUED-2)

Locations	LA				
Variables	Ave	SD	Median	Min	Max
$\log_{10}(E.\ coli)$	3.60	0.60	3.69	-1	4.66
$\log_{10}(Enterococci)$	2.86	0.54	2.87	1.41	4.32
$\log_{10}(Somcoli)$	3.31	0.43	3.36	0	4.00
$\log_{10}(Malcoli)$	1.97	0.47	1.98	0	3.88
DO (mg/L)	7.78	1.68	7.68	2.77	13.81
pH	6.80	0.37	6.78	5.91	7.9
Turbidity (NTU)	12.01	7.95	10.60	1.69	45.02
Conductivity (mmho/cm)	809	291	730	315	1682
Water temp (°C)	20.00	4.30	21.20	6.5	26.1
Solar radiation (W/m^2)	3.64	2.83	2.77	0.01	11.09
Last CSO (hour)	151	167	85.28	0	980
CSO magnitude (gallone)	5E09	1E10	7E08	8E03	3E10
CSO duration (hour)	49.29	72.92	19.80	0.03	282.12
CSO interval (hour)	4E07	7E07	7E07	8E03	2E08
Last rain (hour)	55.78	62.95	37.00	0	353
Rain magnitude (inch)	3.49	7.10	1.01	0.01	25.27
Rain duration (hour)	14.38	16.45	7.00	1	68
Rain interval (hour)	0.16	0.19	0.08	0.003	0.97
Locations	CP				
Variables	Ave	SD	Median	Min	Max
$\log_{10}(E.\ coli)$	3.56	0.59	3.58	-1	4.63
$\log_{10}(Enterococci)$	2.70	0.41	2.70	1.63	3.80
$\log_{10}(Somcoli)$	3.15	0.38	3.16	1.48	3.91
$\log_{10}(Malcoli)$	1.75	0.49	1.79	0	3.07
DO (mg/L)	6.88	1.70	6.49	4.83	12.45
pH	6.80	0.35	6.77	5.79	7.71
Turbidity (NTU)	16.17	10.66	12.61	5.1	51.36
Conductivity (mmho/cm)	807	221	793	441	2080
Water temp (°C)	24.16	5.09	24.50	15.1	37.6
Solar radiation (W/m^2)	5.23	2.79	5.02	0.31	10.77
Last CSO (hour)	175	157	133	0	611
CSO magnitude (gallon)	4E09	9E09	3E08	1E05	3E10
CSO duration (hour)	61.53	90.08	28.02	0.07	282.12
CSO interval (hour)	3E07	5E07	6E07	2E05	2E08
Last rain (hour)	52.72	49.32	36.00	0	222
Rain magnitude (inch)	0.84	0.88	0.44	0.01	2.75
Rain duration (hour)	6.96	5.89	5.00	1	25
Rain interval (hour)	0.13	0.16	0.09	0.006	0.87

Appendix A (Continued)

TABLE LX

WATER QUALITY PARAMETERS (CONTINUED-3)

Locations	RP				
Variables	Ave	SD	Median	Min	Max
$\log_{10}(E.\ coli)$	3.05	0.60	2.88	2.07	4.19
$\log_{10}(Enterococci)$	2.67	0.51	2.47	1.69	4.10
$\log_{10}(Somcoli)$	2.82	0.46	2.73	2	3.88
$\log_{10}(Malcoli)$	1.58	0.96	1.49	0	4.10
DO (mg/L)	7.10	1.68	6.65	4.92	11.25
pH	7.08	0.31	7.10	6.21	7.62
Turbidity (NTU)	29.17	17.40	22.87	7.69	62.88
Conductivity (mmho/cm)	1155	712	1048	478	2410
Water temp (°C)	22.69	3.10	23.49	17.3	28.6
Solar radiation (W/m^2)	4.39	2.03	4.29	0.92	9.47
Last CSO (hour)	104	115	73.81	0	469
CSO magnitude (gallone)	8E09	1E10	1E09	2E06	3E10
CSO duration (hour)	65.15	79.04	28.02	2.92	209.4
CSO interval (hour)	7E07	9E07	4E07	5E05	2E08
Last rain (hour)	74.2	60.52	64	0	162
Rain magnitude (inch)	4.23	8.46	0.79	0.01	25.27
Rain duration (hour)	14.93	16.80	9	1	54
Rain interval (hour)	0.18	0.24	0.88	0.01	0.87
Locations	WO				
Variables	Ave	SD	Median	Min	Max
$\log_{10}(E.\ coli)$	2.07	0.71	2.12	-1	3.43
$\log_{10}(Enterococci)$	1.64	0.60	1.62	0.17	3.54
$\log_{10}(Somcoli)$	2.07	0.53	2.14	1	3.72
$\log_{10}(Malcoli)$	0.75	0.60	0.78	0	2.73
DO (mg/L)	6.67	1.31	6.53	4.48	9.74
pH	6.89	0.46	6.89	5.74	7.9
Turbidity (NTU)	23.31	16.72	19.32	8.5	133
Conductivity (mmho/cm)	919	239	897.50	336	1364
Water temp (°C)	24.87	3.49	24.80	15.9	31.4
Solar radiation (W/m^2)	4.98	3.47	4.36	0.05	12.14
Last CSO (hour)	459	415	307.63	0	2057
CSO magnitude (gallon)	4E08	1E09	2E06	1E05	5E09
CSO duration (hour)	15.75	32.63	0.73	0.25	121.07
CSO interval (hour)	6E06	1E07	5E06	5E04	4E07
Last rain (hour)	80.17	76.30	54.50	0	282
Rain magnitude (inch)	2.97	6.77	0.33	0.005	27.94
Rain duration (hour)	11.80	15.05	6.00	1	61
Rain interval (hour)	0.12	0.14	0.06	0.003	0.46

Appendix A (Continued)

TABLE LXI

WATER QUALITY PARAMETERS (CONTINUED-4)

Locations	NAM				
Variables	Ave	SD	Median	Min	Max
$\log_{10}(E.\ coli)$	3.10	0.73	3.01	-1	5.23
$\log_{10}(Enterococci)$	2.42	0.58	2.36	1.23	4.46
$\log_{10}(Somcoli)$	2.95	0.53	2.92	1.78	4.82
$\log_{10}(Malcoli)$	1.44	0.55	1.30	0	3.27
DO (mg/L)	6.36	1.67	6.03	3.36	11.76
pH	6.78	0.49	6.76	5.2	8.2
Turbidity (NTU)	15.90	6.08	14.41	8.93	45.17
Conductivity (mmho/cm)	744.95	204.02	771	392	1474
Water temp (°C)	22.83	5.80	22.4	5.6	36.9
Solar radiation (W/m^2)	4.96	3.03	4.40	0	10.72
Last CSO (hour)	223.83	197.44	209.08	0	983.3
CSO magnitude (gallon)	5E08	2E09	1E08	1E05	2E10
CSO duration (hour)	28.46	54.06	11.20	0.07	282.12
CSO interval (hour)	2E07	5E07	2E07	5E05	2E08
Last rain (hour)	49.68	50.61	37.00	0	223
Rain magnitude (inch)	1.02	2.53	0.43	0.01	25.27
Rain duration (hour)	8.13	8.29	5.00	1	54
Rain interval (hour)	0.12	0.17	0.07	0.003	0.87
Locations	RM				
Variables	Ave	SD	Median	Min	Max
$\log_{10}(E.\ coli)$	2.72	1.01	2.96	-1	4.26
$\log_{10}(Enterococci)$	1.76	0.69	1.78	0.01	3.66
$\log_{10}(Somcoli)$	2.67	0.40	2.70	1.70	3.56
$\log_{10}(Malcoli)$	1.21	0.56	1.23	0	2.56
DO (mg/L)	7.59	2.34	7.06	5.09	13.43
pH	6.98	0.43	6.84	5.94	7.74
Turbidity (NTU)	15.82	10.96	12.14	6.5	54.54
Conductivity (mmho/cm)	830.9	237.03	854	478	1285
Water temp (°C)	24.04	4.50	23.80	17.7	33.8
Solar radiation (W/m^2)	4.73	3.66	3.84	0.04	10.88
Last CSO (hour)	374.63	291.24	308.13	0	1366
CSO magnitude (gallon)	6E08	2E09	2E06	1E05	5E09
CSO duration (hour)	19.64	38.46	0.73	0.25	121.07
CSO interval (hour)	8E06	1E07	5E06	5E04	5E07
Last rain (hour)	77.13	80.30	46.50	0	281
Rain magnitude (inch)	3.83	8.65	0.31	0.01	27.94
Rain duration (hour)	12.97	18.09	6.00	1	61
Rain interval (hour)	0.14	0.16	0.06	0.004	0.46

CITED LITERATURE

1. Pond, K.: Water Recreation and Disease. Plausibility of Associated Infections: Acute Effects, Sequelae and Mortality. Alliance House, 12 Caxton Street, London SW1H 0QS, UK, International Water Association (IWA) Publishing, 2005.

2. Shuval, H.: Estimating the global burden of thalassogenic diseases: human infectious diseases caused by wastewater pollution of the marine environment. Journal of Water Health, 1(2):53 – 64, 2003.

3. U.S.EPA: Evaluation of multiple indicator combinations to develop quantifiable relationships. EPA-822-R-10-004, 2010.

4. Boehm, A. B., Fuhrman, J. A., Mrse, R. D., and Grant, S. B.: Tiered approach for identification of a human fecal pollution source at a recreational beach: Case study at avalon bay, catalina island, california. Environmental Science and Technology, 37:673 – 680, 2003.

5. Noblet, J. A., Young, D. L., Zeng, E. Y., and Ensari, S.: Use of fecal steroids to infer the sources of fecal indicator bacteria in the lower santa ana river watershed, california: sewage is unlikely a significant source. Environmental ScienceTechnology , 38(22):6002–6008, 2004. PMID: 15573599.

6. He, L.-M.: The missing link in bacteria tmdls: Natural growth. Technical report, Water Environmental Foundation, County of San Diego/AmTech International, San Diego, CA 92172, 2006.

7. Dwight, R. H., Fernandez, L. M., Baker, D. B., Semenza, J. C., and Olson, B. H.: Estimating the economic burden from illnesses associated with recreational coastal water pollution - a case study in orange county, california. Journal of Environmental Management, 76:95 – 103, 2005.

8. Gobel, P., Dierkes, C., and Coldwey, W. G.: Storm water runoff concentration matrix for urban areas. Journal of Contaminant Hydrology, 91(1-2):26 – 42, 2007.

9. Ellis, J. B. and Yu, W.: Bacteriology of urban runoff: The combined sewer as a bacterial reactor and generator. Water Science and Technology, 31(7):303 – 310, 1995.

10. U.S.EPA: Fecal bacteria. <http://water.epa.gov/type/rsl/monitoring/vms511.cfm>, February 2010. [Online; accessed 05/12/2011].

11. Medema, G., van Asperen, I., and Havelaar, A.: Assessment of the exposure of swimmers to microbiological contaminants in fresh waters. Water Science and Technology, 35(11-12):157 – 163, 1997. ¡ce:title¿Health-Related Water Microbiology 1996¡/ce:title¿ ¡xocs:full-name¿Selected Proceedings of the IAWQ 8th International Symposium on Health-related Water Microbiology 1996¡/xocs:full-name¿.

12. Cheung, W., Chang, K., Hung, R., and Kleevens, J.: Health effects of beach water pollution in hong kong. Epidemiological Infections, 105:139 – 162, 1997.

13. Balarajan, R., Raleigh, V. S., Yuen, P., Wheeler, D., Machin, D., and Cartwright, R.: Health risks associated with bathing in sea water. British Medical Journal, 303:1444 – 1445, 1991.

14. Wade, T. J., Pai, N., Eisenberg, J. N., and Jr., J. M. C.: Do u.s. environmental protection agency water quality guidelines for recreational waters prevent gastrointestinal illness? a systematic review and meta-analysis. Environmental Health Perspectives, 111(8), 2003.

15. Marion, J., Lee, J., Lemeshow, S., and Buckley, T.: Association of gastrointestinal illness and recreational water exposure at an inland U.S. beach. Water Research, 44(16):4796 – 4804, 2010.

16. Wang, M.-C., Liu, C.-Y., Shiao, A.-S., and Wang, T.: Ear problems in swimmers. Journal of the Chinese Medical Association, 68(8):347 – 352, 2005.

17. Wong, M., Kumar, L., Jenkins, T. M., Xagoraraki, I., Phanikumar, M. S., and Rose, J. B.: Evaluation of public health risks at recreational beaches in Lake Michigan via detection of enteric viruses and a human-specific bacteriological marker. Water Research, 43(4):1137 – 1149, 2009.

18. Medema, G. J., van Asperen, I. A., and Havelaar, A. H.: Assessment of the exposure of swimmers to microbiological contaminants in fresh waters. Water Science and Technology, 35(11-12):157 – 163, 1997.

19. Outdoor Industry Association Outdoor Foundation: A special report on paddlesports. <http://www.outdoorindustry.org/research.php?action=detail&research_id=79>, 2009. [Online; accessed 04/26/2011].

20. U.S. Department of Homeland Security and U.S. Coast Guard: Recreational boating statistics 2009. <http://www.uscgboating.org/assets/1/workflow_staging/Publications/394.PDF>, July 2010. [Online; accessed 04/26/2011].

21. U.S.EPA: Chicago area waterway system. <http://wiki.epa.gov/watershed2/index.php/Chicago_Area_Waterway_System>, February 2010. [Online; accessed 11/16/2010].

22. ILEPA: Chicago-area-waterways. <http://www.epa.state.il.us/mailman/listinfo/chicago-area-waterways>. [Online; accessed 11/16/2010].

23. Friends of the Chicago River Foundation: Why disinfect. <http://www.chicagoriver.org/upload/Why_Disinfect_brochure_r4b.pdf>, 2009. [Online; accessed 04/29/2011].

24. Prairie Research Institute: Illinois state water survey. <http://www.isws.illinois.edu/data.asp>, 2011. [Online; accessed 05/26/2011].

25. Rubin, D. B.: Inference and missing data. Biometrika, 63:581–592, 1976.

26. Schafer, J. L.: Multiple imputation: a primer. Statistical Methods In Medical Research, 8(1):3–15, 1999.

27. Rubin, D.: Multiple Imputation of Nonresponse in Surveys. New York, Wiley, 1987.

28. Little, R. J. and Rubin, D. B.: Statistical Analysis with Missing Data. Hoboken, NJ, Wiley, 2 edition, 2002.

29. Collins, L. M., Schafer, J. L., and Kam, C.-H.: A comparison of inclusive and restrictive strategies in modern missing data procedures. Psychological Methods, 6:330–351, 2001.

30. Enders, C. K.: A primer on the use of modern missing-data methods in psychosomatic medicine research. Psychosomatic Medicine, 68:427–436, 2006.

31. Sartori, N., Salvan, A., and Thomaseth, K.: Multiple imputation of missing values in a cancer mortality analysis with estimated exposure dose. Computational Statistics and Data Analysis, 49(3):937–953, 2005.

32. Allison, P. D.: Multiple imputation for missing data: a cautionary tale. Sociological Methods and Research, 28:301–309, 2000.

33. Yuan, Y. C.: Multiple Imputation for Missing Data: Concepts and New Development. SAS Institute Inc., 1700 Rockville Pike, Suite 600, Rockville, MD 20852, 1.0 edition.

34. Schafer, J. L.: Analysis of incomplete multivariate data. New York, NY, Chapman and Hall, 1997.

35. Richman, M. B., Trafalis, T. B., and Adrianto, I.: J3.9 multiple imputation through machine learning algorithms.

36. Vapnik, V. N.: Statistical learning theory. New York, NY, Springer Verlag, 1998.

37. Haykin, S.: Neural Networks: a comprehensive foundation. New York, NY, Prentice Hall, 2 edition, 1999.

38. Zhou, X., Eckert, G., and Tierney, W.: Multiple imputation in public health research. Statistics In Medicine, 20(9-10):15–30, May 2001.

39. Chen, H., Quandt, S. A., Grzywacz, J. G., and Arcury, T. A.: A distribution-based multiple imputation method for handling bivariate pesticide data with values below the limit of detection. Environ Health Perspect, 119(3), 11 2010.

40. Whitman, R. L. and Nevers, M. B.: Summer *E. coli* patterns and responses along 23 chicago beaches. Environmental Science and Technology, 42(24), 2008.

41. Nevers, M. B. and Whitman, R. L.: Efficacy of monitoring and empirical predictive modeling at improving public health protection at chicago beaches. Water Research, 45:1659–1668, 2011.

42. Boehm, A. B., Grant, S. B., Kim, J. H., Mowbray, S. L., McGee, C. D., Clark, C. D., Foley, D. M., and Wellman, D. E.: Decadal and shorter period variability of surf zone water quality at huntington beach, california. Environmental ScienceTechnology , 36(18):3885–3892, 2002.

43. Schafer, J. L. and Graham, J. W.: Missing data: Our view of the state of the art. Psychological Methods, 7(2):147 – 177, 2002.

44. Domingo, J. W. S., Bambic, D. G., Edge, T. A., and Wuertz, S.: Quo vadis source tracking? towards a strategic framework for environmental monitoring of fecal pollution. Water Research, 41(16):3539 – 3552, 2007. Identifying Sources of Fecal Pollution.

45. Xie, Y. and Berkowitz, C. M.: The use of positive matrix factorization with conditional probability functions in air quality studies: An application to hydrocarbon emissions in houston, texas. Atmospheric Environment, 40(17):3070 – 3091, 2006.

46. Chung, J., Wadden, R. A., and Scheff, P. A.: Development of ozone-precursor relationships using voc receptor modeling. Atmospheric Environment, 30(18):3167 – 3179, 1996.

47. Friend, A. J., Ayoko, G. A., and Guo, H.: Multi-criteria ranking and receptor modelling of airborne fine particles at three sites in the pearl river delta region of china. Science of The Total Environment, 409(4):719 – 737, 2011.

48. Henry, R. C. and Lewis, C. W.: Review of receptor model fundamentals. Atmospheric Environment, 18(8):1507–1515, 1984.

49. Malm, W. C. and Gebhart, K. A.: Source apportionment of sulfur and light extinction using receptor modeling techniques. Journal of the Air and Waste Management Association, 47(3):250–258, 1997.

50. Gebhart, K. A., Malm, W. C., and Flores, M.: A preliminary look at source-receptor relationships in the texas-mexico border area. Journal of the Air and Waste Management Association, 50(5):858–868, 2000.

51. Yannopoulos, P. C.: Sulfur dioxide dispersion and source contribution to receptors of downtown patras, greece. Environmental Science and Pollution Research International, 14(3):172–175, 2007.

52. U.S.EPA: Receptor modeling. <http://www.epa.gov/scram001/receptorindex.htm>, May 2010. [Online; accessed 05/12/2011].

53. Macias, E. S. and Hopke, P. K.: Atmospheric Aerosol : Source and Air Quality Relationships, volume 167. Washington, D.C., American Chemical Society, 1981.

54. Chelani, A. B., Gajghate, D. G., and Devotta, S.: Source apportionment of pm10 in mumbai, india using cmb model. Bulletin of Environmental Contamination and Toxicology, 81(2):190–195, 2008.

55. U.S.EPA: Chemical mass balance (CMB) model. <http://www.epa.gov/ttn/scram/receptor_cmb.htm>, May 2010. [Online; accessed 08/02/2010].

56. Li, A., Jang, J.-K., and Scheff, P. A.: Application of EPA CMB8.2 model for source apportionment of sediment PAHs in Lake Calumet, Chicago. Environmental Science and Technology, 37:2958–2965, 2003.

57. Saada, C. M.: Source Reconciliation of Volatile Organic Contaminants In Ground Water. Master's thesis, Illinois Institute of Technology, December 1988.

58. Coulter, C. T.: EPA-CMB8.2 Users Manual. U.S. EPA, Research Triangle Park, NC 27711, December 2004.

59. Edge, T. A. and Schaefer, K. A.: Microbial source tracking in aquatic ecosystems: The state of the science and an assessment of needs. Technical Report 7, National Water Research Institute, Burlington, Ontario, 2006.

60. Wilson, T. L., Stahnke, G., Schnabel, W., and Duddleston, K.: Seasonal Variations in Fecal Coliform Bacteria in a Cold Regrion Stream. World Water and Environmental Resources Congress, 2005.

61. Wilkesa, G., Edgeb, T., Gannonc, V., Jokinenc, C., Lyauteyd, E., Medeirose, D., Neumannf, N., Rueckerg, N., Toppd, E., and Lapen, D. R.: Seasonal relationships among indicator bacteria, pathogenic bacteria, Cryptosporidium oocysts, Giardia

cysts, and hydrological indices for surface waters within an agricultural landscape. Water Research, 43:2209–2223, 2009.

62. Lipp, E. K., Kurz, R., Vincent, R., Rodriguez-Palacios, C., Farrah, S. R., and Rose, J. B.: The effects of seasonal variability and weather on microbial fecal pollution and enteric pathogens in a subtropical estuary. Estuaries, 24(2):266–276, 2001.

63. Kleinbaum, D. G., Kupper, L. L., Nizam, A., and Muller, K. E.: Applied Regression Analysis and Other Multivariable Methods, 4ed. 10 Davis Drive, Belmont, CA 94002, Thomson Higher Education, 2008.

64. U.S.EPA: National water quality inventory: Report to congress, 2002 reporting cycle: Findings, rivers and streams, and lakes, ponds and reservoirs. Technical report, U.S. Environmental Protection Agency, 2002.

65. Ram, J. L., Thompson, B., Turner, C., Nechvatal, J. M., Sheehan, H., and Bobrin, J.: Identification of pets and raccoons as sources of bacterial contamination of urban storm sewers using a sequence-based bacterial source tracking method. Water Research, 41(16):3605 – 3614, 2007. Identifying Sources of Fecal Pollution.

66. Dorsey, J. H., Carter, P. M., Bergquist, S., and Sagarin, R.: Reduction of fecal indicator bacteria (fib) in the ballona wetlands saltwater marsh (los angeles county, california, usa) with implications for restoration actions. Water Research, 44(15):4630 – 4642, 2010.

67. Wikipedia: Storm drain. `<http://en.wikipedia.org/wiki/Storm_drain>`, September 2011. [Online; accessed 09/20/2011].

68. Loehlin, J. C.: Latent variable models, an introduction to factor, path, and structural analysis. Mahwah, NJ, Lawrence Erlbaum Associations, 3 edition, 1998.

69. Curran, P. J., West, S. G., and Finch, J. F.: The robustness of test statistics to nonnormality and specification error in confirmatory factor analysis. Psychological Methods, 1(1):16 – 29, 1996.

70. Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., and Strahan, E. J.: Evaluating the use of exploratory factor analysis in psychological research. Psychological Methods, 4(3):272 – 299, 1999.

71. Suhr, D. D.: Principal component analysis vs. exploratory factor analysis, 2003.

72. Shrestha, S. and Kazama, F.: Assessment of surface water quality using multivariate statistical techniques: A case study of the fuji river basin, japan. Environmental Modeling and Software, 22:464–475, 2007.

73. M, V. and B, S.: Assessment of surface water quality using multivariate statistical techniques: a case study of behrimaz stream, turkey. Environmental Monitoring and Assessment, 159(1-4):543–553, December 2009.

74. Kannel, P. R., Kanel, S. R., Lee, S., and Lee, Y.-S.: Chemometrics in assessment of seasonal variation of water quality in fresh water systems. Environmental Monitoring and Assessment, May 2010.

75. Li, S. and Zhang, Q.: Risk assessment and seasonal variations of dissolved trace elements and heavy metals in the upper han river, china. Journal of Hazardous Materials, 181(1-3):1051 – 1058, 2010.

76. Rijal, G., Petropoulou, C., Tolson, J. K., DeFlaun, M., Gerba, C., Gore, R., Glymph, T., Granato, T., O'Connor, C., Kollias, L., and Lanyon, R.: Dry and wet weather microbial characterization of the chicago area waterway system. Water Science and Technology, pages 1847–1855, 2009.

77. Dorevitch, S.: Chicago health, environmental exposure, and recreational study (cheers). Technical report, University of Chicago at Illinois, 2010.

78. Cabelli, V. J., Dufour, A. P., Levin, M. A., McCabe, L. J., and Haberman, P. W.: Relationship of microbial indicators to health effects at marine bathing beaches. American Journal of Public Health, 69(7):690 – 699, 1979.

79. Fleisher, J. M., Fleming, L. E., Solo-Gabriele, H. M., Kish, J. K., Sinigalliano, C. D., Plano, L., Elmir, S. M., Wang, J. D., Withum, K., Shibata, T., Gidley, M. L., Abdelzaher, A., He, G., Ortega, C., Zhu, X., Wright, M., Hollenbeck, J., , and Backer, L. C.: The beaches study: Health effects and exposures from non-point source microbial contaminants in subtropical recreational marine waters. International Journal of Epidemiology, 39:1291 – 1298, 2010.

80. Converse, R. R., Blackwood, A. D., Kirs, M., Griffith, J. F., and Noble, R. T.: Rapid qpcr-based assay for fecal bacteroides spp. as a tool for assessing fecal contamination in recreational waters. Water Research, 43(19):4828 – 4837, 2009. Cross-validation of detection methods for pathogens and fecal indicators.

81. Bzdusek, P. A., Christensen, E. R., Li, A., and Zou, Q.: Source apportionment of sediment pahs in lake calumet, chicago: Application of factor analysis with nonnegative constraints. Environmental Science and Technology, 38(1):97 – 103, 2004.

82. Dorevitch, S., Panthi, S., Huang, Y., Li, H., Michalek, A. M., Pratap, P., Wroblewski, M., Liu, L., Scheff, P. A., and Li, A.: Water ingestion during water recreation. Water Research, 45(5):2020 – 2028, 2011.

83. Abdi, H.: Factor Rotations in Factor Analyses. The University of Texas at Dallas, The University of Texas at Dallas, Richardson, Texas 75083.

84. Goldberg, R. J.: Proc Factor: How to Interpret the Output of a Real-World Example. Guildline Research/Atlanta, Inc., 3675 Crestwood Parkway, N.W., Suite 520, Duluth, GA 30136.

85. Tanriverdi, C., Alp, A., Demirkiran, A. R., and Uckardes, F.: Assessment of surface water quality of the ceyhan river basin, turkey. Environmental Monitoring and Assessment, 167:175–184, June 2009.

86. Li, A., Jang, J.-K., and Scheff, P. A.: Application of epa cmb8.2 model for source apportionment of sediment pahs in lake calumet, chicago. Environmental Science and Technology, 37:2958–2965, 2003.

87. Baccarlli, A., Pfeiffer, R., Consonni, D., Pesatori, A. C., Bonzini, M., Jr, D. G. P., Bertazzi, P. A., and Landi, M. T.: Handling of dioxin measurement data in the presence of non-detectable values: Overview of available methods and their application in the seveso chloracne study. Chemosphere, 60(7):898–906, 2005.

88. Zhang, Z., Tao, F., Du, J., Shi, P., Yu, D., Meng, Y., and Sun, Y.: Surface water quality and its control in a river with intensive human impacts-a case study of the xiangjiang river, china. Journal of Environmental Management, In Press, Corrected Proof:–, 2010.

89. Converse, R. R., Piehler, M. F., and Noble, R. T.: Contrasts in concentrations and loads of conventional and alternative indicators of fecal contamination in coastal stormwater. Water Research, 45:5229 – 5240, 2011.

VITA

NAME: Chiping Nieh

EDUCATION: Ph.D., Public Health Sciences, University of Illinois at Chicago, 2011

M.S., Environmental Sciences and Engineering, The University of North Carolina at Chapel Hill, 2002

B.S., Forestry, National Taiwan University, 1999

EMPLOYMENT: Department of Environmental and Occupational Health Sciences, University of Illinois at Chicago, Chicago Illinois, Research Assistant, August 2007 - August 2011

Division of Environmental Health and Occupational Medicine, National Health Research Institute, Miaoli Taiwan, Research Associate, August 2006 - July 2007

Factor Engineering and Technology Corporation, Taipei Taiwan, Project Manager, September 2005 - July 2006

Environmental Sciences and Engineering, The University of North Carolina at Chapel Hill, Chapel Hill North Carolina, Laboratory Technician, September 2003 - August 2005

PRESENTATIONS:

Chiping Nieh, Sam Dorevitch, Li Liu, Peter Scheff, "An Assessment of Water Quality Using Factor Analysis," EPA National Beach Conference 2011, Miami Florida, 2011

Chiping Nieh, Sam Dorevitch, Li Liu, Peter Scheff,"An Evaluation of Multiple Imputation of Missing Values In Surface Water Measurements," Environmental Health Conference 2011, Salvador Brazil, 2011

Chiping Nieh, Serap Erdal, "A Risk Assessment of Acrylamide Intake from Drinking Coffee," Society for Risk Analysis Annual Meeting 2008, Boston Massachusetts, 2008

Chiping Nieh, Kuen-Yuh Wu, "Factors Associated With Sensitivity of Input Parameters in the Assessment of Health Risk from Dietary Intakes of Carbendazim," Society for Risk Analysis Annual Meeting 2007, San Antonio Texas, 2007

Chiping Nieh, Kuen-Yuh Wu, "A Probabilistic Risk Assessment of Carbendazim from Dietary Intake of Taiwanese Consumers Using 2D Monte Carlo Simulation and Bootstrap Methods," Conference of Industrial Hygiene and Occupational Medicine 2007, Kaohsiung Taiwan, 2007

Kuen-Yuh Wu, Chiping Nieh, Su-Yin Chiang, "A Probabilistic Risk Assessment of Carbamate Residues in Fruits and Vegetables for Consumers in Taiwan," Society for Risk Analysis Annual Meeting 2006, Baltimore Maryland, 2006

Chiping Nieh, Lori A. Todd, "An Evaluation of the Accuracy of Measuring Chemical Emissions by Using Environmental CAT Scanning System," Environmental Analytical Chemistry Conference 2006, Taichung Taiwan, 2006

PUBLICATION: Chiping Nieh, Chyi-Rong Chiou, "An Introduction of Forest Health Monitoring System in the USA," The Journal of Taiwan Forestry, 26(3), 2000

REPORTS: Daphne L. Stoner, William F. Bauer, Ryan Murray, Lori Todd, Kathleen Mottus and Chiping Nieh, 2009, "The Search for Life on Worlds Around Other Stars: A Spectroscopic Analysis of Volatile Biogenic Signatures Emanating from Geothermal Environments," Final Report Grant No. NNG04GM37G

Aneja, V. P., S. Pal Arya, I. Rumsey, Deug-Soo Kim, Wayne Robarge, David Dickey, Len Stefanski, Lori Todd, K. Mottus, K. Bajwa, H. Semunegus, S. Goetz, W. Stephens, Chiping Nieh, 2005, "Evaluation of Environmentally Superior Technologies for Am-

monia Project OPEN," Final Report, Animal and Poultry Waste Management Center, North Carolina State University, Raleigh, NC, p. 69

AWARDS:

President's Award, National Taiwan University, Taipei Taiwan, 1998

Graduate Student Council Travel Award, UIC, Chicago Illinois, 2008

Rodney P. Musselman Travel Award, UIC School of Public Health, Chicago Illinois, 2011

Graduate Student Council Travel Award, UIC, Chicago Illinois, 2011

AFFILIATIONS:

Member, Golden Key International Honour Society

Member, American Industrial Hygiene Association

Using Mass Balance, Factor Analysis and Multiple Imputation to Assess Health Effects of Water Quality

Chiping Nieh, Ph.D.
Department of Environmental and Occupational Health Sciences
University of Illinois at Chicago
Chicago, Illinois (2012)

Dissertation Chairperson: Peter A. Scheff

This dissertation explores the use of three analytical methods to improve the utility of microbial water quality data, collected on the Chicago River from 2007 to 2009, in predicting health risk among water users. The Multiple Imputation(MI) method was applied to fill in microbial missing values and the ability of the method to reduce bias was evaluated, chemical mass balance model and exploratory factor analysis were then utilized to identify sources of fecal contamination in the river system. Sources Identified as contributing to fecal contamination were used in predicting health risk.

The results showed that by introducing a 2% bias to the parameter estimates, the MI method was able to recover 24% of missing data. However, in order to fill in 36% of missing values, 33% of bias was introduced. Chemical mass balance model and exploratory factor analysis both identified the water reclamation plant, combine sewer overflows (CSOs), and the precipitation as sources of fecal contamination in the river system. However, no association between pollutant sources and health risk were observed.