Multiple Linear Regression Models for Estimating Microbial Load in the Drinking Water Source Case from Glomma River, Norway

Fasil Ejigu Eregno

Norwegian University of Life Sciences
Department of Mathematical Sciences and Technology
Master Thesis 30 credits 2013

Multiple Linear Regression Models for Estimating Microbial Load in a Drinking Water Source Case from the Glomma River, Norway

A thesis submitted in partial fulfillment of the requirements for the Master of Science degree in Environment and Natural Resources - Specialization Sustainable Water and Sanitation, Health and Development

By
Fasil Ejigu Eregno

Supervised By
Ass.Prof Arve Heistad

December, 2013

Department of Mathematical Sciences and Technology (IMT)
Norwegian University of Life Sciences (UMB)

Abstract

The application of integrated study of water quality and statistics for environmental modelling is considered as a powerful analytical tool that has been thrived significantly during recent years. The present study was conducted to identify the significant physico-chemical factors that affects the raw water quality, and to study statistical interrelationships amongst them. Multiple linear regression models were developed to estimate microbial load in the raw water source, using data from the NRV drinking water treatment plant published from 1999 to 2012 and also from Norwegian school of veterinary science through VISK project. The study was conducted based on indicator microbial load which contain Total viable count "Kimtall", *Coliform bacteria*, *Escherichia coli*, *Clostridium perfringens*, and *Intestinal Enterococci*. In addition, microbial pathogen load of Noro virus, and Adeno virus were also incorporated. The explanatory variables examined for regression analysis were monitored properties of raw water and hyro-climatic data from the catchment which include; river discharge, raw water temperature, rainfall, pH, turbidity, conductivity, colour, and total organic carbon. Each indicator and pathogenic microbial loads have its own unique set of selected explanatory variables. The statistical significance tests were applied to the coefficients of the multiple linear regression models, and they are found to be significant. The regression equations were evaluated using measures of variability, including adjusted R^2, which ranges from 38.0 % for Adeno virus concentration to 50.0 % for *Ecoli* concentration. The results revealed that the regression analysis provide useful mean for rapid monitoring of microbial raw water quality based on the physico-chemical parameters.

Acknowledgements

I am very grateful that I was given the opportunity to pursue my master degree in Norwegian University of Life Science (UMB), and funds from VISK project to carry out this study and write a master thesis about modelling of microbial quality of source water.

First and foremost I would like to thank my superb supervisor Prof. Arve Heistad (Department of Mathematical Sciences and Technology, Norwegian University of Life Sciences) for his invaluable comments and continuous guidance throughout the study and for making all these possible. I am also grateful to Dr. Razak Seidu (Department of Mathematical Sciences and Technology, Norwegian University of Life Sciences) for his guidance, closer supervision and encouragement throughout the study. My thanks are extended to Vegard Nilsen (PhD student at Department of Mathematical Sciences and Technology, Norwegian University of Life Sciences) for his expertise advice and critical comments.

My gratitude also goes to Lena Solli Sal (Project Manager / Operations Engineer at NRV / NRA IKS) for her collaboration in providing water quality data from NRV drinking water treatment plant. I wish to gratefully acknowledge Ricardo Grøndahl-Rosado for his collaboration in providing raw water viral load data set. I also thank Svein Taksdal (Head of hydro-informatics Section at NVE) for providing river discharge data.

My special thanks are due to my beloved family, my wife, Mebrat Gebreslassie, who had always been with me when I need help more than ever, my children, Amanuel, Michias, and Yohana who always gives me pleasure and strength. I would like to thank my parents and friends for their continuous encouragement during my study.

Declaration

I, Innocent Fasil Ejigu do hereby declare to Norwegian University of Life Science that, this Thesis is my original work and that it has never been submitted for a degree award in any other University.

Signature…….....

Date.....

All Right Reserved

Table of Contents

List of Tables and Figures

List of Symbols and Abbreviations

AIC	Akaike's Information Criteria
ANOVA	Analysis of variance
MSE	Mean Square Error terms of residual
MSR	Mean Square error terms of Regression
NRV	Nedre Romerike Vannverk
R^2	coefficient of determination
RMSE	Root Mean Square Error
SBC	Schwarz Bayesian Criteria
SST	Total Sum of Squares
SSE	Sum of Squared Errors,
TVC	Total Viable Count "Kimtall"
VIF	Variance Inflation Factor
VISK	Reduced vulnerability to waterborne viral infection

1. INTRODUCTION

1.1 Background

Surface water is widely used as a source for drinking water production. There is a wide range of microbial and chemical constituents of drinking water that can cause either acute or chronic detrimental health effects. Besides, water of poor quality can also be harmful from an economic perspective, as resources have to be directed towards improving the water supply system. For these reasons, there is growing pressure to improve water treatment and water quality management at catchment scale in order to ensure safe drinking water at reasonable costs (Astrom et al. 2007b; Won et al. 2013).

Pathogens present in surface waters originate from both point and diffuse sources and concentrations may vary considerably over time. Point sources for pathogens may include municipal wastewater discharges and heavily polluted tributaries within a river system. Diffuse sources, on the other hand, include urban, agricultural and forestry runoffs with microbial impact from livestock and wild animals in the catchment area. Furthermore, the microbial load to the raw water within a catchment is influenced by natural factors, such as climatological parameters (rain, sunlight and temperature), hydrology and topography (Kinzelman et al. 2004; Mills & Thurman 1994).

To produce high-quality drinking water from surface water, the contaminants in the raw water such as physical, chemical and microbial contaminants must be removed by the water treatment process. The performance of a water treatment plant is highly related to the characteristics of the raw-water entering the plant. To optimize the treatment processes and thus provide good quality potable water in an economical manner, the ability to predict the raw-water quality over time is desired by the water treatment industry. This would allow advanced warning of changes in raw-water quality which require alternation of process conditions (Astrom et al. 2007a; Han et al. 2012; Sedmak et al. 2005).

Analytical tools must be developed to properly evaluate raw water quality, adapt management practices and predict water quality improvement or deterioration at different catchment scales. In this regard, an integrated study of water quality and statistics for environmental modelling has grown significantly during recent decades. However, fewer systematic studies have been undertaken to model and predict the microbial raw water quality based on available physic-chemical parameters to assess the level of health risks related to drinking water production

and to improve catchment management practices (Kubeck et al. 2009; Zhang & Stanley 1997).

Among modelling approaches, multiple linear regression analysis is a statistical tool used to examine relationships among variables. It provides a method for quantifying the impact of changes in one or more explanatory variables (known as independent variables) on a variable of interest (known as the dependent variable). Regression analysis is widely used in the field of econometrics, finance, sociology, hydrology, biology, psychology, pharmacology, and engineering, among other fields of study (Fedotovai et al. 2013; Hasani & Shanbeh 2010; Moustris et al. 2012; Noller & Whitehouse 1982; Noorossana et al. 2010; Seidou & Ouarda 2007). In this paper, we perform a multiple linear regression analysis and discuss a number of applications in the microbial water quality context.

1.2 Objectives of the study

This project aims to improve modelling of microbial load of source water by taking into account the physico-chemical parameters. The main objectives of this research are:

1. To identify the specific physico-chemical factors most associated with the specific indicator microorganisms and / or microbial pathogen load in the raw water.
2. To build and evaluate, for each indicator microorganisms and microbial pathogens, multiple regression models that predict microbial load of raw water, using physico-chemical factors as independent variables.

1.3 Structure of the thesis

To overcome the proposed objectives, the present thesis is structured as follows. Following brief background information, Part 1 outlines the objectives of the study. Part 2 serve as a general review of microbial water quality, source of contamination, monitoring and modelling issues. Part 3 reports the methodology used to achieve the designed goal. The results of the study have also been discussed more concisely and critically in Part 4. Finally in Part 5 which is the concluding chapter of the thesis have been highlighted.

2. LITRATURE REVIEW

2.1 Microorganisms in drinking water sources

Drinking water comes from surface water and ground water sources. Large-scale water supply systems tend to rely on surface water resources, and smaller water systems tend to use ground water. Surface water includes rivers, lakes, and reservoirs. On the other hand, ground water is pumped from wells that are drilled into aquifers. Usually surface water has to undergo many more purification steps than groundwater to become suited to drink (Bociort et al. 2012; Davies & Mazumder 2003).

The most common and widespread health risk associated with drinking water sources are contamination, either directly or indirectly through human, animal and occasionally bird faeces and with the microorganisms contained in their faeces. Contamination problems also arise from improperly designed, failing, or overloaded waste water treatment systems, including septic systems from private homes, and leaking sanitary sewer pipes. Floodwater commonly contains high levels of bacteria from numerous sources. (Bociort et al. 2012). An understanding of microbial quality of source waters is essential, because it facilitates selection of the highest quality water source for drinking-water supply, and provides a basis for establishing treatment requirements to meet health based targets. The occurrence of pathogens and indicator organisms in raw water sources depends on a number of factors, including intrinsic physical and chemical characteristics of the catchment area and the magnitude and range of human activities and animal sources that release pathogens to the environment. In surface waters, potential pathogen sources include point sources, such as municipal sewerage and urban storm water overflows, as well as non-point sources, such as contaminated runoff from agricultural areas and areas with sanitation through onsite septic systems and latrines. Other sources are wildlife and direct access of livestock to surface water bodies. Many pathogens in surface water bodies will reduce in concentration due to dilution, settling and die-off due to environmental effects (thermal, sunlight, predation, etc.) (Obasohan et al. 2010; Payment et al. 2000).

In a bid to mitigate such risks to human health by contaminated surface waters, monitoring, assessing, and managing microbiological quality of surface waters is an unending process. Such assessment and monitoring of the microbiological quality of surface waters involve identifying the main sources of fecal microorganisms by analysing river water samples for traditional faecal indicator bacteria; Escherichia coli, intestinal enterococci, and spores of

Clostridium perfringens, and in some cases the test targets specific pathogen (Nnane 2011). The pathogenic organisms of concern include bacteria, viruses and protozoa. The diseases they cause vary in severity from mild gastroenteritis, to severe and sometimes fatal diarrhoea, dysentery, hepatitis, cholera, typhoid fever and campylo-bacteriosis (Farkas et al. 2013).

The multiple barrier approach to providing safe drinking water includes source water protection, treatment, and maintenance of distribution system integrity. Development of watershed management strategies relies on an understanding of the impact of watershed activities and land uses on receiving water quality. Controlling the risks related to these pathogens is a permanent challenge for the water industry. The supply of safe drinking-water involves the use of multiple barriers to prevent the entry and transmission of pathogens. The effectiveness of these multiple barriers should be monitored by a programme based on operational characteristics and testing for microbial indicators of faecal contamination and in some circumstances actual pathogens (Plummer & Long 2007). In addition to the constantly evolving range of pathogens to consider, assessing and managing such risks requires the integration of information issued by a wide range of disciplines.

2.2 Sources of microbial contaminants and its preventive measures

The first step in protecting a public water supply is the development of a watershed or wellhead protection program. Controlling or eliminating microbial sources before they contaminate a water supply will go a long way toward simplifying treatment and reducing costs associated with a contaminated supply. The following are sources of microbial contamination within a water supply protection area and suggested protection measures aimed at reducing the risk they pose to drinking water (Canada 2006; Okoh et al. 2007).

2.2.1 Sewage Disposal Systems

Wastewater collection and treatment systems vary from community to community depending on the population size and local needs. Such systems may separate the storm and sanitary flows, or have a combined sewer system, or both. Wastewater collection and treatment systems are responsible for collecting and treating residential, commercial and industrial wastewater. All of the practices and procedures used to collect and treat wastewater have the potential to pollute surface and subsurface drinking water sources. Failing sewage disposal systems represent the major source of microbial contamination from human waste. Contamination of drinking water sources by sewage can occur from raw sewage overflow, septic tanks, leaking sewer lines, land application of sludge and partially treated waste water.

Sewage itself is a complex mixture and can contain many types of contaminants. Seepage overflow into drinking water sources can cause disease from the ingestion of microorganisms (Ritter et al. 2002).

2.2.1.1 Raw Sewage Overflow

Storm water systems in urban areas are sometimes combined with sanitary sewer systems en route to sewage treatment plants. Excessive storm water can cause this joint system to overflow. In this event, excess flow will be directed into waterways untreated, resulting in sewage contamination. Urban runoff is usually collected by a separate storm sewer system and discharged directly into waterways. Combined systems are cheaper, but the potential to harm health is higher. Some systems have diversions to accommodate heavy flow (Even et al. 2007; Walker 1994).

2.2.1.2 Septic Tanks

Septic tanks are enclosures that store and process wastes where no sewer system exists, such as in rural areas or on boats. Treatment of waste in septic tanks occurs by bacterial decomposition. The resulting material is called sludge. Large portions of the population are still served by septic systems as opposed to public waste treatment facilities. Contamination of water from septic tanks occurs under various conditions (Cheung & Venkitachalam 2004; Khwaja et al. 1999):

- *Poor placement of septic leach fields* can feed partially treated waste water into a drinking water source. Leach fields are part of the septic system for land based tanks and include an area where waste water percolates through soil as part of the treatment process.
- *Badly constructed percolation systems* may allow water to escape without proper treatment.
- *System failure* can result in clogging and overflow to land or surface water.
- *High density placement of tanks*, as in suburban areas, can result in regions containing very high concentrations of waste water. This water may seep to the land surface, run-off into surface water or flow directly into the water table.

There are also site specific environmental factors around the tank and leach field such as soil properties, water table location, subsurface geology, climate, and vegetation which may affect the quality and quantity of released waste water.

2.2.1.3 Leakage from Sewer Lines

Effluent that leaks from sewer lines is generally untreated raw sewage. It may contain industrial waste chemicals. When leaking sewer lines are located deep underground below the biologically active portion of the soil, the sewage can enter groundwater directly. This can result in the introduction of chlorides, microorganisms, organics, trace metals and other chemicals that may cause disease and foul tastes or odours in drinking water. Sewer leaks can occur from tree root invasion, soil slippage, seismic activity, loss of foundation due to washout, flooding and sewage back up, among other events. High pressure systems will push leaks to the soil surface where they can be easily detected by sight or odor. Systematic inspection of sewer lines, exclusion of hazardous waste, and adherence to modern construction and maintenance specifications are necessary preventative measures for protection of groundwater sources from sewer leaks (Eiswirth & Hotzl 1997).

2.2.1.4 Land Application of Partially Treated Waste Water and Municipal Sludge

Sludge is the residue of the chemical, biological, and physical treatment of municipal and industrial wastes. It can be applied to land as fertilizer or as fill. Land application is an alternative to incineration, which causes air pollution. Sludge usually contains concentrated organic matter, nitrogen, inorganic salts, heavy metals, and bacteria. It is a common practice to use partially treated waste water for fertilization, irrigation, and water supply recharge as an alternative to direct discharge into waterways. Waste water is also commonly stored in wells, holes, trenches, open pits and lagoons. Movement and percolation of waste water through the soil biologically and physically removes biodegradable substances, pathogenic organisms, and inorganic substances (Gerba & Smith 2005; Okoh et al. 2007). The effectiveness of this treatment depends upon:

- *Processing or turnover time:* Waste water must spend a sufficient amount of time on or within the soil to allow for filtration and biological processes to degrade the waste. If sufficient time is not allowed for these treatment processes to bring down contaminant levels before introducing waste water to a water system, contamination will occur.
- *Excess waste water and high concentrations of contaminants in the waste water:* High concentrations of waste can take much longer to treat, especially when the consistency reaches that of a slurry or sludge. On the other hand, irrigation of soil with large quantities of waste water will saturate the soil and overload the biological degradation

process. Excess untreated waste water can run off or percolate down to groundwater, causing contamination of drinking water supplies.

- *Level of biological processing:* Lack of appropriate microbial activity can slow the degradation process or provide insufficient treatment. Bacteria which break down wastes without the use of oxygen, known as anaerobic bacteria, are very important in the process of breaking down nitrogen containing substances. Aerobic bacteria, which use oxygen, break down organic waste. Some of the breakdown products include water, carbon dioxide, methane gas, nitrates and other small organic and inorganic substances.

In order to prevent microbial contamination of drinking water sources by sewage disposal system, the following measures are recommended

- Implement proper planning for sewage systems within the watershed.
- Ensure septic systems are inspected and serviced on a regular basis.
- Promote public education on how to care for a septic system.

2.2.2 Agriculture

Non-point sources of pollution from agricultural endeavours have been identified as the greatest contributors to water quality degradation. In order for transmission of agricultural pathogens to humans to occur through contaminated water the pathogen must be excreted by livestock, must reach the waterway in a viable form, must remain viable and virulent in the environment, and the concentration of the pathogen must be sufficient to cause infection when encountered by humans. Runoff carrying animal waste from barnyards, manure storage areas, dairy farms, poultry farms, pig farms, pastures, and the land application of manure is a significant source of microbial contamination (Baudisova 2009; Edge et al. 2012; Gerba & Smith 2005).

The best management practices include storing liquid manure in sealed bottom facilities, applying manure to fields only when ground is thawed, following appropriate application rates and timing, maintaining buffer strips between agricultural fields and waterways, fencing animals away from waterways, installing subsurface drainage tiles around agricultural fields, and preventing runoff from farmyards (Baudisova 2009).

2.2.3 Storm water Runoff

One of the overriding issues associated with the delivery of microbes to surface waters is nonpoint source pollution, and more specifically, storm water runoff from sub urban area. Rainwater and snowmelt flow over the land picking up pollutants and deposit them into water supplies. Runoff can also pick up microbial contaminants from suburban environments such as pet waste on sidewalks (Geldreich 1989; He et al. 2010; Karlaviciene et al. 2009; Sidhu et al. 2013).

- Minimize impervious surfaces within your watershed.
- Install catch basins and settling basins to slow down flows and filter out contaminants.
- Use landscaping techniques that conserve water and limit runoff such as native plants, low maintenance grasses, shrubs, rock gardens, etc.
- Require the proper removal and disposal of pet waste.

2.2.4 Wildlife

Wildlife is an integral part of a balanced watershed. However, birds and mammals can introduce microorganisms into a water supply either through direct contact or from watershed runoff. *Giardia*, *cryptosporidium*, *salmonella*, *campylobacter*, and *Escherichia coli* (*E.coli*) are the most commonly identified microorganisms found in mammals and birds. Wildlife commonly associated with microbial contamination of drinking water supplies include: deer, beavers, muskrats, gulls, and geese (Bishop et al. 2000; Cimenti et al. 2007).

The following protection measures should not be implemented without a good understanding of the nuisance wildlife population in question. These protection measures should not be considered as general practice but should be carefully deployed in specific areas of a water supply protection area, for example, near an intake or in areas where a nuisance wildlife population is concentrated (Ritter et al. 2002).

- Monitor wildlife populations in and around water supplies.
- Keep up a daily human presence along the shoreline.
- Employ scare techniques such as pyrotechnics.
- Modify habitat (shoreline fencing, mowing, landscaping changes, and tree branch pruning to reduce bird roosting).
- Prohibit the public from feeding wildlife, especially waterfowl.

- Reduce food sources such as palatable plant species.
- Keep beavers and muskrats from building dams/dens by installing fencing or drainage devices.
- Consider permitted trapping or hunting.

2.3 Microbial water quality Monitoring

Monitoring microbial water quality has been conducted for more than a century by measuring indicator bacteria that occupy human intestinal systems, primarily fecal coliforms, *Escherichia coli*, and some *Enterococci*. Technological advances described in provide new opportunities for revising these monitoring procedures. Our increased understanding of microbiology at the molecular level allows existing indicators to be measured using faster and cheaper methods. These advances also provide cost-effective opportunities for measuring new indicators or combinations of indicators, and in some cases, pathogens themselves (Devereux et al. 2006).

2.3.1 Indicator microorganism

The number and variety of microbial agents that might be present in source water is considerable. The routine monitoring for all the possibilities is either impossible or impractical. The solution to the problem has been the use of indicator microorganisms that would be present when potential pathogen containing material was present. Indicator organisms are microorganisms whose presence in water indicates probable presence of pathogens (disease-causing organisms). Ideally, such microorganisms are non-pathogenic, occur consistently in pathogen-contaminated water, do not multiply in waters, are reliably detectable even at low concentrations, rapidly detected, easily enumerated, have survival characteristics that are similar to those of the pathogens of concern, and are present in greater numbers than and have similar survival times to pathogens (Scott et al. 2002). It should be emphasized that the presence of indicator bacteria does not mean the water contains pathogenic microorganisms but rather the potential exists for the presence of pathogens since the indicator bacteria point to the presence of fecal material in the sample. In addition, the number of pathogens that might be associated with the concentration of the indicator will be a function of the disease incidence in the community at the time the fecal material was disposed. The indicators microorganisms used to analyse water quality are Total viable count "Kimtall", Coliforms, *Escherichia coli*, Enterococci, and *Clostridium perfringens* were chosen because of their efficacy at predicting pathogen presence, and have higher resistance

to environmental stresses and disinfection. Definition of some indicator microorganisms that are included in this study is as follows (folkehelseinstitutt 2004; Hirata et al. 1991);

2.3.1.1 Total viable count "Kimtall"

Waters of all kinds invariably contain a variety of microorganisms derived from various sources such as soil and vegetation and estimation of the overall numbers provide useful information for the assessment and surveillance of water quality. Total Viable Count (TVC) gives a quantitative idea about the presence of microorganisms such as bacteria, yeast and mold in the water sample. In Norway, the method refers to "Kimtall" and the colony count at 22 °C is a measure of bacteria, yeast and mold that naturally belongs in soil and water and the count actually represents the number of colony forming units.

2.3.1.2 Coliform bacteria

Coliform bacteria are organisms that are present in the environment and in the feces of all warm-blooded animals and humans. Coliform bacteria will not likely cause illness. However, their presence in drinking water indicates that disease-causing organisms (pathogens) could be in the water system. Most pathogens that can contaminate water supplies come from the feces of humans or animals. If coliform bacteria are found in a water sample, water system operators work to find the source of contamination and restore safe drinking water. There are three different groups of coliform bacteria; each has a different level of risk. Total coliform, fecal coliform, and *E. coli* are all indicators of microbial water quality. The total coliform group is a large collection of different kinds of bacteria. Fecal coliforms are types of total coliform that mostly exist in feces. *E. coli* is a sub-group of fecal coliform. Some of these bacteria can grow during decomposition of plant residues in the soil, and some of the plant material in water. Generally the growth of these bacteria in the soil and water are best at temperature below 40 °C. The analysis of coliform bacteria is often takes place at 37 °C.

2.3.1.3 *Escherichia coli*

Escherichia coli (E. coli) bacteria normally live in the intestines of people and animals. It is gram-negative, facultative anaerobic, rod-shaped bacterium that is commonly found in the lower intestine of warm-blooded organisms. Most *E. coli* are harmless and actually are an important part of a healthy human intestinal tract. However, some *E. coli* are pathogenic, meaning they can cause illness, either diarrhea or illness outside of the intestinal tract. The types of *E. coli* that can cause diarrhea can be transmitted through contaminated water or food, or through contact with animals or persons. Still other kinds of *E. coli* are used as

markers for water contamination, which are not themselves harmful, but indicate the water is contaminated. It is the most appropriate group of coliforms to indicate faecal pollution from warm-blooded animals.

2.3.1.4 *Clostridium perfringens*

Clostridium perfringens is a bacterium that grows in the absence of oxygen; it is gram-positive, spore-forming and anaerobic bacterium. It is included in the feces of humans and animals, but in much smaller quantity. These spores survive very long in waters. If a watercourse or groundwater source has been applied feces from humans or animals, the spores will always be detected. Most of these bacteria have natural habitat in soil and sediment in the water, but can cause disease in humans and animals that get them out. Some of them can also grow in foods and cause illness. Spores can withstand more adverse environment, heat and disinfectants than the active (vegetative) bacteria do.

2.3.1.5 Intestinal enterococci

Intestinal Enterococci: are a subgroup of the larger group of organisms defined as faecal streptococci, comprising species of the genus Streptococcus. These bacteria are Gram-positive and relatively tolerant of sodium chloride and alkaline pH levels. They are facultative anaerobic and occur singly, in pairs or as short chains. Faecal streptococci including intestinal enterococci all give a positive reaction with Lancefield's Group D antisera and have been isolated from the faeces of warm-blooded animals. The subgroup intestinal enterococci consist of the species *Enterococcus faecalis, E. faecium, E. durans and E. hirae*. This group was separated from the rest of the faecal streptococci because they are relatively specific for faecal pollution. However, some intestinal enterococci isolated from water may occasionally also originate from other habitats, including soil, in the absence of faecal pollution.

2.4 Microbial water quality modelling

Due to regional and national legislation on water quality and to protect human health, the microbial pollution of catchments is an issue that requires increased attention and analysis. However, the management of microbial pollution sources at catchment scale is challenging (Jamieson et al. 2004). Analysis tools must be developed to properly evaluate alternate management practices and to predict water quality improvements at the catchments scale. Microbial water quality models can be useful tools to simulate and predict the levels, distributions, and risks of microbial pollutants in a given catchment scale and water body. The modeling results from these models under different pollution scenarios are very important

components of environmental impact assessment and can provide a basis and technique support for environmental management agencies to make right decisions (Pullar & Springer 2000).

The wide variety of waterborne pathogens that contaminate water and the lack of quantitative data concerning their origin and distribution within drinking water catchments have made the development of predictive models of pathogen loads from catchments difficult (Ferguson et al. 2005). A comprehensive understanding of the problem requires that watershed factors, including climatic conditions, hydrologic parameters, and site-specific parameters be considered in combination with anthropogenic factors (Coffey et al. 2007).

Available models for waterborne pathogens were evaluated and assessed based on a number of set criteria including: type of model (qualitative or quantitative); treatment of input variables (stochastic or deterministic); use of input data (vector or raster); ability to incorporate various input factors; ability to produce output facilities; and overall model functionality. Specific criteria including land use, meteorological conditions, and soil/geological characteristics were regarded as key risk factors for source water catchment contamination with microbial pathogens and model ability to adequately account for these were considered as important individual parameters when assessing available models (Coffey et al. 2007).

3. MATERIALS AND METHODS

3.1 Glomma River basin

The Glomma River (Fig 1) is Norway's largest river. It is located in South Eastern Norway where it covers 41,200 km^2 (13% of Norway's total area). The north-western parts consist of high mountain areas. The eastern part is covered by forest, whereas the central and southern parts comprise large agricultural areas. In total the agricultural area covers 5.8% of the catchment. The Glomma river basin contains Lake Mjøsa, the Norway's largest lake, which has a surface area of 350 km^2. The river mean annual flow at Solbergfoss (outlet of Lake Øyeren, the lowermost reservoir) is 700 m^3/s. The flow normally varies during the year from 150 to 3500 m^3/s. The river Glomma catchment comprises approximately 675,000 inhabitants. There are 8 cities, in which half of the population lives. Hydropower production is an important water use. In the Glomma catchment there are 45 hydropower stations and 26 hydropower reservoirs (Grizzetti B. 2007).

3.2 Data set

This study is based on the records of five microbial raw water quality parameters namely, total viable count "Kimtall" (TVC), *clostridium perfringens*, *intestinal enterococci*, *Escherichia coli*, and *coliform bacteria*, whose concentration were monitored at Nedre Romerike Vannverk (NRV) drinking water treatment plant in Furuhaugli Mountain at Strømmen, Norway. The report includes weakly records of raw water microbial load for *Escherichia coli*, and *coliform bacteria* from 1999 to 2013, for *intestinal enterococci* from 2002 to 2013, and for total viable count "Kimtall" and *clostridium perfringens* from 2005 to 2013. However, some records are missing and during analysis, the missing values treated as a missing data (not filled with mean or neighborhood values). In addition to these, 16 months record of virus concentration from the same raw water source were taken by Norwegian School of Veterinary Science through the Reduced Vulnerability to Waterborne Viral Infection (VISK) project and incorporated in this study. The record include Adeno virus (85 observations), Noro virus G1 (Genome-1, 71 observations), Noro virus G2 (Genome-2, 62 observations).

The selections of explanatory variables are based both on the theory and availability of data. Since the microbial pathogen concentration in the raw water reflects the overall conduciveness of the environment for the indicator and pathogenic microorganisms, it can be

explained by the physico-chemical condition of the environment, according to the theoretical basis (Crowther et al. 2001). First, in order to reflect the aspect of the environment, raw water temperature, rainfall, pH, turbidity, electrical conductivity, colour and total organic carbon are selected to represent the physico-chemical indicators of the environment. Secondly, in order to track the source area association with the microbial load, five tributary river discharge gauging station records also included. All regression analysis and graphical presentations in this study were performed by Addinsoft's XLSTAT 2012 Statistical Software.

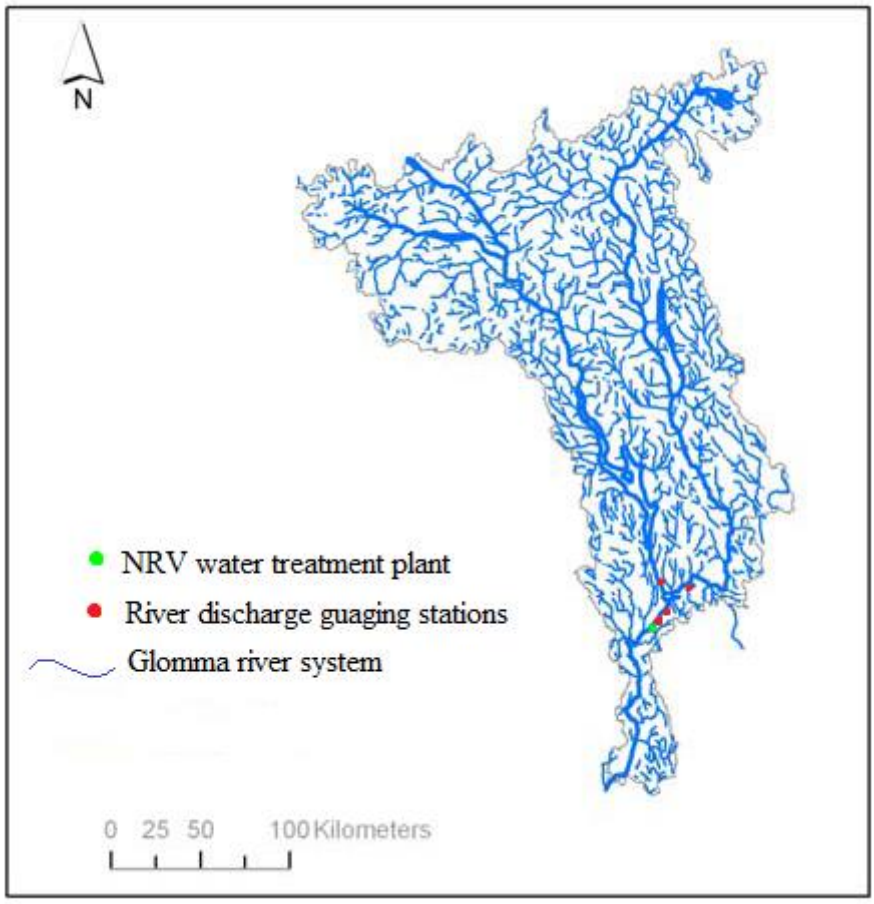


Figure 1 Study catchment showing Glomma River and main tributaries, discharge gauging stations, and NRV water treatment plant (Base map source: (Grizzetti B. 2007)).

3.3 Multiple Linear Regression Analysis

Descriptive statistics was used to describe the basic features of the data set in the study. Correlation analysis was used to examine the relations between microbial pathogen load and environmental and physico-chemical water-quality variables. A linear correlation coefficient (Pearson's r) was used to determine the degree to which variables were related to covariates. The more the coefficient differed from 1 or -1 (close to zero), the weaker the relation.

Multiple linear regression models are used to study the linear relationship between a dependent variable and several independent variables by fitting a linear equation to observed data samples (Coelho-Barros et al. 2008). The generic form of the linear regression model is

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_k x_{ki} + \varepsilon_i, \qquad \text{i =1,2,...N} \qquad (1)$$

Where y is the dependent variable, x_1, x_2..., x_k are the independent or explanatory variables, and i index the n sample observations, the term ε is a random error term. The fitting is performed by minimizing the sum of the squares of the vertical deviations from each data point to the line that best fits for the observed data (Agirre-Basurko et al. 2006; Ferraro & Giordani 2012; Kovdienko et al. 2010). We have employed a stepwise regression procedure to select the independent variables that would result in the best possible model, while at the same time ensuring statistical significance of the results. The t-statistics was used to test whether a particular variable contributes significantly to the regression model or not so as to eliminate statistically insignificant variables. The level of significance (α) for the inclusion of a variable in the model was 0.05. For the coefficient b_j of the j variable, H_0: $b_j = 0$ and H_a: $b_j \neq 0$. This t statistic can be formed as

$$t = \frac{b_j}{S_{bj}} \qquad (2)$$

where S_{bj} is the standard deviation of the respective coefficient b_j (Vounatsou & Karydis 1991). The *F*-ratio, which is computed from the mean squared terms in the Analysis of variance (ANOVA) table, estimates the statistical significance of the regression equation. The *F*-ratio is given by

$$F = \frac{MSR}{MSE} \qquad (3)$$

where MSR mean square error of regression and MSE mean square error of the residuals (Kufs 1992; Pugh et al. 2001).

3.4 Evaluation of the models

To evaluate the models we used statistical performance measures, which is included: coefficient of determination (R^2), Adjusted R^2 (R^2_{adj}), mean square error (MSE), root mean square error (RMSE), Akaike's Information Criteria (AIC), and Schwarz Bayesian Criteria (SBC). The definitions of the statistical measures of the goodness of fit used herein are the following:

$$R^2 = 1 - \frac{SSE}{SST} \quad (4)$$

$$R^2_{adj} = 1 - \frac{(n-i)SSE}{(n-k)SST} \quad (5)$$

$$MSE = \frac{SSE}{n-k} \quad (6)$$

$$RMSE = \sqrt{\frac{SSE}{n-k}} \quad (7)$$

$$AIC = n*\ln(\frac{SSE}{n}) + 2k \quad (8)$$

$$SBC = n*\ln(\frac{SSE}{n}) + k\ln n \quad (9)$$

Where SSE is the sum of squared errors, SST is total sum of squares, *n* is number of observations, k is the number of independent variables, ln is natural logarithm (Archer & Lemeshow 2006; Bedrick & Crandall 2010; Fagerland & Hosmer 2013; Kieseppa 2001; Naidu et al. 2012; Shih 1998; Stone 1979; Yang et al. 2011).

3.5 Checking Multiple Linear Regression Assumptions

In order to use the proposed multiple regression analysis, it is necessary to test and verify that the proposed equation satisfies the assumptions. Assumptions of multiple linear regression tested in this study to validate the proposed multiple regression analysis are: (1) homoscedasticity (Constant variance), nonautoregression (randomness of residuals), nonstochastic (errors are uncorrelated with the individual predictors), normality of the error distribution, were examined by plotting of the residuals against predicted values (2) multicollinearity among predictor variables were tested by Variance Inflation Factor (VIF) described in

$$VIF_J = \frac{1}{1 - R^2_{J|Others}} \quad (10)$$

Where $R^2_{J|Others}$ is multiple coefficient of determination between x_{ij} and all x_i (Ukoumunne et al. 2002)

4. RESULTS AND DISCUSSION

Multiple linear regression analysis is one of the modelling techniques that enable us to depict relationships between microbial raw water quality and physico-chemical properties by fitting a linear equation to the observed data set. In this study, an attempt has been made to establish multiple linear regression equations to provide a prediction of microbial load in the raw water based on the physico-chemical parameters.

Analyses for the presence of waterborne pathogens are extremely difficult and complicated because some pathogens cannot be cultured in the laboratory, or may be injured after exposure to stressful environments. As a result, indicator microorganisms are widely used to detect possible contamination. The study was conducted based on indicator microbial load which contain Total viable count "Kimtall", *coliform bacteria*, *Escherichia coli*, *clostridium perfringens*, and i*ntestinal enterococci.* In addition, direct monitored microbial pathogens load, namely, Noro virus, and Adeno virus were also incorporated.

The summary of descriptive statistics of the results of the analysis is presented in Table 1, indicating the mean, standard deviation, variance, skewness, kurtosis, minimum, 1st quartile, median, 3rd quartile, and maximum value. Total viable count "Kimtall" recorded the highest mean value of 1062 per ml while *clostridium perfringens* the list value of 6.6 per 100 ml. The descriptive statistical result shows that the variation of records for Total viable count and i*ntestinal enterococci* was high and the distribution of i*ntestinal enterococci* was skewed as compare with the other microbial record data. The raw water temperature in the plant ranged from 0.9 to 21.5 ^{o}C, while the pH, turbidity, conductivity, colour and total organic carbon varied from 5.7 to 7.8, 0.1 to 570 NTU, 1.3 to 9.2 mS/m, 3 to 87 mg pt/l, and 1 to 8.8 mg C/l respectively. A wide range of turbidity can be explained by the variation in runoff generated from different land use with a high tendency of washing microbial pathogens from different sources.

Table 1 Descriptive statistics of explanatory variables and raw water microbial load used for modelling

Variable	N	Mean	StDev	Variance	Skewness	Kurtosis	Min	Q1	Median	Q3	Max
Rånåsfoss (m^3/s)	411	705	375	140644	1.16	1.10	136.4	425.7	592.9	897.3	2451.2
Blaker (m^3/s)	341	646.7	325.2	105780	1.51	3.74	98.1	425.8	567.9	789.2	2471.9
Funnefoss o.vann (m^3/s)	547	367.0	190.7	36364	0.84	0.91	125.3	191.2	336.2	502.3	1243.7
Ertesekken ndf. (m^3/s)	492	355.1	200.9	40386	1.29	1.36	63.3	207.8	301.0	441.0	1110.5
Vorma (m^3/s)	385	272.6	244.7	59901	1.13	1.22	61.7	153.0	216	280.3	1153.4
Raw water Temperature (^oC)	315	8.4	5.8	34	0.35	-1.2	0.9	2.7	7.4	13.4	21.5
Rainfall (mm)	462	1.13	1.86	3.45	1.73	1.98	0.0	1.1	2.1	3.8	8.5
pH	531	7.1	0.3	0.10	-1.3	2.96	5.7	6.9	7.1	7.2	7.8
Turbidity (NTU)	530	4.6	25.7	662.1	20.36	443.5	0.1	1.1	1.9	3.4	570
Conductivity (mS/m)	527	4.2	0.8	0.69	0.26	4.68	1.3	3.9	4.3	4.6	9.2
Colour (mg Pt/l)	546	29.4	12.7	162.6	1.26	1.59	3.0	21.0	5.0	35.0	87.0
Total Organic Carbon (mg C/l)	287	4.1	1.3	1.78	0.80	0.60	1.0	3.0	3.8	4.9	8.8
Total viable count - v/22°C (count/ml)	298	1062	1764	3110893	3.9	20.2	1.0	200	420	1100	14000
clostridium perfringens (count/100ml)	302	6.6	6.8	46.6	3.1	16.6	1.0	1.0	5.0	9.0	59.0
i*ntestinal enterococci* (count/100ml)	456	71.2	938.5	880797	20.7	437.3	1.0	2.0	7.0	19.0	1986
Escherichia coli (count/100ml)	547	41.6	46.6	2168	4	34.2	1.0	10.0	30.0	55.0	579
coliform bacteria (count/100ml)	547	243.3	374.2	140023	5.2	35.1	1.0	78.0	160	260	4106
Adeno virus (count)	85	85.6	157.1	24669	3.5	14.5	0.09	4.0	26.6	100	977.8
Noro virus (g1) (count)	71	26.5	35.5	1260	2	3.6	0.23	4.8	11.9	28.5	148.8
Noro virus (g2) (count)	62	102.1	134	17945	1.7	2.3	0.18	11.4	38.9	155.7	525

Correlation analysis was used to examine the relations between physico-chemical variables and microbial water quality variables. A linear correlation coefficient (Pearson's r) was used to detect the degree of association that exists between the variables. In this study, the numerical values of the correlation coefficient, r for microbial water quality parameters and physico chemical variables are tabulated in Table 2. Highly positive correlation between the response variable and the predictor variables are found between i*ntestinal enterococci* and turbidity ($r = 0.45$, $p<0.01$), *Escherichia coli* and turbidity ($r = 0.52$, $p<0.01$), *clostridium perfringens* and conductivity ($r = 0.41$, $p<0.01$), total viable count "Kimtall" and colour ($r = 0.36$, $p<0.01$), *coliform bacteria* and turbidity ($r = 0.26$, $p<0.01$), Adeno virus and conductivity ($r = 0.47$, $p<0.01$), Noro virus G1 and conductivity ($r = 0.54$, $p<0.01$), and Noro virus G2 and conductivity ($r = 0.49$, $p<0.01$). The negative correlation between river discharge and microbial water quality ranges from -0.01 to -0.32 and could be explained by the dilution effect of the discharge volume. Also, negative correlations were observed between microbial water quality and raw water temperature that ranges from -0.06 to -0.40. One can explain that the lowest temperature is more favourable for microbial pathogen growth than highest temperature for the observed temperature range. The highest correlation among the predictor variables was observed between total organic carbon and colour ($r = 0.78$), river discharge and raw water temperature (r ranges from 0.61 to 0.84), river discharge and conductivity (r ranges from -0.28 to -0.58), pH and conductivity ($r = 0.51$). In this modelling, only one of the highly correlated explanatory variables was considered in order to avoid the replication of the same tendency predictor variable.

Logarithmically transformed variables in a regression model is a very common means of transforming a highly skewed variable into one that is more approximately normal so as to improve the overall multiple linear regression model. In this study, all microbial pathogen load response variables data sets were transformed into $\text{Log}_{(10)}$ after they had been tested without transform with unsatisfactory. In the modelling of the microbial load response variable, twelve predictor variables were accounted for: river discharge from different tributaries gauging stations of Glomma River, namely, Rånåsfoss, Blaker, Funnefoss, Ertesekken ndf, Vorma; and also raw water temperature, rainfall, pH, turbidity, conductivity, colour, total organic carbon.

Table 2 Correlation coefficients (r) among explanatory variables and raw water microbial load

	Rån	Bla	Fun	Ert	Vor	Tem	Rain	pH	Tur	Con	Colo	T.Ca	Kim	C.Pe	I. En	Eco	C.ba
Rånåsfoss	1																
Blaker	0.93	1															
Funnefoss	0.83	0.83	1														
Ertesekken ndf	0.76	0.68	0.61	1													
Vorma	0.79	0.76	0.72	0.81	1												
Temperature	0.61	0.79	0.75	0.69	0.84	1											
Rainfall	0.42	0.33	0.20	0.29	0.39	0.29	1										
pH	0.05	-0.32	-0.35	-0.29	-0.15	0.21	0.11	1									
Turbidity	-0.01	0.18	0.39	0.17	0.25	0.02	0.07	-0.13	1								
Conductivity	-0.28	-0.55	-0.58	-0.51	-0.49	-0.16	-0.19	0.51	0.16	1							
Colour	0.22	0.34	0.26	0.19	0.21	-0.16	0.01	-0.16	0.04	-0.23	1						
Total OR. Carbon	0.27	0.33	0.23	0.37	0.29	0.06	0.03	-0.24	0.10	-0.28	0.78	1					
TVC “Kimtall”	-0.03	-0.05	-0.12	-0.09	-0.12	-0.15	-0.19	0.24	0.20	0.21	0.36	0.17	1				
C. perfringens	-0.19	-0.01	0.03	-0.12	-0.22	-0.28	-0.02	0.22	0.23	0.41	0.34	0.06	0.60	1			
Int. enterococci	-0.04	-0.11	0.19	-0.03	-0.16	-0.19	-0.01	-0.08	0.45	0.13	0.22	-0.07	0.50	0.44	1		
Escherichia coli	-0.19	-0.09	-0.01	0.02	0.04	-0.40	-0.11	0.11	0.52	0.34	0.21	0.07	0.48	0.54	0.53	1	
Coliform bacteria	-0.08	-0.02	-0.16	0.09	-0.18	-0.06	-0.10	0.11	0.26	0.23	0.13	0.05	0.39	0.33	0.26	0.55	1
Adeno virus	-0.29	-0.09	-0.11	-0.19	-0,24	-0.16	-0.24	-0.23	-0.04	0.47	0.01	0.02	-	-	-	-	-
Noro virus (g1)	-0.20	0.11	-0.08	-0.18	-0,30	-0.27	-0.10	-0.10	-0.32	0.54	0.12	0.12	-	-	-	-	-
Noro virus (g2)	-0.23	0.19	-0.13	-0.16	-0.32	-0.32	-0.17	0.04	-0.36	0.49	0.06	0.09	-	-	-	-	-

In determining what model would be appropriate in predicting the microbial pathogen load in the raw water, the interaction of the response variable with all predictor variables was considered. A stepwise regression method was applied to select the best possible fitted multiple linear regression model having all the variables of interest already in the processes of selection. In order to test the significance of each interaction of predictor variables, t-test was carried out to test the null hypothesis that the interaction term being tested has no effect on the model against the alternative hypothesis that the interaction term has an effect on the model. Then the t-value was calculated for each parameter estimate, and if the probability associated with each t-value is over an alpha level of 0.05 (standard arbitrary p-value chosen in statistics), then the interaction term is insignificant and the variable is not considered in the model. The t-test eliminates the least significant interaction variable and leaves the model with significant variables that have more association with the response variable. The t-test results show that all regression coefficients are significant (P-value < 0.05). The least square regression coefficients, the standard errors, the t-values and the level of significance for rejecting null hypothesis for each selected variable are given in Tables 3. From these relationships, it is inferred that the regression analysis has led to the formulation of the following multiple linear regression equations for each microbial pathogen load in the raw water:

- Log Kimtall = -4.807 + 0.871*pH + 0.011*Funnefoss + 0.717*Conductivity + 0.050*Colour
- Log Clostridium perfringens = -2,68 -0.003*Rånåsfoss + 0.837*Turbidity + 1.944*Conductivity + 0.077*Colour
- Log Escherichia coli = 1,633 - 0.078*Raw water temperature + 0.029*Turbidity + 0.489*Conductivity + 0.014*Colour
- Log Coliform bacteria = 0.133 - 0.010*Turbidity + 0.434*Conductivity + 0.011*Colour
- Log Intestinal Enterococci = -2.428 - 0.033*Raw water temperature - 0.034*Turbidity + 0.977*Conductivity + 0.028*Colour
- Log Adeno virus = 12.027 - 1.840*pH - 0,132*Rain fall + 0,449*Conductivity
- Log Noro virus (g1) = 5.543 - 1.023*pH + 0.554*Conductivity
- Log Noro virus (g2) = 0.046 - 0.326*Turbidity + 0.421*Conductivity - 0.029*Raw water temperature

Table 3 Coefficients of regression

Response Variable	Predictors	Coefficient	Standard error	t	Pr > |t|
TVC "Kimtall"	Constant	-4,807	1,635	-2,941	0,004
	pH	0,871	0,270	3,220	0,001
	Funnefoss o (m^3/s)	0,011	0,001	2,776	0,006
	Conductivity (mS/m)	0,717	0,110	6,515	< 0,0001
	Colour (mg Pt/l)	0,050	0,006	8,564	< 0,0001
Clostridium perfringens	Constant	-2,683	5,748	-0,467	0,642
	Rånåsfoss (m^3/s)	-0,003	0,001	-2,682	0,008
	Turbidity (NTU)	0,837	0,156	5,347	< 0,0001
	Conductivity (mS/m)	1,944	1,134	1,714	0,45
	Colour (mg Pt/l)	0,077	0,032	2,379	0,019
Escherichia coli	Constant	1,633	0,500	3,267	0,001
	R.water_temprature (oC)	-0,078	0,009	-9,074	< 0,0001
	Turbidity (NTU)	0,029	0,012	2,368	0,019
	Conductivity (mS/m)	0,489	0,095	5,146	< 0,0001
	Colour (mg Pt/l)	0,014	0,004	3,707	0,000
Coliform bacteria	Constant	0,133	0,192	0,690	0,491
	Turbidity (NTU)	-0,010	0,004	-2,264	0,025
	Conductivity (mS/m)	0,434	0,041	10,679	< 0,0001
	Colour (mg Pt/l)	0,011	0,002	4,589	< 0,0001
Intestinal enterococci	Constant	-2,428	0,563	-4,309	< 0,0001
	R. water_temprature (^oC)	-0,033	0,017	-1,952	0,043
	Turbidity (NTU)	-0,034	0,008	-4,261	< 0,0001
	Conductivity (mS/m)	0,977	0,117	8,354	< 0,0001
	Colour (mg Pt/l)	0,028	0,008	3,637	0,000
Adeno Virus	Constant	12,027	4,319	2,785	0,007
	pH	-1,840	0,650	-2,830	0,006
	Rain fall (mm)	-0,132	0,037	-3,597	0,001
	Conductivity (mS/m)	0,449	0,108	4,175	< 0,0001
Noro_G1	Constant	5,543	2,624	2,112	0,039
	pH	-1,023	0,353	-2,900	0,005
	Conductivity (mS/m)	0,554	0,099	5,596	< 0,0001
Noro_G2	Constant	0,046	0,769	0,060	0,953
	Turbidity (NTU)	-0,326	0,070	-4,666	< 0,0001
	Conductivity (mS/m)	0,421	0,130	3,232	0,002
	R.water_temprature (^oC)	-0,029	0,015	-1,925	0,049

From the ANOVA (Table 4), we can see that F value ranges from 15.617 (Adeno virus) to 63.466 (E coli) and significant at $p < .0001$ for all models. This provides evidence of the existence of a linear relationship between the response (microbial pathogen load) and the explanatory variables (physico-chemical factors). This means that, the regression model we have constructed is well determined by the factors.

The other important topic that needs to be discussed in this modelling process is multicollinearity, the problem when one independent variable is correlated with another independent variable that results in an imprecision in the calculated parameter estimates. The problem of multicollinearity can be handled by looking at variance inflation factors (VIF). Those independent variables with VIF > 10 (standard VIF value chosen in statistics), are considered as having a problem of multicollinearity. If less multicollinearity is not significant enough and ignored. Since Table 5 shows that the VIF for all variables are less than 10, we can reasonably assume that our explanatory variables are not too strongly correlated so that it might increase our confidence in understand that how our individual variables affect our response variable.

Table 4 ANOVA for regression

Response Variable	Source	DF	Sum of squares	Mean squares	F	Pr > F
Total viable count "Kimtall"	Regression	4	315,616	78,904	55,442	< 0,0001
	Residual	302	429,802	1,423		
	Total	306	745,418			
Clostridium perfringens	Regression	4	1196,010	299,002	19,961	< 0,0001
	Residual	112	1677,649	14,979		
	Total	116	2873,658			
Escherichia coli	Regression	4	116,722	29,180	63,466	< 0,0001
	Residual	245	112,647	0,460		
	Total	249	229,368			
Coliform bacteria	Regression	3	15,078	5,026	40,605	< 0,0001
	Residual	131	16,215	0,124		
	Total	134	31,293			
Intestinal enterococci	Regression	4	110,886	27,722	22,072	< 0,0001
	Residual	123	154,485	1,256		
	Total	127	265,371			
Adeno Virus	Regression	3	15,149	5,050	15,617	< 0,0001
	Residual	70	22,634	0,323		
	Total	73	37,783			
Noro_G1	Regression	2	6,936	3,468	24,053	< 0,0001
	Residual	60	8,652	0,144		
	Total	62	15,588			
Noro_G2	Regression	3	12,945	4,315	19,039	< 0,0001
	Residual	55	12,465	0,227		
	Total	58	25,410			

The most commonly used criterion to evaluate model performance is coefficient of determination (R^2); however R^2 only tell us how good the model fits with the data used to build the models not beyond the extent of the data set. The R^2 vale in this study ranges from 0.40 to 0.51 (Table 6) and it indicates how much of the variability in microbial load in the raw

water is explained by the independent variables used in the model. The other criteria is adjusted R^2 that also account for the number of explanatory terms that are used in the model. The Mean Square Error (MSE) and Root Mean Square Error (RMSE) measure the residual error which gives an estimate of the mean difference between observed and modeled values of microbial load are relatively low and increase our confidence in the capability of the model.

Table 5 VIF values for multicollinearity test

Response Variable	Statistic	VIF
Total viable count "Kimtall"	pH	1,766
	Funnefoss o.vann	1,208
	Conductivity (mS/m)	1,788
	Colour (mg Pt/l)	1,176
Clostridium perfringens	Rånåsfoss	1,602
	Turbidity (NTU)	1,584
	Conductivity (mS/m)	1,959
	Colour (mg Pt/l)	1,648
Escherichia coli	Raw water Temperature (^oC)	1,341
	Turbidity (NTU)	1,311
	Conductivity (mS/m)	1,557
	Colour (mg Pt/l)	1,484
Coliform bacteria	Rånåsfoss	1,647
	Raw water Temperature (^oC)	1,942
	pH	2,915
	Turbidity (NTU)	1,729
	Conductivity (mS/m)	3,316
	Colour (mg Pt/l)	1,066
Intestinal enterococci	Raw water Temperature (^oC)	2,118
	Turbidity (NTU)	2,859
	Conductivity (mS/m)	4,588
	Colour (mg Pt/l)	1,543
Adeno Virus	ph	1,225
	rain	1,128
	conduct	1,175
Noro_G1	ph	1,041
	conduct	1,041
Noro_G2	turbidity	1,235
	conduct	1,592
	Raw water temperature (^oC)	1,743

The combination of Akaike's Information Criteria (AIC) and Schwarz Bayesian Criteria (SBC) values, coefficient of determination (R^2) and adjusted R^2 values enable us to evaluate the best model performance. The smaller the difference between AIC and SBC values with a combination of the R^2 and adjusted R^2 close to one indicates that the constructed multiple regression model is an appropriate method for microbial pathogen load prediction (Aertsen et

al. 2010). The low difference between AIC and SBC in most models in this study indicates the adequacy of the models in terms of prediction of microbial load based on the independent variables.

Figure 1 shows the graph plotting for observed microbial pathogen load and predicted microbial pathogen load with 95 % confidence interval. Some observations from overall observations were out of the upper and lower boundary range of 95% confidence interval. This is due to great difference between observed and predicted values for some of the observations points. Otherwise, as it is observed from the graphs, most of the points are within the confidence interval. This proved that these models are able to predict microbial pathogen load with reasonable precision.

Finally, the residuals were plotted as a function of the predicted values as illustrated in Fig. 2. Analysing the residuals, there is no pattern in the residuals of each model. This means that there is no left over information in the residuals that the model did not account for. And also it can be seen from the plots that the residuals are attributed evenly above and below zero this means we have nearly constant variance and therefore the models are deemed valid to describe the explanatory variables data set.

Table 6 Goodness of fit statistics of the regression models

Statistics	TVC Kimtall	*Clostridium perfringens*	*E coli*	*Coliform bacteria*	*Intestinal enterococci*	Adeno Virus	Noro virus G1	Noro virus G2
R^2	0.42	0.42	0.51	0.48	0.42	0.40	0.45	0.51
Adjusted R^2	0.41	0.40	0.50	0.47	0.40	0.38	0.43	0.48
MSE	1.42	14.98	0.46	0.12	1.26	0.32	0.14	0.23
RMSE	1.19	3.87	0.68	0.35	1.12	0.57	0.38	0.47
AIC	113.29	321.6	-189.3	-278.11	34.07	-79.66	-119.08	-83.72
SBC	131.93	335.4	-171.7	-266.49	48.33	-70.44	-112.65	-75.41

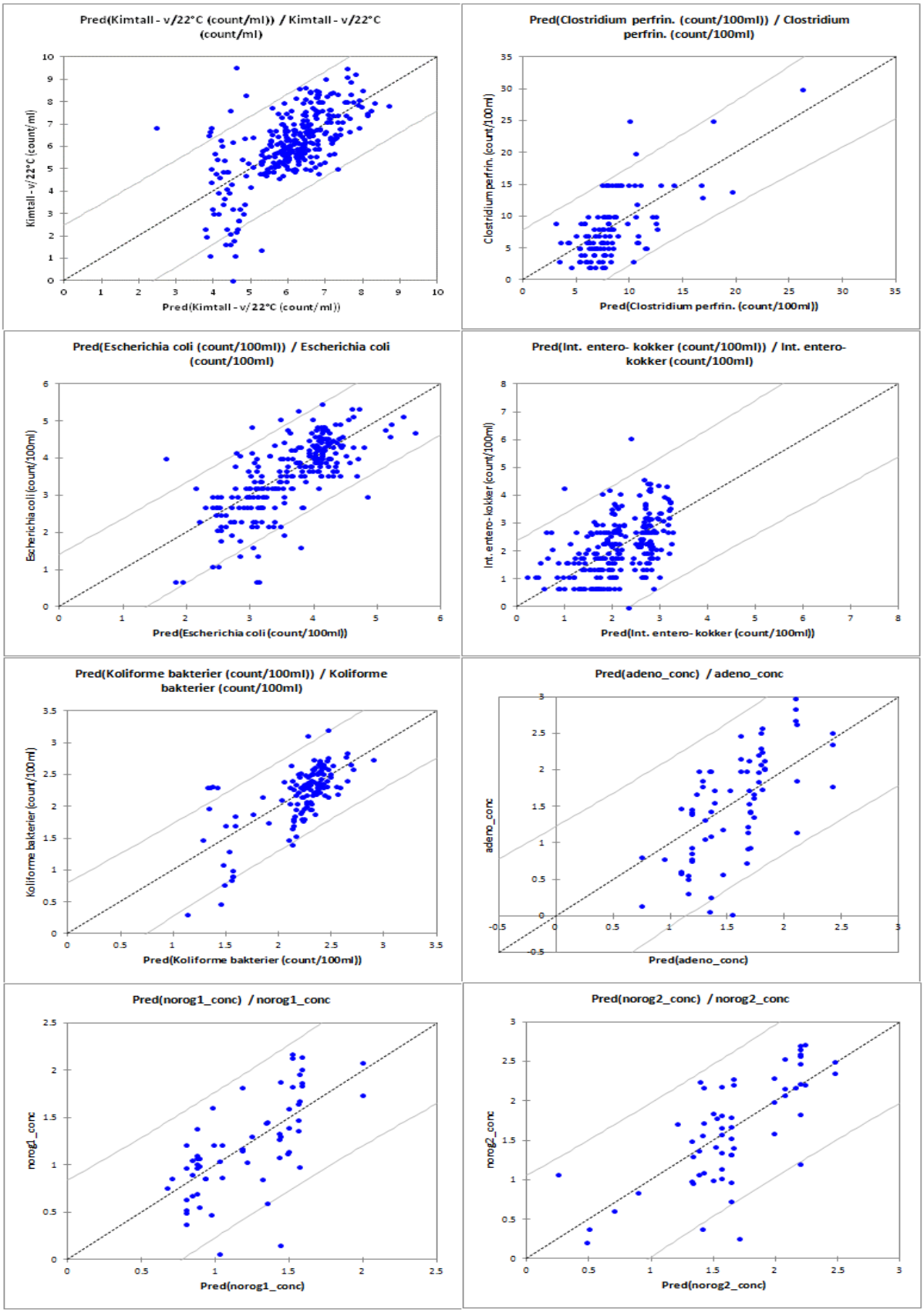


Figure 2 Microbial water quality index predicted versus actual observation (95 % CI)

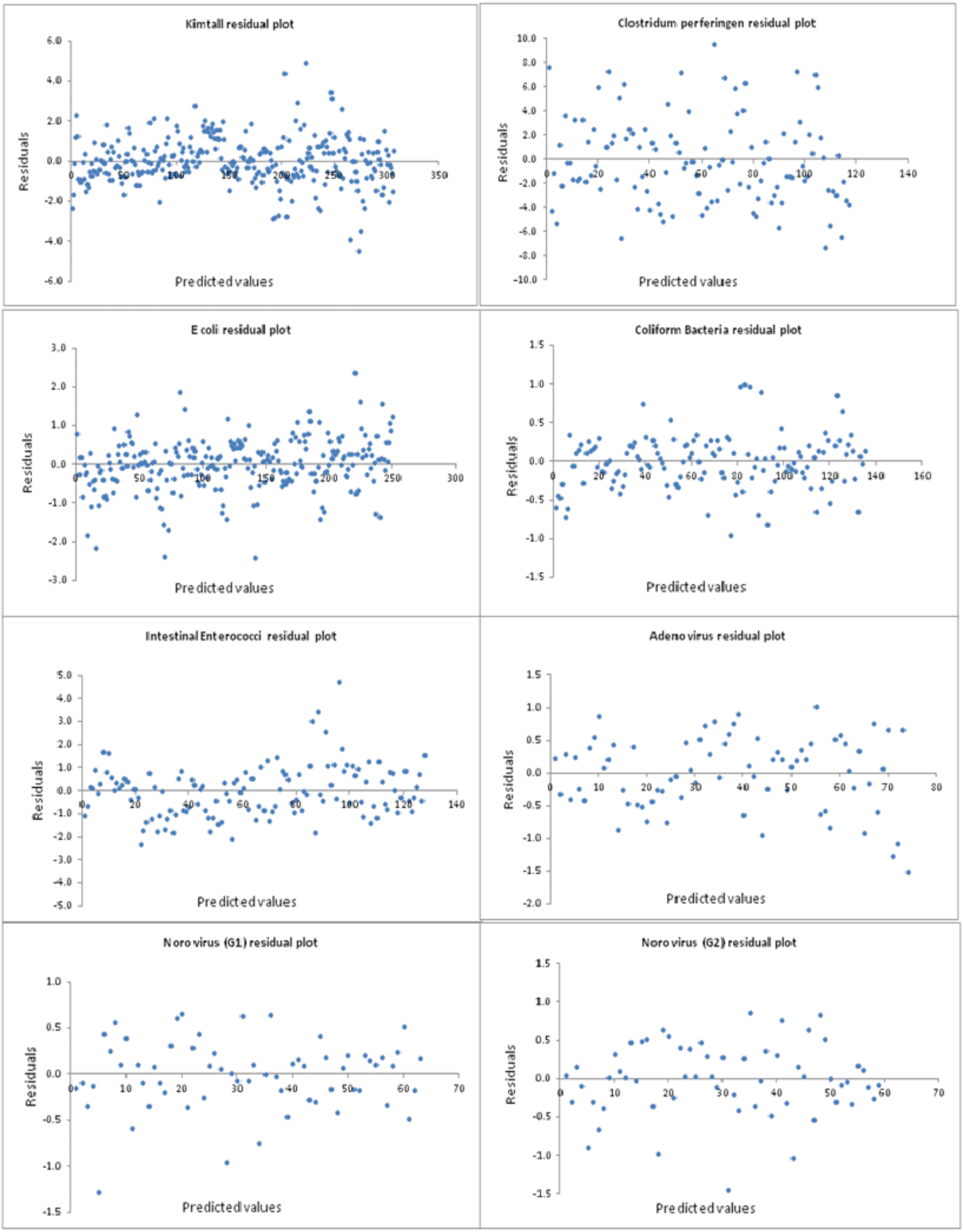


Figure 3 Residuals versus predicted values

CONCLUSION

We have demonstrated that when intensive and regular microbial water quality monitoring become very essential, then, we can estimate the concentration of microbial pathogen in the raw water only by observing a few explanatory factors that will save our time, money, and resources. Hence, this may be an important, economic method for places which are found to be difficult in monitoring all microbial water quality parameters and also when the result is required for quick decision making in the water treatment plant. While not perfect, such systems provide an excellent coarse level tool for regional or even watershed scale river management practices such as visualizing the extent and trend of microbial pathogen load; or developing management or regulatory standards.

The overall aim of the research was to gain an understanding of the factors affecting microbial pathogen load in the raw water through the development and application of a multiple linear regression model. The results indicated that for each microbial pathogen load, different physico-chemical variables could explain from 40 percent to 51 percent of the variation of microbial concentration.

Our models intentionally contained independent variables representing degree of microbial pathogen load in the raw water. The developed linear regression models are simple and provide best fits to the data set. However, the models' predictive accuracy can be less than desired and they have several obvious weaknesses: 1) the quality of the data set; 2) possibly lack of linear relationship between the factors and the dependent variable; and 3) these could be important factors not accounted in the models. This might be the first time that drinking water treatment plants have examined their microbial pathogen load data set in association with different physico-chemical factors in a fairly detailed manner in the river basin. As data sources and modelling approaches improve through time, these modelling tools will become more and more accurate and valuable.

REFERENCES

Aertsen, W., Kint, V., van Orshoven, J., Ozkan, K. & Muys, B. (2010). Comparison and ranking of different modelling techniques for prediction of site index in Mediterranean mountain forests. *Ecological Modelling*, 221 (8): 1119-1130.

Agirre-Basurko, E., Ibarra-Berastegi, G. & Madariaga, I. (2006). Regression and multilayer perceptron-based models to forecast hourly O-3 and NO2 levels in the Bilbao area. *Environmental Modelling & Software*, 21 (4): 430-446.

Archer, K. J. & Lemeshow, S. (2006). Goodness-of-fit test for a logistic regression model fitted using survey sample data. *Stata Journal*, 6 (1): 97-105.

Astrom, J., Petterson, S., Bergstedt, O., Pettersson, T. J. R. & Stenstrom, T. A. (2007a). Evaluation of the microbial risk reduction due to selective closure of the raw water intake before drinking water treatment. *Journal of Water and Health*, 5: 81-97.

Astrom, J., Pettersson, T. J. R. & Stenstrom, T. A. (2007b). Identification and management of microbial contaminations in a surface drinking water source. *Journal of Water and Health*, 5: 67-79.

Baudisova, D. (2009). Microbial pollution of water from agriculture. *Plant Soil and Environment*, 55 (10): 429-435.

Bedrick, E. J. & Crandall, W. K. (2010). Model Selection Criteria for Loglinear Models. *Australian & New Zealand Journal of Statistics*, 52 (4): 439-449.

Bishop, C. A., Struger, J., Barton, D. R., Shirose, L. J., Dunn, L., Lang, A. L. & Shepherd, D. (2000). Contamination and wildlife communities in stormwater detention ponds in Guelph and the Greater Toronto area, Ontario, 1997 and 1998. Part I - Wildlife communities. *Water Quality Research Journal of Canada*, 35 (3): 399-435.

Bociort, D., Gherasimescu, C., Berariu, R., Butnaru, R., Branzila, M. & Sandu, I. (2012). Research on the Degree of Contamination of Surface and Groundwater used as Sources for Drinking Water. *Revista De Chimie*, 63 (11): 1152-1157.

Canada, H. (2006). *Bacterial Waterborne Pathogens — Current and Emerging Organisms of Concern*. Ottawa, Ontario (accessed: April 22).

Cheung, K. C. & Venkitachalam, T. H. (2004). Assessment of contamination by percolation of septic tank effluent through natural and amended soils. *Environmental Geochemistry and Health*, 26 (2-3): 157-168.

Cimenti, M., Hubberstey, A., Bewtra, J. K. & Biswas, N. (2007). Alternative methods in tracking sources of microbial contamination in waters. *Water Sa*, 33 (2): 183-194.

Coelho-Barros, E. A., Simoes, P. A., Achcar, J. A., Martinez, E. Z. & Shimano, A. C. (2008). Methods of Estimation in Multiple Linear Regression: Application to Clinical Data. *Revista Colombiana De Estadistica*, 31 (1): 111-129.

Coffey, R., Cummins, E., Cormican, M., Flaherty, V. O. & Kelly, S. (2007). Microbial exposure assessment of waterborne pathogens. *Human and Ecological Risk Assessment*, 13 (6): 1313-1351.

Crowther, J., Kay, D. & Wyer, M. D. (2001). Relationships between microbial water quality and environmental conditions in coastal recreational waters: The Fylde coast, UK. *Water Research*, 35 (17): 4029-4038.

Davies, J. M. & Mazumder, A. (2003). Health and environmental policy issues in Canada: the role of watershed management in sustaining clean drinking water quality at surface sources. *Journal of Environmental Management*, 68 (3): 273-286.

Devereux, R., Rublee, P., Paul, J. H., Field, K. G. & Santo Domingo, J. W. (2006). Development and applications of microbial ecogenomic indicators for monitoring water quality: Report of a workshop assessing the state of the science, research needs and future directions. *Environmental Monitoring and Assessment*, 116 (1-3): 459-479.

Edge, T. A., El-Shaarawi, A., Gannon, V., Jokinen, C., Kent, R., Khan, I. U. H., Koning, W., Lapen, D., Miller, J., Neumann, N., et al. (2012). Investigation of an Escherichia coli Environmental

Benchmark for Waterborne Pathogens in Agricultural Watersheds in Canada. *Journal of Environmental Quality*, 41 (1): 21-30.
Eiswirth, M. & Hotzl, H. (1997). The impact of leaking sewers on urban groundwater. *Groundwater in the Urban Environment - Vol I*: 399-404.
Even, S., Mouchel, J. M., Servais, P., Flipo, N., Poulin, M., Blanc, S., Chabanel, M. & Paffoni, C. (2007). Modelling the impacts of Combined Sewer Overflows on the river Seine water quality. *Science of the Total Environment*, 375 (1-3): 140-151.
Fagerland, M. W. & Hosmer, D. W. (2013). A goodness-of-fit test for the proportional odds regression model. *Statistics in Medicine*, 32 (13): 2235-2249.
Farkas, A., Dragan-Bularda, M., Muntean, V., Ciataras, D. & Tigan, S. (2013). Microbial activity in drinking water-associated biofilms. *Central European Journal of Biology*, 8 (2): 201-214.
Fedotovai, O., Teixeira, L. & Alvelos, H. (2013). Software Effort Estimation with Multiple Linear Regression: Review and Practical Application. *Journal of Information Science and Engineering*, 29 (5): 925-945.
Ferguson, C. M., Croke, B., Ashbolt, N. J. & Deere, D. A. (2005). A deterministic model to quantify pathogen loads in drinking water catchments: pathogen budget for the Wingecarribee. *Water Science and Technology*, 52 (8): 191-197.
Ferraro, M. B. & Giordani, P. (2012). A multiple linear regression model for imprecise information. *Metrika*, 75 (8): 1049-1068.
folkehelseinstitutt, N. (2004). Vannforsyningens ABC.
Geldreich, E. E. (1989). Drinking-Water Microbiology - New Directions toward Water-Quality Enhancement. *International Journal of Food Microbiology*, 9 (4): 295-312.
Gerba, C. P. & Smith, J. E. (2005). Sources of pathogenic microorganisms and their fate during land application of wastes. *Journal of Environmental Quality*, 34 (1): 42-48.
Grizzetti B., B. F., Bianchi M., Barkved L., Berge D., Campbell D., Dan Kim N., Gooch G., Lana Renoult N., Nesheim I., Machado M., Manasi S., Rieu-Clarke A., Stålnacke, P. and Tjomsland T. (2007). Managing data in Integrated Water Resources Management projects: the STRIVER case: European Commission Joint Research Center. Institute for Environment and Sustainability. Rural, Water and Ecosystem Resources Unit (JRC-EC).
Han, M., Zhao, Z. W., Cui, F. Y., Gao, W., Liu, J. & Zeng, Z. Q. (2012). Pretreatment of contaminated raw water by a novel double-layer biological aerated filter for drinking water treatment. *Desalination and Water Treatment*, 37 (1-3): 308-314.
Hasani, H. & Shanbeh, M. (2010). Application of multiple linear regression and artificial neural network algorithms to predict the total hand value of summer knitted T-shirts. *Indian Journal of Fibre & Textile Research*, 35 (3): 222-227.
He, J. X., Valeo, C., Chu, A. & Neumann, N. F. (2010). Characterizing Physicochemical Quality of Storm-Water Runoff from an Urban Area in Calgary, Alberta. *Journal of Environmental Engineering-Asce*, 136 (11): 1206-1217.
Hirata, T., Kawamura, K., Sonoki, S., Hirata, K., Kaneko, M. & Taguchi, K. (1991). Clostridium-Perfringens, as an Indicator Microorganism for the Evaluation of the Effect of Waste-Water and Sludge Treatment Systems. *Water Science and Technology*, 24 (2): 367-372.
Jamieson, R., Gordon, R., Joy, D. & Lee, H. (2004). Assessing microbial pollution of rural surface waters - A review of current watershed scale modeling approaches. *Agricultural Water Management*, 70 (1): 1-17.
Karlaviciene, V., Svediene, S., Marciulioniene, D. E., Randerson, P., Rimeika, M. & Hogland, W. (2009). The impact of storm water runoff on a small urban stream. *Journal of Soils and Sediments*, 9 (1): 6-12.
Khwaja, A. A., Lisa, M., Boustani, M., Jaffar, M. & Masud, M. K. (1999). An assessment study of septic tank based sewage disposal system on the quality of underground water. *Journal of the Chemical Society of Pakistan*, 21 (2): 141-145.
Kieseppa, I. A. (2001). Statistical model selection criteria and Bayesianism. *Philosophy of Science*, 68 (3): S141-S152.
Kinzelman, J., McLellan, S. L., Daniels, A. D., Cashin, S., Singh, A., Gradus, S. & Bagley, R. (2004). Non-point source pollution: Determination of replication versus persistence of Escherichia

coli in surface water and sediments with correlation of levels to readily measurable environmental parameters. *Journal of Water and Health*, 2 (2): 103-114.

Kovdienko, N. A., Polishchuk, P. G., Muratov, E. N., Artemenko, A. G., Kuz'min, V. E., Gorb, L., Hill, F. & Leszczynski, J. (2010). Application of Random Forest and Multiple Linear Regression Techniques to QSPR Prediction of an Aqueous Solubility for Military Compounds. *Molecular Informatics*, 29 (5): 394-406.

Kubeck, C., van Berk, W. & Bergmann, A. (2009). Modelling raw water quality: development of a drinking water management tool. *Water Science and Technology*, 59 (1): 117-124.

Kufs, C. T. (1992). Statistical Modeling of Hydrogeologic Data .1. Regression and Anova Models. *Ground Water Monitoring and Remediation*, 12 (2): 120-130.

Mills, M. S. & Thurman, E. M. (1994). Reduction of Nonpoint-Source Contamination of Surface-Water and Groundwater by Starch Encapsulation of Herbicides. *Environmental Science & Technology*, 28 (1): 73-79.

Moustris, K. P., Nastos, P. T., Larissi, I. K. & Paliatsos, A. G. (2012). Application of Multiple Linear Regression Models and Artificial Neural Networks on the Surface Ozone Forecast in the Greater Athens Area, Greece. *Advances in Meteorology*.

Naidu, G. M., Balasiddamuni, P., Giri, D., Ismail, S., Rao, C. L. K. & Reddy, C. S. (2012). Model Selection Criteria. *International Journal of Agricultural and Statistical Sciences*, 8 (1): 335-345.

Nnane, D. E. (2011). Sustainable microbial water quality monitoring programme design using phage-lysis and multivariate techniques. *Science of the Total Environment*, 409 (24): 5188-5195.

Noller, D. G. & Whitehouse, G. E. (1982). Multiple Linear-Regression - a Microcomputer Application. *Industrial Engineering*, 14 (6): 26-&.

Noorossana, R., Eyvazian, M., Amiri, A. & Mahmoud, M. A. (2010). Statistical Monitoring of Multivariate Multiple Linear Regression Profiles in Phase I with Calibration Application. *Quality and Reliability Engineering International*, 26 (3): 291-303.

Obasohan, E. E., Agbonlahor, D. E. & Obano, E. E. (2010). Water pollution: A review of microbial quality and health concerns of water, sediment and fish in the aquatic ecosystem. *African Journal of Biotechnology*, 9 (4): 423-427.

Okoh, A. I., Odjadjare, E. E., Igbinosa, E. O. & Osode, A. N. (2007). Wastewater treatment plants as a source of microbial pathogens in receiving watersheds. *African Journal of Biotechnology*, 6 (25): 2932-2944.

Payment, P., Berte, A., PrTvost, M., MTnard, B. & Barbeau, B. (2000). Occurrence of pathogenic microorganisms in the Saint Lawrence River (Canada) and comparison of health risks for populations using it as their source of drinking water. *Canadian Journal of Microbiology*, 46 (6): 565-576.

Plummer, J. D. & Long, S. C. (2007). Monitoring source water for microbial contamination: Evaluation of water quality measures. *Water Research*, 41 (16): 3716-3728.

Pugh, E. W., Papanicolaou, G. J., Justice, C. M., Roy-Gagnon, M. H., Sorant, A. J. M., Kingman, A. & Wilson, A. F. (2001). Comparison of variance components, ANOVA and regression of offspring on midparent (ROMP) methods for SNP markers. *Genetic Epidemiology*, 21: S794-S799.

Pullar, D. & Springer, D. (2000). Towards integrating GIS and catchment models. *Environmental Modelling & Software*, 15 (5): 451-459.

Ritter, L., Solomon, K., Sibley, P., Hall, K., Keen, P., Mattu, G. & Linton, B. (2002). Sources, pathways, and relative risks of contaminants in surface water and groundwater: A perspective prepared for the Walkerton inquiry. *Journal of Toxicology and Environmental Health-Part a-Current Issues*, 65 (1): 1-142.

Scott, T. M., Rose, J. B., Jenkins, T. M., Farrah, S. R. & Lukasik, J. (2002). Microbial source tracking: Current methodology and future directions. *Applied and Environmental Microbiology*, 68 (12): 5796-5803.

Sedmak, G., Bina, D., MacDonald, J. & Couillard, L. (2005). Nine-year study of the occurrence of culturable viruses in source water for two drinking water treatment plants and the influent and effluent of a wastewater treatment plant in Milwaukee, Wisconsin (August 1994 through July 2003). *Applied and Environmental Microbiology*, 71 (2): 1042-1050.

Seidou, O. & Ouarda, T. B. M. J. (2007). Recursion-based multiple changepoint detection in multiple linear regression and application to river streamflows. *Water Resources Research*, 43 (7).

Shih, J. H. (1998). A goodness-of-fit test for association in a bivariate survival model. *Biometrika*, 85 (1): 189-200.

Sidhu, J. P. S., Ahmed, W., Gernjak, W., Aryal, R., McCarthy, D., Palmer, A., Kolotelo, P. & Toze, S. (2013). Sewage pollution in urban stormwater runoff as evident from the widespread presence of multiple microbial and chemical source tracking markers. *Science of the Total Environment*, 463: 488-496.

Stone, M. (1979). Model Selection Criteria of Akaike and Schwarz. *Journal of the Royal Statistical Society Series B-Methodological*, 41 (2): 276-278.

Ukoumunne, O. C., Gulliford, M. C. & Chinn, S. (2002). A note on the use of the variance inflation factor for determining sample size in cluster randomized trials. *Journal of the Royal Statistical Society Series D-the Statistician*, 51: 479-484.

Vounatsou, P. & Karydis, M. (1991). Environmental Characteristics in Oligotrophic Waters - Data Evaluation and Statistical Limitations in Water-Quality Studies. *Environmental Monitoring and Assessment*, 18 (3): 211-220.

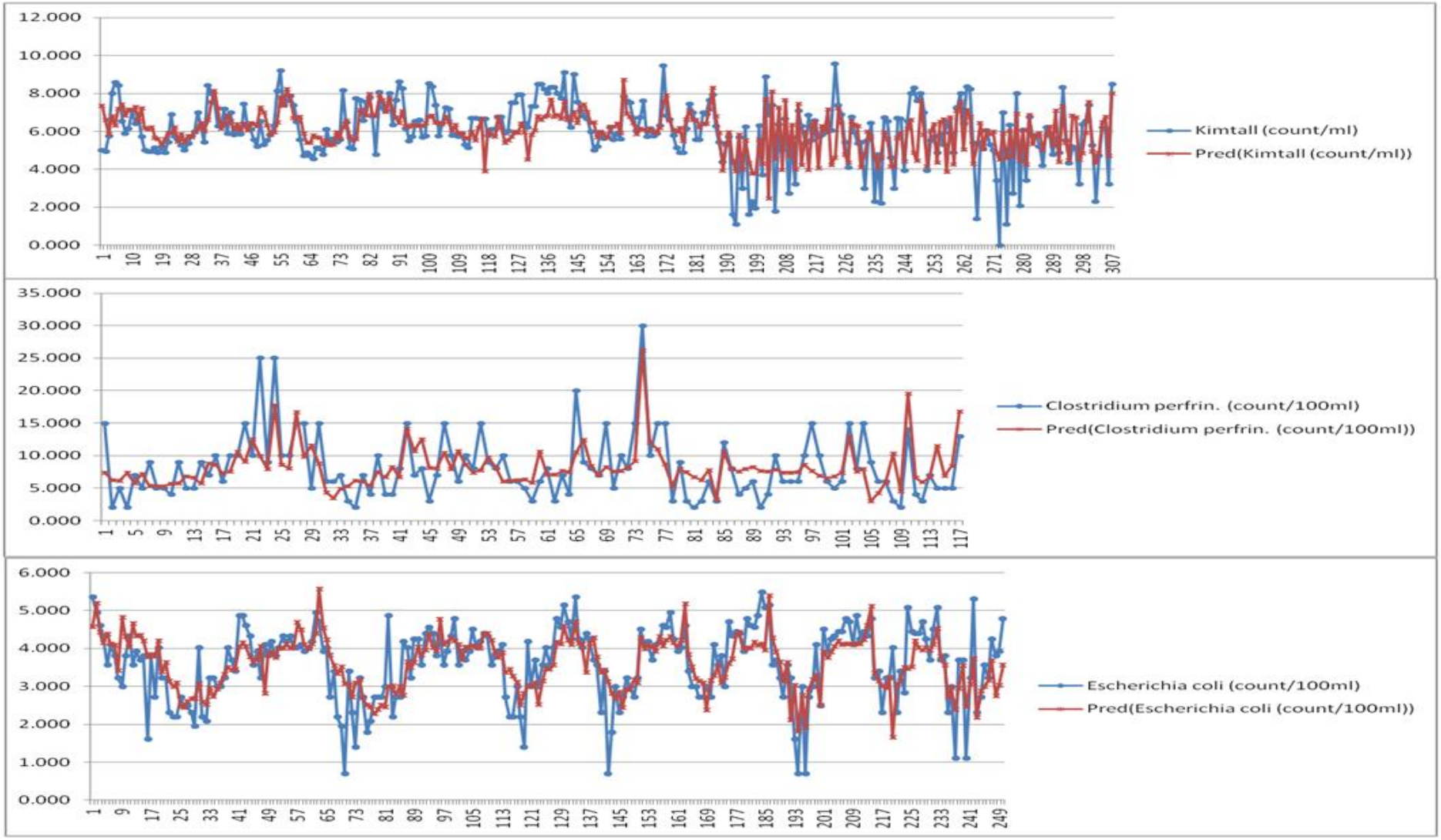
Walker, K. P. (1994). Urban Runoff and Combined Sewer Overflows. *Water Environment Research*, 66 (4): 305-309.

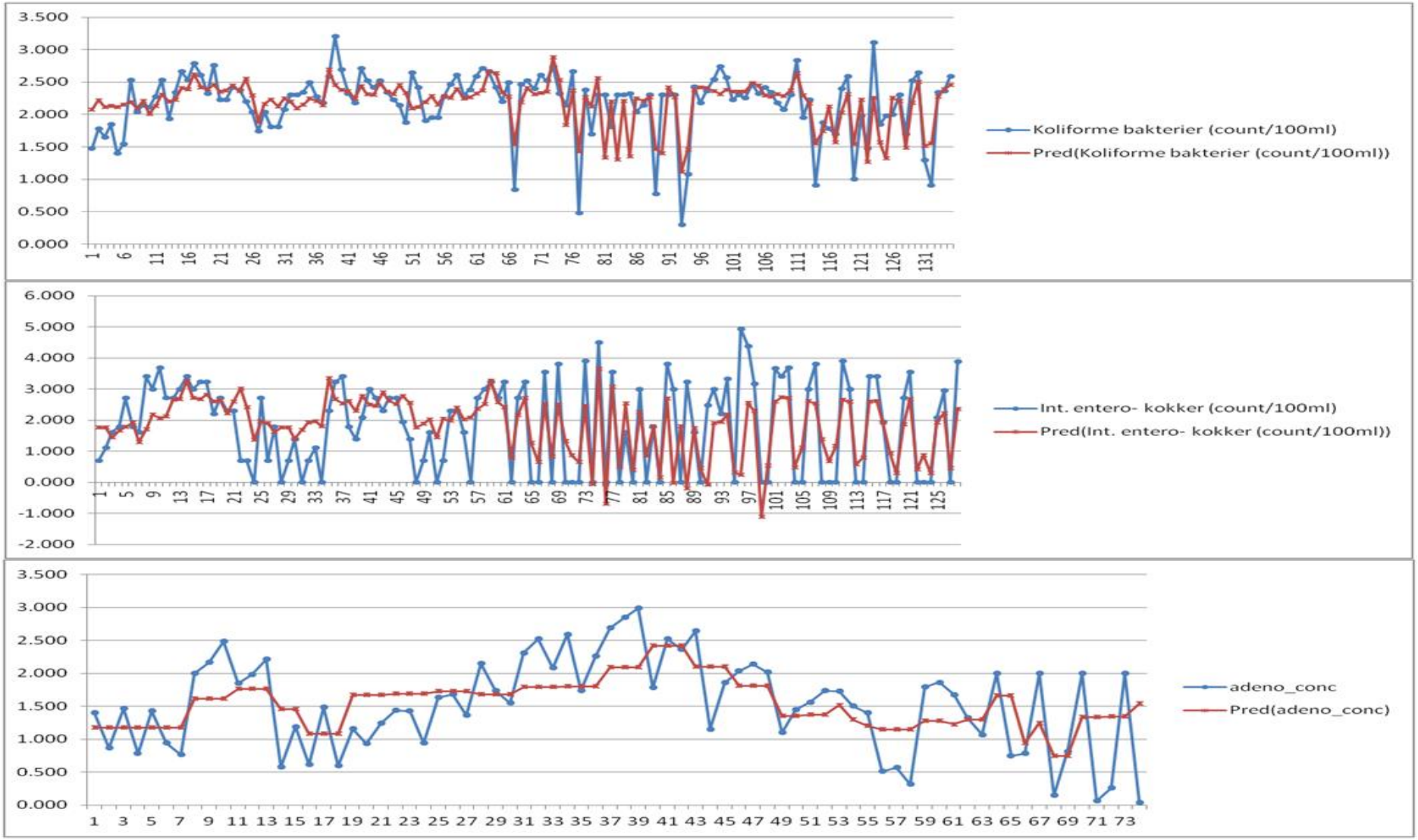
Won, G., Kline, T. R. & LeJeune, J. T. (2013). Spatial-temporal variations of microbial water quality in surface reservoirs and canals used for irrigation. *Agricultural Water Management*, 116: 73-78.

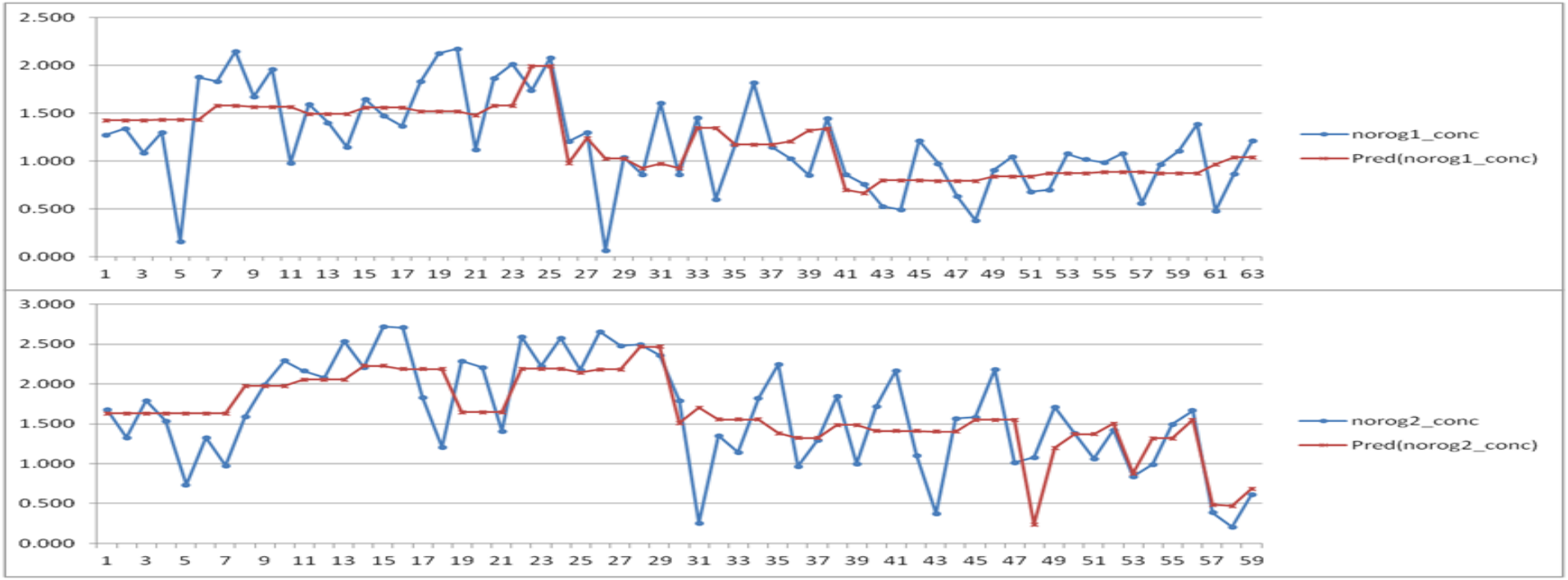
Yang, Y. P., Xue, L. G. & Cheng, W. H. (2011). The Empirical Likelihood Goodness-of-Fit Test for a Regression Model with Randomly Censored Data. *Communications in Statistics-Theory and Methods*, 40 (3): 424-435.

Zhang, Q. & Stanley, S. J. (1997). Forecasting raw-water quality parameters for the North Saskatchewan River by neural network modeling. *Water Research*, 31 (9): 2340-2350.

Appendix 1: Validation of the model: predicted and measured values of different microbial load of raw water




Koliforme bakterier (count/100ml)
Pred(Koliforme bakterier (count/100ml))
Int. entero- kokker (count/100ml)
Pred(Int. entero- kokker (count/100ml))
adeno_conc
Pred(adeno_conc)


2.500
2.000
1.500
1.000
0.500
0.000
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39 41 43 45 47 49 51 53 55 57 59 61 63
norog1_conc
Pred(norog1_conc)
3.000
2.500
2.000
1.500
1.000
0.500
0.000
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39 41 43 45 47 49 51 53 55 57 59
norog2_conc
Pred(norog2_conc)