Norwegian University of Life Sciences
Faculty of Veterinary Medicine and Biosciences
Department of Chemistry, Biotechnology and
Food Science

Master Thesis 2015
60 credits

Comparison of Multivariate Methods to Predict the Quality of Drinking Water in Norway

Siddhartha Dhungana

Comparison of multivariate methods to predict the quality of drinking water in Norway

Siddhartha Dhungana

A Dissertation
Presented to the Faculty
of Norwegian University of Life Sciences
in Candidacy for the Degree
of Masters of Bioinformatics and Applied Statistics

Recommended for Acceptance
by the Department of
IKBM
Supervisor: Ellen Sandberg (NMBU), Trygve Almøy (NMBU),
Carl Fredrik Nordheim (NIPH)

May 2015

Abstract

Water quality in the Water Distribution System (WDS) varies over time. The quality of water in the Water Distribution System (WDS) is measured through Heterotrophic Plate Count (HPC) as an indicator organisms. Parameters such as color, pH, turbidity, conductivity, temperature, organic matters as well as the components of water distribution network system such as generic pipes and their ages, lubricants and storage tanks are linked with water quality. For multivariate modelling of these parameters data were collected from Norwegian Institute of Public Health (NIPH) as yearly average of HPC including physical, chemical and microbial water quality parameters.

Multivariate statistical methods have been applied to predict the quality of drinking water in water distribution system. Model such as Multiple Linear Regression (MLR), Principal Component Regression (PCR) and Partial Least Square Regression (PLSR) methods are adopted to identify the factors that affect the HPC in water distribution network system and consequently the quality of the water. Due to large number of insignificant variables a subset model was chosen using the criteria of Mallow's C_p and $Adj - R^2$. The fitted models were validated through Leave One Out (LOO) cross validation method. Best subset model was performed well on both training and test data set but still suffered from multicollinearity. As an alternative approach PLSR model with three latent components which is predicted closer than PCR model with seven components. The number of components are chosen through prediction error during cross validation.

Key words: Heterotrophic Plate Count, MLR, PCR, PLSR, Cross Validation

Acknowledgements

I would like to thanks my supervisors Ellen Sandberg, Trygve Almøy and Carl Fredrik Nordheim for their guidance, encouragement and valuable suggestions. In addition, I am very greatful with Ellen Sandberg for her creative and motivating counseling. I am also obliged to Carl Fredrik for his help on dataset preparation and assistance on introductory part of this thesis.

I am extremely thankful and indebted to Raju Rimal for his assistance and time on *R* programming and LaTeXwriting. Further, I want to thank Norwegian Institute of Public Health for allowing me to use their water quality data. I am also greatful to Dr. Vidar Lund for his insightful comments and helpful advice. I would like to thanks all the teachers, staffs in the bio-statistics group of NMBU for their direct and indirect help during my study.

Finally, special thanks goes to my parents for their patients, love and motivation.

To my Grandmother.

Contents

List of Tables

List of Figures

Abbreviations

CFU	Colony Formings Units
HPC	Heterotrophic Plate Count
WDS	Water Distribution System
OLS	Ordinary Least Square
MLR	Multivariate Linear Regression
PCA	Principal Component Analysis
PCR	Principal Component Regression
PLS	Partial Least Square
VIF	Variance Inflation Factor
PVC	Polyvinyl chloride
PEL	Polyethylene
GUP	Galvanized Pipe
RMSE	Root Mean Square Error
RMSEP	Root Mean Square Error of Prediction
RMSECV	Root Mean Square Error of Cross-validation
PRESS	Prediction Sum of Squares
BIC	Bayesian Information Criterion

Chapter 1

Introduction

1.1 Introduction

Quality drinking water is defined as water that is clear, free from odor and taste and free from harmful substances of any kind and generally wholesome. To obtain drinking water quality, the water supply system has according to Norwegian regulations to have two hygienic barriers. A protected water source and catchment area is regarded as one barrier and water treatment including disinfection is the second. Only when the water source is protected ground water of good quality, the food authority can decide that water treatment is unnecessary. From the above one can conclude that raw water is treated when necessary in treatment plants and made ready to distribute with a minimum standard complying with drinking water regulations. However, the distribution of water with good quality from the treatment plant can be affected by passing through long water distribution networks and that can be a great challenge for the water utilities. Among the different water quality parameters, Prophetic Plate Count, here abbreviated HPC, (in Norwegian

"Kimtall") is considered as one of the main indicators of water quality parameters in Norway. One of the technique to identify the quality of water in the distribution network is to monitor the levels of HPC. Increased levels of HPC is not necessarily a health risk, but it indicates microbial growth and the possible contamination of the distribution network.

A Water Distribution Systems (WDS) consists of water mains, pumps and control valves and reservoirs such as water towers and distribution pipes. Information on the types of piping materials, age of pipes, volume of storage tanks, number of manholes, number of leakage repairs, episodes of disrupted services, etc. is collected yearly by the Norwegian Food Control Authority. Good water is connected with physical, chemical and microbial characteristics of water. Physical characteristics consists of odor, taste, color, turbidity and pH. Microbial quality consists of the water with accepted level of bacteria such as E.coli, Fecal Coli-forms, Total Coliforms and HPC. These parameters are collected from a sampling points of WDS and analyzed in a laboratory system. The parameters can be interrelated to each others in WDS and their collective study can make a sense of water quality distribution and monitoring.

1.2 Objective

The main objective of this thesis is

1. To find the relationship between the type of material used in distribution system and physio-chemical and microbiological water quality parameters.

2. To analyze HPC using different multivariate statistical methods.

3. To compare multivariate statistical models for predicting HPC and finding the best model using cross validation method.

1.2.1 Overview of Methodology

The study is about the comparison of different statistical methods using the Norwegian water quality data.

- Establish contact with the Norwegian Institute of Public Health (NIPH) to secondary data (the data reported by the water utilities to the Norwegian Food Control Authority (NFCA)). The NFCA regularly transfer waterworks data to the NIPH waterworks registry.
- Analyze the collected data to find the relationship between different water quality parameters and piping materials using statistical tool such as Multiple Linear Regression (MLR), Principal Component Regression (PCR) and Partial Least Square Regression (PLSR).
- Identifying a model that best describes HPC through their comparison.

1.3 Water Production and supply in Norway

Norway has an abundance source of fresh water supply. Surface water is the most important source of drinking water in Norway which supplies nearly 90% of the population. This is higher than other Scandinavian countries. In Denmark and Iceland 90% of the people are served by groundwater whereas this ratio for Sweden and Finland is only about 40 to 50%. In Norway waterworks are responsible for water production and distribution together with the maintenance and monitoring of water quality parameters. There are almost 1616 registered waterworks in Norway, of which 1200 serves less than 1000 people and only 5 waterworks serves more than 100000 people (Liliane Myrstad, 2011). Each waterworks are serving at least 50 people or 20 households. As the water supply in Norway is dominated by small waterworks it can be challenge for them to obtain sufficient resources for operation and adequate maintenance for treatment plant and distribution systems.

1.4 Water Distribution Systems

The main purpose of WDS is to supply a sufficient amount of drinking water with good water quality. The entire distribution system connected with different components such as service water reservoirs, distribution network, storage tanks, pump stations and system monitoring and control. Treated water from treatment plants has to be delivered to consumers by means of pipes known as distribution network. Kawamura (2000) divided distribution network into two parts called Trunk mains and Distribution mains. Trunk mains are used to transport the larger volumes of water with high pressure from treatment plant to storage tanks, while distribution mains carry the water form storage tank to the houses. The later system includes pump-stations and system monitoring as well.

In drinking water distribution system of Norway, material such as metals, cements and plastics are common. Among them plastics material are widely used. There are some other types of pipe as well but their contribution to the total length of the pipelines is only less than 1 percent. The length of pipelines is approximately $49200km$, excluding the individual service lines to water consumption sectors. Plastic materials contributes more than 50% of the total installed pipes while steel and iron pipes (34%) are still popular. Other variables are the information about the storage tanks and their volume, pumping stations, water pipe leakage repair and the planned and unplanned disruptions. Regular cleaning of pipes, emergency maintenance and leakage repairs are performed during disruption. The water production and consumption process is explained in figure - 1.1.

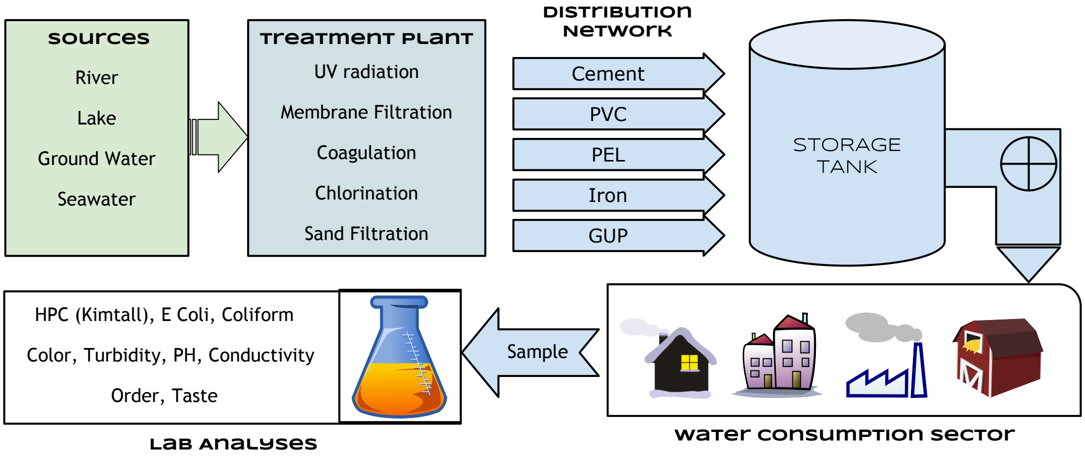


Fig 1.1: Procedure of water treatment process

1.5 Water Quality Variation in Distribution System

Water distribution system (WDS) is targeted to supply enough amount of quality drinking water. However the quality of water is subjected to substantial changes during transport through long distribution systems (Momba et al., 2000). WDS are considered as biological and chemical reactors with transported water where quality changes with time and places (LeChevallier, Welch, and Smith, 1996). Biological changes refers to the regrowth of bacteria in the presence of biofilm inside the inner wall of pipe. Biofilm refers a group of microorganism forming a layer on a inner wall of pipe within an aquatic environment. The biofilm formation and microbial diversity inside the pipe will be influenced by different parameters including fluctuation of temperature due to seasonal change, type of pipe material used for the distribution systems and concentration of biodegradable compounds as a energy source for microbial growth (VAN DER KOOIJ and Zoeteman, 1978). However, pipe surface itself may influence the activity of biofilm composition. Biofilms developed more quickly on iron pipe surfaces than on plastic polyvinyl chloride (PVC) pipes, no matter that adequate corrosion control was applied (Norton and LeChevallier, 2000).

1.5.1 Heterotrophic Plate Count (HPC)

Waters of all kinds contain a variety of microorganisms. Microorganisms (bacteria, molds and yeasts) that uses organic carbon as an energy source for growth are called heterotrophs. Majority of bacteria found in drinking water distribution systems are considered heterotrophs. Heterotrophic Plate Count (HPC) is a test method which estimates total no of culturable microorganisms present in a volume of water. Several other terms that have been used to describe this group of bacteria in water include “standard plate count”, “Plate Count”, “Total Bacterial Count”, “Water Plate Count”,“Colony Count”(Allen, Edberg, and Reasoner, 2004). In Norway it is abbreviated as “Kimtall” and used to measure the overall bacteriological quality of drinking water in water distribution systems. In Norway, there is no threshold value for HPC however if the value exceed 100 CFU/ml the cause should be investigated.

Generally the water authority will expect that HPC bacteria concentration below 10 cfu/ml in finished drinking water but within the drinking water distribution the bacterial regrowth leads to the increase in the density of HPC bacteria. Moreover, the high density can be influenced by the bacterial quality of the finished water entering the system, temperature, residence time, presence or absence of a disinfectant residual, construction materials, surface-to-volume ratio, flow conditions, the availability of nutrients for growth and in chlorinated systems, the chlorine/ammonia ratio and the activity of nitrifying bacteria(Allen, Edberg, and Reasoner, 2004,Payment, Sartory, and Reasoner, 2003,VAN DER KOOIJ and Zoeteman, 1978). However the different method of measuring HPC, and the different types of culture media may have different amount of HPC measurement.

1.6 HPC as a water quality indicator parameter

The microbiological water quality in distribution system can be assessed by measuring the amount of HPC bacteria. HPC testing has a long history in water management. At the end of 19 century HPC test were employed to proper functioning of treatment process and there after the indirect indicator of water safety. In many countries HPC measurements are used (WHO) et al., 2002 as a tools for

- monitoring the effectiveness of water treatment process

- obtaining supplemental information on HPC levels that may interfere with coliform detection on water samples collected for regulatory compliance monitoring
- assessing changes in finished water quality during distribution and storage and distribution system cleanliness
- assessing microbial growth on material used in the construction of potable water treatment and distribution systems
- measuring of numbers of regrowth organisms that may or may not have hygienic significance
- monitoring and performance of filtration and disinfection processes

1.7 Public health aspect of HPC bacteria

Heterotrophic population consists of a broad range of bacteria and yeast. At an international meeting of experts in Geneva, Switzerland, it was concluded that heterotrophic bacteria in drinking water is not a health concern to the general public. However, some bacteria present in a heterotrophic population are opportunistic pathogens that could infect individuals with weakened immune systems. "Heterotrophic bacteria belonging to the following genera have been associated with opportunistic infections: Acinetobacter, Aeromonas, Chryseobacterium (Flavobacterium), Klebsiella, Legionella, Moraxella, Mycobacterium, Serratia, Pseudomonas, and Xanthomonas. These organisms have been mainly associated with nosocomial (hospital acquired) infections, including wound infections, urinary tract infections, post-operative infections, respiratory infections, and infections in burn patients". which is also called as hospital acquired infections such as wound infections, respiratory infections, post operative infections (Allen, Edberg, and Reasoner, 2004).

1.8 Factors affecting water quality within the Distribution System

Microorganisms will grow in water at certain temperature and surfaces in contact with water as biofilms. This biofilm provide a habitat for microorganism inside the pipe, In addition microorganism also have the ability to colonize within the distribution system. Moreover, the rate of colonization will be different with the different types of pipe material used in the distribution. Momba and Makala (2004) found the correlation between type of pipes and bacterial amount within the water distribution system. Water distribution pipes with rough surface have higher potential for bacterial regrowth (Kooij, 2003;Ridgway and Olson, 1981). In addition, other water contact materials such as pump lubricants, pipe coating and plumbing system can also support the growth.

Apart from piping materials the after growth and regrowth of bacteria must be taken into consideration. After growth refers to the growth of bacteria occurring naturally in distribution systems whereas regrowth is the ability of bacteria to recover from treatment process and then multiply within the distribution system. The factors such as bacterial quality of the finished water entering the system, temperature, presence or absence of disinfectant residuals and the availability of nutrients for growth and activity of nitrifying bacteria can affect water quality within the distribution system. The effect of these factor can be summarized in following four points.

1.8.1 Loss of Disinfectant Residuals

Disinfection is a process of removing disease-causing microorganism by means of chemical process such as using chloramines. Some large waterworks use disinfectant residuals to ensure microbiological quality of water and to protect distributed water from re-contamination and regrowth. The loss of disinfectant residual resulted from line breaks and cross-connections can weaken the barrier against microbial contamination and encourage the growth of pathogens.

1.8.2 Pipe surface and water contact material

Type of pipe and roughness of its surfaces which are specific characteristic of distribution system affects the dynamics of microbial growth. Water distribution pipes with rough surface support higher biofilm densities and thus higher potential for bacteria regrowth (Colbourne et al., 1984). Furthermore pipe material themselves can be a factor for growth. Pedersen (1990) reported bacterial population in PVC pipe is lower than those in steel pipe and other generic types of pipe.

In addition, water contact material such as pump lubricants, pipe coating, pipe gaskets can play a positive role for bacteria regrowth in WDS. It is generally accepted that as the pipe is getting older the deposition and pipe sediment in WDS became common and consequently provides a nutritional source for bacteria in connection with the available compound in water such as iron, potassium and manganese.

1.8.3 Organic matter

Organic Carbon present in drinking water either naturally or due to the chemical used in the treatment plants. The total organic carbon is divided into two parts as a) Biodegradable organic dissolved carbon(BODC) b) Assimilable organic carbon (AOC). The first one represents the metabolic activities of bacteria while the other one measures the bacterial growth potential. In the bacterial regrowth process the available carbon is consumed by bacteria for regrowth in distribution network (Kooij, 2003).

1.8.4 Environmental factors

Some environmental factors such as temperature, pH and dissolved oxygen influences the growth of bacteria within WDS. Increasing temperature is always positive for bacterial growth and thus the seasonal changes can alter the metabolism of microorganism. In the distribution system, when the water is warm, bacterial growth is rapid so water temperature is considered to be one of the important factors for affecting microbial growth (WHO). Some bacteria grow within a narrow temperature range where others are able to growing wider range of temperature. LeChevallier, Welch, and Smith (1996) have found significant bacterial growth in a water system at temperature $0 - 5$ and $> 20°C$. Similarly pH can influence microbial growth. Corrosion of iron pipe material can add alkalinity and raise pH value as well. It is obvious that corrosion process also consumes available oxygen from water.

1.9 Water quality modeling

Water quality data are not normally distributed and linear correlation fails to describe the exact relation of different water quality variables. No single technique is sufficient to find the significance of HPC to other water related variables. There are numerous research projects that have been conducted to predict the HPC bacteria but the consideration of predictor variables is limited. As the water is distributed from the same pipe throughout the year,testing the influence of pipe materials and age of the pipes should also be considered. This study will help researchers to increase their understanding of the microbial growth dynamics in drinking water distribution networks. By establishing cause of relationships between bacterial growth and water quality, one can be able to construct a statistical model to predict water quality changes. The complex nature of relationship between quality parameters can only be achieved by multivariate statistical tools. The multivariate treatment of water quality data assist to extract possible influencing factors that cause the variation in water quality. Furthermore, the idea would help water authorities to make effective water safety plans.

Chapter 2

Methodology

2.1 Methodology

Multivariate statistical regression techniques gives a tool for empirical modeling of the data matrix. The purpose of empirical modeling is to obtain a model that can describe the underlying behavior of the selected variables. The improvement on data collection system and modern technology has resulted that model based on least squared method can lead to imprecise parameter estimation either due to presence of more variable or due to the number of observation is less than number of predictor variables or multicollinearity among the variables. To overcome these difficulties the multivariate projection method such as Principal Component Regression(PCR) and Partial Least Square Regression (PLSR) has been used. Both methods can handle the situation above by capturing the underlying characteristics of variables in terms of few number of principal components or latent variables which are the combinations of selected original variables.

2.2 Notation

In this dissertation bold faced lower case letters $\boldsymbol{y}$ are vectors and upper case letters $\boldsymbol{X}$ are matrices. Similarly the index $i = 1, \ldots, m$ denote observations and index $j = 1, \ldots, n$ denote the predictors. For regression approach $\boldsymbol{X}$ denote the predictors matrix and $\boldsymbol{y}$ for the response vector.

2.3 Least Square Regression Method

Suppose that $\boldsymbol{X} = [x_1, \ldots, x_n]$ be n predictor and $\boldsymbol{y}$ be the $(m \times 1)$ response variables.Assuming linear relation exist between $\boldsymbol{y}$ and $\boldsymbol{X}$ and hence least square equation will be

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon} \tag{2.1}$$

where,$\epsilon(m \times 1)$ be the error of observations measured in the direction of $\boldsymbol{y}$ axis,

The main feature of least square method is to estimate the parameter β such that the norm of the ϵ is minimized (Johnson, Wichern, et al., 1992).

$$\sum_{i=1}^{m} \epsilon_i^2 = \boldsymbol{\epsilon}'\boldsymbol{\epsilon} = \sum_{i=1}^{m} (y_i - x_i^T \beta)^2 \tag{2.2}$$

$$\sum_{i=1}^{m} \epsilon_i^2 = \boldsymbol{\epsilon}'\boldsymbol{\epsilon} = \sum_{i=1}^{m} (y_i - x_i^T \beta)^2 = (\boldsymbol{y} - \boldsymbol{X}\beta)^T((\boldsymbol{y} - \boldsymbol{X\beta})) \tag{2.3}$$

By differentiating with respect to $\boldsymbol{\beta}$ the minimum of the square occurs at values of $\hat{\boldsymbol{\beta}}$ that satisfy the normal equation 2.3. So,

$$\boldsymbol{X^T X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}^t\boldsymbol{y} \tag{2.4}$$

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X^T X})^{-1}\boldsymbol{X}^t\boldsymbol{y} \tag{2.5}$$

Here it is assumed that X has a full rank and $\boldsymbol{X^T X}$ matrix contains the variance co-variance matrix for centered data. The solution of β depends on the data matrix $\boldsymbol{X^T X}$. Suppose r is a rank of $\boldsymbol{X}$.if $r < n$ the least square solution is not unique.

2.3.1 Properties of OLS estimator

The OLS estimator defined in 2.5 has the following properties.

1. It is an unbiased estimate.

   Mathematically,

$$\begin{aligned} E(\hat{\beta}) &= E\left[(\boldsymbol{X^T X})^{-1}\boldsymbol{X}^t y\right] \\ &= E\left[(\boldsymbol{X^T X})^{-1}\boldsymbol{X}^t(\boldsymbol{X}\beta + \epsilon)\right] \\ &= (\boldsymbol{X^T X})^{-1}\boldsymbol{X^T X}\boldsymbol{\beta} \\ E(\hat{\beta}) &= \beta \end{aligned}$$

Also,

$$\begin{aligned} V(\hat{\beta}) &= E\left[(\hat{\beta}-\beta)(\hat{\beta}-\beta)^T\right] \\ \hat{\beta}-\beta &= (\boldsymbol{X^TX})^{-1}\boldsymbol{X}^t y-\beta \\ &= (\boldsymbol{X^TX})^{-1}\boldsymbol{X}^t(\boldsymbol{X}\beta+\epsilon)-\beta \\ &= (\boldsymbol{X^TX})^{-1}\boldsymbol{X}^t\epsilon \end{aligned}$$

Now,

$$\begin{aligned} V(\hat{\beta}) &= E\left[\boldsymbol{X^TX})^{-1}\boldsymbol{X}^t\epsilon\epsilon^t\boldsymbol{X}^t(\boldsymbol{X^TX})^{-1}\right] \\ &= (\boldsymbol{X^TX})^{-1}\boldsymbol{X}^t E\left[\epsilon\epsilon^t\right]\boldsymbol{X}^t(\boldsymbol{X^TX})^{-1} \\ V(\hat{\beta}) &= \sigma^2(\boldsymbol{X^TX})^{-1} \end{aligned}$$

2. It provides unbiased estimates of the elements of β which have the minimum variance. Such estimator is called Best Linear Unbiased Estimator (BLUE).

2.3.2 Linear Model assumption

Linear regression model holds the following assumptions.

1. The response variable $\boldsymbol{y}$ is a linear functions of a set of predictor variables.

2. The errors ϵ_i are independent

3. The errors ϵ_i have equal variance

4. The errors ϵ_i are normally distributed.

2.4 Problem in least Square Method

If the data matrix $\boldsymbol{X}$ is not a full rank some linear combination of $\boldsymbol{X}$ tends to zero. It means that the inverse $\boldsymbol{X'X}$ doesn't exist and diagonal value of $\boldsymbol{X'X}$ will be large. This leads to larger estimated variance for β_i and insignificant β estimates as well (Johnson, Wichern, et al., 1992). The situation is also called multicollinearity.

To overcome this multicollinearity problem two approaches has been purposed. One possibility is to use only a subset of a predictor variables where a subset is chosen so that the model doesn't have multicollinearity. The subset predictor can be achieved by stepwise regression procedure. In some cases, the selection of explanatory variable is a direct solution of multicollinearity however, in many cases, even in the absence of collinearity among predictor variables reducing dimensionality problem is often beneficial. The other method is to use a dimension reduction technique such as PCR and PLS. The Variance Inflation factor (VIF) can be used to check the collinearity among predictor variables. VIF values above 10 shows the strong multicollinearity among the variables (Chatterjee and Hadi, 2013) used in model fitted using equation 2.1.

$$VIF = \frac{1}{1 - R_j^2} \tag{2.6}$$

Where R_j^2 coefficient of determination for model fitted with x_j as response and all other $x_k, k = 1, \ldots, j-1, j+1, \ldots m$ as predictor.

2.5 Principal Component Regression

2.5.1 Principal component analysis

PCA is commonly defined on text books such as (Bishop et al. (1995),Jolliffe (2005),Martens (1992),Mardia, Kent, and Bibby (1979)). PCR is the application of least square regression of y on a selected set of principal components which are the linear combination of original variables. Hence PCR is based on the results from PCA. The objective of PCA is to achieve parsimony and reduce dimensionality by extracting the smallest number of components that account the most of the variation in the original multivariate data. This method is based on the characteristics of eigenvalues and eigenvectors.

2.5.2 Mathematical Expression

Consider the data set with n variables and m observation then the first principal component z_1 can be written as $z_1 = w_{11}X_1 + w_{12}X_2 + \ldots + w_{1n}X_n$

where w's are called weights or loadings of the components defined in such a way that $w_{11}^2 + w_{12}^2 + \ldots + w_{1n}^2 = 1$ similarly second principal component z_2

$$z_2 = w_{21}X_1 + w_{22}X_2 + \ldots + w_{2n}X_n$$

with $w_{21}^2 + w_{22}^2 + \ldots + w_{2n}^2 = 1$ if there are n variables there are n principal components and each component is a linear combination of set of n original variables. i.e

$$
\begin{aligned}
z_1 &= w_1^{'}X = w_{11}x_1 + w_{12}x_2 + ... + w_{1n}x_n \\
z_2 &= w_2^{'}X = w_{21}x_1 + w_{22}x_2 + ... + w_{2n}x_n \\
&\vdots \\
z_m &= w_n^{'}X = w_{m1}x_1 + w_{m2}x_2 + ... + w_{mn}x_n
\end{aligned}
$$

Here the random variable $\boldsymbol{X}$ has co-variance matrix $\boldsymbol{S}$ with eigenvalues $\lambda_1, \ldots \lambda_n$. Also the eigen values are in $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$

In matrix notation

$$\boldsymbol{W} = \begin{bmatrix} w_1' & w_2' & \ldots & w_n' \end{bmatrix}' \tag{2.7}$$

Since the principal component depends upon the co-variance/correlation matrix of $\boldsymbol{X}$ hence $z_i = WX_i$.

2.5.3 Principal Component Regression

Principal Component Regression (PCR) is a method of regressing dependent variable on the linear combination of independent variable and thus the linear combination are called principal components. Consider a standard regression model defined on

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon} \tag{2.8}$$

here it is assume that predictor variable are standardized so that $\boldsymbol{X}'\boldsymbol{X}$ represents the correlation matrix. The value of PCs for each component will be

$$\boldsymbol{Z} = \boldsymbol{X}\boldsymbol{A} \tag{2.9}$$

where $\boldsymbol{A}$ is $p \times p$ orthogonal matrix, so $\boldsymbol{X}\beta$ can be written in another form as

$$\boldsymbol{X}AA'\boldsymbol{\beta} = \boldsymbol{Z}\gamma \tag{2.10}$$

Now the original equation becomes

$$\boldsymbol{y} = \boldsymbol{Z}\gamma + \boldsymbol{\epsilon} \tag{2.11}$$

2.6 Partial least square Regression

This is the modern method of constructing predictive models when the data matrix is large and the variables are colinear. PLS or also called ”projection to latent structures” is a method developed by Herman Wold (1975). The theoretical portion of PLS is based on a book by Varmuza and Filzmoser (2009).

PLS is a general technique that generalizes the important features of MLR and PCR. When Y is a vector and $\boldsymbol{X}$ is full rank then ordinary least square regression could be good enough for analytical purposes, but if $\boldsymbol{X}$ is singular the normal regression process is no longer feasible. This difficulties would be handled by partial least square techniques. PLS technique extract factors from both X and Y such that co-variance between the extracted factor is maximized.the process is connected with the linear decomposition of X and Y such that $X = TP^t + E_x$ and

$Y = UQ^t + E_y$, where

$$T = \text{X-score and } U = \text{Y-score}$$

$$P = \text{X-loadings and } Q = \text{Y-loadings}$$

$$E_x = \text{X-residuals and } E_y = \text{Y-residuals}$$

The PLS algorithm automatically predicts Y using the extracted Y-scores (U). The X-scores in (T) are linear combinations of the X variables and Y-scores in (U) are the linear combinations of Y variables.

Suppose t_j, u_j, p_j and q_j denote the j^{th} columns of T, U, P and Q respectively, where ($j = 1, \ldots, a$).

2.6.1 PLS computational procedure using NIPALS algorithms

Consider the general form of PLS1 algorithm. Suppose $\boldsymbol{X}$ and y are mean centered data matrix and vector respectively. Since PLS1 algorithm is start with the initialization as $j = 1$,$\boldsymbol{X_1} = \boldsymbol{X}$, $y_1 = y$. The whole process is proceed to finding g latent variables.

1. Compute the weight w_j as

$$w_j = \frac{\boldsymbol{X_j'} y_j}{\left\| \boldsymbol{X}_j' y_j \right\|} \tag{2.12}$$

The weights are normalized to length 1 and this gives the direction of large variations in x-values accompanied by corresponding y-values.

2. Compute the score vector t_j as a linear combination of columns of $\boldsymbol{X}$ with weights w_j i.e

$$t_j = \boldsymbol{X}_j w_j \tag{2.13}$$

3. Compute the loading vector p_j by regressing the columns of $\boldsymbol{X}$ on t_j

$$p_j = \frac{\boldsymbol{X}_j' t_j}{t_j' t_j} \tag{2.14}$$

4. Compute the loading vector q_j by regressing y on t_j

$$q_j = \frac{t_j' y_j}{t_j' t_j} \tag{2.15}$$

5. Calculate

$$\boldsymbol{X}_{j+1} = \boldsymbol{X}_j - t_j p_j' \tag{2.16}$$

$$y_{j+1} = y_j - t_j q_j \tag{2.17}$$

Here $\boldsymbol{X}_{j+1}$ represents the residuals after regressing $\boldsymbol{X}_j$ on t_j and y_{j+1} represent the residuals after regressing y on t_j.

6. Stop if $j = g$, otherwise if other component needed suppose $j = j + 1$ and return to step 1. After computing g iteration the new relation will be

$$\boldsymbol{X} = TP' + \boldsymbol{X}_{g+1} \tag{2.18}$$

$$y = TQ + y_{g+1} \tag{2.19}$$

2.6.2 Prediction on Partial Least Square

The final fitted PLS regression model for predicted response $\hat{Y}$ of the form

$$\hat{Y} = X\beta + E$$

where $\beta = W(P'W)^{-1}Q'$ and $P = X'T(T'T)^{-1}$

2.7 Comparison between OLS, PCR, PLSR

Advantages of using PLSR and PCR over OLS.

1. The regression variable T are linearly independent so that problem of multicollinearity is addressed.

2. Only the most important latent variables T are included thus that the risk of modeling noise in the data is reduced.

3. PCR captures the variability presented in the X matrix only by maximizing the length of each score vector t.

4. PLS captures the variability presented in both X and Y by maximizing the co-variance between t and u

2.8 Model selection and assessment

Regression model makes sense when the model meets the specified criteria and then can be used for prediction purposes. This can be done through model selection and assessment. The first one is concerned with selection of best model through its performance within the given data set. Model assessment, on the other hand, estimates the model prediction error after the model selection procedure is valid.

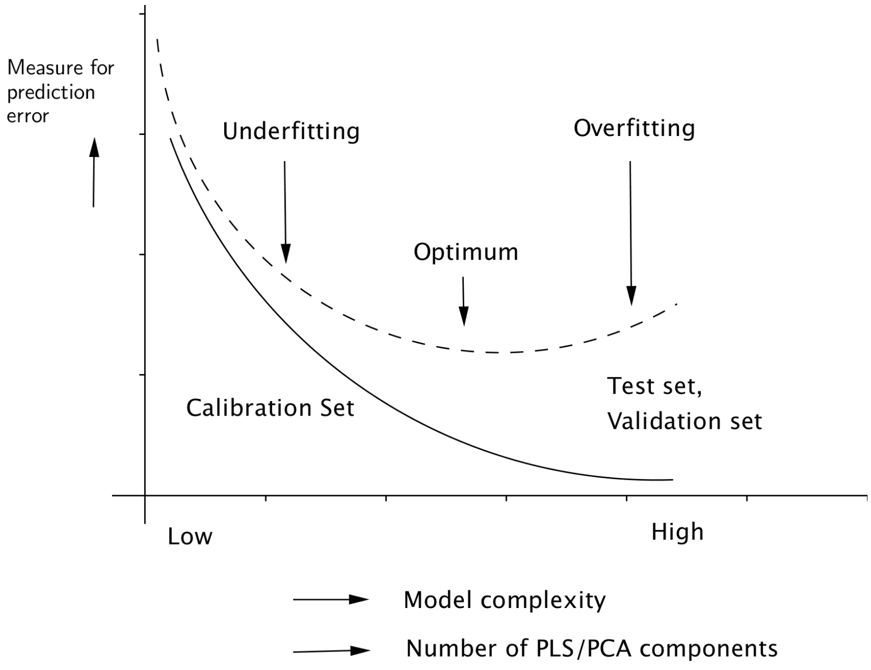


Fig 2.1: Model Complexity versus Prediction error for calibration set and Test set

In some cases the selected model performs well for future dataset however in many cases the regression model is often suffered from over-fitting and under-fitting. The more complex model is capable to fit the calibration set with low prediction error. i.e. a highly complicated model can fit almost all dataset perfectly but the model can not perform that well in case of observations that are not included in the model. The figure 2.1 shows the model complexity and is adopted from the book *Introduction to multivariate statistical analysis in chemometrics* by Varmuza and Filzmoser.

2.8.1 Performance with number of variables

Variable selection method intended to find the optimal number of variables that can predict the response adequately. On one hand, the model contains few number of predictor variables this may lead to poor prediction performance, on the other hand, larger number of predictor may results overfiting. Before selecting variables the model assumptions should be fulfilled. Variable selection methods access following criterion for selecting best subset model from a full model.

Adjusted R^2

For a given n no of variables with m observation the adjusted R^2 defined by

$$\text{Adj-}R^2 = 1 - \frac{m-1}{m-n-1}(1-R^2) \tag{2.20}$$

Where R^2 is called coefficients of determination. In this criteria a model with larger Adj-R^2 value is preferable.

Akaike's Information Criterion (AIC)

This is commonly used method for variable selection using stepwise regression or best subset regression. AIC is given by

$$AIC = m\log(\frac{(RSS)}{m}) + 2n \tag{2.21}$$

under this criteria a model with small AIC value is preferable.

Bayes Information Criterion (BIC)

$$BIC = m \log(\frac{(RSS)}{m}) + n \log n \tag{2.22}$$

Here also smaller value of BIC is preferable.

Mallow's Cp

This is a stopping rule for subset selection method purposed by (Mallows (1973))

$$Cp = \frac{RSS}{s^2 - m + 2n}$$

where s^2 is the estimate error variance for full model. A model with smallest C_p would be preferred.

2.9 Cross Validation

Leave One Out Criteria

Cross-validation is the modern statistical techniques, that is commonly used for assessing the goodness of fit and predictive ability of statistical model. Common way of validation technique consists of the division of whole data set into two parts called training data set and test data set. First analysis is performed on training data and then the test set is used for validation. Validation techniques depends on the way of partition of data set among which leave one out (LOO) cross validation is one of them. In this method one observation is held out as a single test data and the remaining $n-1$ observation as training data set. A regression is performed on training data set and the held out observation is predicted using this model.

RMSE

The root mean square error (RMSE) gives the idea about the fit of the model to the data set used. Mathematically,

$$RMSE = \sqrt{\frac{\sum_{i=1}^{m}(y_i - \hat{y}_i)^2}{m}}$$

RMSECV

RMSECV is contrast to RMSE which is a measure of model's ability to predict new samples. RMSECV is related to the PRESS values.

$$RMSECV = \sqrt{\frac{PRESS}{m}}$$

PRESS

The Prediction sum of Square (PRESS) is a validation method and used to compare regression model as well as the predictive ability of a model. Mathematically,

$$\text{PRESS} = \sum_{i=1}^{n}(y_i - \hat{y}_{i(i)})^2$$

The smaller the PRESS value the better the model's predictability is.

Chapter 3

Results and Discussion

3.1 Data Organization

Construction of data matrix such as processing, coding, missing data removal, transformation, scaling all were made using R statistical package. An $m \times n$ data matrix was created by considering available water quality variable. The selection of the variables are based on an availability of data and their theoretical relationship with HPC. Few distribution variables have highly scattered values and log transformation were taken to minimize the skewness problem. To make a better understanding on analysis interaction terms of some variables were introduced as well. For the analytical purpose of PCR and PLS the column centering and scaling was performed. Scaling of variable gives the equal footing relative to their variation in data. Finally the logarithm values are transformed back for post modeling computation purposes.

3.2 Data Analysis and Discussion

Considering X as the data matrix consisting of 173 observation and 38 variables including interaction terms as well.The water quality parameters,including chemical physical and microbiological variables were considered over 10 years form 1998 to 2008 by waterworks. The variable used in this analysis were coded as in table 3.1 and table 3.2. The water quality data were measured monthly, weakly and in some cases daily throughout the year. A yearly average for each of the water quality parameters were used in this analysis. Further, the pipeline system represent the total installed pipeline in meter by respective waterworks. Those waterworks that had inadequate variable recordings were excluded from the study.

Variable.Code	Unit.of.measurement	Variable.Name
HPC	cfu/ml	Heterotrophic Plate Count
Col	mg/	Color
Ph	pH	Ph
Tur	FTU(FNU)	Turbidity
Irn	mg/l Fe	Iron
Alu	mg/l Al	Aluminium
TOC	mg/l TOC	Total Organic Carbon
Temp	Celcius	Temperature
Cond	mS/m	Conductivity
Cal	mg/l Ca	Calcium
Sod	mg/l Na	Sodium
Alk	Mmol/l	Alkalinity
Mang	mg/l Mn	Manganese
ReCh	mg/l Cl	Residual Chlorine
COD	mg/l O	Chemical Oxygen Demand
Nita	mg/l N	Nitrate
Niti	mg/l N	Nitrite
Amonia	mg/l N	Amonium

Table 3.1: Water quality parameter and their code

Variable.Code	Unit.of.measurement	Variable.Name
PVC	meter	Polyvinyl Chloride Pipe
PEL	meter	Polyethylene Pipe
GUP	meter	Galvanized plastic pipe
Cem	meter	Cement Pipe
Iron	meter	Iron pipe
PiRe	meter	Pipe Repair
VoTa	cubic meter	Volume of Tank
PlDi	time * person affected	Planned Disruption
UPDi	time * person affected	Unplanned Disruption
B1910	meter	Pipe Before 1910
B1940	meter	Pipe Before 1940
B1970	meter	Pipe Before 1970
A1970	meter	Pipe After 1970
A2001	meter	Pipe After 2001

Table 3.2: Distribution network variables and their code

3.3 Descriptive statistics

The descriptive measure of statistics such as mean, standard deviation, minimum value, maximum value and skewness are present in table-3.3 to identify the nature of variable.

variable	n	Min	Max	mean	sd	skewness
HPC	173	0.00	120.00	11.38	17.68	3.60
Irn	173	0.00	0.40	0.05	0.06	2.16
ReCh	173	0.00	0.65	0.05	0.06	7.10
TOC	173	0.00	5.80	2.39	1.36	0.26
Col	173	0.13	35.00	10.66	8.21	0.83
Cal	173	0.00	38.38	14.59	9.23	0.10
Cond	173	2.13	35.00	11.01	7.37	1.04
COD	173	0.50	10.35	3.01	1.98	1.74
Alk	173	0.00	2.60	0.61	0.49	1.50
Sod	173	0.00	43.00	5.10	4.64	3.73
Mang	173	0.00	0.36	0.02	0.05	6.44

Temp	173	0.00	18.50	6.59	1.97	1.38
pH	173	5.30	8.50	7.33	0.81	-0.86
Tur	173	0.04	4.09	0.42	0.61	3.48
Alu	173	0.00	0.71	0.08	0.09	3.95
Niti	173	0.00	0.40	0.03	0.09	4.01
Nita	173	0.03	0.70	0.23	0.15	1.75
Amonia	173	0.00	0.58	0.02	0.09	6.33
Irp	173	0.00	1550747.00	717249.31	545771.46	0.14
PVC	173	0.00	104410.00	21505.98	26505.01	1.62
PEL	173	0.00	55159.00	18846.83	18341.43	0.38
GUP	173	0.00	10900.00	4407.32	4698.20	0.38
A2001	173	0.00	79000.00	16299.58	21124.24	1.27
UPDi	173	0.00	81000.00	19154.18	31123.76	1.11
VoTa	173	215.00	282900.00	128130.32	90299.67	-0.26
PlDi	173	0.00	336532.00	28925.41	69867.01	3.49
PiRe	173	0.00	14897.00	3743.34	3784.11	0.99
Cem	173	0.00	55000.00	23364.65	23335.22	0.29
B1910	173	0.00	154520.00	49671.04	51099.69	1.02
A1970	173	0.00	456000.00	301769.97	179366.68	-0.83
B1940	173	0.00	429487.00	141500.54	156568.06	1.05
B1970	173	0.00	618982.00	284253.27	210853.30	0.26

Table 3.3: Descriptive statistics of water quality variable

From table 3.3 variables have zero values as their lower bound and there was a large variation within the observation on variables as well. The water quality parameters such as color, calcium, conductivity, sodium, temperature seems to have high standard deviation. These variables changes considerably in the drinking water. Similarly, most of the water distribution network variables value ranges from zero to some thousand meter. Most of the waterworks has installed all types of generic pipes but few of them consider either plastic or iron pipes only. Here zero represents the uninstalled pipe by the waterworks. Logarithm transformation was taken on the variables `HPC`, `Irp, PVC, PEL, GUP, A2001, UPDi, VoTa, PlDi,` `PiRe, Cem, B1910, A1970, B1940, B1970` adding 1 in each of the observation due to large number of zero values.

Correlation analysis was used to test the relation between physical, chemical and distribution network variable. Pearson correlation coefficient r matrix was calculate and test result are presented in Appendix D.1. No significant correlation was observed. However, conductivity, color, pH, calcium, iron, iron pipe, PVC, GUP and volume of storage tank shows moderate correlation $\pm(0.3 - 0.5)$ with HPC whereas age of pipes and other water quality parameter has a weak correlation with it. Basically, the observed lower correlation value only have a little practical importance.

3.4 Multiple linear Regression

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Regression	37	168.83	4.56	13.87	0.0000
Residuals	135	44.41	0.33		

Table 3.4: ANOVA for Regression

The test statistic F on table 3.4 is 13.87 and found to be significant with $P-value < 0.000$. The results in tableC shows that some regression coefficients are significant ($P - value < 0.05$) while some are highly insignificant even though R^2 was found to be 0.79.

In linear regression adding more variable in the right hand side gives the better R^2 value but can lead to over-fitted model. The over-fitted model describes only random error instead of the underlying relationship. The model also becomes unable to perform in the future prediction. However the $Adj-R^2$ of 0.73 somewhat provide proof of good fitted model however, most of the predictor variable are still a statistically insignificant.

Large number of insignificant variable may be a result of the collinearity among the predictor variable. The Variance Inflation Factor (VIF) in figure 3.1 was calculated for each of the explanatory variable where the values above 10 suggest the problem of collinearity among predictor variables.

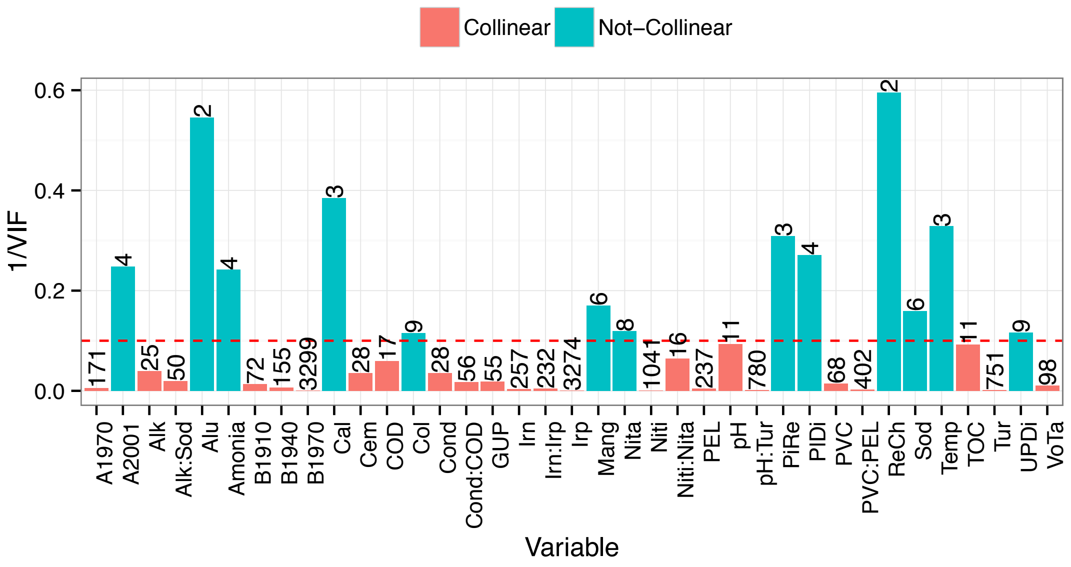


Fig 3.1: Variance Inflation Factor (VIF) for lienar model. The numbers above the bars represents the VIF value for respective variables.

From the chart in figure 3.1 among 37 predictor variable only 13 variable are non collinear while rest are highly correlated. This problem of multicolinear variable may lead to imprecise prediction and can be often addressed through variable selection procedure and the dimension reduction technique.

3.4.1 Variable selection and Subset Regression procedure

A subset selection method were applied and the variable were selected according to BIC, $Adj - R^2$ and rss criteria. In subsets procedure backward elimination, forward selection and exhaustive (forward and backward) methods were used. Appendix C.3 present the selected model under different criteria. Each model used different predictor variables to explain the variation in HPC. The fitted 12 different models were much more sophisticated than one obtained on full multiple regression model.

Subset selection regression results was shown in appendix C.2 . The variation on HPC was explained by 28 predictors variables including water quality and distribution network along with interaction between the variables. R^2 is 0.78 means 78% variation of HPC was explained by the model and rest of the variation is noise. Also all the assumption of regression model was checked. residual plot follows the normal distribution as all the residuals fall roughly in a straight line. Model selection criteria and number of selected variable are also present in figure 3.2.

A model were fitted using different criteria and RMSEP and predicted R^2 present in table 3.5. From the table, subset model selected using exhaustive method with maximum adjusted R^2 has least RMSEP and maximum R^2 predicted among all the subset models. Although the models can be selected as better model than other and it results large number of significant variables (Appendix C.2), it still suffer from multicollinearity problem (Figure 3.3).

Table 3.5: RMSEP and R2 predicted for subset linear models

Method	Criteria	RMSEP	R2prd

backward	adjr2	0.66	0.58
forward	adjr2	0.68	0.55
exhaustive	adjr2	0.67	0.56
backward	bic	0.66	0.54
forward	bic	0.75	0.34
exhaustive	bic	0.67	0.49
backward	rss	0.76	0.51
forward	rss	0.76	0.51
exhaustive	rss	0.76	0.51
backward	cp	0.65	0.58
forward	cp	0.67	0.56
exhaustive	cp	0.65	0.58

Under the criteria best subset fitted model was found as the model with lowest $RMSEP$ 0.65 and highest R^2prd 0.58 and the model can be written as

$$\begin{aligned}\texttt{HPC} = 12.67 - 48.23 \times \texttt{Irn} + 2.5 \times \texttt{ReCh} + 0.03 \times \texttt{Col} - 0.01 \times \texttt{Cal} - 0.15 \times \texttt{Cond}\\ - 0.54 \times \texttt{COD} - 0.9 \times \texttt{Alk} - 0.08 \times \texttt{Sod} + 8.28 \times \texttt{Mang} - 0.06 \times \texttt{Temp}\\ - 0.25 \times \texttt{pH} - 4.14 \times \texttt{Tur} + 2.59 \times \texttt{Nita} - 0.15 \times \texttt{Irp} - 0.18 \times \texttt{PVC}\\ + 0.12 \times \texttt{PEL} + 0.2 \times \texttt{GUP} - 0.06 \times \texttt{A2001} + 0.08 \times \texttt{UPDi} - 0.46 \times \texttt{VoTa}\\ + 0.04 \times \texttt{PlDi} - 0.09 \times \texttt{Cem} - 1.64 \times \texttt{Niti} + 3.96 \times \texttt{Irn:Irp} + 0.03 \times \texttt{Cond:COD}\\ + 0.12 \times \texttt{Alk:Sod} + 0.59 \times \texttt{pH:Tur} - 25.97 \times \texttt{Nita:Niti} \quad (3.1)\end{aligned}$$

3.5 Principal Component Analysis

Principal component analysis has been carried out to find the hidden relation between water quality parameters. Since the variables were in different scale a

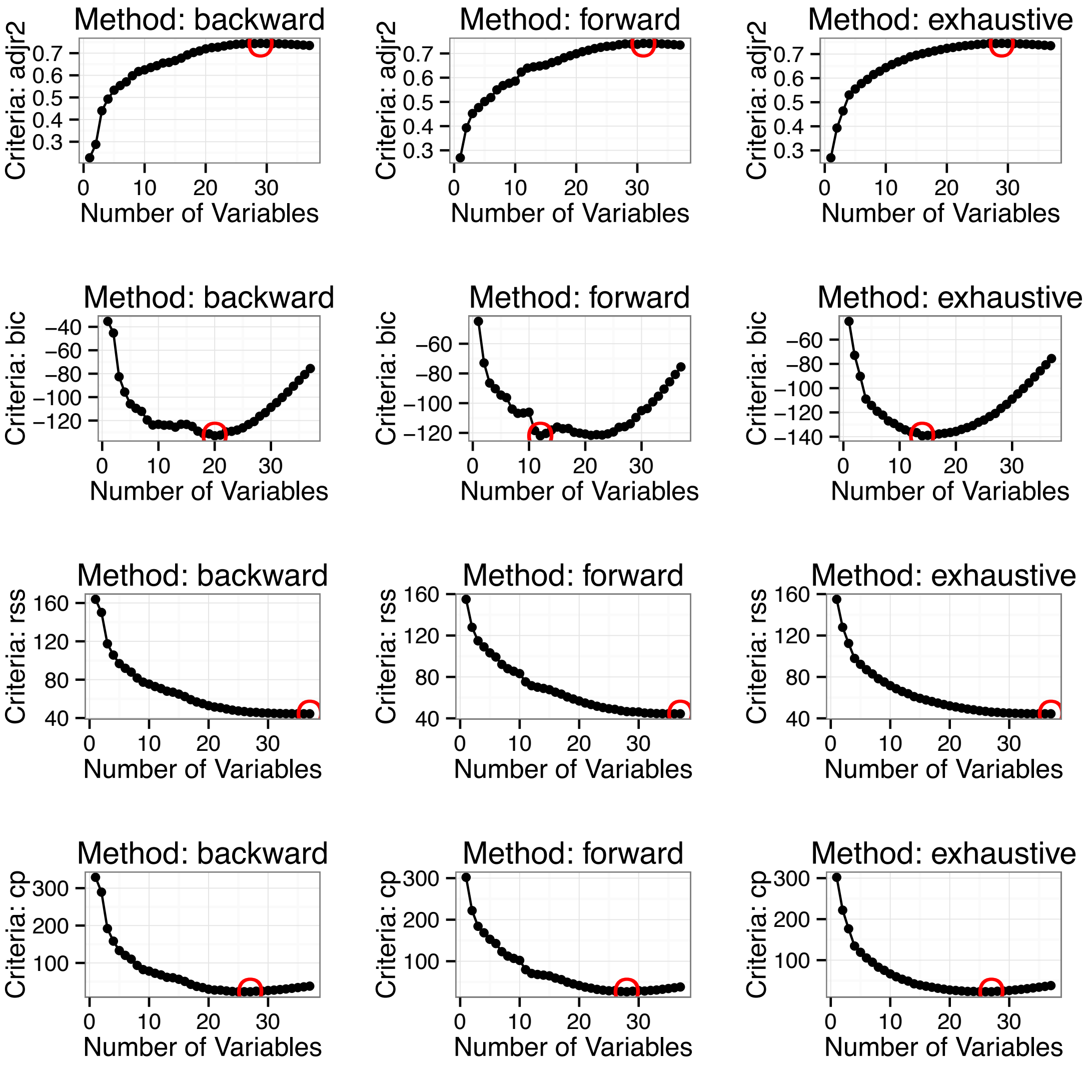


Fig 3.2: Variable Selection with different Creiteria

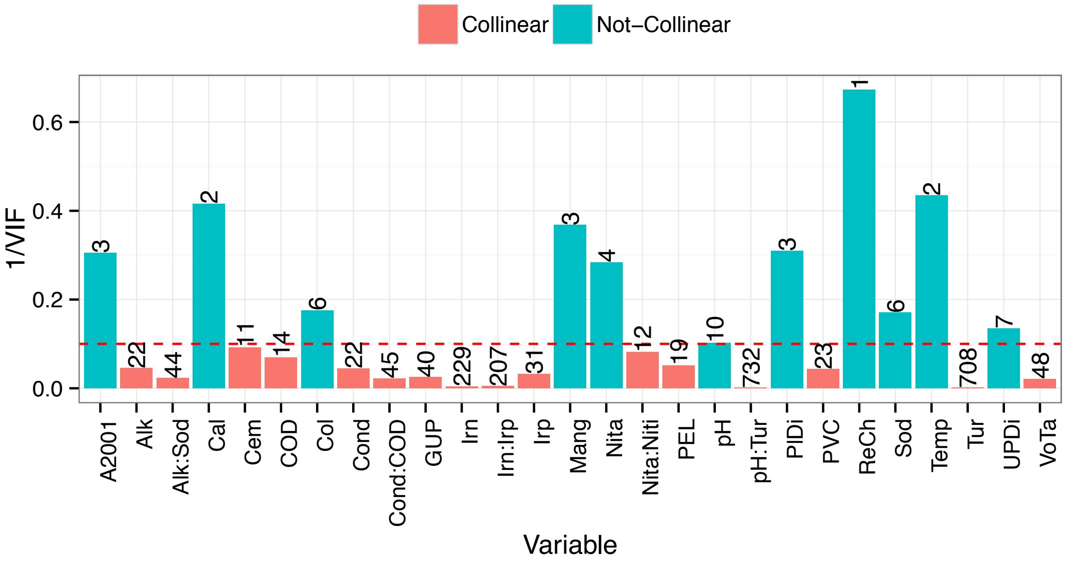
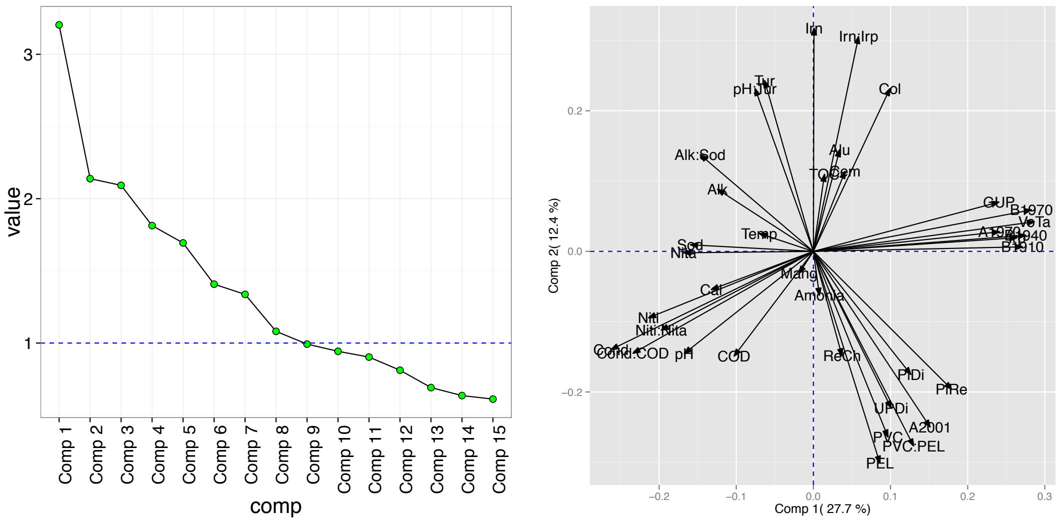


Fig 3.3: VIF for chosen submodel selected from backward methods with minimum Mallow's Cp

correlation matrix was used as suggested by Karpuzcu, Senes, and Akkoyunlu (1987). The result shows that 8 principal components explain 82% of the total variation. The number of components were chosen on the basis of a criteria given by (Kaiser, 1960), i.e eigenvalues greater or equal to 1. In other words, these 8 components explain more variance than the variable itself. Moreover, figure3.4a supports the fact since the curve at 9 components contain an elbow explaing 85% of total variation.

Further, the loading plot in fig-3.4b visualize the underlying similar characteristics within variables. The group of variable in lower right corner are related to water transportation system and their maintenance. All these variables have negative effect on second principal component and positive effect on first princi-



(a) Scree plot of PCA Model

(b) Loading plot of PCA Model

Fig 3.4: Principal Component Plot

pal components. Similarly, old pipes and storage tanks related have high positive effect on first principal component and are grouped on right edge of the plot.

From the loading table in C.4, a relationship between variable on first principal component according to their weights can be written in a functional form as,

$$\begin{aligned} Z_1 = 0.00 \times \texttt{Irn} + 0.04 \times \texttt{ReCh} + 0.01 \times \texttt{TOC} + 0.10 \times \texttt{Col} - 0.13 \times \texttt{Cal} - 0.26 \times \texttt{Cond} \\ - 0.10 \times \texttt{COD} - 0.12 \times \texttt{Alk} - 0.16 \times \texttt{Sod} - 0.02 \times \texttt{Mang} - 0.07 \times \texttt{Temp} - 0.17 \times \texttt{Ph} \\ - 0.06 \times \texttt{Tur} + 0.03 \times \texttt{Alu} - 0.21 \times \texttt{Niti} - 0.17 \times \texttt{Nita} + 0.01 \times \texttt{Amonia} + 0.26 \times \texttt{Irp} \\ + 0.10 \times \texttt{PVC} + 0.09 \times \texttt{PEL} + 0.24 \times \texttt{GUP} + 0.15 \times \texttt{A2001} + 0.10 \times \texttt{UPDi} + 0.29 \times \texttt{VoTa} \\ + 0.13 \times \texttt{PlDi} + 0.18 \times \texttt{PiRe} + 0.04 \times \texttt{Cem} + 0.27 \times \texttt{B1910} + 0.24 \times \texttt{A1970} + 0.28 \times \texttt{B1940} \\ + 0.28 \times \texttt{B1970} + 0.06 \times \texttt{Irn:Irp} - 0.23 \times \texttt{Cond:COD} \\ + 0.13 \times \texttt{PVC:PEL} - 0.15 \times \texttt{Alk:Sod} - 0.08 \times \texttt{pH:Tur} \\ - 0.20 \times \texttt{Nita:Niti} \end{aligned}$$

This linear combination of the variables captures almost 27% of the variance present in the data-set.

However, (Liu, Lin, and Kuo (2003)) classified the loading values as greater than 0.75 shows strong relation to the component between 0.5 to 0.75 as moderate whereas the value below 0.5 denote the week relation. According to this criteria all the variables used in the analysis have the weak relationships to the extracted principal component. Whatever the criteria our intention is to avoid collinearity problem.

3.6 Principal Component Regression

Principal component regression model was fitted based on the selected principal component from PCA as explained in section-3.5. From the result of PCR in table-3.6, eight principal components which have explained more than 80% of the total variation on predictor have only explained %52 variation in response. If all the components are considered, same amount of variation can be captured as in Multiple Linear Regression, however more noise get modeled during the process.

Table 3.6: Percent Variance Captured by Regression Model Using PCR on Reponse and Predictors

comp	Xvar	HPC	comp	Xvar	HPC
Comp 1	27.74	6.92	Comp 20	97.99	57.89
Comp 2	40.10	29.61	Comp 21	98.43	57.94
Comp 3	51.93	33.25	Comp 22	98.80	58.73
Comp 4	60.82	35.25	Comp 23	99.09	58.74
Comp 5	68.57	41.58	Comp 24	99.33	59.16
Comp 6	73.93	42.74	Comp 25	99.49	61.83
Comp 7	78.76	52.38	Comp 26	99.63	65.69
Comp 8	81.92	52.46	Comp 27	99.74	67.50

Comp 9	84.59	54.75	Comp 28	99.84	67.52
Comp 10	86.99	55.31	Comp 29	99.90	67.69
Comp 11	89.19	55.52	Comp 30	99.93	68.40
Comp 12	90.98	55.71	Comp 31	99.96	74.39
Comp 13	92.27	55.94	Comp 32	99.98	75.58
Comp 14	93.36	56.06	Comp 33	99.99	75.73
Comp 15	94.37	56.06	Comp 34	99.99	77.62
Comp 16	95.34	56.53	Comp 35	100.00	78.28
Comp 17	96.19	57.05	Comp 36	100.00	79.17
Comp 18	96.88	57.35	Comp 37	100.00	79.17
Comp 19	97.51	57.62			

A fitted linear relation between response and predictor variable using eight principal components can be written in functional form as,

$$\begin{aligned}\texttt{HPC} = 0.12 &+ 0.02 \times \texttt{ReCh} - 0.02 \times \texttt{TOC} + 0.15 \times \texttt{Col} - 0.06 \times \texttt{Cal} - 0.08 \times \texttt{Cond}\\ &+ 0.08 \times \texttt{COD} - 0.06 \times \texttt{Alk} + 0.03 \times \texttt{Sod} + 0.07 \times \texttt{Mang} + 0.07 \times \texttt{Temp}\\ &- 0.14 \times \texttt{pH} + 0.04 \times \texttt{Tur} + 0.07 \times \texttt{Alu} + 0.06 \times \texttt{Niti}\\ &- 0.03 \times \texttt{Nita} + 0.08 \times \texttt{Amonia} + 0.05 \times \texttt{Irp} - 0.12 \times \texttt{PVC} - 0.08 \times \texttt{PEL}\\ &+ 0.1 \times \texttt{GUP} - 0.04 \times \texttt{A2001} - 0.01 \times \texttt{UPDi} + 0.05 \times \texttt{VoTa} + 0 \times \texttt{PlDi}\\ &- 0.04 \times \texttt{PiRe} + 0.06 \times \texttt{Cem} + 0.05 \times \texttt{B1910} + 0.06 \times \texttt{A1970} + 0.03 \times \texttt{B1940}\\ &+ 0 \times \texttt{B1970} + 0.14 \times \texttt{Irn:Irp} + 0.03 \times \texttt{Cond:COD} - 0.09 \times \texttt{PVC:PEL} - 0.03 \times \texttt{Alk:Sod}\\ &+ 0.02 \times \texttt{pH:Tur} + 0.06 \times \texttt{Niti:Nita}\end{aligned} \tag{3.2}$$

3.7 Partial Least Square Regression

Partial least square regression were performed in the data matrix. This is another method to deal with the collinearity problem. Unlike PCR, PLS extract the factor by considering both the effects of X and Y. Here principal factor are extracted in such a way that the co-variance between X score and Y score are maximized.

Table 3.7: Percent Variance Captured by Regression Model Using PLS

comp	Xvar	HPC	comp	Xvar	HPC
Comp 1	19.98	41.39	Comp 20	95.62	76.92
Comp 2	38.17	52.73	Comp 21	96.20	77.09
Comp 3	45.38	57.25	Comp 22	96.80	77.31
Comp 4	51.36	59.26	Comp 23	97.17	77.70
Comp 5	56.11	61.81	Comp 24	97.57	78.08
Comp 6	64.41	63.47	Comp 25	97.90	78.41
Comp 7	68.69	66.22	Comp 26	98.68	78.48
Comp 8	75.89	67.27	Comp 27	99.04	78.66
Comp 9	78.07	69.84	Comp 28	99.38	78.77
Comp 10	82.18	70.70	Comp 29	99.55	78.87
Comp 11	84.24	71.58	Comp 30	99.77	78.92
Comp 12	85.93	72.39	Comp 31	99.85	78.99
Comp 13	87.78	73.08	Comp 32	99.89	79.04
Comp 14	89.36	73.90	Comp 33	99.97	79.09
Comp 15	90.64	74.92	Comp 34	99.99	79.15
Comp 16	91.68	75.90	Comp 35	100.00	79.17
Comp 17	93.05	76.35	Comp 36	100.00	79.17
Comp 18	94.33	76.58	Comp 37	100.00	79.17
Comp 19	95.00	76.77			

With nine factors are extracted from pls model. Seventy percent of the response variation was already explained while 77% predictor variation was explained with nine latent factor. Table 3.7 presented the percentage variation explained by all factor. The variance in the table represents the cumulative variance for each of the component.

AppendixC.5 gives the factor loading for each of the measures. From the results the distribution network related variables such as type of pipe, age of pipe and storage tanks all have a positive impact on first latent factor. Second component is mainly composed of organic material and Nitrogen related compound. Factor there is the chlorine and chemical oxygen demand and total organic carbon. The variables with strong positive factor loadings are residual chlorine, iron, total organic carbon, nitrate, ammonia, age of pipe, type of pipe and turbidity in the extracted 9 latent components which are presented in appendix as well.

3.8 Cross Validation

Model validation is necessary to check the predictive ability of a model and to avoid over-fitting or under-fitting. In this thesis leave one out (LOO) cross validation techniques was used. With this criteria the model is fitted first by omitting one observation and then the fitted model is used to predict the omitted observation. The process is repeated until all the observations have been omitted once.

3.9 Comparison of OLS, PCR and PLS model

Before any model comparison, it is desirable to see how well the model fit the data. Figure-3.5 shows that the fitted values are very close to the original values, however linear models and its subset has better fit than other two. Since, this fit is only on the calibration (training) data set containing those observations which are already included during model fitting. Therefore, it is necessary to see the performance of the fitted model in the case of new observation a cross-validation is performed. Models are compared on the basis of the prediction error both on the training data set and during cross-validation.

The prediction performance of a model can be determined by root mean square error of prediction (RMSEP). However, its prediction behavior in the case of new observation can be measured by RMSECV. In least square regression it seems that by considering 27 predictor variables the percentage variation explaind in HPC is 78%. Almost similar variation (more than 75%) in HPC is explained by only 13 components of PLS model (3.7).

In the figure 3.6 it seems that best subset model have the least $RMSEP$ value during calibration and cross-validation but it has taken 26 predictor variables for the prediction purposes. However the model is suffered from multicollinearity problem so the coefficient estimates might have been distorted. To avoid this problem, PCR or PLS model is recommended as it is free from multicollinearity problem.

It is important to use optimal number of components in both PCR and PLS model for better prediction during calibration as well as cross-validation. From the validation plot for these model in figure-E.1, RMSECV starts increasing from 7 and 3 components of PCR and PLS model respectively. These components can be considered as an optimal components for these model to perform better in the case of new observations. For an in-depth comparison, the plot in figure-3.6 shows that PLS model with 3 components perform better than the PCR model with 7 components. Hence, the first one can be considered as a selected model for the data considered in this thesis. The fitted regression model can be expressed as a function form as,

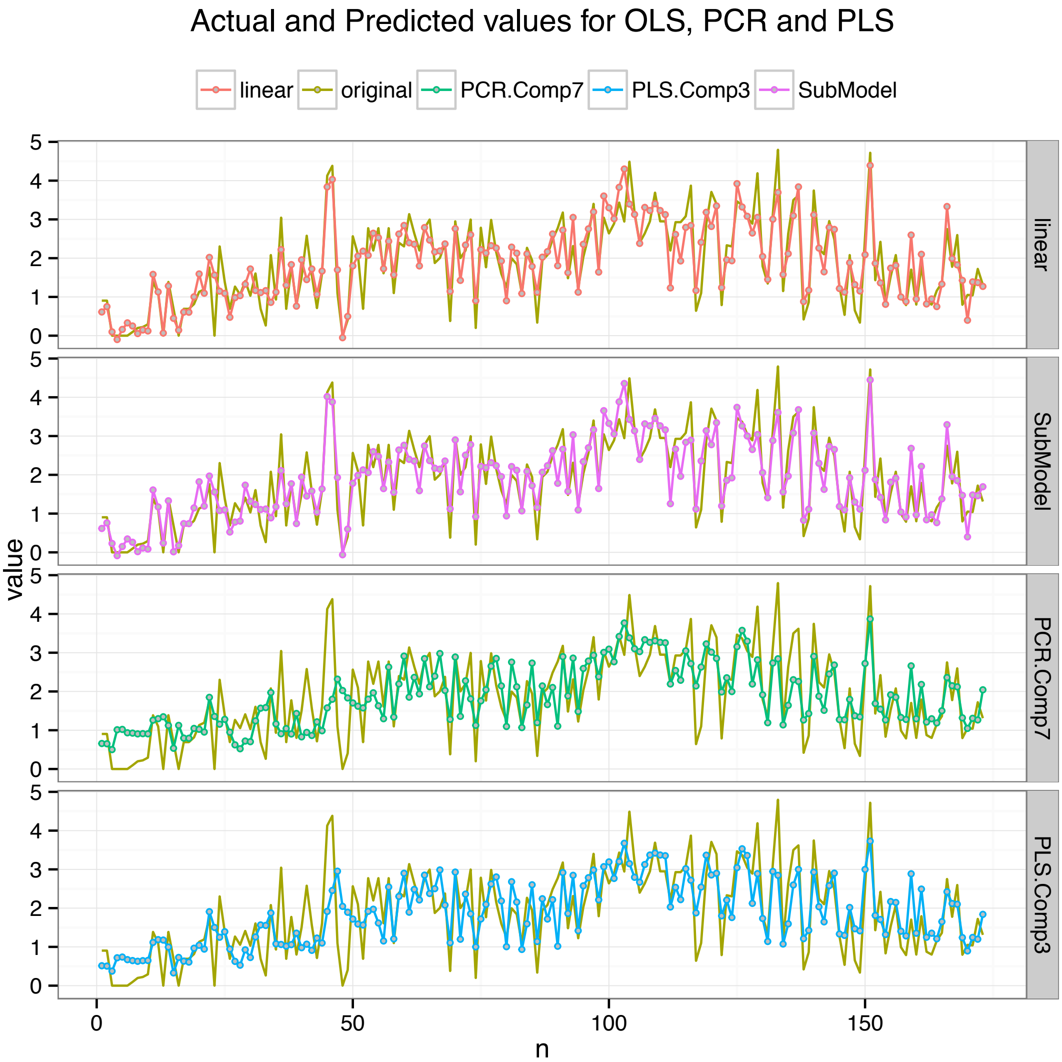


Fig 3.5: Actual and predicted values for OLS, PCR and PLS model

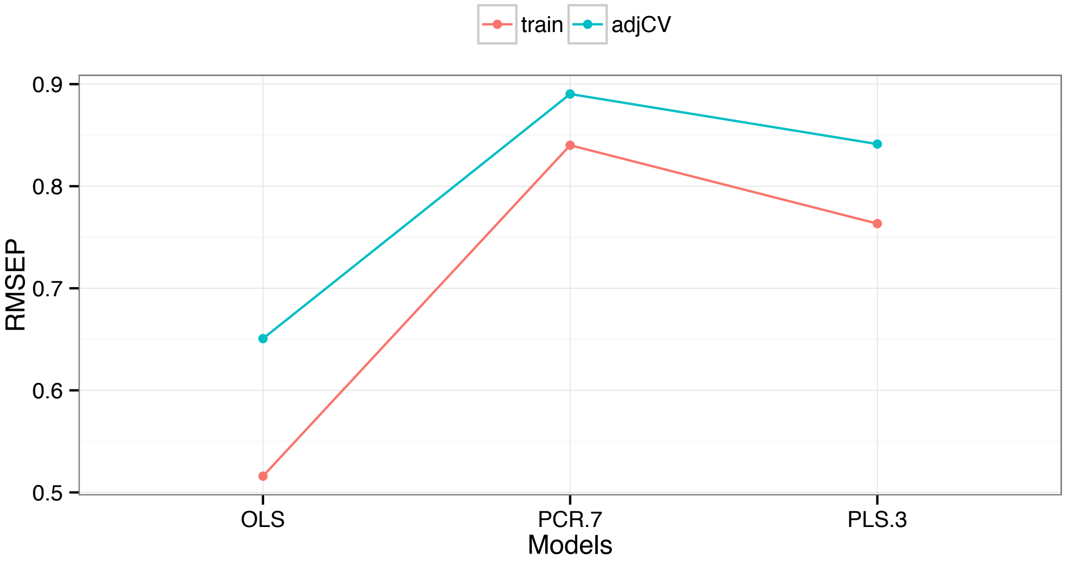


Fig 3.6: RMSEP plot for selected OLS, PCR and PLS models

$$\begin{aligned}
\texttt{HPC} = {} & 0.07 + 0.1 \times \texttt{ReCh} - 0.05 \times \texttt{TOC} + 0.18 \times \texttt{Col} - 0.09 \times \texttt{Cal} \\
& - 0.14 \times \texttt{Cond} + 0.09 \times \texttt{COD} - 0.03 \times \texttt{Alk} + 0.01 \times \texttt{Sod} + 0.09 \times \texttt{Mang} \\
& + 0.03 \times \texttt{Temp} - 0.17 \times \texttt{pH} + 0.04 \times \texttt{Tur} + 0.14 \times \texttt{Alu} + 0.11 \times \texttt{Niti} \\
& - 0.02 \times \texttt{Nita} + 0.06 \times \texttt{Amonia} + 0.06 \times \texttt{Irp} - 0.14 \times \texttt{PVC} - 0.04 \times \texttt{PEL} \\
& + 0.12 \times \texttt{GUP} - 0.07 \times \texttt{A2001} + 0.01 \times \texttt{UPDi} + 0.02 \times \texttt{VoTa} + 0 \times \texttt{PlDi} \\
& - 0.05 \times \texttt{PiRe} + 0.08 \times \texttt{Cem} + 0.02 \times \texttt{B1910} + 0.06 \times \texttt{A1970} + 0.01 \times \texttt{B1940} \\
& - 0.02 \times \texttt{B1970} + 0.12 \times \texttt{Irn:Irp} + 0.04 \times \texttt{Cond:COD} - 0.08 \times \texttt{PVC:PEL} - 0.02 \times \texttt{Alk:Sod} \\
& + 0.02 \times \texttt{pH:Tur} + 0.01 \times \texttt{Niti:Nita}
\end{aligned} \tag{3.3}$$

Table 3.8: RMSEP values for 15 components from PCR and PLS model

	PLS			PCR		
Comp	train	CV	adjCV	train	CV	adjCV
0	1.110	1.117	1.117	1.110	1.117	1.117
1	0.850	0.898	0.898	1.071	1.087	1.087
2	0.763	0.842	0.841	0.932	0.956	0.954
3	0.726	0.835	0.834	0.907	0.945	0.945
4	0.709	0.858	0.858	0.893	0.934	0.934
5	0.686	0.905	0.904	0.849	0.896	0.896
6	0.671	0.917	0.916	0.840	0.890	0.890
7	0.645	0.932	0.931	0.766	0.813	0.812
8	0.635	0.927	0.926	0.765	0.831	0.831
9	0.610	0.905	0.904	0.747	0.813	0.813
10	0.601	0.898	0.897	0.742	0.815	0.815
11	0.592	0.893	0.892	0.740	0.830	0.830
12	0.583	0.907	0.906	0.739	0.840	0.840
13	0.576	0.895	0.894	0.737	0.845	0.845
14	0.567	0.891	0.890	0.736	0.855	0.854

Chapter 4

Conclusion

1. Although all the values are in accceptable level specified by water authorities, HPC has moderate correlation with the other quality parameters such as color, pH, conductivity, calcium and iron. Further, the effect of interaction between some variables give idea of non-linear relation of HPC with other variables as well.

2. The multicollineartiy among the predictor variables was addressed completely by the use of latent variable methods suchs as principal component regression (PCR) and partial least square regression (PLSR).

   PLS and PCR results shows that older pipe, and the component of the water distribution network system are significant factors for the changes in HPC values. Older pipe no matter which types, is one of the significant factors for deterioration of water quality.

3. Since few variables are found significant in linear model with full set of variables, a subset model is selected using criteria such as Mallow's Cp,

Maximum adjusted R^2 and minimum BIC. Model selected from exhaustive method with minimum Mallow's Cp is considered better among others.

4. On the basis of predictability and model fit, OLS, PCR and PLS methods are compared. Selected subset model have least RMSEP during both calibration and cross-validation, however being unable to handle multicollinearity, PCR and PLS models are opted.

5. Models are validated using Leave one out cross-validation method from with PLS model with three components is found better among other models and have closer prediction.

6. The study is based on the availability of data sets and mainly affected by missing value problem. Among 6000 observation only 173 observation with average values of the water quality parameters were used. Prior to making any practical decisions, the analysis with extended data with more observations is suggested.

Bibliography

[(WH+02] World Health Organization (WHO) et al. "Heterotrophic plate count measurement in drinking water safety management". In: *Report of a meeting of specialists in Geneva, Switzerland, April.* 2002, pp. 24–25.

[AER04] Martin J Allen, Stephen C Edberg, and Donald J Reasoner. "Heterotrophic plate count bacteriawhat is their significance in drinking water?" In: *International journal of food microbiology* 92.3 (2004), pp. 265–274.

[Aug12] Baptiste Auguie. *gridExtra: functions in Grid graphics. R package version 0.9. 1.* 2012.

[Bis+95] Christopher M Bishop et al. "Neural networks for pattern recognition". In: (1995).

[CH13] Samprit Chatterjee and Ali S Hadi. *Regression analysis by example.* John Wiley &amp; Sons, 2013.

[Col+84] JS Colbourne et al. "Water fittings as sources of Legionella pneumophila in a hospital plumbing system". In: *The Lancet* 323.8370 (1984), pp. 210–213.

[Dah09] David B Dahl. *xtable: Export tables to LaTeX or HTML. R package version 1.5-6.* 2009.

[Fox+09] John Fox et al. "CAR: Companion to applied regression, R Package version 1.2-16". In: *Online at http://cran. r-project. org/web/packages/car/index. html (accessed on August 2012)* (2009).

[Hla13] Marek Hlavac. *stargazer: LaTeX code and ASCII text for well-formatted regression and summary statistics tables. R package version 4.5. 3.* 2013.

[Jol05] Ian Jolliffe. *Principal component analysis.* Wiley Online Library, 2005.

[JW+92] Richard Arnold Johnson, Dean W Wichern, et al. *Applied multivariate statistical analysis.* Vol. 4. Prentice hall Englewood Cliffs, NJ, 1992.

[Kai60] Henry F Kaiser. “The application of electronic computers to factor analysis.” In: *Educational and psychological measurement* (1960).

[Kaw00] Susumu Kawamura. *Integrated design and operation of water treatment facilities*. John Wiley & Sons, 2000.

[KN12] Lukasz Komsta and Frederick Novomestky. “Moments: moments, cumulants, skewness, kurtosis and related tests”. In: *R package version 0.13* (2012).

[Koo03] D Van der Kooij. “Managing regrowth in drinking water distribution systems”. In: *Heterotrophic plate counts and drinking-water safety. IWA Publishing, London, United Kingdom* (2003), pp. 199–232.

[KSA87] M Karpuzcu, S Senes, and A Akkoyunlu. “Design of monitoring systems for water quality by principal component analysis and a case study”. In: *Proceedings, Int. Symp. on Environmental Management: Environment87*. 1987, pp. 673–690.

[LLK03] Chen-Wuing Liu, Kao-Hung Lin, and Yi-Ming Kuo. “Application of factor analysis in the assessment of groundwater quality in a blackfoot disease area in Taiwan”. In: *Science of the Total Environment* 313.1 (2003), pp. 77–89.

[LM09] T Lumley and A Miller. “Leaps: regression subset selection. R package version 2.9”. In: *See http://CRAN. R-project. org/package= leaps* (2009).

[LM11] Bjørg Einan Liliane Myrstad Carl Fredrik Nordheim. *Report from the Drinking Water Register. Drinking Water Status (2003 and 2004)*. Vannrapport 116. Folkehelseinstituttet, 2011.

[LSL14] Kristian Hovde Liland, Solve Sæbø, and Maintainer Kristian Hovde Liland. “Package ‘mixlm’”. In: (2014).

[LWS96] Mark W LeChevallier, Nancy J Welch, and Darrell B Smith. “Full-scale studies of factors related to coliform regrowth in drinking water.” In: *Applied and Environmental Microbiology* 62.7 (1996), pp. 2201–2211.

[Mal73] Colin L Mallows. “Some comments on C p”. In: *Technometrics* 15.4 (1973), pp. 661–675.

[Mar92] Harald Martens. *Multivariate calibration*. John Wiley & Sons, 1992.

[MKB79] Kantilal Varichand Mardia, John T Kent, and John M Bibby. *Multivariate analysis*. Academic press, 1979.

[MM04] Maggy NB Momba and N Makala. "Comparing the effect of various pipe materials on biofilm formation in chlorinated and combined chlorine-chloraminated water systems". In: *Water SA* 30.2 (2004), pp. 175–182.

[Mom+00] MNB Momba et al. "An overview of biofilm formation in distribution systems and its impact on the deterioration of water quality". In: (2000).

[MW07] Björn-Helge Mevik and Ron Wehrens. "The pls package: principal component and partial least squares regression in R". In: *Journal of Statistical Software* 18.2 (2007), pp. 1–24.

[NL00] Cheryl D Norton and Mark W LeChevallier. "A pilot study of bacteriological population changes through potable water treatment and distribution". In: *Applied and Environmental Microbiology* 66.1 (2000), pp. 268–276.

[Ped90] Karsten Pedersen. "Biofilm development on stainless steel and PVC surfaces in drinking water". In: *Water Research* 24.2 (1990), pp. 239–243.

[Pin+09] Jose Pinheiro et al. "the R Core team (2009) nlme: Linear and Nonlinear Mixed Effects Models. R package version 3.1-96". In: *R Foundation for Statistical Computing, Vienna* (2009).

[PSR03] P Payment, DP Sartory, and DJ Reasoner. "The history and use of HPC in drinking-water quality management". In: *Heterotrophic plate counts and drinking-water safety* (2003), pp. 20–48.

[Rev14] William Revelle. "psych: Procedures for personality and psychological research". In: *Northwestern University, Evanston. R package version* 1.1 (2014).

[Rim14] Raju Rimal. "Evaluation of Models for predicting the average monthly Euro versus Norwegian krone exchange rate from financial and commodity information". Thesis. Aas, Akershus, Norway: Norwegian University of Life Sciences, 2014.

[Rip11] Brian Ripley. "MASS: support functions and datasets for Venables and Ripley's MASS". In: *R package version* (2011), pp. 7–3.

[RO81] HF Ridgway and BH Olson. "Scanning electron microscope evidence for bacterial colonization of a drinking-water distribution system." In: *Applied and Environmental Microbiology* 41.1 (1981), pp. 274–287.

[VDKZ78] D VAN DER KOOIJ and BCJ Zoeteman. "Water quality in distribution systems". In: *Proc. IWSA 12th Congress, Kyoto.* 1978.

[VF09] Kurt Varmuza and Peter Filzmoser. *Introduction to multivariate statistical analysis in chemometrics*. CRC press, 2009.

[War+12] GR Warnes et al. *gdata: Various R programming tools for data manipulation (2010). R package version 2.8. 1*. 2012.

[WC09] Hadley Wickham and Winston Chang. *ggplot2: An implementation of the Grammar of Graphics. R package version 0.8. 3*. 2009.

[Wei13] Taiyun Wei. *corrplot: visualization of a correlation matrix. R package version 0.60*. 2013.

[WF14] Hadley Wickham and R Francois. "dplyr: A Grammar of Data Manipulation". In: *URL http://CRAN. R-project. org/package= dplyr. R package version 0.2* (2014).

[Wic09] Hadley Wickham. "plyr: Tools for splitting, applying and combining data". In: *R package version 0.1* 9 (2009), p. 651.

[Wic12] Hadley Wickham. "reshape2: Flexibly reshape data: a reboot of the reshape package". In: *R package version* 1.2 (2012).

[Xie13] Yihui Xie. "knitr: A general-purpose package for dynamic report generation in R". In: *R package version* 1.7 (2013).

Appendix A

R packages used

Name	Version	Title
`MASS`(Ripley, 2011)	7.3-35	Support Functions and Datasets for Venables and Ripley's MASS
`car`(Fox et al., 2009)	2.0-22	Companion to Applied Regression
`pls`(Mevik and Wehrens, 2007)	2.4-3	Partial Least Squares and Principal Component regression
`xtable`(Dahl, 2009)	1.7-4	Export tables to LaTeX or HTML
`grid`(Auguie, 2012)	3.1.2	The Grid Graphics Package
`gridExtra`(Auguie, 2012)	0.9.1	functions in Grid graphics
`knitr`(Xie, 2013)	1.8	A General-Purpose Package for Dynamic Report Generation in R
`leaps`(Lumley and Miller, 2009)	2.9	regression subset selection
`gdata`(Warnes et al., 2012)	2.13.3	Various R programming tools for data manipulation
`plyr`(Wickham, 2009)	1.8.1	Tools for splitting, applying and combining data
`dplyr`(Wickham and Francois, 2014)	0.3.0.2	A Grammar of Data Manipulation

Name	Version	Title
`ggplot2`(Wickham and Chang, 2009)	1.0.0	An implementation of the Grammar of Graphics
`reshape2`(Wickham, 2012)	1.4	Flexibly reshape data: a reboot of the reshape package.
`mixlm`(Liland, Sæbø, and Liland, 2014)	1.0.7	Mixed Model ANOVA and Statistics for Education
`stargazer`(Hlavac, 2013)	5.1	LaTeX/HTML code and ASCII text for well-formatted regression and summary statistics tables
`moments`(Komsta and Novomestky, 2012)	0.14	Moments, cumulants, skewness, kurtosis and related tests
`psych`(Revelle, 2014)	1.4.8.11	Procedures for Psychological, Psychometric, and Personality Research
`corrplot`(Wei, 2013)	0.73	Visualization of a correlation matrix
`graphics`(Pinheiro et al., 2009)	3.1.2	The R Graphics Package
`grDevices`(Pinheiro et al., 2009)	3.1.2	The R Graphics Devices and Support for Colours and Fonts
`utils`(Pinheiro et al., 2009)	3.1.2	The R Utils Package
`datasets`(Pinheiro et al., 2009)	3.1.2	The R Datasets Package
`methods`(Pinheiro et al., 2009)	3.1.2	Formal Methods and Classes
`base`(Pinheiro et al., 2009)	3.1.2	The R Base Package

Appendix B

R Codes and Functions

```

## ----frontMatter, child="frontMatter.Rnw"--------------------------------


## ----LoadingPkgs, echo=FALSE, message=FALSE, warning=FALSE, results='hide
    '----
req.package<-c("MASS", "car", "pls", "xtable", "grid", "gridExtra", "knitr", "
    leaps",  "gdata", "plyr", "dplyr", "ggplot2", "reshape2", "mixlm", "
    stargazer", "moments", "psych", "corrplot")
lapply(req.package, require, character.only=TRUE, quietly = T, warn.conflicts =
     F)




## ----setup, include=FALSE, cache=FALSE, echo=TRUE------------------------
opts_chunk$set(fig.path='Includes', fig.align='center')
render_listings()
setwd('~/Dropbox/Thesis/FinalThesis/')
Sys.setenv(TEXINPUTS=getwd(),
           BIBINPUTS=getwd(),
           BSTINPUTS=getwd())
data.path<-path.expand(file.path(dirname(getwd()), "FinalThesis","Datasets", "
    mrgData.xlsx"))
codebook.path<-path.expand(file.path(dirname(data.path), "CodeBook.xlsx"))
subMdls.path<-path.expand(file.path(dirname(data.path), "SubMdls.RData"))
abv.path<-path.expand(file.path(dirname(data.path), "abbri.xlsx"))


## ----readFun, child="Includes/FunctionsAndDataPrep.Rnw"------------------


## ----dataPrep, echo=FALSE------------------------------------------------
```

```r
## Loading Dataset
mrgData<-read.xls(data.path, sheet = 2)
codeBook<-read.xls(codebook.path, sheet=1)
names(codeBook)<-c("Variable Code","Unit of Measurements","Variable Name")

## Data Preperation -------------
log.var<-c("Irp", "PVC", "PEL", "GUP", "A2001", "UPDi", "VoTa", "PlDi", "PiRe",
     "Cem", "B1910", "A1970", "B1940", "B1970")
x.var<-c('Irn','ReCh','TOC','Col','Cal','Cond','COD','Alk','Sod','Mang','Temp',
    'pH','Tur','Alu','Niti','Nita','Amonia','Irp','PVC','PEL','Cem','GUP','UPDi
    ','PlDi','VoTa','PiRe','B1910','B1940','B1970','A1970','A2001')
int.cpl<-list(c("Irn", "Irp"), c("Cond", "COD"), c("PVC","PEL"),c("Alk","Sod"),
    c("pH","Tur"),c("Niti", "Nita"))
int.var <- sapply(seq_along(int.cpl), function(x){
    paste(int.cpl[[x]], collapse=":")
})
y.var<-"HPC"

tmp.x<-mrgData[,x.var[!x.var %in% log.var]]
tmp.lx<-apply(mrgData[, log.var], 2, log1p)
tmp.y<-mrgData[,y.var]
tmp.int<-data.frame(t(ldply(seq_along(int.cpl), function(x){
  apply(mrgData[, int.cpl[[x]]], 1, prod)
})), row.names = NULL)
names(tmp.int)<-c("irn.irp", "cond.cod","pvc.pel", "alk.sod", "ph.tur" ,"niti.
    nita")
o<-c(157,47,42,111,135,30)

mrgdata<-data.frame(HPC=log1p(tmp.y), tmp.x, tmp.lx)[-o,]
mrgdata1 <- data.frame(HPC=log1p(tmp.y), tmp.x, tmp.lx, tmp.int)[-o,]
x.var<-c(names(mrgdata)[-1], int.var)


## ----functions, echo=FALSE-----------------------------------------------
## Submodel Fitting Function ----------------------
makeFormula<-function(x.var, y.var){
  formula<-paste(y.var, paste(x.var, collapse="+"), sep="~")
  return(formula)
}
subFit<-function(x.var, y.var, dataset, nbest=1, nvmax=NULL, method="backward",
     criteria='cp'){
  ## Lodaing Packages -------------------------------------------------
  require("leaps")
  require("ggplot2")
  require("plyr")
  ## ------------------------------------------------------------------

  subModel <- regsubsets(as.formula(makeFormula(x.var, y.var)),
                         data=dataset, nbest = nbest, nvmax = nvmax,
                        method = method)
  mdl.which<-summary(subModel)$which
  mdl.criteria<-summary(subModel)[[criteria]]
  nvar<- as.numeric(rownames(mdl.which))
  criteria.df<-data.frame(n=nvar, criteria=mdl.criteria)
  cm<-match.fun(ifelse(criteria %in% c("rsq", "adjr2"), "max", "min"))
  which.cm<-which(mdl.criteria==cm(mdl.criteria))
  which.crt.df<-data.frame(n=which(mdl.criteria==cm(mdl.criteria)), criteria=cm(mdl.criteria))
  ## Variable vs Criteria Plot -----------------------------------------
  plt<-ggplot(criteria.df, aes(n, criteria))+geom_line()+geom_point()
  plt<-plt+geom_point(data=which.crt.df, aes(n, criteria), shape="O", color="red", size=6)
  plt<-plt+theme_bw()
  plt<-plt+labs(x="Number of Variables", y=paste("Criteria:", criteria))
  plt<-plt+ggtitle(paste("Method:", method))
  ## ------------------------------------------------------------------

  ## Fitting Models ---------------------------------------------------
  which.var<-names(which(mdl.which[which.cm, ]))[-1]
  formula<-paste(y.var, paste(which.var, collapse="+"), sep="~")
  mdl.ft<-lm(makeFormula(which.var, y.var), data=dataset)
  ## ------------------------------------------------------------------

  return(list(plt, mdl.ft))
}

mdl.cv<-function(model, split=1){
  dataSet<-model$model
  formula<-model$terms
  x.var<-colnames(attr(formula, 'factors'))
  y.var<-rownames(attr(formula, 'factors'))[1]

  segment<-split(1:nrow(dataSet), ceiling(1:nrow(dataSet)/split))
  mdl<-list()
  predVec<-rep(NA, nrow(dataSet))
  errVec<-rep(NA, nrow(dataSet))

  for(i in seq_along(segment)){
    dataset<-dataSet[-segment[[i]],]
    testset<-dataSet[segment[[i]],]
    mdl[[i]]<-lm(formula, dataset)
    predVec[segment[[i]]]<-predict(mdl[[i]], newdata=testset[,-1])
    errVec[segment[[i]]]<-testset[,y.var]-predVec[segment[[i]]]
  }
  rmse.cv<-sqrt(1/nrow(dataSet)*sum(errVec^2))
  r2pred<-1-sum(errVec^2)/sum((predVec-mean(dataSet[,y.var]))^2)
  invisible(list(Model=mdl, Predicted=predVec, Error=errVec, rmsep=rmse.cv, r2pred=r2pred))
```

```r
}
mdlFit.mthd<-c("backward", "forward", "exhaustive")
mdlFit.crt<-c("adjr2", "bic", "rss", "cp")
mf.mc<- expand.grid(mdlFit.mthd, mdlFit.crt)


## ----mdlFit, echo=FALSE--------------------------------------------------
## Linear Model
lm.model<-lm(makeFormula(x.var, y.var),data=mrgdata)

## PCR Model
pcr.model<-pcr(as.formula(makeFormula(x.var, y.var)), data=mrgdata, scale=T,
    validation="LOO")

## PLS Model
pls.model<-plsr(as.formula(makeFormula(x.var, y.var)), data=mrgdata, scale=T,
    validation="LOO", method=’oscorespls’)


## ----subModels, echo=FALSE, eval=FALSE-----------------------------------
## mdlFit<-list()
## SubMdls<-laply(mdlFit.mthd, function(x){
##   laply(mdlFit.crt, function(y){
##     subMdlFit<-subFit(x.var, y.var, mrgdata, method = x, criteria = y)
##     mdlFit$plots<-subMdlFit[[1]]
##     mdlFit$model<-subMdlFit[[2]]
##     mdlFit$cvRslt<-mdl.cv(mdlFit$model, split = 1)
##     return(mdlFit)
##   })
## })
## dimnames(SubMdls)<-list(c(mdlFit.mthd), c(mdlFit.crt), c(’Plots’,’Models’, ’
    CV’))


## ----subModelsLoad, echo=FALSE-------------------------------------------
load(subMdls.path)
bssMdl <- update(SubMdls[[’exhaustive’, ’cp’, ’Models’]], .~.+Niti, data=
    mrgdata)


## ----subModelExtracts, echo=FALSE----------------------------------------
plts<-lapply(mdlFit.mthd, function(x){
  lapply(mdlFit.crt, function(y){
    SubMdls[[x,y,’Plots’]]
  })
})

## Making Coefficients Table
subMdlsNames<- paste(mf.mc$Var1, mf.mc$Var2, sep=".")
```

```
selectedVars<-unlist(lapply(mdlFit.mthd, function(x){
  lapply(mdlFit.crt, function(y){
    names(SubMdls[[x,y,’Models’]]$coef[-1])
  })
}), recursive = FALSE)

## Cross validation table
valdVal<-mdply(mf.mc, function(Var1, Var2){
      cbind(RMSEP=SubMdls[[Var1, Var2, ’CV’]]$rmsep,
                         R2pred=SubMdls[[Var1, Var2, ’CV’]]$r2pred)
   })

colnames(valdVal)<-c("Method","Criteria","RMSEP","R2prd")

## Make Equation Function
makeEqn <- function(mdl, comp = NULL){
  if(class(mdl) == ’lm’){
    coefVec <- mdl$coef
  } else if(class(mdl) == ’mvr’){
    coefVec <- mdl$coef[,,comp]
  } else{
    stop(’Please input correct Model!!’)
  }
  return(paste(’\texttt{HPC}’,
                 gsub(’\\+ -’, ’ - ’, paste(round(coefVec[1], 2), paste(paste(
    round(coefVec[-1], 2), paste(’\texttt{’,names(coefVec)[-1],’}’, sep=""),
    sep=" \times "), collapse= " + "), sep=" + ")), sep=" = "))
}




## ----AbvSymb-include, child="Includes/AbvSymb.Rnw", eval=TRUE------------


## ----abvUsedPrint, echo=FALSE, results=’asis’----------------------------
abvList <- read.xls(abv.path, sheet = 1, header=FALSE)
abvXtbl<-xtable(abvList, align = ’llX’)

print.xtable(abvXtbl, include.rownames = FALSE,
      tabular.environment = "tabularx",
      width = "\\textwidth",
      floating=FALSE,
      booktabs = TRUE,
      add.to.row = list(pos = list(0),command = "\\hline \\endhead "),
      sanitize.text.function = function(x){x},
      caption.placement = "top",
      table.placement = ’htbp’,
      include.colnames = FALSE)
```

```


## ----include1, child="Includes/Include-1.Rnw", eval=TRUE-----------------





## ----include2, child="Includes/Include-2.Rnw", eval=TRUE-----------------





## ----Include3, child="Includes/Include-3.Rnw", eval=TRUE-----------------



## ----varTable, echo=FALSE, results='asis', eval=TRUE---------------------
codebook<-read.xls(codebook.path, sheet=1)
varTabl<-xtable(codebook[1:18,], caption = "Water quality parameter and their
    code", label = "tbl:codbook")
print(varTabl, include.rownames = FALSE, floating = FALSE, tabular.environment
    = "longtable")


## ----disTable, echo=FALSE, results='asis', eval=TRUE---------------------
codebook<-read.xls(codebook.path, sheet=1)
disTabl<-xtable(codebook[19:32,], caption="Distribution network variables and
    their code", label = "tbl:codbok")
print(disTabl, include.rownames=FALSE,floating = FALSE, tabular.environment = "
    longtable")


## ----sumryTabl, echo=FALSE, results='asis', eval=TRUE, message=FALSE, warning
    =FALSE----
desData<-ddply(melt(mrgData[-o, names(mrgdata)]), c("variable"), summarise,
      n=length(value),
      Min=min(value),
      Max=max(value),
      mean=mean(value),
      sd=sd(value),
      skewness=skewness(value)
)
sumryTable<-xtable(desData, caption = "Descriptive statistics of water quality
    variable",label="tbl:sumtab")
print(sumryTable, include.rownames=FALSE, floating=FALSE, tabular.environment =
     "longtable")


## ----anoreg, echo=FALSE, results='asis'----------------------------------
```

```
table<-xtable(anova_reg(lm.model), caption="ANOVA for Regression",label="tbl:anovareg")
print(table,floating=FALSE,tabular.environment="longtable")


## ----vifplot,echo=FALSE,results='asis', fig.cap="Variance Inflation Factor (VIF) for lienar model. The numbers above the bars represents the VIF value for respective variables.", fig.height=4,fig.pos='htb'----
vif.lm<-data.frame(Variable=names(vif(lm.model)), VIF=vif(lm.model), row.names = NULL)
# viftbl<-cbind(vif.lm[1:17,],vif.lm[18:34,])
# print(xtable(viftbl,label="tbl:VIF"),include.rownames=F,floating=FALSE,tabular.environment="longtable")
vifPlot<-ggplot(vif.lm, aes(Variable, 1/VIF))+geom_bar(stat="identity", aes(fill=ifelse(1/VIF<0.1, "Collinear", "Not-Collinear")))+theme_bw()+theme(axis.text.x=element_text(angle=90, hjust=1), legend.title=element_blank(), legend.position="top")+geom_hline(yintercept=0.1, color="red", linetype=2)
vifPlot <- vifPlot + geom_text(aes(label=round(VIF)), angle=90, hjust=0, size=4)
print(vifPlot)


## ----echo=FALSE---------------------------------------------------------
## Making Coefficients Table
mdlFit.mthd<-c("backward", "forward", "exhaustive")
mdlFit.crt<-c("adjr2", "bic", "rss", "cp")
mf.mc<- expand.grid(mdlFit.mthd, mdlFit.crt)
subMdlsNames<- paste(mf.mc$Var1, mf.mc$Var2, sep=".")
selectedVars<-unlist(lapply(mdlFit.mthd, function(x){
  lapply(mdlFit.crt, function(y){
    names(SubMdls[[x,y,'Models']]$coef[-1])
  })
}), recursive = FALSE)


## ----plotsubfit,echo=FALSE,results='asis', fig.cap='Variable Selection with different Creiteria'----
# making plot
plts<-lapply(mdlFit.crt, function(y){
  lapply(mdlFit.mthd, function(x){
    SubMdls[[x,y,'Plots']]
  })
})

plts$ncol=3
do.call(grid.arrange, unlist(plts, recursive = FALSE))


## ----complot,echo=FALSE,results='asis'----------------------------------
```

```
#making table
valdVal<-mdply(mf.mc, function(Var1, Var2){
      cbind(RMSEP=SubMdls[[Var1, Var2, 'CV']]$rmsep,
                         R2pred=SubMdls[[Var1, Var2, 'CV']]$r2pred)
   })

colnames(valdVal)<-c("Method","Criteria","RMSEP","R2prd")
print(xtable(valdVal, caption = 'RMSEP and R2 predicted for subset linear
    models',   label = 'tbl:vldTbl'),floating=FALSE,include.rownames = FALSE,
    tabular.environment="longtable", caption.placement='top')



## ----vifSubset, echo=FALSE, fig.cap='VIF  for chosen submodel selected from
    backward methods with minimum Mallow\'s Cp', fig.height=4, fig.pos='htb
    '----
SubVIF.mat <- vif(SubMdls[['backward', 'cp', 'Models']])
SubVIF.mat <- melt(SubVIF.mat)
SubVIF.mat$Var <- rownames(SubVIF.mat)
rownames(SubVIF.mat) <- NULL
plt <- ggplot(SubVIF.mat, aes(Var, 1/value))
plt <- plt + geom_bar(stat='identity', aes(fill=ifelse(1/value < .10, "
    Collinear", "Not-Collinear")))
plt <- plt + theme_bw() + theme(axis.text.x=element_text(angle=90, hjust=1),
                                legend.title=element_blank(),
                                legend.position="top")
plt <- plt + geom_hline(yintercept=0.1, color="red", linetype=2)
plt <- plt + geom_text(aes(label=round(value)), angle=90, hjust=0, size=4)
plt <- plt + labs(y = '1/VIF', x = 'Variable')
plt



## ----egnValPlot, echo=FALSE, fig.subcap = c('Scree plot of PCA Model','
    Loading plot of PCA Model'), out.width='0.48\\textwidth', fig.show='hold',
    fig.cap='Principal Component Plot'----
eigenvalues<-apply(pcr.model$scores, 2, sd)[1:15]

#screeplot
fun<-melt(eigenvalues)
fun$comp<-factor(rownames(fun), levels = rownames(fun))
rownames(fun)<-NULL
ggplot(fun, aes(comp, value, group=1))+geom_line()+geom_point(shape=21, fill='
    green', size=3)+theme_bw()+theme(axis.text.x=element_text(angle=90, hjust
    =1))+geom_hline(yintercept=1, color="blue", linetype="dashed")+theme(text=
    element_text(size=20))

#biplot
pcrL <- data.frame(pcr.model$loadings[ ,1:2])
pcrL$vars <- rownames(pcrL)
rownames(pcrL) <- NULL
```

```
ggplot(pcrL, aes(Comp.1, Comp.2)) + geom_text(aes(label = vars)) +
  geom_hline(yintercpet = 0, color = 'blue', linetype = 2) +
  geom_vline(xintercpet = 0, color = 'blue', linetype = 2) +
  geom_segment(aes(x = 0, y = 0, xend = Comp.1, yend = Comp.2),
               arrow = arrow(length = unit(0.25, 'cm'), type = 'closed', angle
    = 20)) +
  labs(x = paste('Comp 1(', round(explvar(pcr.model)[1],1), '%)'),
       y = paste('Comp 2(', round(explvar(pcr.model)[2], 1), '%)'))


## ----sumrypcr,echo=FALSE,results='asis'----------------------------------
Xvar<-round(cumsum(explvar(pcr.model)),4)
HPC<-apply(fitted(pcr.model),3,var)/var(mrgdata$HPC)*100
pc.sum<-data.frame(Xvar,HPC)
pc.sum$comp<-rownames(pc.sum)
pc.sum<-pc.sum[,c(3,1:2)]
rownames(pc.sum)<-NULL
pcr.sum<-cbind(pc.sum[1:19,],pc.sum[20:38,])
print(xtable(pcr.sum,caption='Percent Variance Captured by Regression Model
    Using PCR on Reponse and Predictors',
                label='tbl:varPCR'),include.rownames=F,floating=FALSE,
      tabular.environment = "longtable",
      caption.placement='top')


## ----sumrypls,echo=FALSE,results='asis'----------------------------------
## Repeat Table Header Row for longtable ########
addtorow          <- list()
addtorow$pos      <- list()
addtorow$pos[[1]] <- c(0)
addtorow$command  <- c(paste("\\hline \n",
                             "\\endhead \n",
                             "\\hline \n",
                             "{\\footnotesize Continued on next page} \n",
                             "\\endfoot \n",
                             "\\endlastfoot \n",sep=""))
## ------------------------ #########

Xvar<-round(cumsum(explvar(pls.model)),4)
HPC<-apply(fitted(pls.model),3,var)/var(mrgdata$HPC)*100
pl.sum<-data.frame(Xvar,HPC)
pl.sum$comp<-rownames(pl.sum)
pl.sum<-pl.sum[,c(3,1:2)]
rownames(pl.sum)<-NULL
plr.sum<-cbind(pl.sum[1:19,],pl.sum[20:38,])
pxtable<-xtable(plr.sum,
                caption='Percent Variance Captured by Regression Model Using
    PLS',
                label='tbl:varPLS')
```

```
print(pxtable,
      include.rownames = F,
      tabular.environment = "longtable",
      floating=FALSE,
      booktabs=TRUE,
      add.to.row = addtorow,
      sanitize.text.function = function(x){x},
      caption.placement = "top",
      table.placement = 'htbp',
      hline.after=c(-1,-1, nrow(pxtable)))


## ----rmsep.plspcr, echo=FALSE-------------------------------------------
pls.rmsep<-ldply(RMSEP(pls.model)$comps[-1], function(x){RMSEP(pls.model)$val[,,x]})
pcr.rmsep<-ldply(RMSEP(pcr.model)$comps[-1], function(x){RMSEP(pcr.model)$val[,,x]})


## ----lspcpr,echo=FALSE, results='asis', fig.cap='Actual and predicted values for OLS, PCR and PLS model'----
pred.lm<-predict(lm.model) # Prediction using Linear Model
pred.sub<-predict(SubMdls[['backward', 'cp','Models']])
pred.pcr<-predict(pcr.model, ncomp=7)[,,]
pred.pls<-predict(pls.model, ncomp=3)[,,]

## Prediction Matrix
pred.mat<-data.frame(cbind(n=1:nrow(mrgdata),
                           original=mrgdata$HPC,
                           linear=pred.lm,
                           SubModel=pred.sub,
                           PCR.Comp7=pred.pcr,
                           PLS.Comp3=pred.pls))

## Make some Prediction
ggplot(melt(pred.mat, 1:2), aes(n, value))+
  facet_grid(variable~.)+
  geom_line(aes(y=original, color="original"))+
  geom_line(aes(color=variable))+
  theme_bw()+theme(legend.position='top', legend.title=element_blank())+
  ggtitle("Actual and Predicted values for OLS, PCR and PLS")+
  geom_point(aes(color=variable), shape=21, size=1.2, fill='grey')


## ----MdlComp, echo=FALSE------------------------------------------------
lin.vld.cv <- SubMdls[['backward', 'cp' ,'CV']]$rmsep
lin.vld.trn <- rmserr(mrgdata[,y.var], predict(SubMdls[['backward', 'cp', 'Models']]))$rmse
lin.vld <- cbind(train=lin.vld.trn, adjCV=lin.vld.cv)
```

```
vld.mat <- rbind(lin.vld, pcr.rmsep[7, -2], pls.rmsep[3, -2])
vld.mat$Models <- c('OLS', 'PCR.7', 'PLS.3')
vld.mat<-melt(vld.mat, 3)


## ----MdlCompPlt, echo=FALSE, fig.cap='RMSEP plot for selected OLS, PCR and
    PLS models', fig.height=4----
vldPlt<-ggplot(vld.mat, aes(Models, value, color=variable, group=variable)) +
  geom_line() +
  geom_point() +
  theme_bw() +
  theme(legend.position = 'top', legend.title=element_blank())+
  labs(y = 'RMSEP')
vldPlt


## ----rmsep,echo=FALSE,results='asis'-------------------------------------
#rmsep table
rmp<-cbind(Comp = 0:(nrow(pls.rmsep)-1), pls.rmsep,pcr.rmsep)
rmp<-rmp[1:15,]
print(xtable(rmp, digits=3,caption="RMSEP values for 15 components from PCR and
     PLS model",label='tbl:rmse'), add.to.row=list(pos=list(-1), command="\\
    hline  & \\multicolumn{3}{c}{PLS} & \\multicolumn{3}{c}{PCR}\\\\"), caption
    .placement = 'top', include.rownames = FALSE)




## ----pkgsUsed, child="Includes/pkgsUsed.Rnw", eval=TRUE------------------


## ----pkgsUsed, echo=FALSE------------------------------------------------
pkgsDesc<-ldply(c(req.package, "graphics", "grDevices", "utils", "datasets", "
    methods", "base"), function(x){
    data.frame(
    `Package Name`=packageDescription(x)$Package,
    `Version`=packageDescription(x)$Version,
    `Title`=packageDescription(x)$Title)
})
citeKey<-c('car2011FJnWS','dplyr2014WHFR','gdata2014WG','ggplot22009WH','
    gridExtra2012AB','knitr2013XY','leaps2009LT','MASS2001WNV','mixlm2014SK','
    pls2013MBH','plyr2011WH','R2014Rcore','reshape22007WH','xtable2014DD','
    stargazer2013hlavac','moments2012komsta','psych2014revelle','
    corrplot2013wei')
ckSrtd<-unlist(lapply(paste("^",pkgsDesc$Package.Name, sep=""), function(x){
    grep(x, x = citeKey, value = TRUE)
}))
ckSrtd<-c(ckSrtd,rep('R2014Rcore', 6))
citeCmd<-paste("\\cite{",ckSrtd,"}", sep="")
```

```
## ----appendixCodeUsed, child="Includes/codeUsed.Rnw", eval=TRUE----------



## ----appendixRes,child="Includes/rresults.Rnw",eval=TRUE-----------------


## ----summaryr, echo=FALSE------------------------------------------------
lm.summary <- summary(lm.model)
# print(xtable(lm.summary,caption="Multiple linear Regression Summary "))
print(lm.summary)


## ----SelectdSubModel, echo=FALSE-----------------------------------------
summary(update(SubMdls[[’exhaustive’, ’cp’, ’Models’]], .~.+Niti, data=mrgdata)
    )


## ----printSubMdls, echo=FALSE, results=’asis’----------------------------
print(xtable(cbind(Models=subMdlsNames,
                   ‘Selected Variables‘=lapply(selectedVars, paste, collapse=",
     ")),
             align = ’rlX’,
             caption = ’Selected vaiables’,label="tbl:submdl"),
      width = ’\\textwidth’,
      tabular.environment = ’tabularx’,
      floating=FALSE,
      hline.after = c(-1,0,0,1:length(subMdlsNames)),
      caption.placement=’top’)


## ----echo=FALSE,results=’asis’, size="footnotesize"----------------------
# loading results with selected 8 component
pcrlod<-round(pcr.model$loadings[,1:8],digits=3)
print(xtable(pcrlod, caption = ’Loading Tables PCR’,label=’tbl:lodPCR’),
    floating=FALSE,tabular.environment="longtable")


## ----plslod,echo=FALSE,results=’asis’, size="footnotesize"---------------
# loading results with selected 8 component
plslod<-round(pls.model$loadings[,1:8],digits=3)
print(xtable(plslod, caption = ’LoadPing Tables PLS’,label=’tbl:lodPLS’),
    floating=FALSE,tabular.environment="longtable")
```

```


## ----appendixPlots, child="Includes/releventPlots.Rnw", eval=TRUE--------


## ----corPlots, echo=FALSE, fig.width=’\\textwidth’, fig.pos=’!ht’--------
corrplot(cor(mrgdata), method = ’ellipse’, type = ’lower’)


## ----diag plot,echo=FALSE------------------------------------------------
par(mfrow=c(2,2))
plot(lm.model,1:4)


## ----diagsplot,echo=FALSE------------------------------------------------
par(mfrow=c(2,2))
plot(SubMdls[[’exhaustive’, ’cp’, ’Models’]],1:4)


## ----comparisonplot,echo=FALSE,comment=NA,size=’small’,fig.align=’center’,
    fig.height=4, fig.cap=’RMSEP plot for PCR and PLS model’----
##validation plot
pcr.rmsep<-ldply(RMSEP(pcr.model)$comps+1, function(x){RMSEP(pcr.model)$val[,,x
    ]})
pls.rmsep<-ldply(RMSEP(pls.model)$comps+1, function(x){RMSEP(pls.model)$val[,,x
    ]})
rmsep.mat<-melt(list(PCR=pcr.rmsep,PLS=pls.rmsep),0)
rmsep.mat<-data.frame(comp=rep(RMSEP(pcr.model)$comps, 3), rmsep.mat)
ggplot(rmsep.mat, aes(comp, value))+geom_line(aes(color=variable))+theme_bw()+
    theme(legend.position="bottom")+ylab("RMSEP")+xlab("Number of Components")+
    ggtitle("Validation plot for PCR and PLS model")+facet_grid(.~L1)+theme(
    legend.title=element_blank())
```

Appendix C

Some R Results

C.1 Multiple Regression Summary

```

Call:
lm(formula = makeFormula(x.var, y.var), data = mrgdata)

Residuals:
     Min       1Q   Median       3Q      Max
-1.56416 -0.28736 -0.00614  0.28775  1.22034

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  12.883816   2.485770   5.183 7.78e-07 ***
Irn         -48.141938  11.393822  -4.225 4.36e-05 ***
ReCh          2.474979   0.993451   2.491 0.013940 *
TOC           0.034209   0.105931   0.323 0.747242
Col           0.026212   0.015659   1.674 0.096466 .
Cal          -0.010614   0.007637  -1.390 0.166853
Cond         -0.145087   0.031630  -4.587 1.02e-05 ***
COD          -0.558906   0.090704  -6.162 7.75e-09 ***
Alk          -0.897209   0.450326  -1.992 0.048349 *
Sod          -0.083998   0.023606  -3.558 0.000516 ***
Mang          8.726631   2.230993   3.912 0.000145 ***
Temp         -0.041000   0.038768  -1.058 0.292129
pH           -0.251952   0.177008  -1.423 0.156929
Tur          -4.229286   1.966653  -2.150 0.033297 *
Alu           0.255400   0.684222   0.373 0.709532
Niti         -5.954323  16.042706  -0.371 0.711104
Nita          2.610701   0.838129   3.115 0.002248 **
Amonia       -0.574468   1.021065  -0.563 0.574629
Irp          -0.379690   0.851574  -0.446 0.656407
PVC          -0.174165   0.079136  -2.201 0.029449 *
PEL           0.300310   0.152650   1.967 0.051198 .
```

```
GUP           0.172865   0.081400   2.124 0.035524 *
A2001        -0.065971   0.019007  -3.471 0.000697 ***
UPDi          0.078052   0.025668   3.041 0.002835 **
VoTa         -0.527945   0.205707  -2.566 0.011364 *
PlDi          0.034492   0.016559   2.083 0.039143 *
PiRe          0.016154   0.020618   0.783 0.434717
Cem          -0.062787   0.046158  -1.360 0.176015
B1910         0.015313   0.079664   0.192 0.847859
A1970         0.300080   0.228178   1.315 0.190702
B1940         0.123059   0.115201   1.068 0.287333
B1970        -0.163310   0.711946  -0.229 0.818916
Irn:Irp       3.959401   0.855353   4.629 8.53e-06 ***
Cond:COD      0.031870   0.005449   5.848 3.56e-08 ***
PVC:PEL      -0.017946   0.018827  -0.953 0.342191
Alk:Sod       0.112931   0.059789   1.889 0.061061 .
pH:Tur        0.599387   0.258191   2.321 0.021758 *
Niti:Nita   -26.672534   5.187499  -5.142 9.36e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

s: 0.5736 on 135 degrees of freedom
Multiple R^2: 0.7917,
Adjusted R^2: 0.7347
F-statistic: 13.87 on 37 and 135 DF,  p-value: < 2.2e-16
```

C.2 Backward subset model chosen with minimum Mallows' Cp and minimum RMSEP

```

Call:
lm(formula = HPC ~ Irn + ReCh + Col + Cal + Cond + COD + Alk +
    Sod + Mang + Temp + pH + Tur + Nita + Irp + PVC + PEL + GUP +
    A2001 + UPDi + VoTa + PlDi + Cem + Niti + Irn:Irp + Cond:COD +
    Alk:Sod + pH:Tur + Nita:Niti, data = mrgdata)

Residuals:
     Min       1Q   Median       3Q      Max
-1.53036 -0.31723 -0.03898  0.27381  1.27127

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  12.669396   1.984749   6.383 2.23e-09 ***
Irn         -48.227050  10.718474  -4.499 1.39e-05 ***
ReCh          2.504438   0.924022   2.710 0.007538 **
Col           0.027944   0.012999   2.150 0.033247 *
Cal          -0.010555   0.007245  -1.457 0.147339
```

```
Cond         -0.145477   0.028804  -5.051 1.31e-06 ***
COD          -0.537448   0.084586  -6.354 2.59e-09 ***
Alk          -0.901174   0.425108  -2.120 0.035732 *
Sod          -0.082238   0.022668  -3.628 0.000396 ***
Mang          8.282068   1.529085   5.416 2.49e-07 ***
Temp         -0.064091   0.033257  -1.927 0.055934 .
pH           -0.245898   0.166986  -1.473 0.143050
Tur          -4.138340   1.882630  -2.198 0.029534 *
Nita          2.588576   0.536104   4.828 3.47e-06 ***
Irp          -0.153365   0.081301  -1.886 0.061256 .
PVC          -0.182320   0.046159  -3.950 0.000122 ***
PEL           0.115147   0.043586   2.642 0.009156 **
GUP           0.198438   0.073370   2.705 0.007663 **
A2001        -0.060670   0.016981  -3.573 0.000481 ***
UPDi          0.079764   0.024553   3.249 0.001443 **
VoTa         -0.462296   0.164675  -2.807 0.005688 **
PlDi          0.035103   0.015371   2.284 0.023851 *
Cem          -0.088498   0.028341  -3.123 0.002167 **
Niti         -1.642486   2.452259  -0.670 0.504067
Irn:Irp       3.959788   0.805718   4.915 2.39e-06 ***
Cond:COD      0.030163   0.005089   5.927 2.18e-08 ***
Alk:Sod       0.117667   0.055377   2.125 0.035307 *
pH:Tur        0.585919   0.246413   2.378 0.018730 *
Nita:Niti   -25.973292   4.617559  -5.625 9.35e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

s: 0.5646 on 144 degrees of freedom
Multiple $R^2$: 0.7847,
Adjusted $R^2$: 0.7429
F-statistic: 18.75 on 28 and 144 DF,  p-value: < 2.2e-16
```

C.3 Subset of linear model using various criteria

	Models	Selected Variables
1	backward.adjr2	Irn, ReCh, Col, Cal, Cond, COD, Alk, Sod, Mang, Temp, pH, Tur, Nita, Irp, PVC, PEL, GUP, A2001, UPDi, VoTa, PlDi, Cem, A1970, B1940, Irn:Irp, Cond:COD, Alk:Sod, pH:Tur, Nita:Niti
2	forward.adjr2	Irn, ReCh, Col, Cond, COD, Alk, Sod, Mang, Nita, PVC, PEL, GUP, A2001, UPDi, VoTa, Cem, Irn:Irp, Cond:COD, Alk:Sod, Nita:Niti

3	exhaustive.adjr2	Irn, ReCh, TOC, Col, Cal, Cond, COD, Alk, Sod, Mang, Temp, pH, Tur, Alu, Niti, Nita, Amonia, Irp, PVC, PEL, GUP, A2001, UPDi, VoTa, PlDi, PiRe, Cem, B1910, A1970, B1940, B1970, Irn:Irp, Cond:COD, PVC:PEL, Alk:Sod, pH:Tur, Niti:Nita
4	backward.bic	Irn, ReCh, Col, Cal, Cond, COD, Alk, Sod, Mang, Temp, pH, Tur, Nita, Irp, PVC, PEL, GUP, A2001, UPDi, VoTa, PlDi, Cem, Irn:Irp, Cond:COD, Alk:Sod, pH:Tur, Nita:Niti
5	forward.bic	Irn, ReCh, TOC, Col, Cal, Cond, COD, Alk, Sod, Mang, Temp, pH, Tur, Niti, Nita, PVC, PEL, GUP, A2001, UPDi, VoTa, PlDi, PiRe, Cem, B1970, Irn:Irp, Cond:COD, PVC:PEL, Alk:Sod, pH:Tur, Niti:Nita
6	exhaustive.bic	TOC, Col, Cond, COD, pH, Niti, PVC, PEL, Cem, Irn:Irp, Cond:COD, Niti:Nita
7	backward.rss	Irn, ReCh, TOC, Col, Cal, Cond, COD, Alk, Sod, Mang, Temp, pH, Tur, Alu, Niti, Nita, Amonia, Irp, PVC, PEL, GUP, A2001, UPDi, VoTa, PlDi, PiRe, Cem, B1910, A1970, B1940, B1970, Irn:Irp, Cond:COD, PVC:PEL, Alk:Sod, pH:Tur, Niti:Nita
8	forward.rss	Irn, ReCh, TOC, Col, Cal, Cond, COD, Sod, Mang, Temp, pH, Tur, Niti, Nita, PVC, PEL, GUP, A2001, UPDi, VoTa, PlDi, Cem, B1970, Irn:Irp, Cond:COD, PVC:PEL, pH:Tur, Niti:Nita
9	exhaustive.rss	Irn, ReCh, Col, Cal, Cond, COD, Alk, Sod, Mang, Temp, pH, Tur, Niti, Nita, PVC, PEL, GUP, A2001, UPDi, VoTa, PlDi, Cem, B1970, Irn:Irp, Cond:COD, PVC:PEL, Alk:Sod, pH:Tur, Niti:Nita
10	backward.cp	Irn, Col, Cond, COD, Mang, PVC, PEL, A2001, UPDi, B1940, Irn:Irp, Cond:COD, PVC:PEL, Niti:Nita
11	forward.cp	Irn, ReCh, TOC, Col, Cal, Cond, COD, Alk, Sod, Mang, Temp, pH, Tur, Alu, Niti, Nita, Amonia, Irp, PVC, PEL, GUP, A2001, UPDi, VoTa, PlDi, PiRe, Cem, B1910, A1970, B1940, B1970, Irn:Irp, Cond:COD, PVC:PEL, Alk:Sod, pH:Tur, Niti:Nita
12	exhaustive.cp	Irn, ReCh, Col, Cal, Cond, COD, Alk, Sod, Mang, Temp, pH, Tur, Nita, Irp, PVC, PEL, GUP, A2001, UPDi, VoTa, PlDi, Cem, Irn:Irp, Cond:COD, Alk:Sod, pH:Tur, Nita:Niti

C.4 PCA Results

	Comp 1	Comp 2	Comp 3	Comp 4	Comp 5	Comp 6	Comp 7	Comp 8
Irn	0.00	0.32	-0.15	0.07	-0.17	0.17	-0.07	0.08
ReCh	0.04	-0.15	-0.14	-0.11	-0.02	0.21	-0.17	0.17
TOC	0.01	0.11	0.06	0.47	-0.02	0.04	0.08	0.11
Col	0.10	0.23	-0.07	0.18	-0.31	0.07	-0.18	0.02
Cal	-0.13	-0.05	-0.20	-0.13	0.07	-0.24	-0.05	0.35
Cond	-0.26	-0.14	-0.12	-0.08	0.01	-0.09	0.03	0.06
COD	-0.10	-0.15	0.11	0.00	-0.45	0.03	-0.17	0.12
Alk	-0.12	0.09	-0.32	-0.08	0.13	-0.24	-0.01	0.26
Sod	-0.16	0.01	-0.12	-0.21	-0.08	-0.01	-0.04	-0.52
Mang	-0.02	-0.03	-0.01	0.22	0.20	-0.15	-0.55	-0.12
Temp	-0.07	0.03	-0.02	0.10	0.14	0.39	-0.32	0.23
pH	-0.17	-0.14	-0.09	-0.01	0.26	-0.18	0.06	0.25
Tur	-0.06	0.24	-0.32	0.04	-0.18	-0.03	0.03	-0.00
Alu	0.03	0.14	0.14	-0.02	-0.11	-0.04	-0.02	0.29
Niti	-0.21	-0.10	0.15	-0.14	-0.27	-0.10	-0.14	0.05
Nita	-0.17	-0.00	0.05	0.24	0.11	0.20	0.01	-0.15
Amonia	0.01	-0.06	-0.05	0.16	0.15	-0.22	-0.58	-0.06
Irp	0.26	0.02	0.01	-0.12	-0.09	-0.24	-0.07	-0.08
PVC	0.10	-0.26	-0.23	0.26	-0.04	-0.04	0.11	-0.11
PEL	0.09	-0.30	-0.15	0.22	-0.22	-0.06	0.04	-0.07
GUP	0.24	0.07	0.06	-0.24	-0.06	0.05	-0.09	0.04
A2001	0.15	-0.25	-0.17	0.01	-0.06	0.20	0.01	0.02
UPDi	0.10	-0.22	-0.20	-0.20	-0.01	0.32	-0.07	0.08
VoTa	0.29	0.04	0.08	-0.05	0.05	-0.01	-0.06	0.10
PlDi	0.13	-0.17	-0.16	-0.23	-0.01	0.24	-0.04	0.14
PiRe	0.18	-0.20	-0.17	0.09	-0.12	-0.06	-0.01	0.09
Cem	0.04	0.11	-0.21	-0.28	0.24	0.22	-0.16	-0.27
B1910	0.27	0.01	0.10	-0.08	-0.06	-0.16	-0.05	0.16
A1970	0.24	0.03	-0.08	-0.13	-0.08	-0.20	-0.09	-0.16
B1940	0.28	0.02	0.10	-0.08	-0.00	-0.19	-0.01	0.07
B1970	0.28	0.06	-0.09	0.01	0.06	-0.14	0.03	-0.11
Irn:Irp	0.06	0.31	-0.10	0.02	-0.20	0.13	-0.11	0.03
Cond:COD	-0.23	-0.14	0.06	-0.14	-0.25	-0.07	-0.13	0.03
PVC:PEL	0.13	-0.28	-0.21	0.23	-0.11	-0.02	0.04	-0.08
Alk:Sod	-0.15	0.14	-0.35	-0.07	-0.03	-0.17	0.04	-0.04
pH:Tur	-0.08	0.23	-0.33	0.04	-0.16	-0.05	0.04	0.01
Niti:Nita	-0.20	-0.11	0.16	-0.13	-0.24	-0.04	-0.14	-0.15

Table C.2: Loading Tables PCR

C.5 Partial Least Square Loading table

	Comp 1	Comp 2	Comp 3	Comp 4	Comp 5	Comp 6	Comp 7	Comp 8
Irn	0.16	0.15	-0.11	-0.04	-0.47	0.44	0.03	-0.06
ReCh	-0.03	-0.06	0.40	0.13	0.11	0.07	-0.25	0.04
TOC	0.04	-0.02	-0.22	0.25	-0.38	0.06	0.28	-0.51
Col	0.25	0.08	0.12	0.06	-0.28	0.37	0.05	-0.26
Cal	-0.20	0.04	-0.04	-0.10	-0.11	0.19	-0.13	0.30
Cond	-0.35	0.14	-0.02	-0.19	-0.09	0.06	-0.11	0.14
COD	-0.08	0.16	0.36	-0.44	0.14	0.03	0.18	-0.30
Alk	-0.15	0.06	-0.12	0.21	-0.11	0.40	-0.31	0.34
Sod	-0.14	0.18	0.08	-0.16	-0.03	0.05	-0.39	0.23
Mang	-0.04	0.01	0.21	0.55	-0.44	-0.18	0.18	0.03
Temp	-0.05	0.09	0.11	0.21	-0.60	-0.11	0.27	0.07
pH	-0.30	-0.01	-0.16	0.07	-0.08	-0.09	-0.03	0.22
Tur	0.02	0.14	-0.07	0.04	-0.33	0.57	-0.45	-0.01
Alu	0.16	0.11	0.04	0.15	0.30	-0.16	-0.30	-0.20
Niti	-0.17	0.27	0.20	-0.33	0.34	-0.10	0.13	-0.03
Nita	-0.19	0.12	-0.10	0.28	-0.15	-0.02	0.28	-0.22
Amonia	-0.03	-0.02	0.27	0.41	-0.49	-0.16	0.12	0.10
Irp	0.29	-0.20	0.10	-0.17	0.11	0.01	-0.07	0.13
PVC	-0.12	-0.29	0.16	0.07	-0.12	0.28	-0.08	-0.28
PEL	-0.10	-0.24	0.30	-0.12	0.01	0.27	0.03	-0.34
GUP	0.32	-0.11	0.13	-0.18	0.14	-0.07	-0.03	0.21
A2001	-0.01	-0.27	0.28	-0.18	-0.16	0.14	0.01	-0.00
UPDi	-0.02	-0.18	0.36	-0.19	-0.09	0.12	-0.12	0.24
VoTa	0.32	-0.22	0.04	-0.01	0.00	-0.13	0.07	0.05
PlDi	0.03	-0.18	0.28	-0.26	-0.06	0.09	-0.09	0.30
PiRe	0.04	-0.28	0.23	-0.17	-0.20	0.15	-0.12	-0.07
Cem	0.07	0.01	0.09	0.28	-0.12	0.10	-0.30	0.52
B1910	0.30	-0.21	0.05	-0.22	0.05	-0.13	0.06	0.05
A1970	0.26	-0.19	0.12	-0.14	0.01	0.12	-0.12	0.20
B1940	0.30	-0.22	0.01	-0.11	0.15	-0.15	-0.00	0.05
B1970	0.27	-0.27	-0.04	0.05	-0.10	0.10	-0.13	0.09
Irn:Irp	0.23	0.12	-0.06	-0.11	-0.39	0.39	0.08	-0.02
Cond:COD	-0.23	0.23	0.20	-0.40	0.15	-0.06	0.06	-0.00

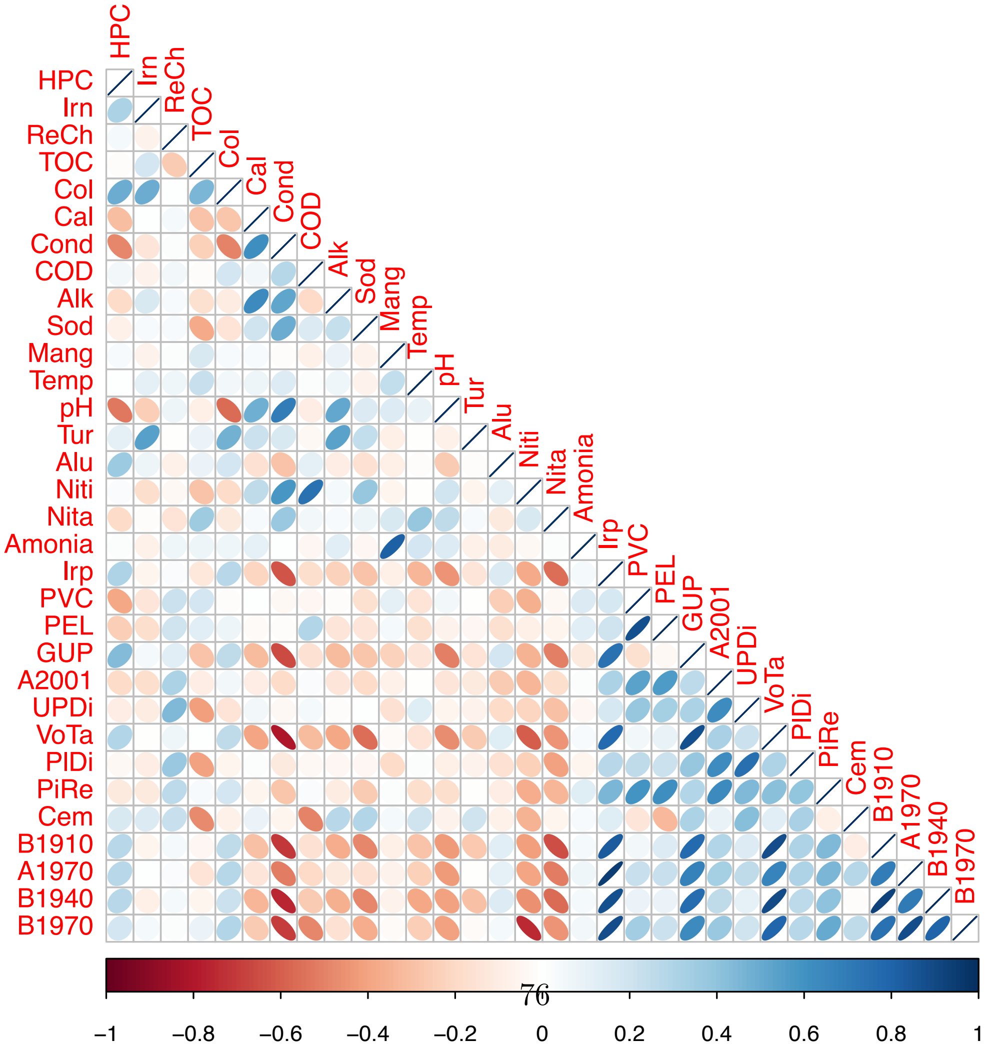
PVC:PEL	-0.07	-0.30	0.26	-0.02	-0.12	0.28	-0.02	-0.26
Alk:Sod	-0.14	0.13	-0.08	0.10	-0.15	0.49	-0.48	0.21
pH:Tur	-0.00	0.14	-0.08	0.05	-0.33	0.57	-0.45	0.01
Niti:Nita	-0.16	0.23	0.16	-0.43	0.12	-0.21	0.12	-0.06

Table C.3: Loading Tables PLS

Appendix D

Some Relevent Plots

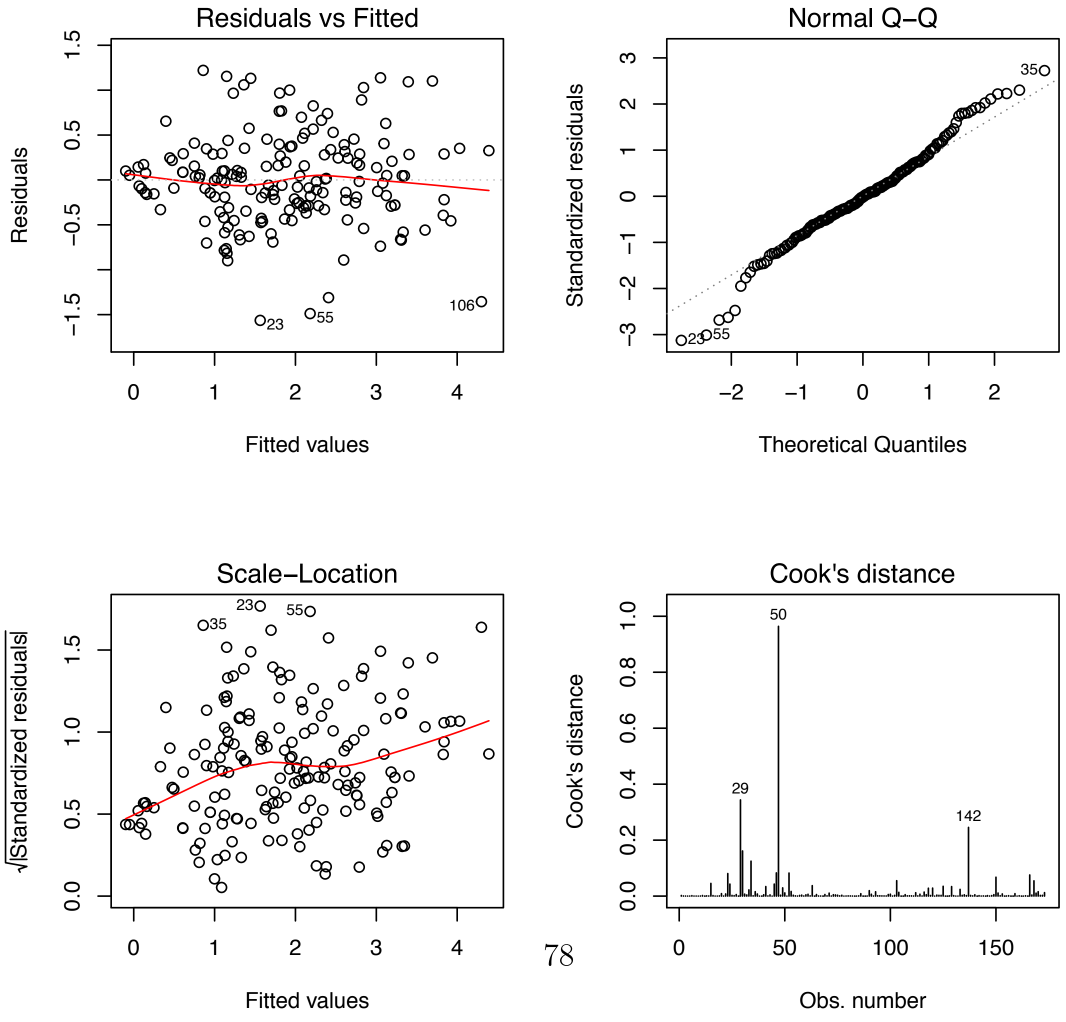
D.1 Correlation between variables



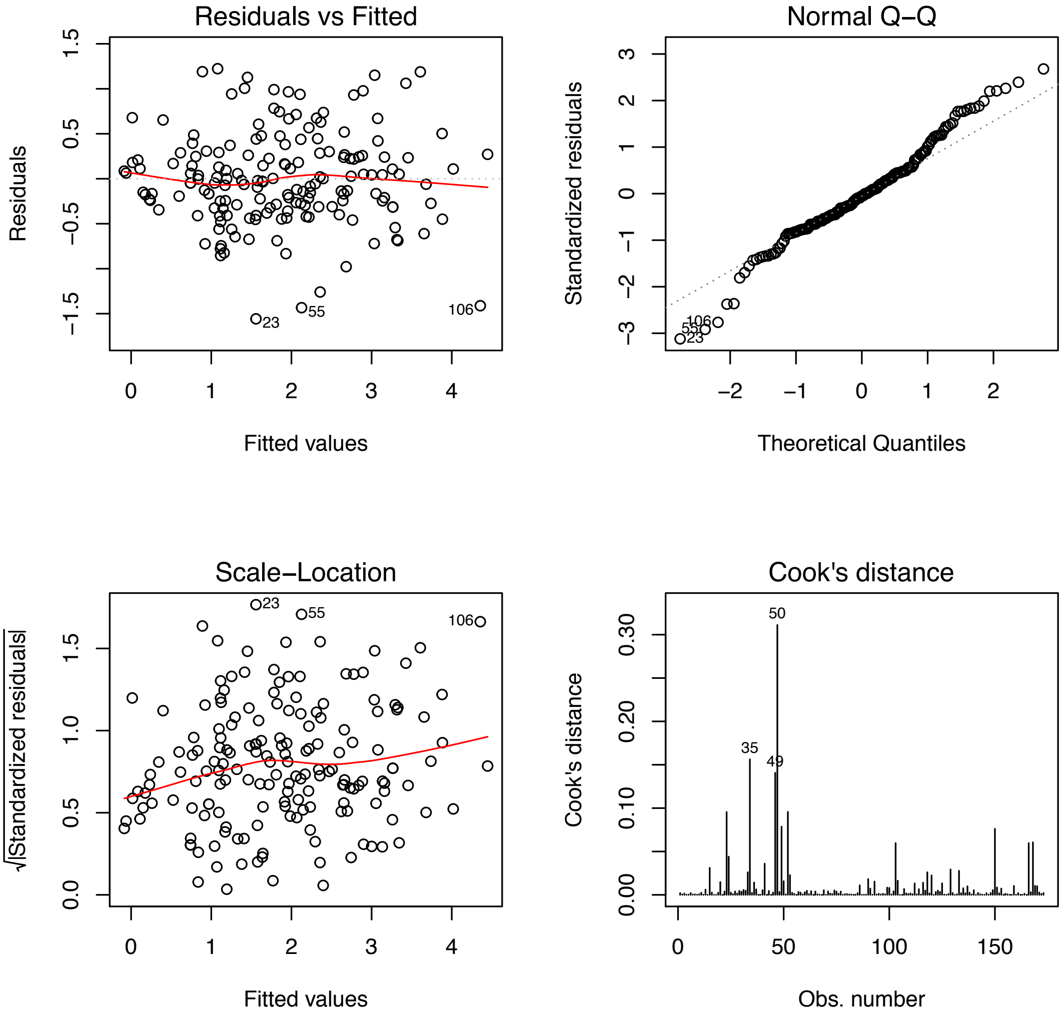
Appendix E

Diagnostic plot

E.1 Diagnostic Plot Multiple Regression model



E.2 Diagnostic Plot Subset Model



E.3 Prediction plot

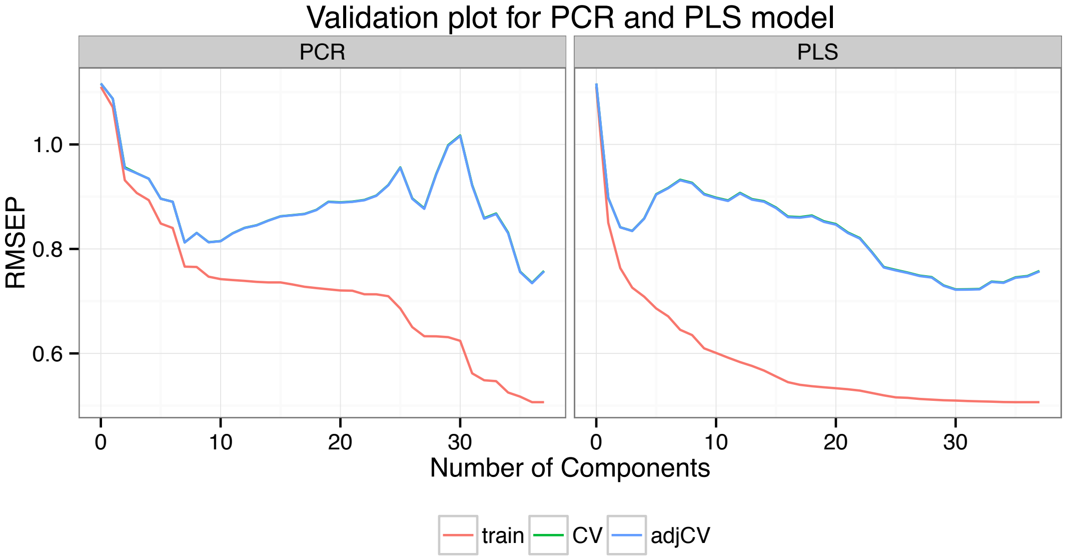


Fig E.1: RMSEP plot for PCR and PLS model

Norwegian University
of Life Sciences

Postboks 5003
NO-1432 Ås, Norway
+47 67 23 00 00
www.nmbu.no