Smart Data Driven Quality Prediction for Urban Water Source Management

Di Wu,[a,1], Hao Wang[b,2,*], Razak Seidu,[a,3]

[a] *Norwegian University of Science and Technology, Ålesund, Norway.*
[b] *Norwegian University of Science and Technology, Gjøvik, Norway.*

Abstract

A water supply system that integrates water source management, treatment and distribution is a critical infrastructure in urban areas all around the world. The water quality is becoming a key factor to evaluate life quality for local residents. However, the urban water quality control is facing more and more challenges from growing human population, industrial and agricultural pollution. Traditional water quality research mostly focused on separate aspects, such as different types of physical, chemical or biological indicators. These works lack of a comprehensive coverage of all aspects, which undermines the accuracy of the predictive models.

In this paper, we build a smart data analysis scheme to analyze and predict the water quality, considering all the water quality standard indicators in a comprehensive environment. Instead of data output from water treatment, we collect the raw water data directly from water sources, which are the origins

[*] Corresponding author
Email address: `hawa@ntnu.no` (Hao Wang)
[1] Department of ICT and Natural Science, Norwegian University of Science and Technology, Ålesund, 6009, NORWAY.
Url: https://www.ntnu.edu/iir/cps#/view/about
Email: di.wu@ntnu.no.
[2] Department of Computer Science, Norwegian University of Science and Technology, Gjøvik, 2815, NORWAY.
Url: https://www.ntnu.edu/idi.
Email: hawa@ntnu.no.
[3] Water and Environmental Engineering Laboratory, Department of Ocean Operations and Civil Engineering, Norwegian University of Science and Technology, Ålesund, 6009, NORWAY.
Url: https://www.ntnu.edu/ihb/water-and-environmental-engineering-lab.
Email: rase@ntnu.no.

of water supply systems. By doing so, we have better coverage of the whole process and better chances of detecting anomalies and risks earlier. We design two models to predict the water quality, especially for the most critical biological indicators. The two models are (1) adaptive learning rate BP neural network (ALBP) and (2) 2-step isolation and random forest (2sIRF) considering different features of these indicators. We applied these models in the practical urban water supply systems of Oslo and Bergen in Norway. The data was collected directly from the water sources in the two cities for over 7 years. The results show that ALBP is theoretically simple and easy to implement. 2sIRF considers the risk distribution and shows higher prediction accuracy. In addition, we perform the correlation analysis of all the indicators and and the importance analysis over different indicators. The domain experts have confirmed that this work is meaningful for future risk control and decision support in urban water supply systems.

Keywords: Water Supply System, Smart Data Analysis, Water Quality Control, Correlation Analysis, Adaptive Learning Rate BP Neural Network (ALBP), 2 step Isolation and Random Forest (2sIRF).
2018 MSC: 00-01, 99-00

1. Introduction

Water covers 71% of the Earth's surface, mostly in the four oceans. But only a small portion (3%) is fresh water. Most of the fresh water is in icecaps, glaciers (69%) and groundwater (30%). All the lakes, rivers and swamps combined merely account for 0.3% of the total reserves [1]. The importance of water as a resource has been acknowledged by all over the world for a long time. Currently, there are still more than 1.2 billion people lacking access to clean drinking water. This has resulted in a new concept known as *Water Stress*. Fresh water provides people's fundamental requirements for agriculture, industry and living, especially in modern cities. The easily accessible freshwater, which is mainly groundwater and surface water bodies are significantly prone to chemical and microbial contamination. Currently, more than 631 million of the global population rely on drinking water sources that are seriously contaminated. In 2015, 91% of the world's population had improved their drinking-water sources, compared with 76% in 1990. An important note is that these figures depend a lot on the various national or regional water quality standards.

The 21st century has been a continuous process of global urbanization. According to the United Nations Department of Economic and Social Affairs (UN-DESA), the majority of the world population now is living in cities, and this proportion will go on to develop with projections of 68% by the year 2050 [2].

It is often said "water is life", but another truth is *Water Quality is Health.* The United Nations General Assembly published in September 2015, the Sustainable Development Goals (SDGs). It presents a collection of 17 global goals as a general vision of achieving a higher level of human health and well-being worldwide by the year 2030. In this document, Clean Water and Sanitation is ranking as Goal 6 [3] in order to raise the global concerns of water quality. However, the need to improve water quality to safeguard public health in urban water supply remains a major challenge.

Water supply is one of the most traditional industries to provide drinking water for city lives. The prevalent urban water supply process can be divided into 3 parts, including: water source management, water treatment and water distribution. Water source management refers to the control of water origins in urban areas. The present water sources in urban areas are mainly ground water, such as lakes, rivers or underground water. This is the first step for water quality control, however often neglected based on geographical or economical factors. Water treatment is the key step in urban supply systems. Current routine monitoring of water quality is carried out at water treatment plants with the purpose of assessing risk levels and tendencies, based on which effective management decisions are taken for the provision of safe drinking water to the public. Water distribution is to carry treated water to the end users by distributed pipes in the city.

The urban water quality is facing more and more challenges from industrial, agriculture and social pollution. Extensive anthropogenic activities in many regions of the world result in compelling water quality deterioration. To addressed this situation, several key directives and guidelines for water quality improvements such as European Union Water Framework Directive [4], Clean Water Act [5] in United States, the World Health Organization (WHO) drinking water quality guidelines [6] have been developed. The valid implementation of these principles depends on the establishment of a robust and verifiable monitoring regime of water supply systems, especially in water source management. Good water quality control in the source management enables earlier detection of risks so the supply system can have longer time to react to the risks.

Water quality refers to the chemical, physical, biological, and sometimes radiological features of water. In the face of these conventional preventive measures, cases of waterborne illnesses resulting from the use of both drinking and recreational water are regularly reported worldwide. It is often evaluated by various water quality indicators. Among these indicators, concentrations of fecal indicator organisms (FIOs) in raw water provide an overview of potential levels of pathogens in the water source, and form the basis for optimizing the following treatment procedures to prevent potential disease outbreaks. In addition, real-time water quality monitoring systems are applied to detect variations in water source quality such that appropriate management strategies and actions can be implemented.

Complex interactions among physical and chemical indicators of water make it difficult to identify precisely the effect of each indicator on the concentrations of the FIOs in water [7]. Moreover, typical water treatment workers and managers may not entirely have skills for analysis and exploration of data collection [8]. Accordingly, showing the patterns, tendencies and relationships between the water quality indicators in simple, convenient way has the potentials to convey the complex nature of the data that can be easily followed and interpreted by water supply workers, managers and researchers. Furthermore, water quality control is based on the trends of these indicators. These trends can affect the whole process of the local water treatment plant.

Most of the national water quality standards rely on biological indicators. They are different for various geographical, industrial or development stage conditions of different regions. Typical biological indicators can include coliform, plecoptera, mollusca, escherichia coli, ephemeroptera, trichoptera, *etc.* In practice, most of them are inspected with sample-based testing. The samples are taken to bacterial culture in the laboratory, and we can obtain the results from several hours to a full day. If we consider the effect for bacteria on human's body, these biological indicators are much more dangerous compared to others. This issue has been an important challenge in water supply industry.

Urban water supply has been included in all the new *Smart City* designs. The new generation of *Smart Water Supply* systems are pictured to integrate various sensors, controllers, cloud computing and data technologies in order to provide safe, stable and sufficient water for the increasing requirements in many enlarging cities. At present, more and more advanced ubiquitous sensing technologies bring efficiency, convenience and new insights for almost all of our daily lives [9]. They offer the ability to detect, transmit and measure many environmental indicators, from delicate ecological and natural resources to urban surroundings. In many smart city designs, water supply systems are also integrated with various sensors in order to manage resources and monitor water quality efficiently. This design makes data to play an important role for our better understanding toward existing systems. In Norway, we deployed many different sensors in the water source areas, including multiple sensors for pH, temperature, conductivity, *etc.* The massive data collected by those low-cost sensors plus the recent data analysis technologies, help us greatly to improve the water quality control process. One of the key issue in water quality control is for biological indicators. Until now, there is very few appropriate microbial sensors to directly measure biological indicators [10].

This paper proposes a smart data-driven framework and related algorithms to support quality prediction in urban water supply systems. It takes the advantages of the historical data resources in the industrial process, and builds feasible mathematical models to predict biological water quality indicators. We design two algorithms for water quality prediction, namely adaptive learning rate BP neural network (ALBP) and 2 step isolation and random forest (2sIRF). For application, we select two typical cities, Oslo and Bergen, from Norway to apply our methods. We present an in-depth analysis of the relationship between physical, chemical and biological indicators. The results are compared technically for prediction accuracy and time efficiency. In addition, we also provide the insight analysis for the urban water supply domain.

To our knowledge, this is the first effort to integrate data-driven technologies with smart water supply system for water quality control. This work is essentially a first step towards developing the necessary tools to improve the comprehension of the interactions among the water quality indicators and how they result in varying the levels of FIOs in raw water.

There are several technical contributions for this research. (1) It takes the advantage of the modern big data technology to solve a traditional water quality control question with low costs and untouched historical data. (2) It builds the connection between easily accessible physical and chemical indicators with biological indicators, which currently are not possible for real time measurements. (3) It provides risk control and decision support for urban water supply systems and our methods have been applied and tested in the real-world applications. This can avoid the questions such as laboratory data reliability and applicability. Regular treatment methods usually include coagulation and flocculation, sedimentation, filtration, and disinfection. These steps in the treatment plants can adjust to the development and prediction of water quality indicators, and improve drinking water quality in the supply systems.

2. Related work

This work is interdisciplinary by bridging data analysis and water quality control. In this section, we compare this work with related work in these fields.

2.1. Smart Water Supply

According to World Health Organization (WHO) [11], drinking water supply is still a worldwide challenging issue. Norway has an extensive coverage of drinking water sources in most of the urban areas. In order to ensure that drinking water quality is sustained, the government adopts strict quality guidelines [12] that comply with the European Water Directive Framework. The Norwegian drinking water regulations nowadays mainly depend on the water control in treatment plants. They require operators in the water treatment plants to consistently monitor various quality indicators in their raw and treated water.

Traditionally, the water quality monitoring approach has been employed to achieve these drinking water requirements. This approach primarily involves the interlinked steps of sample collection, transportation/institution analysis, laboratory analysis reporting [13]. It is particularly a cumbersome process and very time-consuming when applied to biological/microbial indicator organisms with respect to public health concerns. For example, following a contamination event in a water supply system, it takes at least 24 hours to obtain results on all the indicator organisms. Such delay in the identification of microbial organisms can inevitably lead to major disease outbreaks. For instance, a waterborne disease outbreak happened in Bergen 2004, partly attributed to the failure of the water utility operators to identify the aetiological agent in the treated water.

To build smart water management systems, Yuan *et al* [14] reviewed the instrumentation, control and automation (ICA) research in the sub-systems of urban water systems.Dogo *et al* [15] designed a framework for water management using blockchain and IoT technologies. Petri *et al* [16] proposed an intelligent analytics system to optimize catchment flow for national fishery. Chang *et al* [17] built a platform to analyze flood inundation with hydroinformatics. Sven *et al* [18] reviewed the data-driven technologies in urban water management systems.

In water quality control, there are some trial work to use data for predictions. Kang *et al* [19] reviewed the possible data-driven technologies in water quality analysis. Holger *et al* [20] designed an ANN to predict salinity in the River

Murray, Australia. Based on the data collected at Astane station in Sefidrood River, Iran, Orouji *et al* designed different models such as ANFIS, Genetic Algorithm and Shuffled Frog Leaping Algorithm to predict chemical indicators (e.g., sodium, potassium, magnesium, *etc* [21][22][23]. Chang *et al* [24] proposed a systematic analysis scheme to analyze and predict the value of NH_3-H for Dahan River basin in Taiwan. Our team in NTNU has also developed an adaptive Neuro-Fuzzy inference system to predict norovirus concentration in drinking water supply [7][25] and frequency analysis[26].

However, smart water management systems are concentrating on the supply and consumption balance for water supply, not directly related with water quality. For quality control research, most of these works are aiming for single value prediction. They have not addressed well the difficulties of data collection and efficiency in industrial process. In addition, the connections between different water quality indicator groups, especially for biological bacteria have been ignored.

2.2. Data-Driven Technologies

The United Nations SDGs for water and sanitation targets for more detailed monitoring and response to understand the coverage and quality of safely managed water sources. Data has to be collected by various sensors in the industrial process. Recently, there have been significant development and applications of sensing technologies in different sectors for the detection and transmission of vital data on key indicators [9][27][28]. At the same time, different types of sensors also attracted a lot of attention in water quality sections. Currently, most water treatment plants have also deployed various sensors that enable plant operators to collect and collate real-time data on a wide range of physical and chemical water quality indicators.

For environment analysis, Wang *et al* [29] has analyzed the air pollution in 2017. More precisely, Luis Andres *et al* [10] in 2018 reviewed different types of sensors to monitor water and sanitation interventions, including satellite remote sensing, aerial vehicles, water quality sensors *et al.* In 2007, Le Dinh *et* *al* [30] designed a remote sensor network for outdoor water quality monitoring in Queensland, Australia. They used real-time sensors to test indicators like temperature, conductivity, salinity, the level of the underground water table, *etc.* In order to meet Water Framework Directive in European Union (EU), O'Flynn *et al* [31] proposed a Smart Coast system to monitor water quality indicators including temperature, phosphate, dissolved oxygen, conductivity, pH, turbidity and water level in Ireland. More recently, Yagur-Kroll *et al* [32][33] introduced general bacterial sensor cells for water quality monitoring. The most commonly deployed real-time sensors in Norwegian water treatment plants are for pH, temperature, turbidity, conductivity, alkalinity and color. There have also been attempts to establish real-time sensors for fatal indicator bacteria. However, the fatal indicator sensors have proven to be significantly flawed in terms of sensitivity and specificity. In order to prevent waterborne disease outbreaks, mathematical modeling approaches are being applied to predict the occurrence and concentrations of fatal indicator bacteria and pathogenic organisms using real-time physical and chemical water quality data.

For data technology itself, as we can see, a big amount of data is gathered by these numerous sensors that need to be analyzed using robust methods [34][35][36]. The challenge for water quality monitoring is to find the essential features from those heterogeneous and unstructured data sets to build the corresponding model. Traditional big data processing techniques including exploratory data analysis (EDA) methods [37], such as principal component analysis (PCA) [38], Singular value decomposition (SVD) [39]. But they cannot integrate domain knowledge. Therefore, in order to find efficiently the useful insight for modeling process, we can also explore the new machine learning algorithms. During the learning process, it is convenient to dynamically integrate with appropriate domain knowledge models. The commonly used machine learning algorithms include logistic regression [40], supporting vector machine [41], as well as reinforcement learning [42]. The new insight we find from the data can support both the monitoring and prediction of water quality.

Most of these methods have different requirements to the problems, such as heuristic knowledge, concrete models, large scale of training sets, *ect.* In this study, we will first analyze our problem, and then adapt appropriate methods according to the requirements.

3. Problem Analysis

3.1. System Description

A water supply system consists of the interlinked steps of source, treatment and distribution. In order to preserve the water quality for end users, all these steps have to be well managed within a well articulated water safety planning framework. Water sources are the origins of supply chain. They play the key role in water quality control. Raw water are collected from the catchment areas such as lakes. The monitoring and control in the catchment area is also an industrial process.

In practice, waste water from residents and industry, natural rainfall will also drop into the water source. Some of these flows will infiltrate into the ground, some will go into the lake as overland flow. In addition, storm water can affect this process by leading the water away from the catchment area. The water source has to be protected from excessive contamination from point and non-point sources of pollution through better catchment management practices. The water treatment processes have to be well managed against potential failures; and the water distribution network has to be secured against potential intrusion of contaminants.

In general the main data that have been used to achieve the objective of effective safety planning of the water supply system include:

- Physical data. Drinking water supply has to monitor physical attributes in water quality for the whole process, including temperature, conductivity, total suspended solids (TSS), transparency, total dissolved solids (TDS), taste of water, *etc.*

- Chemical data. Standard chemical water quality indicators include pH, biochemical oxygen demand (BOD), total hardness (TH), heavy metals, nitrate, orthophosphates, pesticides, surfactants, *etc.*
- Biological data. Biological indicators contain different bacteria figures, such as ephemeroptera, plecoptera, mollusca, trichoptera, escherichia coli (E.coli), coliform bacteria, *etc.*
- Environmental data. Environment data includes the whole process of water supply. It includes geographic information system (GIS), weather, hydrology, soil, ecology, *etc.*

Data used in this study were collected from two cities in Norway, Oslo and Bergen. The locations of these two cities and their water sources are shown in Figure 1. In Oslo, the Maridalsvannet lake serves as the main water source to supply the majority population living inside the city. Raw water from the lake flows to Oset Water Treatment plant (WTP). This plant has a capacity of 3.9 x 105 m^3/day and provides drinking water to about 90% of the citizens of Oslo. In Bergen, the main water source is Svartediket lake which lies in the east of urban area. It is an artificial lake in Hordaland. These two lakes in Oslo and Bergen are both shallow. This means they are prone to contamination from anthropogenic activities within its water catchment area. In this study, weekly raw-water samples are taken from these lakes and analyzed for physical-chemical and fecal indicator organisms.

The current water quality monitoring regime employed by these two water source operators are collecting samples every week. The samples in Oslo and Bergen are taken from the main three inflow points individually, as the red points shown in Figure 1. All samples are transported to an accredited laboratory where they are analyzed for relevant water quality indicators. We checked the data sets from these two lakes and found the water quality indicators are selected differently. Oslo Maridalsvannet lake has taken a complete records, including 11 indicators, as follows. Note that data from Bergen does not include the indicators shown with *.

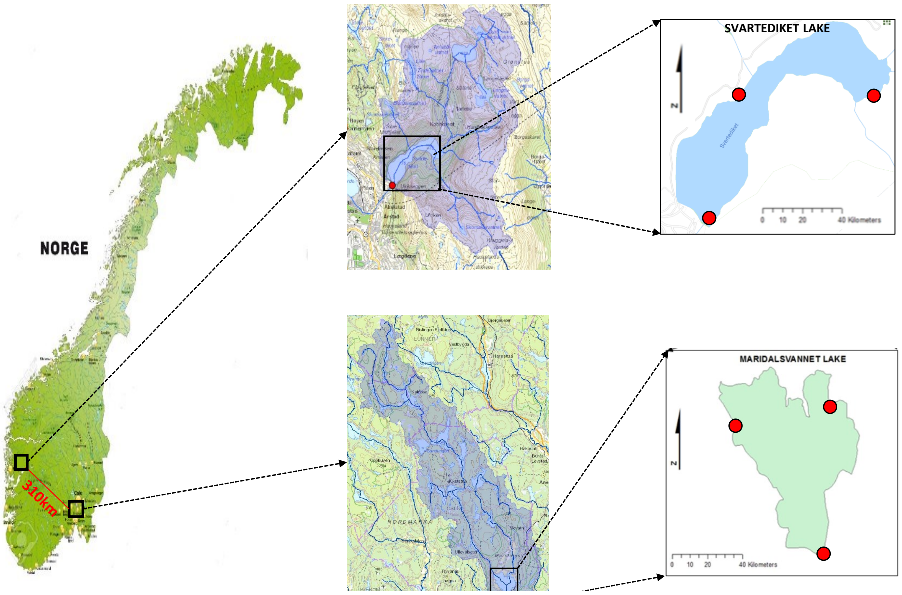


Figure 1: Maridalsvatnet Lake and Svartediket Lake are two main drinking water sources for the city Oslo and Bergen in Norway.

- Physical indicators:
  - Water temperature* (°C)
  - Conductivity (mS/m)
  - Color (mgPt/l)
  - Turbidity (FNU)
- Chemical indicators:
  - pH
  - Alkalinity* (mmol/l)
- Biological indicators:
  - Coliform bacteria (cfu/100 ml)
  - Escherichia coli bacteria (cfu/100 ml)
  - Intestinal enterococci bacteria (cfu/100 ml)

  - Clostridium perfringens bacteria* (cfu/100 ml)

- Other indicator:
  - Time (MM/DD/YYYY).

The Geographic Information System (GIS) data for these two locations are important in some regional analysis as well. Here we give as follows:

- Maridalsvannet: 59.9806°N, 10.7718°E.
- Svartediket: 60.3860°N, 5.3667°E.

3.2. Technical Challenges

In order to predict water quality tendencies and analyze the mechanisms behind these data resources, we are facing several challenges:

- Data Sparsity: the pool of available data is often very large if you consider both in location and time domains. In practice, however, overlap between two conditions (such as the same time, same location) is often very small or none based on two main reasons. First, operators that take the samples do not follow the standard procedure (incomplete indicator collections, and missing data). Second, data standard has been changed over last years (indicators have been added or removed). These make the data set sparse. These are very common situations in water quality data collections.

- Data Synchronization: current sensing technologies can support real-time data collection over most of the physical and chemical indicators for water quality. Biological indicators are directly related with our health. But the tests for different bacteria usually take longer time, from several hours to 1 day. These make the data set difficult to synchronize different indicators.

- Risk Analysis: the final objective of drinking water quality control is to improve health. Some types of bacteria can cause significant disease outbreaks. When they broadcast in the drinking water distribution system, the consequences can be irreversible. Time and accuracy are the two major questions for risk detection and control in water quality.

4. Data-Driven Framework

In this paper, we propose a smart data driven framework to analyze and predict water quality, as shown in Figure 2. In this framework, the whole process can be divided into five parts.

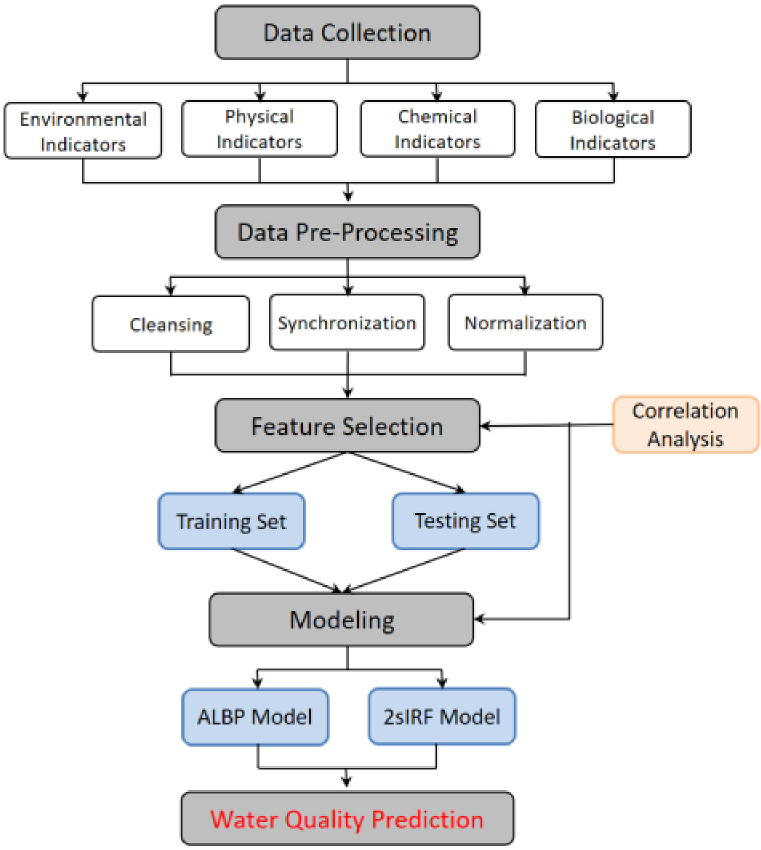


Figure 2: Smart Data Driven Framework for Water Quality Control

325 Data collection involves the collection of data from the sensor network and laboratory test results in the water source monitoring systems. The concrete indicators follow the local regulations and are limited by the sensor types. The data pre-processing usually involves transforming raw data into a computable format. Here we have three steps in pre-processing.

- Cleansing is based on the fact that data collection for water quality indicators is often loosely controlled, resulting in out-of-range values, impossible data combinations, missing values, etc. So, we have to detect and remove those meaningless values using anomaly detection filters.

- Synchronization is needed because different indicators are collected in a hybrid frequency environment. The data needs interpolation and spectrum analysis to be synchronized together.

- Normalization is the result of the quality indicators in water system usually have different units. It is meaningless to analyze them on the same level.

After the data is prepared, we need to find the key factors from different dimensions of water quality indicators by primary correlations analysis, probability distribution and generate training and testing data sets. Correlation Analysis is used to provide some heuristic knowledge for feature selection. In the beginning, we will use direct Pearson correlation distance to find initial relationships. One thing worth to notice is with our model 2sIRF, it can generate the correlations as a side results of water quality indicator predictions. We will compare the two different results.

The selected features will be separated as training and testing data sets. Within each set, the data is divided into input and output. Input data include physical, chemical and environmental indicators. Output data include various biological indicators.

With the training set, we will generate two prediction models from input and output. Considering the multiple dimensions of nonlinear properties in the water quality indicators, we build adaptive learning rate BP neural network (ALBP) and 2 step isolation and random forest (2sIRF) models.

The final step is to take these models and predict water quality indicator tendencies in the future and test them in the testing set for prediction accuracy. Especially for sensitive biological bacteria, the models can provide early warning to the following procedures. Also, in practice, the models need to be evolved with evaluation.

5. Smart Data Driven Water Quality Prediction Model

5.1. Feature Correlation Analysis

Prior to each data-driven water quality modeling exercise, features of the data set are examined through descriptive analysis in order to find some possible heuristic knowledge. This often comprises a summarized quantitative description of the raw data set. Conventional descriptive statistical measures include data distributions, measures of central tendencies, and measures of dispersion. In addition, the relationships between features can be determined through correlation analysis.

In this study, we apply Pearson's correlation coefficient to explore preliminary relationships among the water quality indicators. This means to provide pilot insight into effects that each of the physical and chemical indicators of the raw water has on the fecal indicator organisms. Pearson's correlation is a measure of the magnitude of a linear relationship between a paired data set. The coefficient can be computed as in Equation 1:

$$r = \frac{\sum XY - \frac{\sum X \sum Y}{N}}{\sqrt{(\sum X^2 - \frac{(\sum X)^2}{N})(\sum Y^2 - \frac{(\sum Y)^2}{N})}} \tag{1}$$

The coefficient r has the following constraint:

$$-1 \leq r \leq 1$$

In this equation, r is the result of coefficient. X and Y are two different types of indicators. N is the number of recordings. Generally, positive values of r indicate the presence of positive linear correlation between the paired data set whereas negative values denote negative linear correlation. Thus, how close the coefficient is to 1 or -1 gives an indication of the strength of the linear relationship between a pair of parameters. When no linear correlation exists between a pair of data variables, r is zero.

5.2. Pre-Processing

Data Pre-Processing in this framework is divided into three parts, as cleansing, synchronization and normalization.

For Data cleansing in this framework, we first use a selection equation to find suspicious data according to the average values of water quality indicators, and then remove the obvious wrong data or correct them by the domain experts (verified at least from the operator and water quality expert).

Assume $x_i(n)$ to be the recordings of water quality indicator i at serial number n. The suspicious data set is selected according to the Equation 2. In this equation, L_i and H_i are low and high thresholds for the difference between the recording $x_i(n)$ and its average value. These thresholds are set by the domain experts.

$$S = \{x_i(n) \quad | \quad L_i < (x_i(n) - \overline{x_i(n)}) < H_i\} \tag{2}$$

Data synchronization we directly use the time information in the data set. We synchronize the data according to the time unit. For example, if the data was collected weekly, then we synchronize the recordings of different indicators in the same week (Monday to Sunday) as the same time. To compare the work, we remove the recordings which are not complete (lack of indicator information).

As we listed in Section 3, different water quality indicators have diversified units. This is because two main reasons. First the indicators represent different concrete practical meanings. Second, even for the same type of indicators in different countries or regions, they can have different units according to the local standards.

Hence normalization is an inevitable step to process the data. In this work, we use Equation 3 to normalize the raw data. In this equation, N_i is the total number of indicator x_i.

$$x_i^{'}(n) = \frac{x_i(n) - \overline{x_i(n)}}{\sqrt{\frac{\sum_{n=0}^{N_i}(x_i(n) - \overline{x_i(n)})^2}{N_i}}} \tag{3}$$

And, we have,

$$\overline{x_i(n)} = \frac{\sum_{n=0}^{N_i} x_i(n)}{N_i}$$

5.3. Adaptive learning Rate BP Neural Network

The concept of Artificial neural networks (ANN) comes from biological neural networks in human's brain. It is one of the most important fundamental framework for modern machine learning algorithms. An artificial neural network is composed of a group of connected artificial neurons. Generally they can be organized by different functional layers (input layer, hidden layer, output layer). In this study, we choose ANN as the basic prediction method for water quality based on several inherent advantages as follows:

- ANN does not ask for prior knowledge. As we explained in Section 3, the physical and chemical interactions between various water quality indicators have not been well understood, so we can not provide a rigorous theoretical model to predict biological indicators.
- ANN has self learning ability. By error back propagation structure and relevant algorithms, ANN can learn in the training process. In practice, we are using ANN as a supervised regression for our prediction problem.
- ANN can support parallel output computation. The water quality predictions are not facing unique indicator. ANN method are not sensitive to the output numbers. So we can adapt ANN models according to different national or regional standards.
- ANN is suitable for highly nonlinear problems, especially on regression. For biological indicator predictions, the problem is typically multi-output nonlinear. ANN has been designed for using massive structured neurons to build nonlinear relationships between inputs and outputs.

In this study, by the virtue of above reasons, we chose ANN as the basic framework to solve our problem. Furthermore, Back Propagation (BP) is a broadly used model in ANN practice. Its main conveniences rely on solid theoretical foundation, versatility and simple learning rules assumption. But classical BP neural network with unique learning rate can slow down the training and easily fall into local minimum value. To avoid these problems, we design an adaptive learning rate BP neural network, as ALBP. It is described as in Algorithm 1.

In this algorithm, we separate our data set to training and testing in order to evaluate the prediction results. Normalization handles the water quality indicators that are collected with different units. The normalization is followed with the Equation 3. In this equation, x_i represents different indicator. n is the sample serial number.

Several questions are important in this ALBP model, including network structure, activation function, prediction function, learning principle and learning rate update function.

There are two main factors to consider for the structure of this model, size of the data set and training efficiency. The structure in our ALBP model follows the classical principle of ANN, which contains input, output and hidden layers. Usually, for the water quality predictions in a specific urban area, according to the local standard and regulation, input and output are fixed. This means the physical and chemical indicators are stable as the system input, the same with the biological indicators as the output. In practice, it is also possible that standards were modified in the history. In that case, the data should be adjusted in the pre-processing stage.

As for the hidden layers, they should be dynamic as a result of the size of the data increases. In principle, more neurons and more layers can improve the prediction accuracy. But it also has the saturation effect. we design this structure which aims to unified with most of the water quality prediction problems as well as taking into account for different local urban requirements.

Activation function is added in ANN in order to better fitting nonlinear functions. Popular activation function include sigmoid, hyperbolic tangent, ReLu, Maxout, *etc.* We use hyperbolic tangent in accordance with the zero average output and less *dead* neurons in the training process. The equation for activation function is in Equation 4. In this equation, y is the input of the perception neuron, and $g(y)$ is the output.

```
Data: Raw Data Collection from Water Sources
Result: Biological indicator predictions
Step 1: Pre-Processing.
1. Initialization;
2. Normalization;
while Each city do
    Step 2: Training Model.
    3. Select the training set;
    4. Separate input and output data set;
    while !$Training\_Stop\_Conditions$ do
        while Each training sample inputs do
            5. Take the inputs as the input layer of ALBP;
            6. Calculate the output of the neurons of each layer;
            7. Take the outputs of the corresponding inputs as expected
              outputs;
            8. Calculate the error between expected outputs and actual
              outputs calculated;
            9. Back propagate the error to the hidden layer;
            10. Change the weights of neurons according to the error
              reduction principle;
            11. Record error;
        end
        12. Update learning rate parameter;
        13. Update $Stop\_Conditions$;
    end
    14. Conclude the trained model;
    15. Select the testing set;
    Step 3: Testing Model
    while Each testing sample inputs do
        16. Calculate the expected output with the trained ALBP
          model;
        17. Calculate the error between expected outputs and actual
          outputs;
    end
    18. Evaluate the model result;
end
```

Algorithm 1: ALBP algorithm for water quality predictions

$$g(y) = tanh(y) = \frac{(e^y - e^{-y})}{(e^y + e^{-y})} \tag{4}$$

We chose normalized exponential function for output layer in this ALBP model. This function is also named as *softmax* function. Softmax is to compress a K-dimensional vector $\mathbf{z}$ of arbitrary real values to a K-dimensional vector $\sigma(\mathbf{z})$ of real values, where each entry is in the range $(-1, 1)$, and all the entries add up to 1. This function is described in Equation 5. In this equation, $\mathbf{z}$ is the vector which include all the output of the last hidden layer. K is the number of the neurons in the last hidden layer.

$$\sigma(\mathbf{z}) = \frac{e^{z_j}}{\sum_{k=1}^{K} e^{z_k}} \qquad j = 1, ..., K. \tag{5}$$

As for the learning strategy, we need to find a cost function to guide the learning process. In this study, in order to improve the training efficiency, we select the cross entropy as the cost function. It is given by Equation 6

$$C = -\frac{1}{n} \sum_{x} [y \ln a + (1 - y) \ln(1 - a)] \tag{6}$$

where C is the total cost. n is the number of samples. x is the training sample. y is the expected output value, a is the prediction output value.

Learning rate in the classical BP model is fixed, but this makes the training easily fall into local optimal values. But for concave function regressions are very common in water quality indicator prediction problems. In this study, we design an adaptive learning rate in order to not only suits for concave functions, but also improve training efficiency. The update function of the learning rate is following Equation 7. In this equation, we use secondary partial derivative of the cost C to expected output. ω and b is the parameters to linearize the derivative to learning rate.

$$Lr(n) = \omega \frac{\partial^2 C_{n-1}}{\partial^2 a_{n-1}} + b \tag{7}$$

5.4. 2 Step Isolation and Random Forest

Random forest (RF) is a well-known ensemble learning method proposed by Tin Kam Ho in 1995 [43]. RF is based on decision tree theory. It is broadly used in both classification and regression. However, the predictions for water quality indicator and values do not take the equal priorities according to their suggesting risks. In this study, we add a step in the beginning to select the important indicators and high risk values. High risk values have much more influences on the drinking water quality for end users. According to our experience, they often appear as anomaly points in the data distribution. We use an isolation forest (IF) to quickly locate them and furthermore predict their values using random forest regression. Isolation Forest algorithm was proposed by Liu *et al* in [44]. It is designed for anomaly detection by finding the shortest depth classifications in most of created decision trees. We chose IF as the classification method considering two main reasons. First, it does not require prior knowledge for classification. Second, it is not sensitive to the number of dimensions of the input and output data. In general, we call this method 2 step isolation and random forest, as 2sIRF. In this method, we generate multiple prediction models simultaneously and conclude the results in order to improve the overall accuracy. Through our applications, we found several distinct advantages of this method.

- The introduction of isolation forest can separate different risk levels. Traditional random forest algorithm regression usually take mean squared error (MSE) as the evaluation function. This has an assumption of all the data are equally important. But in water quality predictions, biological indicators are more critical in the peak values. So in this step, we classify these data first.

- In regular RF regression, The usage of randomness and multiple decision trees can the avoid over-fitting. Over-fitting often appears with many algorithm in the training process. It usually uses over strict assumption to get uniform convergence. This can generate large error in prediction. Compare with regular decision tree, RF has greatly reduced over-fitting.

- Regular RF is designed for high dimensional data set, in addition, RF does not require directly feature selection process. In regular RF algorithm, the training set will be randomly chosen for features. The dimensions of each selection can also be customized for preliminary knowledge. As a result, RF can provide is importance order for variables for the prediction output. So, it can integrate correlation analysis and feature selection process.

- This method 2sIRF is easy for parallel computing, because both isolation forest and random forest are creating various decision trees. These multiple prediction models are running independently. This property is helping to fasten the training process. It is useful in big data set.

- 2sIRF is straightforward to implement.

Considering the above reasons and the requirements of our problem, we design a new 2sIRF to predict water quality for biological indicators. First to classify high risk and low risk data from the training set using an isolation forest algorithm. And intake them as the input in the second step, use the random forest algorithm to train them for independent regressions as different groups. The model will be evaluated by the testing data set. The work flow of this method is depicted in Algorithm 2.

In this algorithm, we adopt easily-measured physical and chemical indicators to predict time-consuming biological indicators in water quality management. Usually, both input indicator and output indicator are not unique. This requires a multiple-input and multiple-output (MIMO) algorithm. Both Isolation Forest and Random Forest naturally satisfy this requirement. Besides, in practice, for water quality prediction, we consider the risks in two levels, high risks and low risks. According to the different water quality standards, usually high risk situations are happening when specific biological indicators or their combinations get high values. The high risk situations are more difficult to predict, both in

```
Data: Raw Data Collection from Water Sources
Result: Biological indicator predictions
1. Initialization;
while Each city do
    2. Separate data into training and testing sets;
    Step 1: Training Model.
    while training set do
        3. Separate data into training_IF_training and
          training_IF_testing sets;
        while training_IF_training do
            4. Train the isolation forest model for high and low risk
              samples;
            5. Generate the model;
        end
        while training_IF_testing do
            6. Evaluate the classification results;
        end
        7. Classify all the samples in the training set;
        while samples in $group_i$ do
            8. Train the random forest regression model for $group_i$
              samples;
            9. Generate $group_i$ regression model;
        end
    end
    Step 2: Testing Model.
    while testing set do
        10. Calculate the error between expected outputs and actual
          outputs;
        11. Evaluate the regression results;
    end
    12. Evaluate feature importance;
    13. Evaluate the model results.
end
```

Algorithm 2: 2sIRF algorithm for water quality predictions

time and accuracy. Therefore we use two steps in this method for water quality prediction.

Both isolation forest and random forest algorithms are based on the decision tree. There are generally several ways to build the tree, such as Iterative Dichotomiser 3 (ID3), C4.5, Classification And Regression Tree (CART). They use different ways to degrade information uncertainty in the tree. For calculation simplicity, we choose CART to build our decision trees. The evaluation of information uncertainty in CART is Gini impurity, Gini impurity is a measure of how often a randomly chosen element from the set would be incorrectly labeled if it was randomly labeled according to the distribution of labels in the subset. It is shown in Equation 8. Suppose our water quality data (both classification and regression) has K classes, $i = 1, 2, ..., K$. Let p_i be the fraction of items labeled with class i in the set. In this equation $I_G(p)$ is the Gini impurity of a specific layer of the decision tree. N_{C_i} is the number of items has been labeled in the class i. N is the number of the items in the whole data set.

$$I_G(p) = 1 - \sum_{i=1}^{K} p_i^2 = 1 - \sum_{i=1}^{K} \left(\frac{N_{C_i}}{N} \right)^2 \tag{8}$$

CART provides the basic decision tree. Next step is to build the forests. The reason for the difficulty in predicting high risk data is they are *few and different*. In the first step of 2sIRF, we classify the training data set as high and low risk sets. Based on their features, we can define them as anomaly points in the data set. We use isolation forest algorithm to find them. This algorithm uses the length of decision tree to detect irregular points. The anomaly score of a data is defined as in Equation 9. In this equation, x is the item, N is the number of the items in the whole data set. $c(N)$ is the average path length of unsuccessful search. $E(h(x))$ is the average path length from a collection of isolation trees.

$$s(x, N) = 2^{-\frac{E(h(x))}{c(N)}} \tag{9}$$

For the items in high risk and low risk groups, we use two regression models to predict them. We are taking this way because for water quality prediction, high risk situations are much more influential for people's health. The accuracy for them have higher priority over low risk situations. If we use a single prediction function for regression accuracy, it will average the effect in high risk situations because there are much fewer data for them. In random forest regression, for high risk group, we use Equation 10 to predict the indicator values. In this equation, $f(\hat{x})$ is the prediction of unknown item $\hat{x}$. $f_b(\hat{x})$ is the prediction of tree b. α_i is an adaptive parameter in biological indicator i. B is the number of trees in the random forest.

$$f(\hat{x}) = max(f_b(\hat{x})) - \alpha_i * \frac{max(f_b(\hat{x})) - min(f_b(\hat{x}))}{B} \tag{10}$$
$$b = 1, 2, ..., B.$$

For low risk samples, we use Equation 11.

$$f(\hat{x}) = \frac{1}{B} \sum_{b=1}^{B} f_b(\hat{x}) \tag{11}$$

6. Experiment and Results

6.1. Experiment Design

Typical municipal drinking water supply systems carry out sampling and laboratory analysis to ascertain the concentrations of physical, chemical and biological indicators of raw water before treatment. But we propose to bring it one step ahead, to the water source management stage. The frequency of the analysis rely on the source types, importance of indicator, equipment constraints, as well as the water quality standards of the local regulatory authorities.

The raw water quality data used in this study was obtained through a Norwegian national water research project to improve water supply. It is collected from two major cities in Norway, Oslo and Bergen. These data sets are based on the lab test results from the water sources of the two cities. The locations of the water sources are shown in Figure 1, in which the red points are sampling locations. Based on the constraints on human resources, the samples are taken in weekly. For each city, one location will be taken randomly from the sampling locations.

The water source from Oslo is Maridalsvannet, which is the biggest lake in the municipality. In Bergen, the water is coming from a lake from the east side of the city, named Svartediket. The time periods for Oslo is from 2009 to 2015, for Bergen is from 2007 to 2015. They are recorded in weekly bases from their water sources. The data set from Oslo contains a typical complete group of indicators as turbidity, conductivity, pH, temperature, color, alkalinity, coliform bacteria, E.coli, intestinal enterococci and clostridium perfringens. However for the data set from Bergen, the indicators are less, temperature, alkalinity and clostridium perfringens are not recorded. So, in this study, we treat them independently. The detailed average values and number of recordings for these two data sets are shown in Table 1. We use 0.0 for the indicators lack of valid recordings. The further normalization, and data concrete biological indicator predictions will be based on these values.

The experiments in this section are conducted on the platforms of Tensorflow 1.13 and Python 3.6. The hardware environment contains a system using Intel(R) Core(TM) i7-6600U CPU 2.80 GHz with 16.0 GB of RAM, 64-bit Operating System.

Indicator	Oslo AVG	Bergen AVG
No. of Recordings	356	467
Temperature	7.11	0.0
Conductivity	2.60	3.52
Color	26.76	21.70
Turbidity	0.45	0.55
pH	6.51	6.20
Alkalinity	0.09	0.0
Coliform	7.78	83.17
Ecoli	0.29	1.81
Intestinal	0.23	0.42
ClPerf	0.42	0.0

Table 1: Statistical Performance of the Raw Data Sets

6.2. Correlations among water quality indicators

Results of the Pearson's correlation analysis between the various water quality indicators for the two cities are shown in Figure 3 and 4. The red stars at upper right corner in every cell show the level of linear correlation between the indicators. The number in the cell represents Pearson coefficient. Bigger it is, stronger is the correlation. These provide different means of extracting information from the correlation matrix. From these two matrix, there are several numbers worth noting.

In Oslo, correlations between conductivity and color (-0.72), as well as between color and alkalinity (0.67) are clearly standing out. In Bergen, the number between conductivity and color (-0.61) is also higher than other correlations. The Oslo matrix shows the Pearson's correlation matrix shows that the water pH is positively correlated with all the observed fecal indicator organisms

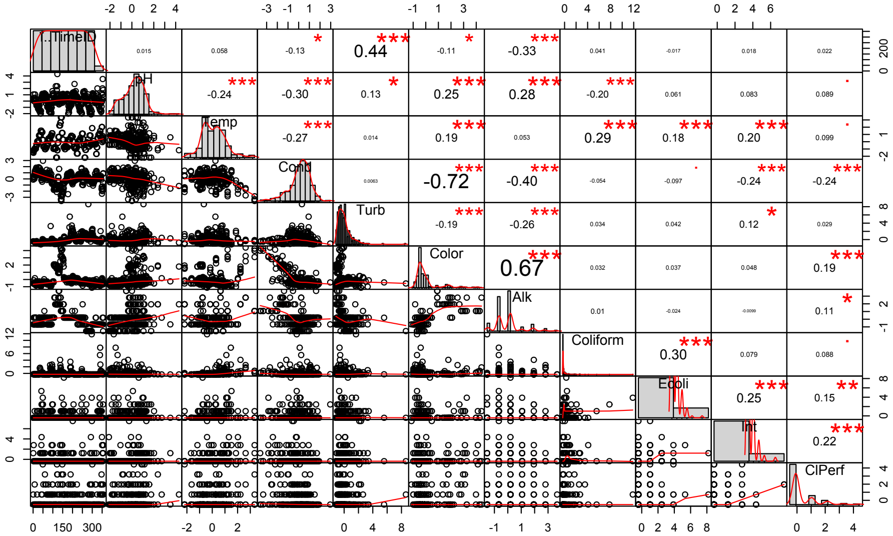


Figure 3: Correlation Analysis of Water Quality Indicators in Oslo

except coliform bacteria. The numbers of PH and the organisms are 0.061, 0.083, and 0.089 respectively for E.coli, intestinal enterococci, and clostridium perfringens respectively. Although the water pH is negatively correlated with coliform bacteria, the strength of the relationship is much higher than in the other organisms.

However, these results can not be seen from Bergen's matrix (Figure 4), in which pH has stronger correlation with intestinal enterococci (-0.15). In Bergen's matrix, color is more influential to the biological indicators, especially for Ecoli (0.36), and intestinal enterococci (0.36).

From here we can see that: (1) The correlation analysis between water quality indicators can provide us some preliminary knowledge over their relationship. These can be used as heuristic knowledge for future water quality indicator predictions. (2) There is no physical or chemical indicators can directly related to predict biological indicators. The reason is the correlation analysis is the linear relationship between two indicators. It does not consider the multiple indicators effect or non linear relationship. So, the correlation analysis can provide

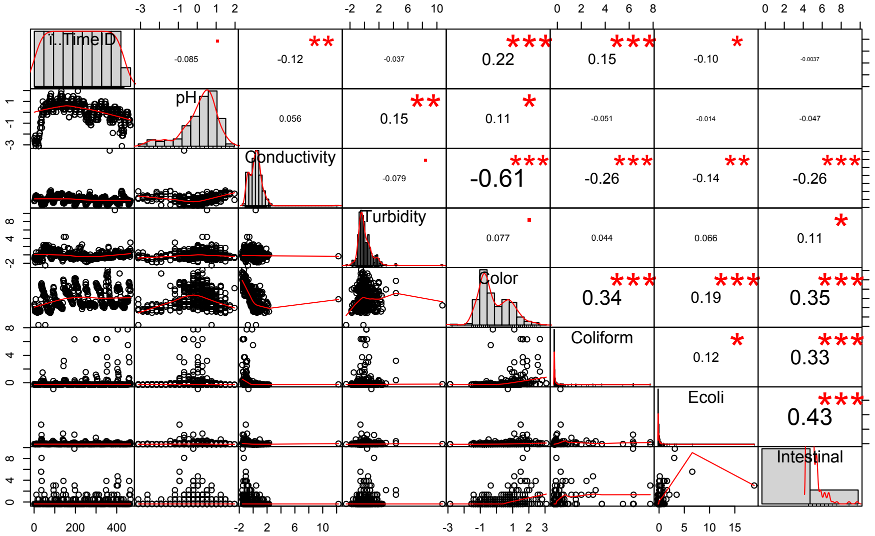


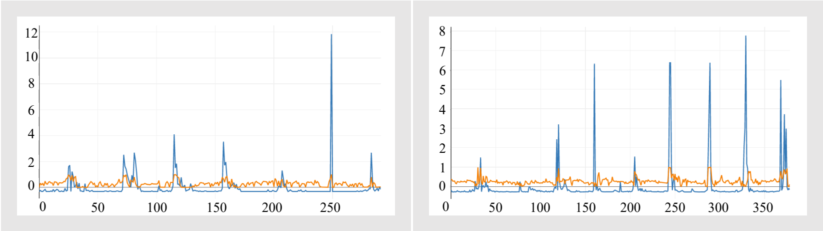
Figure 4: Correlation Analysis of Water Quality Indicators in Bergen

some basic knowledge, but can not provide enough information to predict water quality indicators.

6.3. ALBP model performance

ANN is designed to fit non-linear object properties. In practice, the more neurons we use, the fitting for non linearity is more accurate. But too complex network can lead to over fitting and prolong the training time. We have experiment for the structure in 3, 4, and 5 layers. In the hidden layers, we tried for 10, 50, 100, 500, 1000, 3000 neurons.

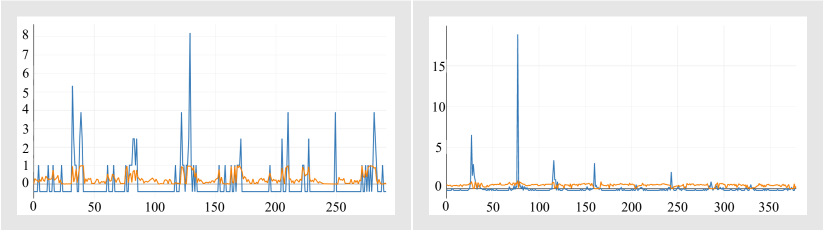
Considering the training efficiency and prediction accuracy, we select the structure with 5 layers, including input, output and 3 hidden layers. Each hidden layer has 1000 neurons. We use 80% of the data (285 recordings in Oslo, 374 recordings in Bergen) as training data and 20% for testing (71 recordings in Oslo, 93 recordings in Bergen). The training iteration takes 1500 times. The performance of our ALBP model is given in the following Figures 5 for Coliform, 6 for Ecoli, and 7 for intestinal enterococci (Int). In these figures, (a) for Oslo and (b) for Bergen. One of the important advantages for ALBP is that it can give the predictions for all the biological indicators at the same time. Actually when we run the experiment, the model can also give the results for clostridium perfringens. For easy comparisons, here we only show the results for the three indicators.



(a) Oslo Coliform Prediction

(b) Bergen Coliform Prediction

Figure 5: Coliform indicator predictions with ALBP in Oslo and Bergen

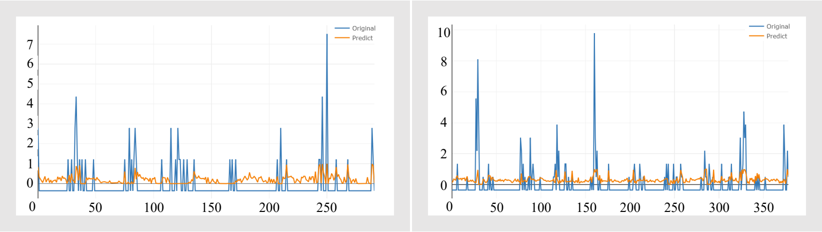


(a) Oslo Ecoli Prediction

(b) Bergen Ecoli Prediction

Figure 6: Ecoli indicator predictions with ALBP in Oslo and Bergen

In these three Figures, blue, orange and green color lines represent the original output of indicator Coliform, Ecoli and intestinal enterococci. Red, purple and brown color lines represent the prediction output of those indicators. From the results we can see that (1) Original out put are with higher degree of oscillation. All the predictions are smoother. (2) The accuracy of prediction is relatively low. Bergen's results are even lower, which could be they have two input indicators missing. (3) The tendencies of the biological indicators can hardly seen from the predictions. (4) Based on the inherent deficit of Neural Network, we can not give the explanations of different indicators on their



(a) Oslo Int Prediction

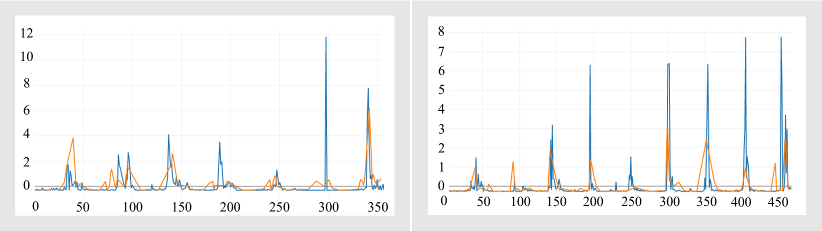
(b) Bergen Int Prediction

Figure 7: Int indicator predictions with ALBP in Oslo and Bergen

influences for the results.

6.4. 2sIRF model performance

We design this 2sIRF model to predict the water quality biological indicators considering the object data sets are non linear, with different risk levels and also the data distribution. We take the heuristic knowledge from the correlation analysis in Section 6.2, and choose temperature for Oslo and color for Bergen as the chief feature for prediction. In this method, we use 80% of the data as training data and 20% for testing. In Oslo, we have 284 recordings for training and 72 for testing. In Bergen, there are 373 recordings for training and 94 for testing.



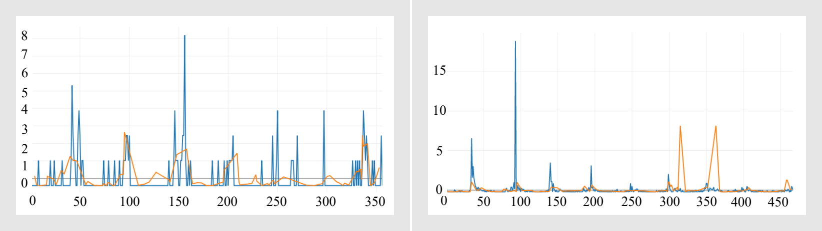
(a) Oslo Coliform Prediction

(b) Bergen Coliform Prediction

Figure 8: Coliform indicator predictions with 2sIRF in Oslo and Bergen

We present our prediction results for Oslo and Bergen in Figure 8, 9, and 10. We give also the results for the three biological indicators, Coliform, Ecoli and Int independently. In these two pictures, blue lines represent original output, and orange lines represent the prediction results.

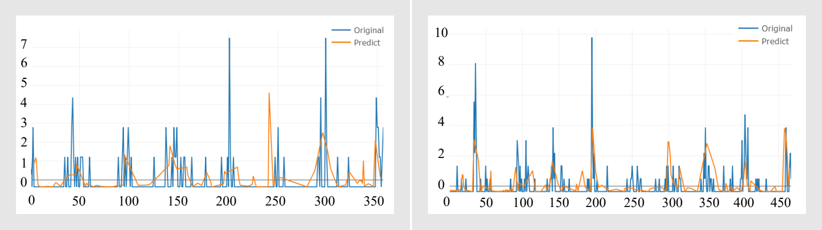
From these results we can see that: (1) 2sIRF can better detect the high risk values. (2) This method can better depict the data oscillation and tendencies. (3) The accuracy for this method is better than ALBP in both cities. In addition, it is worth to note that this method contribute on the high risk values. This is more important to provide efficient and accurate decision support for industrial operators.



(a) Oslo Ecoli Prediction

(b) Bergen Ecoli Prediction

Figure 9: Ecoli indicator predictions with 2sIRF in Oslo and Bergen



(a) Oslo Int Prediction

(b) Bergen Int Prediction

Figure 10: Int indicator predictions with 2sIRF in Oslo and Bergen

7. Discussion and Insight

All the water quality indicators have been considered in previous works to be independent because currently there is no systematic theory to prove the relationships between them. Therefore, it is difficult to predict one indicator (or several ones) from others with precise models. This work uses the historical data resources to find the possible hidden correlations behind them by analysis. In this section, we evaluate the results from the two case studies, including prediction accuracy and time for training. Besides, we compare the importance order of indicators in prediction, using Pearson's correlation coefficient and our 2sIRF models. From here we list some insights we found for the domain of water quality control.

7.1. Model Performance Evaluations

To evaluate the two model performances, we use prediction accuracy and training time as references. For prediction, we use root mean square error (RMSE) in Equation 12 and Mean Absolute Error (MAE) in Equation 13 to evaluate prediction results. This equation gives a general error measurement for all the results. $E_r(x)$ is the RMSE value. $E_m(x)$ is the MAE value. N is the number for predicted indicators. N_i is the number of data recordings of indicator i in the test set. x_{ij} is the real output value and $\hat{x_{ij}}$ is the prediction value.

$$E_r(x) = \sqrt{\frac{\sum_{j=1}^{N_i}(\hat{x_{ij}} - x_{ij})^2}{N_i}} \tag{12}$$

$$E_m(x) = \frac{\sum_{j=1}^{N_i}|\hat{x_{ij}} - x_{ij}|}{N_i} \tag{13}$$

Here, in order for further comparisons, we use the classical ANN and Random Forest algorithms as references. The initial parameters used for them are the same with our method. We compare the accuracy with RMSE and MAE. The results is shown in Figure 11. This result tell us that (1) In general, improved algorithms are outperforming their classical algorithms (ALBP to ANN, 2sIRF to RF). (2) From the city perspective, Bergen has lower performance results than Oslo. This can be caused by Bergen contains fewer indicators. (3) For MAE, 2sIRF has a better performance in most of the cases.

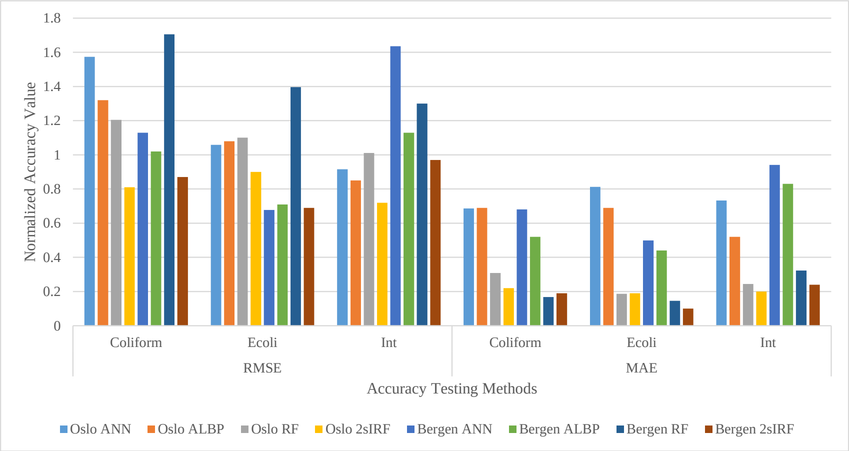


Figure 11: Accuracy Evaluations for ALBP & 2sIRF

Another important measurement is training time, this is meaningful for real-time prediction with further data updates. The training times are shown in Figure 12. This result proves that the use of data driven method can greatly improve the risk warning efficiency in urban water supply systems. Compared to the current culture test for biological indicators, these data models have decreased the prediction time to minute level. Among the values from different methods, we found 2sIRF is better in both cities.

7.2. Insights for water quality prediction

7.2.1. Correlation Analysis

The complex interactions between these water quality indicators are not understood properly. In this study, we are using the historical data as a resource to learn the relationships between them. Besides the results of Pearson's correlation coefficient in Section 5.1, as a byproduct from 2sIRF model, we also get analysis results in biological indicator predictions.

We compare the results in Figure 13. This figure gives the impact order for each biological indicator to all the other physical and chemical indicators, as in Oslo and Bergen. We have several interesting findings as follows:

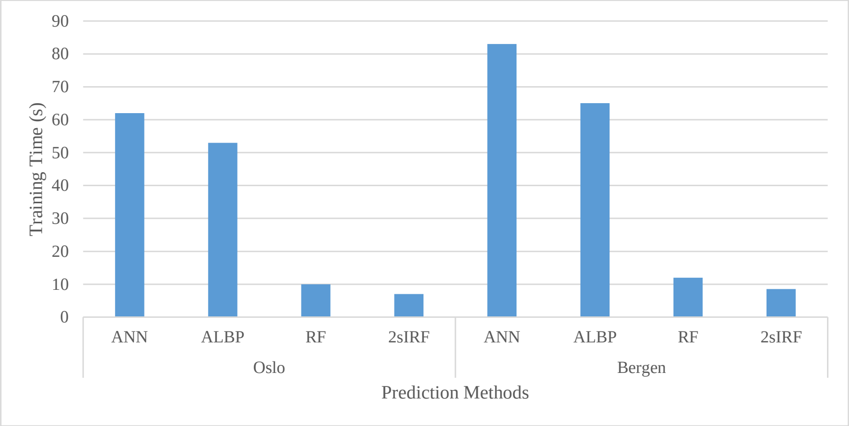


Figure 12: Training Time Evaluations for ALBP & 2sIRF

- Between the two cities, it does not exist a universal principle of reliance for specific biological indicator. in general, for the biological indicators in Bergen, conductivity, color and turbidity are the three leading indicators for the prediction. But in Oslo, we found temperature and pH play more important position (except for intestinal enterococci). This could be based on the temperature in Bergen has not been recorded. Accordingly, we suggest to the manager of Bergen that temperature can be a better choice in data recording. The prediction for intestinal enterococci is more difficult than others.

- Between indicators, by value, we saw that the correlation between color and conductivity are obvious in both models. For this we can consider to remove one recording from the monitoring process. In addition, we can see alkalinity is a weak indicator (not recorded in Bergen, and low importance in Oslo), but this is a potential indicator for tendency prediction. So, it can not be removed from the data recordings, we will explain our findings in the next paper.

- Between the two models, the leading indicators in the prediction are rel-

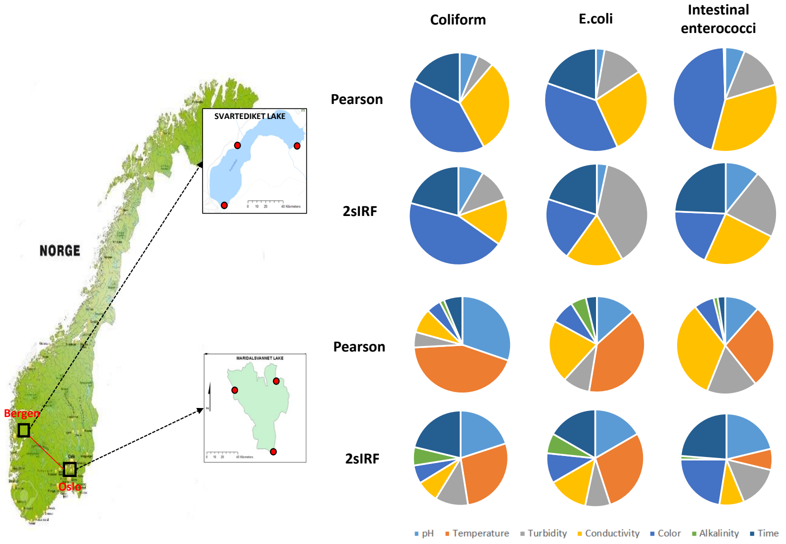


Figure 13: Water indicators importance evaluation

atively similar. But the leading effects are exaggerated by Pearson correlation analysis. However, we can not get the accuracy conclusion from this comparison work. More work need to be done based on more data collections from more locations and longer time.

- Time effect, from the Figure 13 we can see all the indicators have another indicator reliance, which is Time. This is important for water quality control to check whether the indicators have some repeating features. This is more visible in 2sIRF models. We need further analysis and explanations from geographical and ecological characters.

7.2.2. Limitations

Based on the difficulties of quality prediction in urban water supply systems, limitations of our methods as ALBP and 2sIRF can be found in the following aspects:

- The ALBP does not show very high accuracy in prediction, the training time is relatively long. However, with the data size grows bigger, the scalability of this method is higher, and easily adapt to more complex architecture as Deep Learning.

- The 2sIRF gives higher prediction accuracy and efficiency. The scalability in water quality indicator is high. However, the size of the data sets will have exponential impact on the time efficiency. In addition, now all the indicator and recordings are treated as independent parameters, the analysis on the time domain is relatively weak.
- The explanability of both of these methods are relatively low.

7.2.3. Domain evaluation

This study has combined the researchers both from information technology and urban water supply systems. To our best knowledge, this is the first trial to provide a comprehensive solution for the water quality control in urban water source management. As for the results, it has greatly improved the time efficiency for biological indicator predictions, the accuracy is acceptable and can provide sufficient support for decision making in water treatment and water quality risk warnings.

8. Conclusion & Future work

Water, as being one of the most important resources on earth, attracted much attention from governmental, academic and industrial organizations. Water quality prediction is a critical step not only in water supply systems, but also meaningful in modern smart city development. With the evolution of advanced sensing technologies, data collections for water quality indicators have become more and more convenient. However, how to efficiently analyze the data and provide early warnings remains a major challenge.

In this study, we propose to integrate smart data-driven technologies for with urban water supply system for water quality prediction. The main contributions of this research are three aspects.

- Smart Data-Driven framework. This paper proposed a framework using the historical data resources from water source management process, aiming to predict water quality evolution by advanced analyzing and modeling techniques. It provides an efficient way for future decision making support of water quality control and risk management.

- Biological indicator prediction. Biological indicators in the water supply industrial system is always the most difficult to measure and collect, based on that most of the microbial indicators have to go through bacterial culture process, it takes much longer time than regular physical and chemical indicators. This paper proposed two models to solve this problem, including adaptive learning rate BP neural network (ALBP) and 2 step isolation and random forest (2sIRF). ALBP is theoretically simple and easy to implement. 2sIRF considers the risk distribution and shows higher prediction accuracy.

- Real-world-environment-oriented. The methods in this paper are experimented and tested in the practical processes at Oslo and Bergen in Norway. This supports the work reliability.

Our future work is planned in several directions:

- To improve prediction accuracy and efficiency. We will develop more modeling techniques for water quality data analysis.

- To explore water quality properties in time domain. We will analyze and develop circulation features of water quality data analysis.

- To connect related data from other water supply sections. We will collect water quality data from urban water treatment and distribution process.

- To provide decision making support for the whole urban water supply systems.

References

[1] P. C. D. E. ALIANÇA, Água e energia: aproveitando as oportunidades de eficientização de água e energia não exploradas nos sistemas de água municipais, Washington: ALLIANCE.

[2] World Health Organization (WHO), Un news, https://www.un.org/development/desa/en/news/population/2018-revision-of-world-urbanization-prospects.html (Oct. 2019).

[3] United Nations (UN), Ensure access to water and sanitation for all, https://www.un.org/sustainabledevelopment/water-and-sanitation/ (Oct. 2019).

[4] H. Bennion, R. Battarbee, The european union water framework directive: opportunities for palaeolimnology, Journal of Paleolimnology 38 (2) (2007) 285–295.

[5] D. A. Keiser, J. S. Shapiro, Consequences of the clean water act and the demand for water quality, The Quarterly Journal of Economics 134 (1) (2018) 349–396.

[6] World Health Organization (WHO), Guidelines for drinking-water quality, https://www.who.int/water_sanitation_health/publications/drinking-water-quality-guidelines-4-including-1st-addendum/en/ (Oct. 2019).

[7] H. Mohammed, I. A. Hameed, R. Seidu, Adaptive neuro-fuzzy inference system for predicting norovirus in drinking water supply, in: 2017 International Conference on Informatics, Health & Technology (ICIHT), IEEE, 2017, pp. 1–6.

[8] G. E. Host, N. R. Will, R. P. Axler, C. J. Owen, B. H. Munson, Interactive technologies for collecting and visualizing water quality data, URISA-WASHINGTON DC- 12 (3) (2000) 39–46.

[9] J. Gubbi, R. Buyya, S. Marusic, M. Palaniswami, Internet of things (iot): A vision, architectural elements, and future directions, Future generation computer systems 29 (7) (2013) 1645–1660.

[10] L. Andres, K. Boateng, C. Borja-Vega, E. Thomas, A review of in-situ and remote sensing technologies to monitor water and sanitation interventions, Water 10 (6) (2018) 756.

[11] World Health Organization (WHO), Drinking water, `https://www.who.int/en/news-room/fact-sheets/detail/drinking-water` (Oct. 2019).

[12] Norsk Helse- og omsorgsdepartementet, Forskrift om vannforsyning og drikkevann, `https://lovdata.no/dokument/SF/forskrift/2016-12-22-1868` (Oct. 2019).

[13] V. Novotny, Water quality: prevention, identification and management of diffuse pollution, Van Nostrand-Reinhold Publishers, 1994.

[14] Z. Yuan, G. Olsson, R. Cardell-Oliver, K. van Schagen, A. Marchi, A. Deletic, C. Urich, W. Rauch, Y. Liu, G. Jiang, Sweating the assets–the role of instrumentation, control and automation in urban water systems, Water research.

[15] E. M. Dogo, A. F. Salami, N. I. Nwulu, C. O. Aigbavboa, Blockchain and internet of things-based technologies for intelligent water management system, in: Artificial Intelligence in IoT, Springer, 2019, pp. 129–150.

[16] I. Petri, B. Yuce, A. Kwan, Y. Rezgui, An intelligent analytics system for real-time catchment regulation and water management, IEEE Transactions on Industrial Informatics 14 (9) (2017) 3970–3981.

[17] L.-C. Chang, F.-J. Chang, S.-N. Yang, I. Kao, Y.-Y. Ku, C.-L. Kuo, I. Amin, M. Z. bin Mat, et al., Building an intelligent hydroinformatics integration platform for regional flood inundation warning systems (2019).

[18] S. Eggimann, L. Mutzner, O. Wani, M. Y. Schneider, D. Spuhler, M. Moy de Vitry, P. Beutler, M. Maurer, The potential of knowing more: A review of data-driven urban water management, Environmental science & technology 51 (5) (2017) 2538–2553.

[19] G. Kang, J. Z. Gao, G. Xie, Data-driven water quality analysis and prediction: A survey, in: 2017 IEEE Third International Conference on Big Data Computing Service and Applications (BigDataService), IEEE, 2017, pp. 224–232.

[20] H. R. Maier, G. C. Dandy, The use of artificial neural networks for the prediction of water quality parameters, Water resources research 32 (4) (1996) 1013–1022.

[21] H. Orouji, O. Bozorg Haddad, E. Fallah-Mehdipour, M. Mariño, Modeling of water quality parameters using data-driven models, journal of environmental engineering 139 (7) (2013) 947–957.

[22] O. Bozorg-Haddad, S. Soleimani, H. A. Loáiciga, Modeling water-quality parameters using genetic algorithm–least squares support vector regression and genetic programming, Journal of Environmental Engineering 143 (7) (2017) 04017021.

[23] N. Mahmoudi, H. Orouji, E. Fallah-Mehdipour, Integration of shuffled frog leaping algorithm and support vector regression for prediction of water quality parameters, Water resources management 30 (7) (2016) 2195–2211.

[24] F.-J. Chang, Y.-H. Tsai, P.-A. Chen, A. Coynel, G. Vachaud, Modeling water quality in an urban river using hydrological factors–data driven approaches, Journal of environmental management 151 (2015) 87–96.

[25] D. Wu, H. Mohammed, H. Wang, R. Seidu, Smart data analysis for water quality in catchment area monitoring, in: 2018 IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and

Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData), IEEE, 2018, pp. 900–908.

[26] D. Wu, H. Wang, H. Mohammed, R. Seidu, Quality risk analysis for sustainable smart water supply using data perception, IEEE Transactions on Sustainable Computing.

[27] X. Wang, L. T. Yang, X. Xie, J. Jin, M. J. Deen, A cloud-edge computing framework for cyber-physical-social services, IEEE Communications Magazine 55 (11) (2017) 80–85.

[28] X. Wang, L. T. Yang, L. Kuang, X. Liu, Q. Zhang, M. J. Deen, A tensor-based big-data-driven routing recommendation approach for heterogeneous networks, IEEE Network 33 (1) (2019) 64–69.

[29] Q. Wang, H.-N. Dai, H. Wang, A smart MCDM framework to evaluate the impact of air pollution on city sustainability: A case study from china, Sustainability 9 (6) (2017) 911.

[30] T. Le Dinh, W. Hu, P. Sikka, P. Corke, L. Overs, S. Brosnan, Design and deployment of a remote robust sensor network: Experiences from an outdoor water quality monitoring network, in: 32nd IEEE Conference on Local Computer Networks (LCN 2007), IEEE, 2007, pp. 799–806.

[31] B. O'Flynn, R. Martinez-Catala, S. Harte, C. O'Mathuna, J. Cleary, C. Slater, F. Regan, D. Diamond, H. Murphy, Smartcoast: a wireless sensor network for water quality monitoring, in: 32nd IEEE Conference on Local Computer Networks (LCN 2007), Ieee, 2007, pp. 815–816.

[32] S. Yagur-Kroll, E. Schreuder, C. J. Ingham, R. Heideman, R. Rosen, S. Belkin, A miniature porous aluminum oxide-based flow-cell for online water quality monitoring using bacterial sensor cells, Biosensors and Bioelectronics 64 (2015) 625–632.

[33] X. Wang, L. T. Yang, H. Li, M. Lin, J. Han, B. O. Apduhan, Nqa: A nested anti-collision algorithm for rfid systems, ACM Transactions on Embedded Computing Systems (TECS) 18 (4) (2019) 32.

[34] V. Mayer-Schönberger, K. Cukier, Big data: A revolution that will transform how we live, work, and think, Houghton Mifflin Harcourt, 2013.

[35] H.-N. Dai, R. C.-W. Wong, H. Wang, Z. Zheng, A. V. Vasilakos, Big data analytics for large-scale wireless networks: Challenges and opportunities, ACM Computing Surveys (CSUR) 52 (5) (2019) 99.

[36] H.-N. Dai, H. Wang, G. Xu, J. Wan, M. Imran, Big data analytics for manufacturing internet of things: opportunities, challenges and enabling technologies, Enterprise Information Systems (2019) 1–25.

[37] M. Komorowski, D. Marshall, J. Salciccioli, Y. Crutain, Exploratory data analysis in secondary analysis of electronic health records (new york (us) (2016).

[38] H. Abdi, L. J. Williams, Principal component analysis, Wiley interdisciplinary reviews: computational statistics 2 (4) (2010) 433–459.

[39] N. Raychev, Measure of entanglement by singular value decomposition, International Journal of Scientific and Engineering Research 6 (7) (2015) 1350–1355.

[40] D. W. Hosmer Jr, S. Lemeshow, R. X. Sturdivant, Applied logistic regression, Vol. 398, John Wiley & Sons, 2013.

[41] S. Tong, D. Koller, Support vector machine active learning with applications to text classification, Journal of machine learning research 2 (Nov) (2001) 45–66.

[42] R. S. Sutton, A. G. Barto, Reinforcement learning: An introduction, MIT press, 2018.

[43] T. K. Ho, Random decision forests, in: Proceedings of 3rd international conference on document analysis and recognition, Vol. 1, IEEE, 1995, pp. 278–282.

[44] F. T. Liu, K. M. Ting, Z.-H. Zhou, Isolation forest, in: 2008 Eighth IEEE International Conference on Data Mining, IEEE, 2008, pp. 413–422.