Quality Risk Analysis for Sustainable Smart Water Supply Using Data Perception

Di Wu, *Member, IEEE*, Hao Wang, *Member, IEEE*, Hadi Mohammed, and Razak Seidu

Abstract—Constructing *Sustainable Smart Water Supply* systems are facing serious challenges all around the world with the fast expansion of modern cities. Water quality is influencing our life ubiquitously and prioritizing all the urban management. Traditional urban water quality control mostly focused on routine tests of quality indicators, which include physical, chemical, and biological groups. However, the inevitable delay for biological indicators has increased the health risk and leads to accidents such as massive infections in many big cities. In this paper, we first analyze the problem, technical challenges, and research questions. Then, we provide a possible solution by building a risk analysis framework for the urban water supply system. It takes indicator data we collected from industrial processes to perceive water quality changes, and further for risk detection. In order to provide explainable results, we propose an Adaptive Frequency Analysis (Adp-FA) method to resolve the data using indicators' frequency domain information for their inner relationships and individual prediction. We also investigate the scalability properties of this method from indicator, geography, and time domains. For the application, we select industrial quality data sets collected from a Norwegian project in four different urban water supply systems, as Oslo, Bergen, Strømmen, and Ålesund. We employ the proposed method to test spectrogram, prediction accuracy, and time consumption, comparing with classical Artificial Neural Network and Random Forest methods. The results show our method better perform in most of the aspects. It is feasible to support industrial water quality risk early warnings and further decision support.

Index Terms—Sustainable water supply, water quality control, data perception, risk evaluation, frequency analysis, scalability

1 Introduction

During the latest years of 21st century, two important phenomena have been emerging: urbanization and information technologies. The United Nations (UN) Department of Economic and Social Affairs (DESA) reports that for the first time ever, the majority of the world's population lives in cities, and this proportion continues to grow with projections of 68 percent by 2050 [1]. Urban water supply systems are the most critical infrastructure all over the world. A *Smart Water Supply* system that integrates sensors, controllers, cloud computing and data technologies, are essential for the development of sustainable smart cities in the future. It is aiming to provide safe, stable and sufficient water for the increasing requirements in many expanding cities. However, the urban water quality is facing serious challenges from industrial, agriculture and social pollution.

To emphasize the importance of water safety in urban supply is nowadays a truism. In 2015, the United Nations Development Programme published the Sustainable Development Goals (SDGs), including Clean Water and Sanitation as Goal 6 [2]. The dwindling supplies of safe drinking water is a major problem impacting every continent, around 2.1 billion people [3]. The concerns of the modern society regarding this issue are reflected in numerous legislative initiatives in this field, such as the European Union Water Framework Directive [4], United States Clean Water Act [5]. The prevalent water supply process can be divided into 3 sections, including water source management, treatment, and distribution.

Traditional water quality control is taken after water treatment. But the current water sources are mainly groundwater and surface water. They are significantly prone to chemical and microbial contamination. The quality control after the water treatment apparently delays the risk detection and reduces the response time to take preventive measures. In Norway, the new national standard for water quality in the source area is in progress [6], [7].

Water quality refers to physical, chemical, and biological characteristics as indicators. Among the water quality indicators, biological indicators have a more direct impact over people's health. Most of the national standards are made on biological indicator levels. Typical indicators include coliform, escherichia coli (Ecoli), intestinal enterococci (Int), clostridium perfringens (ClPerf), etc. Further treatment actions are made according to the test results [8]. Coliform itself is not usually causing serious illness, but their presence is a signal to indicate other active pathogenic organisms presentation. Some special types of Ecoli are the reason for water poisoning. Int is more dangerous to cause urinary tract infections, bacterial endocarditis, diverticulitis, and meningitis. The tests of biological indicators are primarily based on the bacterial

- *D. Wu is with the Big Data Lab, Department of ICT and Natural Science, Norwegian University of Science and Technology, Lesund 6009, Norway. E-mail: di.wu@ntnu.no.*
- *H. Wang is with the Department of Computer Science, Norwegian University of Science and Technology, Gjøvik 2815, Norway. E-mail: hawa@ntnu.no.*
- *H. Mohammed and R. Seidu are with the Water Lab, Department of Ocean Operations and Civil Engineering, Norwegian University of Science and Technology, Lesund 6009, Norway. E-mail: {hadi.mohammed, rase}@ntnu.no.*

Manuscript received 9 Sept. 2018; revised 3 June 2019; accepted 15 July 2019. Date of publication 5 Aug. 2019; date of current version 8 Sept. 2020. (Corresponding author: Hao Wang.)
Digital Object Identifier no. 10.1109/TSUSC.2019.2929953

culture in the laboratory. This process can take up to 24-48 hours. Compare to the effectual time on the human body, the danger is much higher than other indicators. In Norway, the giardia outbreak in Bergen 2004 affected more than 2500 people including young children due to the bacteria test delay results. Therefore, we have a severe requirement for early risk detection in smart water supply systems.

There have been some trial work for water quality control based on data. In 2018, Hounslow [9] interpreted multiple water quality indicators. In 2015, Yagur-Kroll et al. [10] showed a group of general bacterial sensor cells for water quality monitoring. There is some research work to use data for water quality prediction. Holger et al. [11] designed an Artificial neural network to predict salinity level for an Australian river named Murray. Based on the data collected at Astane station in Sefidrood River, Iran, Orouji and his colleagues designed a series of models as ANFIS, GA and Shuffled FLA to predict water quality chemical indicators (sodium, potassium, magnesium, etc) in [12], [13], [14]. Chang et al. [15] proposed a systematic analysis framework to predict NH_3-H levels for Dahan River in Taiwan, China. However, their work is generally on individual quality indicator and ignored the inner relationship between them.

Today the advanced ubiquitous sensing technologies cut across many areas of modern research, industry and daily life [16]. They offer the ability to detect, transmit and measure more environmental indicators. A sustainable smart water supply system adopts various sensors in order to manage resources and monitor water quality efficiently. In this process, data becomes an important tool to improve our understanding of existing systems. By observing data itself, through the appropriate methods, we can perceive the changes in our water supply system. In practice, we applied 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 improve the water quality control process.

At present, zettabytes of data are collected by these numerous sensors [17], [18]. At the same time, stronger data analysis tools have been developed. Water quality indicators are typical spatiotemporal variables. The analysis can be divided into correlation analysis and numerical prediction analysis. Early works with correlation analysis include Hardoon et al. [19] used Kernel Correlation Analysis method for web page images and associated texts. For multiple variables, Principal component analysis (PCA) is often the first choice. Jolliffe et al. [20] reviewed classical PCA and newly developed methods such as Robust PCA, Adaptive PCA, etc. Luo et al. [21] applied tensor model in correlation analysis for gait recognition. But they did not consider the correlations in the time domain. As for spatiotemporal data analysis, most of the recent work is facing very huge data sets. For example, Gudmundsson et al. [22] surveyed the player's trajectories in team-sports with respect to behavior and prediction. Lecun et al. [23] proposed the pioneer concept for Deep Learning to deal with spatiotemporal data. Liu et al. [24] analyzed 3D human actions with modern LSTM method. Laptev et al. [25] detects anomalies in the industrial platform data. However, their work has to rely on large training sets, which we cannot provide currently in water supply systems. In addition, the explanation with those methods cannot support the requirements for industrial use.

In this paper, we introduce our preliminary experience in Norway. First, we analyze the problem, challenges and research questions. Second, based on water quality data collected from water supply systems, we propose a framework for water quality analysis with data perception. Third, we provide an adaptive frequency analysis method for risk detection and prediction. This method is scalable in multiple domains, including water quality indicators, geography and time. Furthermore, by application, we select industrial quality data sets collected from a national project in 4 different Norwegian city water supply systems, as Oslo, Bergen, Strømmen and Ålesund. We show our preliminary findings of the frequency property relationship between water quality indicators, as well as risk detection, prediction and evaluation analysis. The results are compared also with classical Artifical Neural Network and Random Forest in their prediction accuracy and time consumption. In addition, scalability in time domain is also analyzed.

There are several visible motivations for this research. First, it takes the advantage of the modern data analysis technologies to solve a water quality control problem in future *Sustainable Smart Water Supply* systems, especially in transferring the knowledge across different indicator, geography and time domains. Second, it copes with the practical water source monitoring process, applies the data directly collected from the industrial process. This avoids questions such as laboratory data reliability and industrial applicability. This is also valuable to the current water supply in urban infrastructure systems. Third, it builds the connection between easily accessible physical and chemical indicators with biological indicators that are critical to water quality risk. Fourth, this work provides the support for further reasoning of decision-making process and analysis over the pollution from industrial and residential activities in the corresponding water source areas.

2 Problem Analysis

2.1 System Description

Water source management is to control the origins of drinking water. In order to improve the water quality for the end users, the control in the water source is naturally a critical step. However, this is often neglected by most water supply systems because of geographical inaccessibility, costly tests or unprofessional operators. The Norwegian standard process for water quality control is to take samples from the water source area twice or four times a month from the several inflow points in the area. After, the samples have to be tested in the lab for all of the water quality indicators. In this work, we collected the data from 4 different cities from Norway, generally from 2007 to 2015. Their locations are shown in Fig. 1.

The water source from Oslo is Maridalsvannet, which is the biggest lake in this municipality. The water from the lake will be sent to the Oset Water Treatment plant (WTP) in the north of Oslo. The primary inflows are Skjærsjøelva and Dausjøelva. The lake has an area of 3.83 km^2 and 149 meters in height. The water serves as the main drinking water source locally and covers approximate 90 percent of Oslo's water consumption. Weekly raw-water samples are taken from the lake and analyzed for physical-chemical and fecal indicator organisms.

Fig. 1. Urban water sources in Norway.

The water source of Bergen is coming from Svartediket lake in the east of Bergen. It is an artificial lake in Hordaland. It covers 0.5 km^2 and 75 m in height. Drinking water is collected at a 28 m depth in Svartediket. After treatment, the clean drinking water is stored in a 15,000 m^3 large water pool inside the mountain. It covers the drinking water requirement for over 70 percent population in Bergen.

Strømmen is taking the freshwater from all the river networks around Nitelva. The biggest lake nearby is Øyeren in the Glomma River watershed. It is located in the southeast of Lillestrøm. The water is transferred to the Nedre Romerike Avløpsselskap/Vannverk (NRV) treatment plant. All the water source area takes the surface of more than 121 km^2, with an average height of 101 m. The rivers around are 0.5 m to 71 m high.

Ålesund is a city with 47,000 citizens. It lies on the west coast of Norway. The drinking water for the dwellers mainly comes from Brusdalsvatnet lake. This lake sits on Uksenya in the community of Ålesund and Skodje in Møre og Romsdal province. It takes the inflow from Spjelkavikelva river. The water is pumped from the lake to a warehouse inside Emblemsfjellet. It has an area of 7.52 km^2 and 26 m above sea level. The lake itself has a volume of 300 million m^3.

These four cities have different water source types as lakes or rivers. The general impact factors for water quality are not the same. For example, Maridalsvannet lake is surrounded with some industrial factories and residents, Svartediket lake is known to have more active bacteria, Øyeren area covers a large surface and easily affected by rainstorms, and Brusdalsvatnet lake is rural and better preserved. This brings diverse difficulties for water quality monitoring, risk detection and prediction.

Norway has adopted stringent drinking water quality guidelines in accordance with the European Water Directive Framework [6]. In which, water quality indicators can be divided into 4 groups, including,

a. Physical data. Drinking water has to verify physical attributes in water quality for the whole supply process.
b. Chemical data. Chemical indicators are the traditional representation of water quality. They provide information on what is impacting on the system as well.
c. Biological data. Biological indicators are direct measures of the health of the fauna and flora in the water supply.
d. Environmental data. Environment data can be a leading impact factor for water quality in some places.

2.2 Challenges & Questions

In order to evaluate the risk from water quality change and analyze the mechanism behind the data resources, we are facing several challenges:

a. Data Sparsity: the pool of available data is often very large. In practice, for water quality indicator samples, the overlaps between two conditions (such as the same time, same location) are often very small or none. This is based on two main reasons. First, the operators who take the samples do not follow the standard procedure (incomplete indicator collections, and data loss). Second, data standard has been changed over last years (indicators have been added or removed). These make the data set sparse.
b. Data Synchronization: current sensing technologies can support real-time data collection over most of the physical and chemical indicators for water quality. However, for biological indicators, which are the key factors for health, the tests usually take much longer time, from several hours to several days. This makes the data set difficult to synchronize.
c. Risk Modeling: the final objective of drinking water quality control is to improve health. Some specific biological indicators as bacteria can cause significant disease outbreaks, such as Ecoli. When they broadcast in the drinking water distribution system, the consequences can be irreversible. The relationship between those biological indicators and drinking water risk needs a new model.

From our trial work in the smart water supply system in Norway, we try to provide a solution to improve water services, starting from water source management and control. Here we generate some research questions.

- Risk Detection and Prediction. Based on the data analysis, can we predict the risk?
- Domain Explanation. Based on the data analysis result, can we provide any domain explanation?
- Evaluation. Based on the prediction results, how can we evaluate different methods?

3 Approach Formulation

3.1 Framework

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

All the raw data is collected from the sensor networks and laboratory tests of water source areas. It covers all the relevant water quality indicators. Data pre-processing usually involves transforming raw data into an analytical format. Cleaning, Synchronization, and Normalization. It has to take into account the raw data which are out-of-range, missing, multi-resolution and with different units. Here is worth to note that the clustering and declustering processes are optional. This is

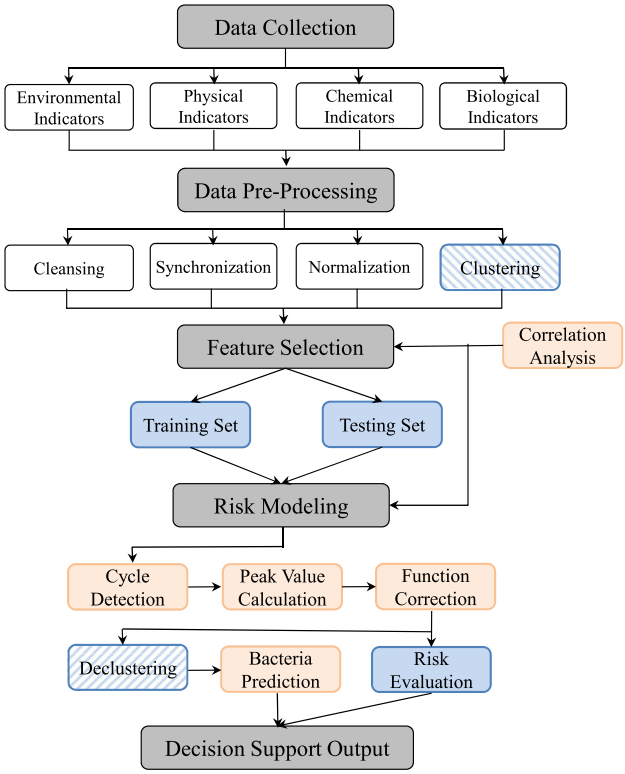

Fig. 2. System framework.

designed to ensure the data can be organized from different perspectives and simple to find hidden patterns. For example, cluster and decluster can consider the time-sensitive features in water quality, as a different time scale, such as days, weeks, months, seasons or years.

After the data is prepared, we need to find the key factors from multiple dimensions of indicators by primary correlations analysis, probability distribution and generate training and testing data sets. The eventual aim of this work is to predict water quality risk. In order to find the risk model, we have investigated with researchers from water quality control. Here the risk evaluation model is further divided into three parts. Cycle detection is to find the hidden cycle for indicator changes in the time domain. Peak value calculation is used to track and evaluate the levels of multiple biological bacteria outbreak. Parameter correction is based on training set adaptation.

Furthermore, we have to decluster the results and predict accurate bacteria indicators, both in tendency and values. These values can map to different risk modes according to practical water source management standards in different countries and regions. Future decision support in water treatment plants can adjust to both prediction and risk mode. Also, in practice, the models need to be evolved with both domain knowledge data set growing.

3.2 Domain Knowledge Analysis

The Norwegian government always gives the highest priority to the drinking water supply for people. We are working as a group for water quality control in the water sources. This team contains the water experts, sampling operators, water treatment plant managers, policy makers, and data researchers. In this project, in order to improve the explanation ability

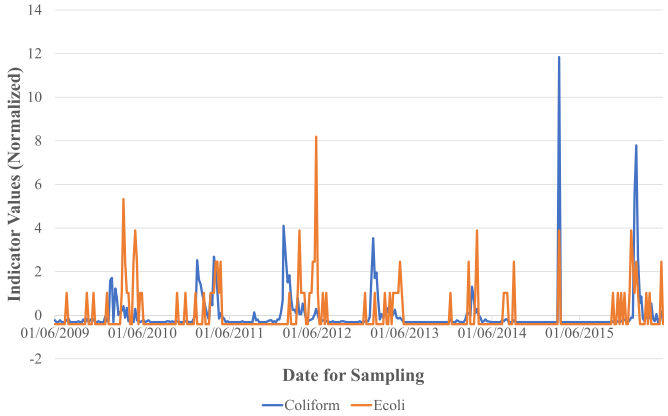

Fig. 3. Original biological indicators from Oslo.

of the results, we try to interpret from the domain knowledge of water quality.

First, we can check an example as the biological indicators from raw data in Oslo, as shown in Fig. 3 to see whether we can predict the data by visualization. As we listed two different bacteria Coliform, Ecoli in this picture, we find it is hardly possible for this task.

Next, we consider the water quality evaluation and risk detection, currently there are several key factors need to be specified:

- Cycle. Cycle detection for water quality is to find the periodic characters for indicator changes in the time domain. Detected cycles in water quality can be beneficial to find predict biological indicators, analyze leading indicators and take preventive measures.
- Peak Values. For water quality biological indicators, the peak values imply infection outbreaks. It is sensitive to quality evaluation. The peak values prediction is critical to water quality classification, development of standards and initialization of early warning mechanism.
- Scalability. Sustainable computing requires computational scalability. In water quality control, we need to deal with generally the scalability of indicators in the time domain.

3.3 Basic Modeling

The original water quality indicators are changing in nonlinear and disordered way. Since we have eliminated the processing with ordinary black box methods, we have to seek for regular data analysis according to their traits. We can not deduce the cycle directly from the visual observation from the data, such as in Fig. 3. However, if we examine the indicators as regular electronic signals, then signal frequency tools can be applied to detect cycles.

We define water quality indicators as:

$$f_i(t), \quad t = t_0, t_1, t_2, \ldots, t_{i_T}. \tag{1}$$

According to the water indicator standards in different countries or regions, i is defined as:

$$i = 1, 2, \ldots, N.$$

For example, in Norway, we have typically 11 collected water indicators. We give the corresponding mapping from

the water quality indicators to the model as follows. But in practice, different cities would select a fraction of them to test and record. Different water quality indicators have diversified units. This is because of two main reasons. First, the indicators represent different practical features. Second, even for the same indicator in different countries or regions, they can have different units according to the local standards.

Formulation Mapping

$f_1(t)$ Temperature (° C).
$f_2(t)$ Conductivity (mS/m).
$f_3(t)$ Turbidity (FNU).
$f_4(t)$ Color (mgPt/L).
$f_5(t)$ pH.
$f_6(t)$ Alkalinity (mmol/L).
$f_7(t)$ Coliform (cfu/100 ml).
$f_8(t)$ Ecoli (cfu/100 ml).
$f_9(t)$ Int (cfu/100 ml).
$f_{10}(t)$ ClPerf (cfu/100 ml).
$f_{11}(t)$ Termotol coliform (cfu/100 ml).

Thus, here we get,

$$f_i(t) = \begin{cases} Physical\ Indicators & 1 \le i \le 4 \\ Chemical\ Indicators & 5 \le i \le 6 \\ Biological\ Indicators & 7 \le i \le 11 \end{cases}. \quad (2)$$

3.4 Cycle Detection

Next step, we design an algorithm to analyze the spectrum properties for all the water quality indicators in order to find the relationships between the indicators and different cities. Traditional methods for water quality analysis mostly concentrated on the indicator changes or for individual prediction.

To our knowledge, our method is the first trial to analyze water quality in the frequency domain. The analysis can help easily to find the indicator cycles and their predictions. Our algorithm is shown in Algorithm 1.

Algorithm 1. Water Quality Frequency Domain Analysis Algorithm

Data: $F_{M\times N\times T}$
Result: $S_{M\times N\times K}$
– Initialization;
– *Clustering to $M^{'}$;
while $m < M$ **do**
– *Clustering to $N^{'}$ or $T^{'}$;
– Normalization;
while $n < N$ **do**
Adp-FFT with $F_{mn} \rightarrow y[k_{mn}]$;
$Sig\_k = k\ in\ max(A[k_{mn}])$;
if $Sig\_k < T/2$ **then**
$S_{mn}[k_{mn}] = y[k_{mn}]$;
else
$S_{mn} = 0$;
end
$S_{mN}[k_m] = S_{mn}[k_m](0 < n < N)$;
end
– $S_{m\times N\times K}$;
end
– *Declustering to M, N, T;
– $S_{M\times N\times K}$;

We list the symbols in this algorithm as follows:

$F_{M\times N\times T}$ Input data set with M cities, N indicators and T recordings.
$S_{M\times N\times K}$ Output data set with M cities, N indicators and K frequencies.
$M^{'}, N^{'}, T^{'}$ Clustering results.
k_{mn} FFT results frequency for city m, indicator n.
$A[k_{mn}]$ FFT results amplitudes for city m, indicator n.
$y[k_{mn}]$ FFT results with frequencies and amplitudes for city m, indicator n.
Sig_k Significant frequency.

In order to cope with the diverse units, normalization is an inevitable step to process the data. In this work, we transform all the water quality indicators of raw data to have a mean of zero and a standard deviation of 1. Some people also call this z-score standardization.

For regular frequency domain analysis, people often use the Fast Fourier Transform (FFT) method. Classical FFT is defined as in Equation (3). In this equation, $y[k]$ of length T is the result of FFT for the indicator sequence $x[t]$ of length T.

$$y[k] = \sum_{t=0}^{T-1} e^{-2\pi j \frac{k_t}{T}} x[t]. \quad (3)$$

As we can see from this equation, the length T is an important parameter in FFT. But in practice, different water quality indicators are difficult to synchronize both for city and indicator domains. In addition, the clustering step in the Algorithm 1 can create changes for T. Thus, here we define a function $T^{'}_{mn}$ as adaptive parameter of T, as in

$$T^{'}_{mn} = C \times \alpha_m \times \sum_{n=1}^{N} \frac{\beta_n T_n}{N}. \quad (4)$$

In this equation, $C \times \alpha_m$ is the adaptive parameter for the clustering effect in the city domain, in which C represents clustering scale among all the cities and α_m as the weight value for each city. For the second part of the equation represent the synchronization effect between different indicators. N is the number of indicator types in one city. For example, in Oslo, we have $N = 10$, but Bergen has $N = 7$. T_n is the recording length of indicator n, β_n is adaptive weight value for indicator n.

So the overall adaptive FFT (Adp-FFT) method, we define as in the Equation (5), in which we considered the clustering and synchronization effect in water quality indicator frequency analysis.

$$y(k_{mn}) = \sum_{t=0}^{T^{'}_{mn}-1} e^{-2\pi j \frac{k_t}{T^{'}_{mn}}} f_{mn}(t). \quad (5)$$

From here we get complete spectrograms of all the indicators. After, we have to find the significant frequency in order to detect the cycles for different quality indicators. To get the significant frequency, first, we use the following equation to find the maximal amplitude in the frequency domain.

$$A_{k_{mn}} = max(\sqrt{(y(k_{mn})_{re})^2 + (y(k_{mn})_{im})^2}). \quad (6)$$

In this Equation 6, $y(k_{mn})_{re}$ and $y(k_{mn})_{im}$ represent the real and imaginary parts of $y(k_{mn})$ in the result of Adp-FFT. $y(k_{mn})$ is the sequence of complex numbers.

We find the corresponding frequency of the amplitude $A_{k_{mn}}$ in the frequency domain is then the significant frequency for city m, indicator n. We will provide more examples in Section 4.2.

3.5 Indicator Prediction

By getting the result of spectrograms for the indicators, our work is not finished. We want to use these results to predict the tendencies of the water quality, especially for biological indicators. Algorithm 2 is designed as follows to perform this function.

Algorithm 2. Water Quality Prediction Algorithm

Data: $S_{M\times N\times K}$
Result: $F_{M\times N\times[T+P_M]}$
– Initialization;
while $m < M$ **do**
  $P_m = P$;
  $H_m = H$;
  while $n < N$ **do**
    – Sort $S_{mn}[k_t]$ according to amplitude $A_{mn}[k_t]$;
    – Select top H_m elements in $S_{mn}[k_t]$;
    – $S_{mnh}[k_t] = S_{mn}[k_t]$ $(0 < NH < H_m)$;
    if $0 < t_p < P_m$ **then**
      – Calculate $A_{mn}[T_{mn} + t_p]$;
      – Calculate $\Phi_{mn}[T_{mn} + t_p]$;
      – Calculate $F_{mn}[T_{mn} + t_p]$;
      – t_p++;
    else
      – $F_{mn}[t + t_p] = 0$;
    end
    – $F_{mn}[T + P_m]$;
  end
  – $F_{mN}[T + P_m]$;
end
– $F_{MN}[T + P_M]$;

We list the additional symbols as follows:

$F_{MN[T+P_M]}$	Output prediction data set with M cities, P_M Prediction range.
H_m	Number of harmonics.
$\Phi_{mn}[T_{mn} + t_p]$	Phase value for prediction at time $t + t_p$ in city m and indicator n.

We use adaptive strategy during the frequency transform process as in Equation (5). In this Algorithm 2, we also adjust our inverse transform Equation (7) as follows:

$$F_{mn}(T_P) = \frac{1}{T_{mn}} \sum_{t=T_{mn}-1}^{T_P} e^{-2\pi j \frac{k_t}{T_{mn}}} y(k_{mn}). \tag{7}$$

In this equation, T_{mn} is defined the same as in Equation (4). Inverse Adp-FFT is used to transform water quality indicators from the frequency domain back to the time domain to see its tendencies. The prediction result can be calculated with the following formula. In this Equation (8), we have $T_{mn} \leq t \leq T_{mn} + P_m$.

$$\begin{aligned} A_{mn}[t] &= \sqrt{(S_{mnh}[k_{mn}]_{re})^2 + (S_{mnh}[k_{mn}]_{im})^2} \\ \Phi_{mn}[t] &= tg^{-1} \frac{S_{mnh}[k_{mn}]_{im}}{S_{mnh}[k_{mn}]_{re}} \\ F_{mn}[t] &= A_{mn}[t] \times \cos(2\pi k \times t + \Phi). \end{aligned} \tag{8}$$

As for our experience, the prediction range P_m, harmony parameter H_m can both affect the accuracy. In practice, we can set up a threshold for accuracy in order to find the optimal solution of P_m and H_m values.

3.6 Scalability

Scalability is an important property to evaluate the algorithms. For this water quality prediction issue, we consider the scalability of our method in three data domains, indicator, geography, and time.

3.6.1 Indicator Domain Scaling

The number of water quality indicators can vary from one to several hundred, depending on the standards in different countries or regions. Even, as for people's requirement for higher quality water, there are gradually new types of indicators keep appearing. Traditional water quality research mostly concentrated on individual indicator observation or prediction. This is partly because it is highly challenging to analyze the complex synergies between the physical, chemical and biological indicators.

In this method, we are trying to find the indicator relationship in the frequency domain. By visualizing the spectrogram of indicators, we can discover their characters in the frequency domain, and search for their resonance effect.

In this algorithm, to scale in the indicator domain is fairly easy by just adding the new indicator recordings into the frequency analysis following Algorithm 1 and find the significant frequency with Equation (6).

3.6.2 Geography Domain Scaling

Geography location is one of the most important factors to affect the water quality, especially for the urban surface water source. Geographical domain scaling is essential for policy making, regional water source quality evaluations, pollution analysis, etc. When we consider the geography domain scaling in practice, there are several aspects can be used for classification, such as:

- Water source type (surface, river, ground, frozen, desalination, etc);
- Water source description (area, depth, discharge, flow velocity, etc);
- Locations (longitude, latitude, height, etc);
- Climate (tropical, dry, mild mid-latitude, cold mid-latitude, polar, etc);
- Main pollution type (organic, inorganic, macroscopic contaminants, etc);
- Residential states (types, population, main activities, etc);
- Agriculture states (planting, farming, fishery, etc);
- Industrial states (factory types, main discharge, etc).

The geography domain scaling can be executed from the perspectives in the above description. In this study, we use

a weighted arithmetic mean function for geography scaling. This means the same water quality indicators in city, region or country can be clustered, as shown

$$F_{m'n}[t] = \frac{\sum_{m=1}^{m'} \omega_{m'n} f_{mn}(t)}{\sum_{m=1}^{m'} \omega_{m'n}}. \quad (9)$$

In this equation, $\omega_{m'n}$ is the weight of water quality indicator n for the new geographical indicator $F_{m'n}$. By adjusting m', we can control the scaling process of the geography domain. Changing different $f_{mn}(t)$ can be customized to observe the data from different geographical perspectives.

3.6.3 Time Domain Scaling

Water quality prediction is beneficial for the whole process of water supply. It provides early warnings and supports early preventive measures. Time domain scaling can contribute to prolonging the warning time. At the same time, it can be helpful to analyze water quality changes in the source area for longer periods (e.g., from second records to year records). In this study, one of the most important reasons we choose frequency domain analysis for water quality data processing is to cope with the time domain scaling issues.

The Algorithm 1 we use for cycle detection has applied our Adp-FFT (Equation (5)) to analyze frequency domain. This method has an inherent time scaling property. So, we can conclude time scaling property for adp-FFT as in the following Equation (10), here we omit the proof process.

$$\begin{aligned} &if \quad \mathbf{Adp-FFT}(f_{mn}(t)) \rightarrow y(k_{mn}) \\ &then \quad \mathbf{Adp-FFT}(f_{mn}(\lambda t)) \rightarrow y(\lambda k_{mn}) = \frac{1}{|\lambda|} y(\frac{k_{mn}}{\lambda}). \end{aligned} \quad (10)$$

By virtue of this good property, we can keep the properties we analyzed in the whole time domain. In this method, we should have $1 < \lambda < T_{mn}$. Because, in practice, on one side, we can not analyze the frequency properties in the smaller time domain that we don't have supported data. On the other side, to group the whole data as one has lost the meaning of analysis. We are going to give more examples of time scaling in Section 4.3.

3.7 Risk Modeling

In the water supply industry, most of the water quality monitoring and control are taken in the treatment plant for easy access reasons. Most countries or regions in the world have made the water quality standards according to this step.

In this paper, we propose the data perception approach for water quality risk early detection and prediction in the water source area. Among all the water quality indicators, biological indicators are directly related to people's health. In the drinking water supply, we concentrate on most the biological indicator changes, especially for their peak values. Peak values normally represent environment alter. This could be a sudden change from weather, industrial or agricultural activities. This is an important alert for water source protection. The peak values of biological indicators require a special process in the treatment plants accordingly.

According to the present published version in [7], we define the risk of water quality with peak values as follows:

$$R_i(t) = \begin{cases} f_i(t) & f_i^{'}(t) = 0 \;\&\&\; f_i^{'}(t-1) > 0 \\ f_i(t) & f_i^{'}(t) = 0 \;\&\&\; f_i^{'}(t+1) < 0 \\ f_i(t) & f_i^{'}(t) \neq 0 \;\&\&\; f_i(t) = max(f_i) \\ 0 & Others \end{cases}. \quad (11)$$

In this definition, $f_i(t)$ is a biological indicator, we choose the peak value based on its first order derivative. If there is no 0 derivative (data set is too small), we choose the max value of the sequence.

4 Application

The application of this approach is based on our water quality project in Norway. In this project, we are working closely with the people coming from the whole water supply process to improve water quality by early warnings. In this team, there are water quality experts, source sampling operators, treatment plant managers, policy makers, and data researchers. In this section, we describe this application and provide our preliminary results with analysis.

4.1 Data Collection & Description

The data we collected for this application is from several industrial drinking water supply systems in Norway.

For *geography domain*, it includes 4 Norwegian cities, Oslo, Bergen, Strømmen, and Ålesund, as we depicted in Section 2.1.

For *indicator domain*, constrained by the synchronization of different cities, we select meaningful indicators as physical: conductivity, turbidity, and color, chemical: pH, and biological: Coliform, Ecoli, and Int.

For *time domain*, it varies in different cities. We got the data as Oslo (2009.01 - 2015.12), Bergen (2007.01 - 2015.12), Strømmen (2008.01 - 2014.12), and Ålesund (2005.01 to 2015.12).

However, the data qualities are quite uneven. In practice, some operators in the lab did not record all the sample results correctly and led to massive missing values. For example, the first issue is the time synchronization between different cities is difficult. The data from Oslo, Bergen and Ålesund was taken once a week, but Strømmen was once every two weeks. The second issue is missing values. Some of the physical and chemical indicators from Ålesund were only recorded 25 times for 11 years; alkalinities all equal to zero; values for Ecoli are over 95 percent zero. In Bergen, they did not record any data for Clostridium perfringen. After discussions with domain experts, these issues can make prediction accuracy fluctuate.

4.2 Implementation Process

We are running our application according to the framework designed in Section 3.1.

In data pre-processing, we have worked with water quality experts to clean the data which are errors, not meaningful and correct the inaccurate values. We synchronized the data according to the recordings from all the 4 cities in order to keep most of the useful values. The normalization process has

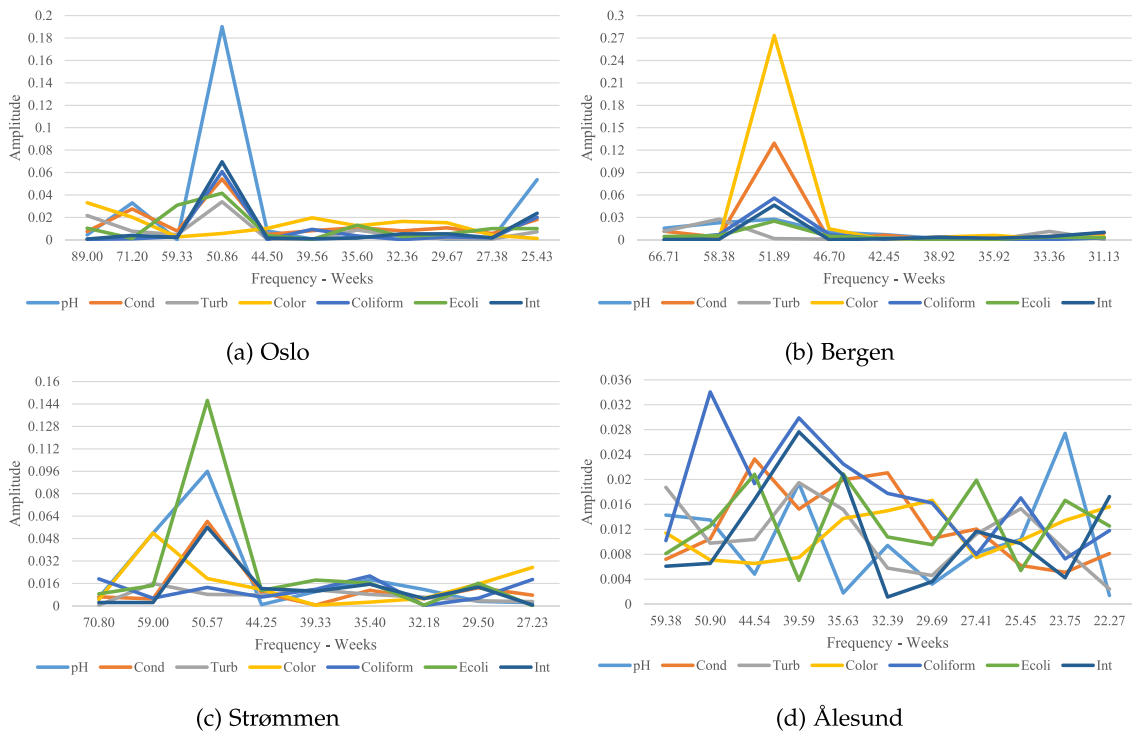

Fig. 4. Spectrogram for weekly water quality indicators.

been followed by our Algorithm 1. In this study, we use the pre-processed weekly data sets to analyze related features for Oslo, Bergen, Strømmen and Ålesund. In addition, we will analyze the scalability of this question in Section 4.3.3.

In feature selection, we also synchronize collected usable water quality indicators for analysis. As for the practical constraints, we selected pH, Conductivity, Turbidity, and Color as input features. Output biological indicators are Coliform, Ecoli, and Int. Training set and testing set have taken according to time. For each indicator, the first 90 percent of recordings are used for training and the rest 10 percent are used for testing.

The risk modeling, prediction and evaluation are based on the models we gave in Sections 3.4, 3.5, and 3.7.

4.3 Results & Analysis

In water research, there is no well-accepted theoretical analysis for the complex interactions among all the water quality indicators. This study takes the assumption as each indicator is independent. But different from other work to analyze each indicator separately, here we can provide a perspective to find the relationships between indicators by frequency analysis. At the same time, we present various evaluations to show the prediction accuracy. In this section, we also show the scalability of our method can serve as a very powerful tool for practical water quality early warning.

4.3.1 *Frequency Domain Analysis*

The correlation analysis is a natural way to find the relationship between different water quality indicators. We have shown our results in our previous paper [26], [27]. From there, we found no obvious results by direct correlation findings between indicators. Frequency domain analysis in this study is meaningful for water quality, in both theory and practice. In our application, we have executed spectrogram analysis in 4 Norwegian cities for all the indicators as weekly values using our Algorithm 1. The results of spectrograms are shown in Fig. 4. Different colors represent different indicators. The x-axes in the sub figures are frequencies, y-axes are amplitude after Adp-FFT. We can see from this figure, in 4 cities, there are some indicators share the same significant frequency.

More precisely, we give significant frequencies for different indicators in the 4 cities in Fig. 5. Different colors to represent different cities. 7 angles show different types of indicators, including 3 biological indicators, as output and 4 physical and chemical indicators as input. Each spoke length gives the value of their significant frequency with the unit as weeks. We can interpret this figure from the following aspects.

a. Many water quality indicators posses the periodic properties, but not all of them. Some indicators do not

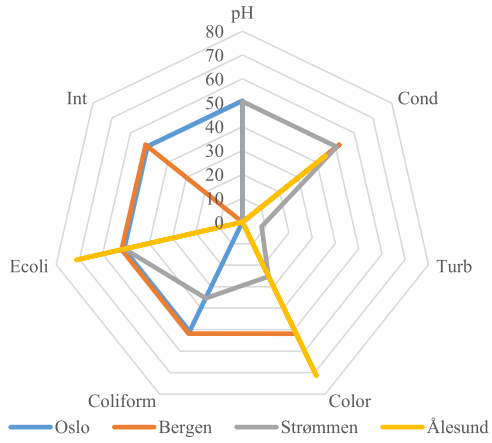

Fig. 5. Weekly indicator significant frequency.

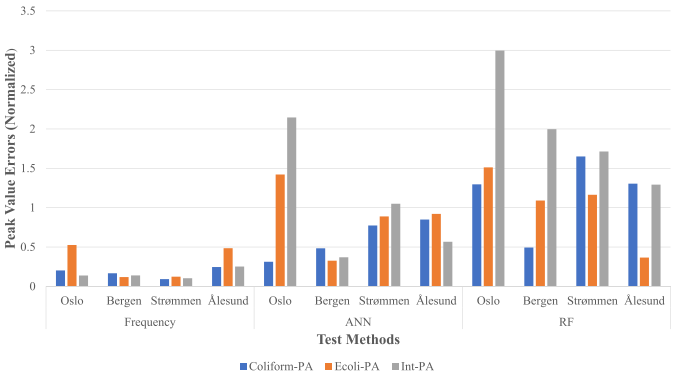

Fig. 6. Prediction accuracy for peak values.

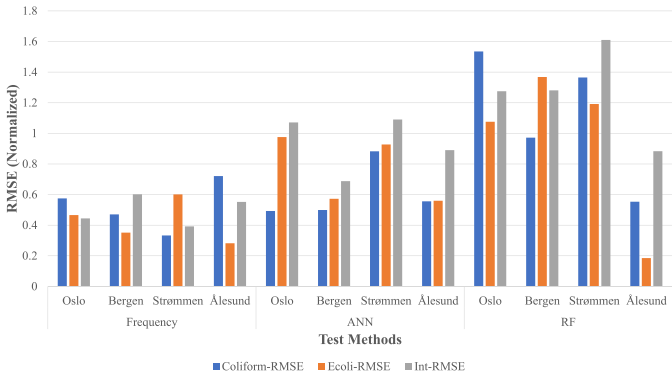

Fig. 7. Prediction accuracy for RMSE measurements.

have significant frequencies, or they are not meaningful in the field. Here we note them as zero. There are various reasons for them. In practice, reasons can be data are not recorded, measures are not standardized, etc. Our results show the *Frequency Zero* indicators are: pH (Bergen, Ålesund), Conductivity (Oslo), Turbidity (Oslo, Bergen, Ålesund), Color (Oslo), Coliform and Int (Ålesund).

b. Inside one city, some quality indicators share the same significant frequency. We are interested in this feature, because potentially, the physical or chemical indicator could provide early risk warning for corresponding biological indicators, because they are much faster to access. For example, in Bergen, Color has the same frequency with Coliform, Ecoli, and Int, as 51.89 weeks. From Fig. 5, we can see in details, Oslo can use pH for all the three indicator predictions (50.86 weeks); Bergen can take Conductivity and Color (51.89 weeks); Strømmen can use pH or Conductivity to predict Ecoli (50.57 weeks); Ålesund can take Color to predict Ecoli (71.26 weeks).

c. Among all the cities, some indicators have similar significant frequencies (concrete value depends on the number of recordings). Our results show that Turbidity does not support meaningful prediction for biological indicators in all the 4 cities from the frequency analysis perspective. Ecoli has similar significant frequencies in 3 cities (Oslo, Bergen and Strømmen). Oslo and Bergen show good frequency connections between indicators as 50.86 weeks and 51.89 weeks. This could potentially be used for different cities collaborative analysis and provide risk early warning.

4.3.2 Risk Prediction

The risk in the water supply system depends highly on biological water quality indicators. The following treatment process will regulate accordingly to the changes of them. Based on our analysis in Section 3.2, peak values of those indicators give important information. We compare our frequency analysis methods with two classical prediction methods, including artificial neural network (ANN) and random forest (RF). We evaluate them from three aspects. First one we calculate the average *prediction accuracy for peak values*. Peak values were selected based on the risk model defined in Section 3.7. Second one we apply Root Mean Square Error (RMSE) for *overall prediction accuracy*. Third one we measure the *computation time* as the efficiency of these methods.

In this experiment, inputs of these methods are physical and chemical indicators, as pH, Conductivity, Turbidity, and Color. Their outputs are biological indicators as Coliform, Ecoli, and Int. We take training and testing sets split as 90 to 10 percent regarding limited recordings.

For ANN method, we use a three-layer back propagation (BP) network structure. Input layer as 4 nodes, 3 nodes in the output layer, and hidden layer for 300 nodes. Hyperbolic tangent (tanh) activation function is chosen considering we have normalized the data sets. Batch size as $N_t/20$ is based on our data size. N_t is the total number of data recordings in different cities. For each data set, we train them for 1000 times.

For RF method, we take into account the results from frequency analysis to choose one input indicator which has the same significant frequency as the heuristic important feature. Initially, we choose 1000 as the number of trees in the forest, and 40 to be the random seed for pseudo-random number generator.

For our Frequency Analysis prediction method, we apply the method we described in Section 3.5. The parameter as the number of harmonics is sensitive to the accuracy, we have made the experiments and draw the chart to analyze their relationships between different water indicators. In this case, we chose 20 as the number of harmonics to be the optimal solution. This part can be further improved by more adaptive strategies.

Fig. 6 shows the prediction accuracy for 3 biological indicator peak values. This is a special evaluation of water quality prediction. The x-axis is the combination of methods and cities and y-axis is the average prediction error. Different colors show different water quality indicators. Because the data sets have been normalized before, so there is no unit for these errors. We can see from here if we compare the three methods, Frequency Analysis performs better than the other two for lower error values. As for the comparison of indicators, Int shows higher error values, Coliform and Ecoli do not show clear synchronization for peak value prediction errors. Between different cities, Oslo shows higher prediction error values among all these three methods.

The general RMSE accuracy comparison is given in Fig. 7. It shows overall accuracy for all the predicting points. Axes are made the same meaning as Fig. 6. Because it takes

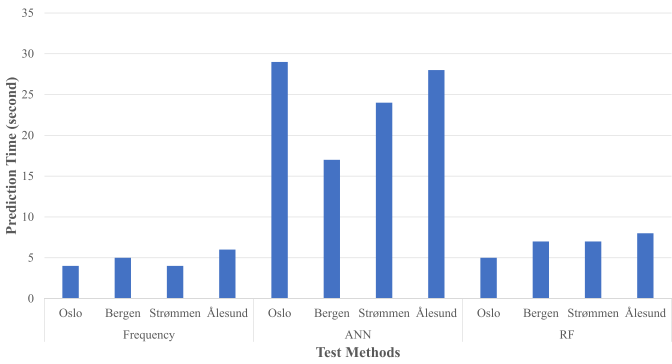

Fig. 8. Time consumption for prediction.

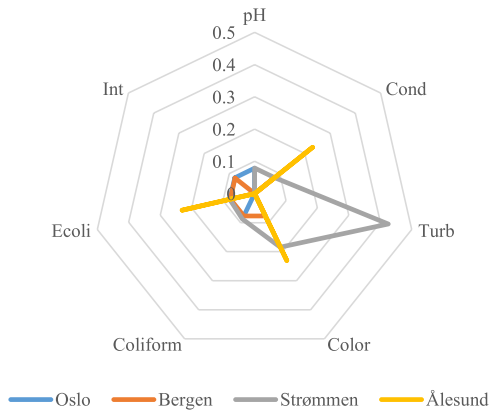

Fig. 9. Significant frequency scalability.

all the points and calculate the average error values, so in general, it is smaller than only peak value errors. The error values do not show a high distinction between different methods. For the average RMSE of different methods, our Frequency Analysis improves more than 10 percent than ANN and RF. The comparisons between indicators and cities do not show substantial similarities in these results.

We also compare the prediction time consumption for different methods. Here we did not test the time for each concrete indicator. Because these methods are all applied in the parallel platform. We have run the experiment 30 times and calculate the average time. The results are shown in Fig. 8. We can see ANN costs more than the other two methods. Frequency Analysis is slightly better than RF.

4.3.3 Scalability Discussion

In Section 3.6, we have discussed theoretical scalability for this method in indicator, geography and time domains. As for we did not collect enough information for more synchronized indicators (indicator domain) and cities (geography domain), in this section, we show the scalability of our method in the time domain. As a reference, we also test our method scalability in prediction accuracy and time consumption.

In order to test the scalability of our method, we add the step to cluster our data in seasons. In Norway, the seasons are generally mild. We use this scalability evaluation to find the connections between indicator frequencies with seasons. In this study, according to the government management principles, we consider seasons according to the time, defined as follows:

- Spring: February to April;
- Summer: May to July;
- Autumn: August to October;
- Winter: November to January.

Scalability is one of the most important profits we get from this method. We have also run our whole application for time scaling. We have solved the water quality prediction for weekly data sets from 4 Norwegian cities in Section 4.3. In order to evaluate the scalability of this method, we will compare the season data results with weekly data sets in the significant frequency, prediction accuracy of peak values and RMSE, and the time consumption. We run the experiment following Sections 4.3.1 and 4.3.2 for Frequency Analysis method on the seasonal data sets. The results are recorded and further divided by the corresponding value for weekly data sets.

We use radar charts to depict our scalability results. Fig. 9 shows the scalability ratio on significant frequencies of indicators. Water indicators are set at the 7 directions, input indicators on the right side, output indicators on the left side. The lengths of the vectors are the ratio values. Different colors represent different cities. Here we see Oslo and Bergen show the linear scalability for all the meaningful indicators. Ålesund has unified sub-linear scalability for its meaningful indicators. As for Strømmen, Turbidity and Color show their unique sub-linear scalability. We attribute this exception to raw data recording errors based on domain analysis. In general, we can say the scalability for this method shows good linear scalability in significant frequency analysis for water quality indicators.

As for the scalability in prediction accuracy, the results are shown in Fig. 10. This radar chart shows the output indicators accuracy in two groups, Peak values on the right side as *PA*, and RMSE on the right side as *RMSE*. Different colors represent different cities. For seasonal data sets, since the recordings are much less than weekly data, so the training sets are limited. This makes the prediction accuracy errors getting much higher. So, in this figure, we see the ratios are in general more than 1. They are sub-linear. From this, we can say there is no general similarity for accuracy scalability.

The time consumption scalability results are in Table 1. We can see with the seasonal data sets, prediction time consumption is overall obviously reduced. But the reduction is sub-linear scalability.

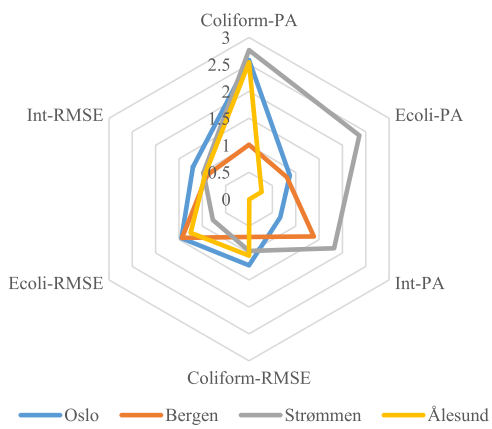

Fig. 10. Accuracy scalability for frequency analysis.

TABLE 1
Time Consumption Scalability

City	Scalability
Oslo	0.55
Bergen	0.59
Strømmen	0.67
Ålesund	0.58

4.3.4 Limitation & Insight Analysis

Limitations of the frequency analysis method can be found in the following aspects:

- This method is difficult to use for the data sets which do not have significant frequency effects. Some water indicators in our urban supply system do not have the meaningful frequencies, the predictions for those have shown high accuracy errors.
- This method analyzes the relationship between different indicators on their frequencies. Every indicator is considered to be independent, this results in higher level complex relations between indicators are ignored.
- The parameters in the prediction, such as the number of harmonics need time to adjust, this extra step can take longer time. We are also looking for new strategies to fix this.

This frequency analysis method for water quality prediction can also bring many new visions for urban water supply systems. We discussed with the domain experts, the insight can be found from several perspectives:

- This work can provide suggestions for IoT integration sensor deployments in water supply systems. For example, we found Color has a strong connection with the biological indicators, so we suggest to put more real-time color sensors all through the water supply process in order to detect the risk.
- Compare with most of the *black box* algorithms, this method can provide explainable relationship analysis between indicators on their frequencies.
- This method can also provide a method to evaluate data quality. Industrial data collections are usually with noise. This method can find obviously inaccurate points by abnormal frequency detection. For example, the seasonal data in the Turbidity of Strømmen is beyond scalability values, we are suspicious for the quality in data collection.
- Urban systems can also be compared with this method, so it provides a collaborative analysis between different cities for the national management level.

5 Conclusion

Water quality is a very critical issue in modern urban life all around the world, especially for *Smart Water Supply* system development. Traditional monitoring and risk control methods are difficult to detect bacteria broadcast on time and provide efficient decision support. In this paper, we propose an approach for water quality risk early warning using data perception. With the application among four different cities in Norway, we have proved the feasibility, accuracy, and efficiency of our approach. The preliminary results evaluated by domain experts are very promising. This work is beneficial in generally three aspects:

- It provides an early warning mechanism from the water source areas using cost-less data analysis techniques. This prolongs the preventive measures response time, and support more decision options in the latter steps of water supply.
- This approach integrates indicator, geography and time domains. It provides a new frequency domain analysis perspective to find the relationship between different indicators and their predictions. At the same time, it embraces scalability for these three domains.
- This work is applied to real industrial water supply systems from 4 different Norwegian cities.

Acknowledgments

The authors would like to thank the management of the four Norwegian city water supply systems, as well as the Research Council of Norway for funding under the KLIMAFORSK programme.

References

[1] S. Franco, V. Gaetano, and T. Gianni, "Urbanization and climate change impacts on surface water quality: Enhancing the resilience by reducing impervious surfaces," *Water Res.*, vol. 144, pp. 491–502, 2018.
[2] T. Hák, S. Janoušková, and B. Moldan, "Sustainable development goals: A need for relevant indicators," *Ecological Indicators*, vol. 60, pp. 565–573, 2016.
[3] World Health Organization (WHO), *Guidelines for Drinking-Water Quality: Recommendations*. Geneva, Switzerland: World Health Organization, 2004.
[4] E. Weinthal, Y. Parag, A. Vengosh, A. Muti, and W. Kloppmann, "The eu drinking water directive: The boron standard and scientific uncertainty," *Eur. Environment*, vol. 15, no. 1, pp. 1–12, 2005.
[5] R. W. Adler, J. C. Landman, and D. M. Cameron, *The Clean Water Act 20 Years Later*. Washington, D.C., USA: Island Press, 1993.
[6] D. Berge, "Overvåking av farrisvannet med tilløp fra 1958–2010," NIVA-rapport: 6175, Res. Rep., 2011. [Online]. Available: https://niva.brage.unit.no/niva-xmlui/handle/11250/215478
[7] I. W. Andersen, "EUs rammedirektiv for vann–miljøkvalitetsnormer for vannmiljøet i møte med norsk rett," *Kart og Plan*, vol. 73, no. 5, pp. 355–366, 2013.
[8] V. Novotny, *Water Quality: Prevention, Identification and Management of Diffuse Pollution*. New York, NY, USA: Van Nostrand-Reinhold Publishers, 1994.
[9] A. Hounslow, *Water Quality Data: Analysis and Interpretation*. Boca Raton, FL, USA: CRC Press, 2018.
[10] S. Yagur-Kroll, E. Schreuder, C. J. Ingham, R. Heideman, R. Rosen, and S. Belkin, "A miniature porous aluminum oxide-based flow-cell for online water quality monitoring using bacterial sensor cells," *Biosensors Bioelectronics*, vol. 64, pp. 625–632, 2015.
[11] H. R. Maier and G. C. Dandy, "The use of artificial neural networks for the prediction of water quality parameters," *Water Resources Res.*, vol. 32, no. 4, pp. 1013–1022, 1996.
[12] H. Orouji, O. Bozorg Haddad, E. Fallah-Mehdipour, and M. Mariño, "Modeling of water quality parameters using data-driven models," *J. Environmental Eng.*, vol. 139, no. 7, pp. 947–957, 2013.
[13] O. Bozorg-Haddad, S. Soleimani, and H. A. Loáiciga, "Modeling water-quality parameters using genetic algorithm–least squares support vector regression and genetic programming," *J. Environmental Eng.*, vol. 143, no. 7, 2017, Art. no. 04017021.
[14] N. Mahmoudi, H. Orouji, and E. Fallah-Mehdipour, "Integration of shuffled frog leaping algorithm and support vector regression for prediction of water quality parameters," *Water Resources Manag.*, vol. 30, no. 7, pp. 2195–2211, 2016.

[15] F.-J. Chang, Y.-H. Tsai, P.-A. Chen, A. Coynel, and G. Vachaud, "Modeling water quality in an urban river using hydrological factors–data driven approaches," *J. Environmental Manag.*, vol. 151, pp. 87–96, 2015.
[16] J. Gubbi, R. Buyya, S. Marusic, and M. Palaniswami, "Internet of things (iot): A vision, architectural elements, and future directions," *Future Generation Comput. Syst.*, vol. 29, no. 7, pp. 1645–1660, 2013.
[17] V. Mayer-Schönberger and K. Cukier, *Big Data: A Revolution that Will Transform how we Live, Work, and Think*. Boston, MA, USA: Houghton Mifflin Harcourt, 2013.
[18] Y. Wu, F. Hu, G. Min, and A. Y. Zomaya, *Big Data and Computational Intelligence in Networking*. Boca Raton, FL, USA: CRC Press, 2017.
[19] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor, "Canonical correlation analysis: An overview with application to learning methods," *Neural Comput.*, vol. 16, no. 12, pp. 2639–2664, 2004.
[20] I. T. Jolliffe and J. Cadima, "Principal component analysis: A review and recent developments," *Philosoph. Trans. Roy. Soc. A: Math. Phys. Eng. Sci.*, vol. 374, no. 2065, 2016, Art. no. 20150202.
[21] Y. Luo, D. Tao, K. Ramamohanarao, C. Xu, and Y. Wen, "Tensor canonical correlation analysis for multi-view dimension reduction," *IEEE Trans. Knowl. Data Eng.*, vol. 27, no. 11, pp. 3111–3124, Nov. 2015.
[22] J. Gudmundsson and M. Horton, "Spatio-temporal analysis of team sports," *ACM Comput. Surveys*, vol. 50, no. 2, 2017, Art. no. 22.
[23] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, 2015, Art. no. 436.
[24] J. Liu, A. Shahroudy, D. Xu, and G. Wang, "Spatio-temporal LSTM with trust gates for 3D human action recognition," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 816–833.
[25] N. Laptev, S. Amizadeh, and I. Flint, "Generic and scalable framework for automated time-series anomaly detection," in *Proc. 21st ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2015, pp. 1939–1947.
[26] H. Mohammed, I. A. Hameed, and R. Seidu, "Adaptive neuro-fuzzy inference system for predicting norovirus in drinking water supply," in *Proc. Int. Conf. Inform. Health Technol.*, 2017, pp. 1–6.
[27] D. Wu, H. Mohammed, H. Wang, and R. Seidu, "Smart data analysis for water quality in catchment area monitoring," in *Proc. 4th IEEE Int. Conf. Smart Data*, 2018, pp. 20–27.

Di Wu received the BS degree from the Department of Automation, Beijing Institute of Technology, China, in 2004, and the PhD degree in pattern recognition and intelligent system from the School of Automation, Beijing Institute of Technology, in 2010. She worked as post-doc/lecturer with the Department of Computer and Communication Engineering, University of Science and Technology Beijing, China. Currently, she is working toward the PhD degree in data science at the Norwegian University of Science and Technology. She has already published more than 17 papers for international journals and conferences. Her research interest covers large scale composable simulation system, artificial life, and data analysis in industrial applications such as water supply systems. She is a member of the IEEE.

Hao Wang received the BEng and PhD degrees, both in computer science and engineering, from the South China University of Technology. He is an associate professor in the Department of Computer Science in Norwegian University of Science & Technology, Norway. His research interests include big data analytics, industrial internet of things, high performance computing, safety-critical systems, and communication security. He has published 100+ papers in reputable international journals and conferences. He served as a TPC co-chair for IEEE DataCom 2015, IEEE CIT 2017, ES 2017, a senior TPC member for CIKM 2019, and reviewer for journals such as *TKDE*, *TII*, *TBD*, *TETC*, *T-IFS*, *IoTJ*, *TCSS*, *TOMM* and *TIST*. He is a member of IEEE IES Technical Committee on Industrial Informatics.

Hadi Mohammed received the BSc and MSc degrees from the Department of Physics, Kwame Nkrumah University of Science and Technology (KNUST) - Kumasi, Ghana, in 2008 and 2013, respectively. He subsequently worked as a physics and mathematics teacher with the Ghana Education Service He is currently working toward the PhD degree in the Department of Marine Operations and Civil Engineering, Norwegian University of Science and Technology (NTNU), Norway. His research interest include hydrological - hydrodynamic and water quality modeling, quantitative microbial risk assessments, and the application of artificial intelligence in water quality management.

Razak Seidu is a professor and leader of the Water and Environmental Engineering Group at the Institute for Marine Operations and Civil Engineering in the Norwegian University of Science and Technology, Norway. His research interest include big data for water safety planning and management, quantitative risk assessment and intelligentsmart water systems. He has served as a technical consultant for several international organizations including the UN-Habitat, World Health Organization (WHO), United Nations Environment Programme (UNEP), and the Stockholm Environment Institute. He is currently a member of the WHO Technical Committee for Wastewater and Microbiology; and has been involved in several water quality related projects with funding from the European Union, Google.org, the Bill and Melinda Gates Foundation, and the Research Council of Norway. Information about his research groups activities can be found at www.watercube.no.