A Tensor Model for Quality Analysis in Industrial Drinking Water Supply System

Di Wu
Department of ICT and Natural Science, Norwegian University of Science and Technology
Ålesund, 6009 Norway
Email: di.wu@ntnu.no

Hao Wang
Department of Computer Science, Norwegian University of Science and Technology
Gjøvik, 2815 Norway
Email: hawa@ntnu.no

Razak Seidu
Department of Ocean Operations and Civil Engineering, Norwegian University of Science and Technology
Ålesund, 6009 Norway
Email: rase@ntnu.no

***Abstract*—Drinking Water Supply (DWS) is one of the most critical and sensitive systems to maintain city operations globally. In Europe, the contradiction between the fast growth of population and obsolete urban water supply infrastructure is even more prominent. The high standard water quality requirement not only provides convenience for people's daily life but also challenges the risk response time in the systems. Prevalent water quality regulations are relying on periodic parameter tests. This brings the danger in bacteria broadcast within the testing process which can last for 24-48 hours.**

In order to cope with these problems, we propose a tensor model for water quality assessment. This model consists of three dimensions, including water quality parameters, locations and time. Furthermore, we applied this model to predict water quality changes in the DWS system using a Random Forest algorithm. For a case study, we select an industrial water supply system in Oslo, Norway. The preliminary results show that this model can provide early warning for water quality risks.

***Keywords*-Tensor; Drinking Water Supply; Water Quality; Early Warning.**

I. Introduction

Drinking Water Supply (DWS) systems are one of the most important infrastructures in our daily lives. According to the report from the World Health Organization (WHO) [1], there are still over 681 million people on earth struggling to receive sufficient clean water. In DWS systems, water quality is a key factor across the whole process, from the water source, treatment, and distributed pipelines. Prevalent water quality is controlled using a series of parameters. They are different from countries or regions based on geographical and development conditions. Typical water quality parameters can be divided into three groups, as physical, chemical and biological parameters. To test the parameters in practice can take from several minutes to 24 hours. The outbreaks of contagious bacteria can be much faster than the testing time and therefore cause serious threats to people's health.

In this paper, we address this problem using the data analysis method. This becomes feasible thanks to the significant improvements in sensing and data analysis technologies. To predict the water quality in DWS systems. We need to build a model to provide the changes in water quality parameters. Because the time consumption from computation is much shorter than regular tests, we can, therefore, provide early warnings to risk detection. There are some trial works before in this field. In 2015, Yagur Kroll *et al* [2] introduced some sensors to monitor biological parameters. For water quality prediction, Holger *et al* [3] designed an artificial neural network for salinity level in an Australian river named Murray. In Iran, Orouji *et al* provided a series of algorithms for chemical parameter predictions, such as in [4] [5] [6]. Chang *et al* [7] proposed a framework to predict NH_3-H for Dahan River. However, their works are focusing on single water quality parameter prediction, without considering the relationship between parameters. In addition, geographical and time factors are ignored.

We propose a tensor model for water quality analysis, taking into account parameter, location and time domains. This model provides a comprehensive understanding of DWS systems for water quality control. Furthermore, we use a Random Forest algorithm to predict these parameters in order to provide early warnings in the industrial process and decision-making support for corresponding actions.

II. Problem Formulation

A. Tensor Model

We build this model for water quality analysis based on the inherent data structures. The data collected from industrial water supply systems usually have three dimensions, as parameter, location and time. Inside each dimension, there are hierarchical levels as well. A high-order tensor representation can describe such a complex data format. According to the requirements from the DWS process, we build a tensor model for further analysis as shown in Figure 1.

This is a three-order tensor model, including Parameter, Location and Time. Parameter order can be divided into concrete water quality parameters, including physical group (*e.g.* Temperature, Conductivity), chemical group (*e.g.* pH,

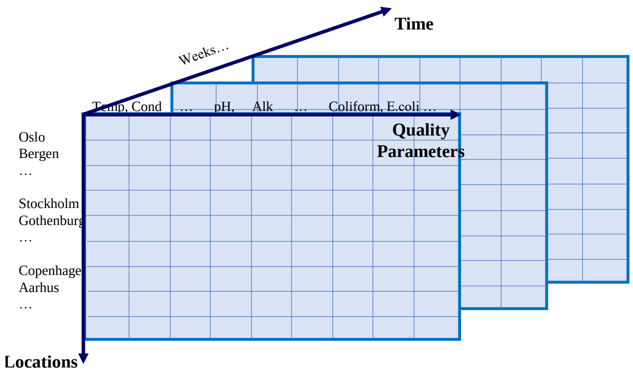


Figure 1: A tensor model for water quality analysis

Alkalinity), and biological group (*e.g.* Coliform, Escherichia coli (E.coli)). Location order is separated into cities\towns according to the requirements. For Time order, the present recordings are based on weekly registrations. Therefore, We use *weeks* to show the water quality evolution. An obvious advantage of this model is easily extending for diverse analysis concerns from both research and industry.

B. Prediction Algorithm

Random forest (RF) is a famous ensemble learning method proposed by Tin Kam Ho in 1995 [8]. It is based on decision tree theory and broadly adopted in both classification and regression process. In this work, we design an algorithm based on RF to predict water quality parameters, as shown in Algorithm 1. In this algorithm, we define the physical and chemical water quality parameters as input and biological parameters as output. This is based on biological parameters normally take much longer time to test (e.g. 24-48 hours) and they are direct threats to human's health. Therefore, with this algorithm, we can take the efficient test results of physical and chemical parameters to evaluate harmful bacteria risks.

This algorithm takes each city's recordings. Firstly separates them into training sets and testing sets. Then train the data to generate a regression model. After this model will be tested and evaluated. The results for each city will be evaluated separately.

Data: Raw Data Collected from Water Sources
Result: Biological parameter predictions
- Normalization;
while *Each city* **do**
  - Separate data into training and testing sets;
  while *training set* **do**
    - Train the random forest regression model for samples;
    - Generate regression model;
  end
  while *testing set* **do**
    - Calculate the error between expected outputs and actual outputs;
    - Evaluate the regression results;
  end
  - Evaluate the city results.
end

Algorithm 1: Random Forest algorithm for water quality predictions

III. Case Study

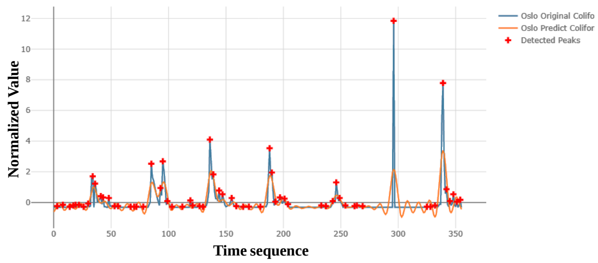
The data we used for this case is collected from Oslo Maridalsvannet lake. It serves as the main water source for this biggest Norwegian city. The overall data set covers water quality parameter testing results on a weekly basis from 2009 to 2015. Here we conduct our experiment for biological parameters for prediction. There are four different bacteria included, as Coliform, Escherichia coli, Intestinal enterococci, and Clostridium perfringens. For the input parameters, we collected Temperature, Conductivity, Turbidity, Color, pH, and Alkalinity as physical and chemical parameters. This complies with the Norwegian National Standard for water quality [9].

The prediction results are shown in Figure 2. This figure shows several interesting findings.

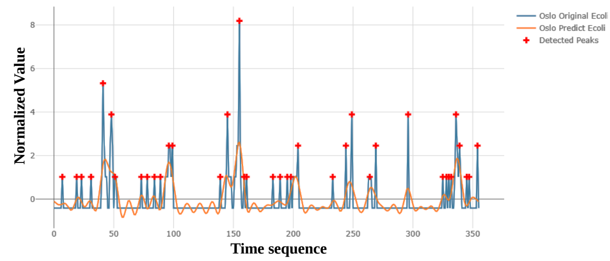
- This model and the corresponding prediction algorithm can generally predict the tendency of biological parameter changes.
- For the peak value predictions, as the red cross points in the figure, the accuracy is not satisfactory.
- We have evaluated the general accuracy using RMSE, it ranges from 0.22 to 0.55. Coliform has better prediction accuracy. Clostridium perfringens is the worst. The reason behind needs to be interpreted by domain experts
- The prediction time efficiency is in average 5.26 seconds. This fulfills the requirements of the DWS systems.

IV. Conclusion

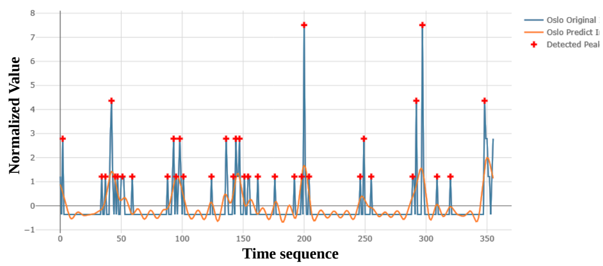
This paper presents a new high-order tensor model for water quality prediction in industrial DWS systems. We first build a three-order tensor model considering water quality parameter, location, and time domains. This model is easily extended on these domains. Secondly, we use a random forest algorithm to predict the most dangerous biological water quality parameters based on easily tested physical and chemical parameters. For a case study, we applied our method in an industrial DWS system in Oslo, Norway. The results show our method is feasible and fulfills the requirements from the domain.



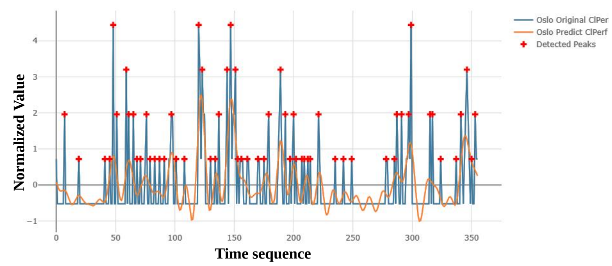
(a) Coliform



(b) Escherichia coli



(c) Intestinal enterococci



(d) Clostridium perfringens

Figure 2: Prediction results of biological parameters in Oslo

Acknowledgment

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

References

[1] W. H. Organization, *Guidelines for drinking-water quality: recommendations*. World Health Organization, 2004, vol. 1.

[2] 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 and Bioelectronics*, vol. 64, pp. 625–632, 2015.

[3] H. R. Maier and G. C. Dandy, "The use of artificial neural networks for the prediction of water quality parameters," *Water resources research*, vol. 32, no. 4, pp. 1013–1022, 1996.

[4] H. Orouji, O. Bozorg Haddad, E. Fallah-Mehdipour, and M. Mariño, "Modeling of water quality parameters using data-driven models," *journal of environmental engineering*, vol. 139, no. 7, pp. 947–957, 2013.

[5] O. Bozorg-Haddad, S. Soleimani, and H. A. Loáiciga, "Modeling water-quality parameters using genetic algorithm–least squares support vector regression and genetic programming," *Journal of Environmental Engineering*, vol. 143, no. 7, p. 04017021, 2017.

[6] 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 management*, vol. 30, no. 7, pp. 2195–2211, 2016.

[7] 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," *Journal of environmental management*, vol. 151, pp. 87–96, 2015.

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

[9] 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.