



ISSN: 2321-2152

IJMECE

*International Journal of modern
electronics and communication engineering*

E-Mail

editor.ijmece@gmail.com

editor@ijmece.com

www.ijmece.com

PREDICTING URBAN WATER QUALITY WITH UBIQUITOUS DATA - A DATA-DRIVEN APPROACH

¹ Mr.K. LAKSHMANA REDDY, ² JAHNAVI DEVARAKONDA

¹Associate Professor, S.V.K.P & Dr K.S. Raju Arts & Science College(A), Penugonda,

W.G.District, Andhra Pradesh, klreddy@gmail.com

²PG, scholar, S.V.K.P & Dr K.S. Raju Arts & Science College(A) Penugonda, W.G.District,

Andhra Pradesh, jahnavidevarakonda1516@gmail.com

ABSTRACT

Urban water quality is of great importance to our daily lives. Prediction of urban water quality help control water pollution and protect human health. However, predicting the urban water quality is a challenging task since the water quality varies in urban spaces non-linearly and depends on multiple factors, such as meteorology, water usage patterns, and land uses. In this work, we forecast the water quality of a station over the next few hours from a data-driven perspective, using the water quality data and water hydraulic data reported by existing monitor stations and a variety of data sources we observed in the city, such as meteorology, pipe networks, structure of road networks, and point of interests (POIs). First, we identify the influential factors that affect the urban water quality via extensive experiments. Second, we present a multi-task multi-view learning

method to fuse those multiple datasets from different domains into an unified learning model. We evaluate our method with real-world datasets, and the extensive experiments verify the advantages of our method over other baselines and demonstrate the effectiveness of our approach.

1.INTRODUCTION

Urban water is a vital resource that affects various aspects of human, health and urban lives. People living in major cities are increasingly concerned about the urban water quality, calling for technology that can monitor and predict the water quality in real time throughout the city. Urban water quality, which serves as “a powerful environmental determinant” and “a foundation for the prevention and control of waterborne diseases” [1], refers to the physical, chemical and biological characteristics of a water body, and several

chemical indexes (such as residual chlorine, turbidity and pH) can be used as effective measurements for the water quality in current urban water distribution systems [2].

With the increasing demand for water quality information, several water quality monitoring stations have been deployed throughout the city's water distribution system to provide the real-time water quality reports in a city. Figure 1 illustrates the water quality monitor stations that have been deployed in Shenzhen, China. Besides water quality monitoring, predicting the urban water quality plays an essential role in many urban aquatic projects, such as informing waterworks' decision making (e.g., pre-adjustment of chlorine from the waterworks), affecting governments' policy making (e.g., issuing pollution alerts or performing a pollution control), and providing maintenance suggestions (e.g., suggestions for replacements of certain pipelines).

Predicting urban water quality, however, is very challenging due to the following reasons. First, urban water quality varies by locations non-linearly and depends on multiple factors, such as meteorology, water usage patterns, land use, and urban structures. As depicted in Figure 1, the water quality indexes (RC) reported by the three stations demonstrate different

patterns. Existing hydraulic model-based approaches try to model water quality from physical and chemical perspective, but such hydraulic model can hardly capture all of those complex factors. Moreover, the parameters in model are hard to get, which make it difficult to extend to other water distribution systems. Second, as all the stations are connected through the pipeline system, the water quality among different stations are mutually correlated by several complex factors, such as attributes in pipe networks and distribution of POIs. Traditional hydraulic model-based approaches build hydraulic model for each station and ignore their spatial correlations, and thus their performance is far from satisfactory. Hence, besides identifying the influential factors, how to efficiently characterize and incorporate such relatedness poses another challenge.

Fortunately, in the era of big data [3] [4] [5], unprecedented data in urban areas (e.g., meteorology, POIs, and road networks) can provide complementary information to help predict the urban water quality. For example, temperature can be an indicator of water quality, with higher temperature indicating better water quality. The possible reason is that the water consumption tends to grow when temperature is high since most people may choose to take a shower,

and the increased water consumption is one major cause that prevents the water quality's deterioration in the distribution systems.

To benefit from the unprecedented data in urban areas, in this paper, we predict the water quality of a station through a data-driven perspective using a variety of data sets, including water quality data, hydraulic data, meteorology data, pipe networks data, road networks data, and POIs. First, we perform extensive experiments and data analytics between the water quality and multiple potential factors, and identify the most influential ones that have an effect on the urban water quality. Second, we present a novel spatio-temporal multi-task multi-view learning (stMTMV) framework to fuse the heterogeneous data from multiple domains and jointly capture each station's local information as well as their global information into an unified learning model [6].

We summarize the contributions as follows:

- _ **Data-driven Perspective:** We present a novel data-driven approach to co-predict the future water quality among different stations with data from multiple domains. Additionally, the approach is not restricted to urban water quality prediction, but also can be applied to other multi-locations

based coprediction problem in many other urban applications.

- _ **Influential Factor Identification:** We identify spatially-related (such as POIs, pipe networks, and road networks) and temporally-related features (e.g., time of day, meteorology and water hydraulics), contributing to not only our application but also the general problem of water quality prediction.

- _ **Unified Learning Model:** We present a novel spatio-temporal multi-view multi-task learning framework (stMTMV) to integrate multiple sources of spatio-temporal urban data, which provides a general framework of combining heterogeneous spatio-temporal properties for prediction, and can also be applied to other spatio-temporal based applications.

- _ **Real evaluation:** We evaluate our method by extensive experiments that use real-world datasets in Shenzhen, China. The results demonstrate the advantages of our method beyond other baselines, such as ARMA, Kalman filter, and ANN, and reveal interesting discoveries that can bring social good to urban life.

The rest of this paper is organized as follows: Section 2 overviews the framework of our method. Section 3 and 4 analyze the correlations between multi-sources of urban data and the water quality.

Section 5 introduces the multi-task multi-view learning method for urban water quality prediction, and Section 6 presents evaluations and visualizations. Section 7 summarizes the related work, followed by the conclusion in the last section.

As an extension of our previous work [6], this journal version claims following contributions: First, we focused on the data driven perspective. Specifically, we included the insight of our methodology as well as the correlation analysis between different data with the urban water quality. The detailed correlation analysis is shown in Section 3 and 4. Second, we refined the task relationship computation in our STMTMV model by figuring out the best configuration over various pipe attributes, which is achieved through the data correlation analysis in Section 5.4.1. Third, we conducted more comprehensive experiments to validate our system. For instance, we added another two popular algorithms (Kalman, ANN) as the time series prediction baselines in Section 6.3. In addition, we compared the performance of our approach with other baselines over each individual station in Section 6.6.

2.LITERATURE SURVEY

The literature on predicting urban water quality with ubiquitous data reflects a

growing recognition of the potential of data-driven approaches to address challenges in water quality monitoring and management. One key aspect highlighted in the literature is the role of sensor networks and IoT devices in collecting real-time data at high spatial and temporal resolutions. These technologies enable continuous monitoring of water bodies, providing insights into dynamic changes in water quality parameters and facilitating early detection of pollution events or anomalies.

Remote sensing imagery, including satellite and aerial data, is another valuable source of information for assessing water quality on a larger scale. Studies have explored the use of remote sensing techniques to estimate water quality indicators such as chlorophyll-a concentration, turbidity, and water clarity, thereby enabling the monitoring of water bodies over extensive geographic areas.

Moreover, the integration of social media data and citizen science initiatives has emerged as a novel approach to supplement traditional monitoring efforts. By harnessing user-generated content and crowdsourced data, researchers can gather valuable information on water-related activities,

pollution incidents, and public perceptions of water quality, contributing to more comprehensive and participatory monitoring systems.

In terms of predictive modeling, machine learning algorithms have demonstrated promising results in predicting water quality parameters based on diverse datasets. These models can capture complex relationships between environmental variables and water quality indicators, allowing for accurate predictions and early warning systems to support decision-making processes.

Despite these advancements, challenges remain, including issues related to data quality, interoperability, and scalability.

Addressing these challenges requires interdisciplinary collaborations among researchers, policymakers, and stakeholders to develop standardized data collection protocols, enhance data sharing mechanisms, and improve model validation techniques.

3.EXISTING SYSTEM

Several studies in the environmental science have been tried to analyze the water quality problems via data-driven based approaches, and those studies covers a

range of topics, from the physical process analysis in the river basin, to the analysis of concurrent input and output time series [64] [65]. The approaches adopted in these studies include instance-based learning models (e.g., kNN) as well as neural network models (e.g., ANN). In general, those data-driven approaches in the environmental science can fall into the following three major categories: Instance-based Learning models (IBL), Artificial Neural Network models (ANN) and Support Vector Machine models (SVM).

Instance-based learning models (IBL) is a family of learning algorithms that model a decision problem with instances or examples of training data that are deemed important to test model [66]. As a typical example of IBL, k-Nearest Neighbors (k-NN) is widely used due to its simplicity and incredibly good performance in practice.

For example, the work introduced by Karlsson et al. [67] addressed the classical rainfall-runoff forecasting problem by k-NN algorithm, and demonstrated promising results. Toth et al. [68] used k-NN to predict the rainfall depths from the history data, and showed the persistent outperformance of k-NN over other time series prediction methods.

As another example, Ostfeld et al. [69] developed a hybrid genetic k-Nearest Neighbor algorithm to calibrate the two-dimensional surface quantity and water quality model. Artificial Neural Network (ANN) is a network inspired by biological neural networks (in particular the human brain), which consists of multiple layers of nodes (neurons) in a directed graph with each layer fully connected to the next one [65]. Neural networks have been widely employed to solve a wide variety of tasks, and can achieve good results. For instance, Moradkhani et al. [70] proposed an hourly streamflow forecasting method based on a radial-basis function (RBF) network and demonstrated its advantages over other numerical prediction methods. Also, the work introduced by Kalin [44] predicted the water quality indexes in watersheds through ANN.

Support Vector Machines (SVMs) are typical supervised learning models that analyze data used for classification and regression [71].

In aquatic studies, it was also extended to solving prediction problems [64]. For instance, Liong et al. [72] addressed the issue of flood forecasting using Support Vector Regression (SVR) which is an extension of SVM. Another work by Xiang et al. [73] utilized a LS-SVM model to deal

with the water quality prediction problem in Liuxi River in Guangzhou.

However, none of these approaches is applied into urban scenarios, which is quite different from our applications. Moreover, those existing approaches process the data from a single source, and can hardly integrate the data from different sources. Thus, their applications in the urban scenarios are restricted.

Disadvantages

- ❖ The system is implemented only Multi-task Multi-view Learning Approaches.
- ❖ Instance-based learning models (IBL) is a family of learning algorithms that model a decision problem with instances or examples of training data that are deemed important to the model.

4.PROPOSED SYSTEM

— Data-driven Perspective: We present a novel data-driven approach to co-predict the future water quality among different stations with data from multiple domains. Additionally, the approach is not restricted to urban water quality prediction, but also can be applied to other multi-locations based coprediction problem in many other urban applications.

— Influential Factor Identification: We identify spatially-related (such as POIs, pipe networks, and road networks) and temporally-related features (e.g., time of

day, meteorology and water hydraulics), contributing to not only our application but also the general problem of water quality prediction. _ Unified Learning Model: We present a novel spatio-temporal multi-view multi-task learning framework (stMTMV) to integrate multiple sources of spatio-temporal urban data, which provides a general framework of combining heterogeneous spatio-temporal properties for prediction, and can also be applied to other spatio-temporal based applications.

_ Real evaluation: We evaluate our method by extensive experiments that use real-world datasets in Shenzhen, China. The results demonstrate the advantages of our method beyond other baselines, such as ARMA, Kalman filter, and ANN, and reveal interesting discoveries that can bring social good to urban life.

Advantages

1) Water quality data: We collect water quality data every five minutes from 15 water quality monitoring stations in Shenzhen City. It comprises residual chlorine (RC), turbidity (TU) and pH. In this paper, we only use RC as the index for water quality, since RC is the most important and effective measurement for water quality in current urban water distribution system.

2) Hydraulic data: Hydraulic data consists of flow and pressure, which are collected

every five minutes from 13 flow sites and 14 pressure sites, respectively.

5.ARCHITECTURE:

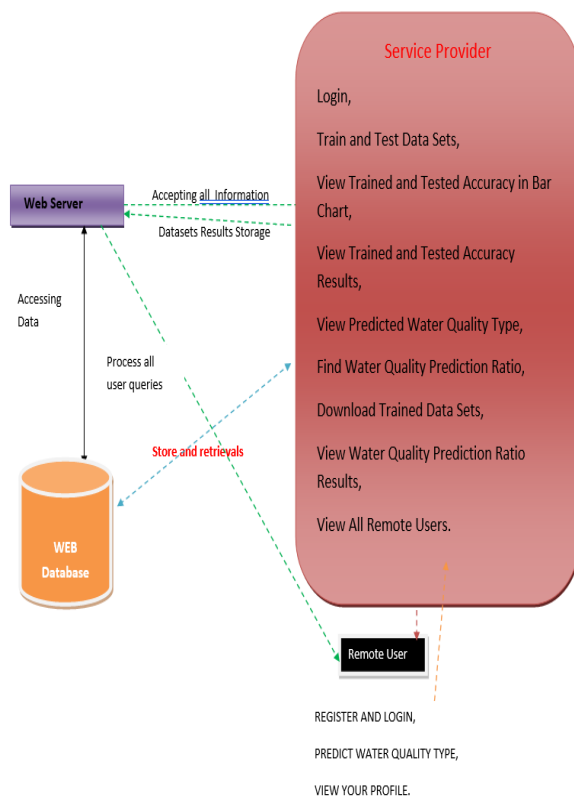
System Architecture mainly consists of 2 modules and database to store all the data. Those are:

- Remote User
- Service provider

The Remote User module can perform the following operation: Register and login, view your profile, Predict Recommendation Type

The Service provider module can perform the following operations:

Login, Browse and Train & Test data sets, View Trained And tested Accuracy in Bar Charts, view Trained and Tested Accuracy Results, View Predicted Water Quality Type Ratio, Download Trained data sets, view Water Quality ratio results.



6.MODULES

Service Provider:

In this module, the Service Provider has to login by using valid user name and password. After login successful he can do some operations such as

Login, Train and Test Data Sets, View Trained and Tested Accuracy in Bar Chart, View Trained and Tested Accuracy Results, View Predicted Water Quality Type, Find Water Quality Prediction Ratio, Download Trained Data Sets, View Water Quality Prediction Ratio Results, View All Remote Users.

View and Authorize Users:

In this module, the admin can view the list of users who all registered. In this, the admin can view the user's details such as, user name, email, address and admin authorizes the users.

Remote User:

In this module, there are n numbers of users are present. User should register before doing any operations. Once user registers, their details will be stored to the database. After registration successful, he has to login by using authorized user name and password. Once Login is successful user will do some operations like REGISTER AND LOGIN, PREDICT WATER QUALITY TYPE, VIEW YOUR PROFILE.

7.OUTPUT SCREENS:

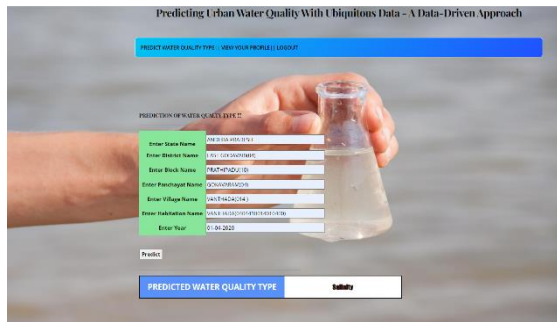
Login Screen:



Register Screen:



Prediction Screen:



Predicting Urban Water Quality With Ubiquitous Data - A Data-Driven Approach

PREDICT WATER QUALITY TYPE - VIEW YOUR PROFILE / LOGOUT

PRODUCTION OF WATER QUALITY TYPE

Enter State Name: MAHARASHTRA
 Enter District Name: PUNE
 Enter Block Name: KATRAJ
 Enter Ward Name: KATRAJ
 Enter Village Name: KATRAJ
 Enter Year: 2022

Predict

PREDICTED WATER QUALITY TYPE: **Healthy**

Admin Login:



Predicting Urban Water Quality With Ubiquitous Data - A Data-Driven Approach

ADMIN LOGIN

Enter Username:
 Enter Password:
 Login

Train & Test Data Screen:



Predicting Urban Water Quality With Ubiquitous Data - A Data-Driven Approach

Train & Test Data Screen

State	District	Block	Ward	Village	Year	Water Quality Type
MAHARASHTRA	PUNE	KATRAJ	KATRAJ	KATRAJ	2022	Healthy

Accuracy Bar Chart Screen:



View Profile Screen:



Predicting Urban Water Quality With Ubiquitous Data - A Data-Driven Approach

VIEW ALL PROFILE DATA

State	District	Block	Ward	Village	Year	Water Quality Type
MAHARASHTRA	PUNE	KATRAJ	KATRAJ	KATRAJ	2022	Healthy

8.CONCLUSION:

This paper presents a novel data-driven approach to forecast the water quality of a station by fusing multiple sources of urban data. We evaluate our approach based on Shenzhen's water quality and various urban data. The experimental results demonstrate the effectiveness and efficiency of our approach. Specifically, our approach outperforms the traditional RC decay model [2] and other classical time series predictive models (ARMA, Kalman) in terms of RMSE metric. Meanwhile, as our approach consists of two components, each of the components demonstrates its effectiveness through extensive experiments and analysis. In particular, the first component is the influential factors identification, which explores the factors that affect the urban water quality via extensive experiments and analysis in Section 3 and 4. The second one is a spatiotemporal multi-view multi-task learning (STMTMV) framework that consists of multi-view learning and multi-task learning. The experiments have shown that STMTMV has a predictive accuracy of around 85% for

forecasting next 1-4 hours, which outperforms the single-task methods (LR) by approximately 11% and the single-view methods (t-view and s-view) by approximately 11% and 12%, respectively. The code has been released at: <https://www.microsoft.com/en-us/research/publication/urbanwater-quality-prediction-based-multi-task-multi-view-learning-2/> In future, we plan to deal with the water quality inference problems in the urban water distribution systems through a limited number of water quality monitor stations.

9. REFERENCES

- [1] W. H. Organization, Guidelines for drinking-water quality, 2004, vol. 3.
- [2] L. A. Rossman, R. M. Clark, and W. M. Grayman, "Modeling chlorine residuals in drinking-water distribution systems," *Journal of environmental engineering*, vol. 120, no. 4, pp. 803–820, 1994.
- [3] Y. Zheng, "Methodologies for cross-domain data fusion: An overview," *IEEE Transactions on Big Data*, vol. 1, no. 1, pp. 16–34, 2015.
- [4] Y. Zheng, L. Capra, O. Wolfson, and H. Yang, "Urban computing: Concepts, methodologies, and applications," *ACM Transactions on Intelligent Systems and Technology*, vol. 5, no. 3, pp. 38:1–38:55, 2014.
- [5] Y. Zheng, H. Zhang, and Y. Yu, "Detecting collective anomalies from multiple spatio-temporal datasets across different domains," 2015.
- [6] Y. Liu, Y. Zheng, Y. Liang, S. Liu, and D. S. Rosenblum, "Urban water quality prediction based on multi-task multi-view learning," in *Proceedings of the International Joint Conference on Artificial Intelligence*, 2016.
- [7] H. Cohen, "Free chlorine testing," <http://www.cdc.gov/safewater/chlorineresidual-testing.html>, 2014, accessed on 5 August 2016.
- [8] B. D. Barkdoll and H. Didigam, "Effect of user demand on water quality and hydraulics of distribution systems," in *Proceedings of the World Water and Environmental Resources Congress*, 2003.
- [9] P. Castro and M. Neves, "Chlorine decay in water distribution systems case study–lousada network," *Electronic Journal of Environmental, Agricultural and Food Chemistry*, vol. 2, no. 2, pp. 261–266, 2003.
- [10] L. W. Mays, *Water distribution system handbook*, 1999.

- [11] L. A. Rossman and P. F. Boulos, "Numerical methods for modeling water quality in distribution systems: A comparison," *Journal of Water Resources planning and management*, vol. 122, no. 2, pp. 137–146, 1996.
- [12] W. M. Grayman, R. M. Clark, and R. M. Males, "Modeling distributionsystem water quality: dynamic approach," *Journal of Water Resources Planning and Management*, vol. 114, no. 3, pp. 295–312, 1988.
- [13] J. Lin, E. Keogh, S. Lonardi, and B. Chiu, "A symbolic representation of time series, with implications for streaming algorithms," in *Proceedings of the ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*, 2003, pp. 2–11.
- [14] G. Luo, K. Yi, S.-W. Cheng, Z. Li, W. Fan, C. He, and Y. Mu, "Piecewise linear approximation of streaming time series data with max-error guarantees," in *Proceedings of the IEEE International Conference on Data Engineering*, 2015, pp. 173–184.
- [15] E. O. Brigham and E. O. Brigham, *The fast Fourier transform*. Prentice-Hall Englewood Cliffs, NJ, 1974, vol. 7.
- [16] C. S. Burrus, R. A. Gopinath, and H. Guo, "Introduction to wavelets and wavelet transforms: a primer," 1997.
- [17] M. W. LeChevallier, T. Evans, and R. J. Seidler, "Effect of turbidity on chlorination efficiency and bacterial persistence in drinking water." *Applied and environmental microbiology*, vol. 42, no. 1, pp. 159–167, 1981.
- [18] L. Monteiro, D. Figueiredo, S. Dias, R. Freitas, D. Covas, J. Menaia, and S. Coelho, "Modeling of chlorine decay in drinking water supply systems using epanet msx," *Procedia Engineering*, vol. 70, pp. 1192–1200, 2014.
- [19] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 68, no. 1, pp. 49–67, 2006.
- [20] W. Zhang, K. Zhang, P. Gu, and X. Xue, "Multi-view embedding learning for incompletely labeled data," in *Proceedings of the International Joint Conference on Artificial Intelligence*, 2013, pp. 1910–1916.
- [21] Y. Liu, L. Zhang, L. Nie, Y. Yan, and D. S. Rosenblum, "Fortune teller: Predicting your career path," in *Proceedings of the Thirtieth AAAI*

Conference on Artificial Intelligence, 2016, pp. 201–207.

[22] S. Sun, “A survey of multi-view machine learning,” Neural Computing and Applications, vol. 23, no. 7-8, pp. 2031–2038, 2013.

[23] C. Xu, D. Tao, and C. Xu, “A survey on multi-view learning,” arXiv preprint arXiv:1304.5634, 2013.

[24] J. Zhang and J. Huan, “Inductive multi-task learning with multiple view data,” in Proceedings of the ACM International Conference on Knowledge Discovery and Data Mining, 2012, pp. 543–551.

[25] L. Zhao, Q. Sun, J. Ye, F. Chen, C.-T. Lu, and N. Ramakrishnan, “Multitask learning for spatio-temporal event forecasting,” in Proceedings of the ACM International Conference on Knowledge Discovery and Data Mining, 2015, pp. 1503–1512.