JOURNAL OF WATER AND LAND DEVELOPMENT
e-ISSN 2083-4535

Polish Academy of Sciences (PAN) Institute of Technology and Life Sciences – National Research Institute (ITP – PIB)

JOURNAL OF WATER AND LAND DEVELOPMENT
DOI: 10.24425/jwld.2023.147223
2023, No. 59 (X–XII): 8–12

The correlation of water quality parameters over wireless sensors generated dataset in the Sitnica River in Kosovo

Figene Ahmedi ✉, Shkumbin Makolli

The University of Prishtina, Faculty of Civil Engineering, Hydrotechnic Department,
Rr. Agim Ramadani, ndërtesa e "Fakultetit Teknik", 10000 Prishtina, Kosovo

RECEIVED 02.03.2023 ACCEPTED 29.05.2023 AVAILABLE ONLINE 10.11.2023

Abstract: In this paper, the regression analysis technique is applied to a large water quality dataset for the Sitnica River in Kosovo. It has been done to assess the correlation between water quality parameters. The data are generated by a wireless sensors network deployed in Sitnica. A regression analysis is applied to four water quality parameters: temperature, dissolved oxygen, pH, and electrical conductivity. The correlation between each pair of parameters has been assessed by using the WEKA software package, which is a popular time-saving tool for data analysis in distinct domains. The data are pre-processed to exclude out-of-range values and then the assessment of correlation for the pairs of parameters is applied. In comparison to other pairs of water quality parameters, the results show that dissolved oxygen and electrical conductivity correlate particularly closely with temperature. Regression equations of these two pairs of parameters may provide inferred information on dissolved oxygen and electrical conductivity about the Sitnica River. Such information may otherwise not be available to resource managers in Kosovo. Moreover, due to its easy to use and availability as an open-source software, WEKA may aid decision-makers on the management providing almost real-time information about surface water quality within the basin. This can be particularly useful especially in the case of continuous observation of water quality and a huge dataset gathered by using wireless sensors.

Keywords: monitoring, parameters' pair, regression analysis, water quality, wireless sensors

INTRODUCTION

Depending on land use in the catchment area, water quality for the intended use is defined by physical, chemical, and biological characteristics of water (Sperling von, 2015; Malik and Szwilski, 2016). Water characteristics change due to rapid urbanisation, industrial development, and large-scale human activity, which actually degrade water resources. The use of such degraded water resources for a particular application is limited. Hence, ensuring water quality is no doubt an important issue to maintain a sustainable living conditions for all (Gorchev and Ozolins, 1984 as cited in Pule, Yahya and Chuma (2017), p. 464). The provision of clean water has been listed as an important goal among the 17 Sustainable Development Goals drafted by the United Nations (UN, 2015). The assessment of water quality can be done by monitoring water resources.

Water quality monitoring through analysed and reported observations and measurements, provides information about catchments and waterways (GWA-DW, 2009). In the past, water quality monitoring programs involved collecting data rather than analysing them to obtain information regarding water quality. According to Ward, Loftis and McBride (1986), this is defined as a "data-rich but information poor" syndrome. To avoid this syndrome, as well as to make monitoring systems more efficient, a balanced combination of data collection (initial product) and information generation (final product) should be provided in relation to water quality monitoring systems (Maasdam, 2000). Generation of information based on data collected is a helpful tool to ensure efficient management of water resources and the protection of aquatic life (Varol, Gökot and Bekleyen, 2010). Water quality monitoring involves the detection of physical, chemical, and biological characteristics of water (Pule, Yahya and

Chuma, 2017), and the identification of spatial changes or trends in water quality over time (DHI Gras Solution, 2000).

Water analysis is performed by taking physical samples of water resources and sample analysis in the laboratory. This is one of the most common methods of water quality monitoring. Recently, advanced technologies are used to facilitate the monitoring of water quality (O'Flynn *et al*., 2007). Water quality monitoring that applies a conventional "stream to a bottle" monitoring technique in centralised laboratories is limited in time and space, whereas the use of advanced wireless sensor networks (WSN) has become increasingly popular among researchers and in the market (Ahmedi *et al*., 2018). Monitoring water quality using wireless sensor networks (WSNs) can provide temporal and spatial data as frequency required to pick up water quality variability, which is missed when using conventional grab sampling approaches (O'Flynn *et al*., 2010).

There are several parameters to consider when observing water quality. Sensor technology is rapidly evolving, and parameters such as electrical conductivity (*EC*), temperature, pH, dissolved oxygen (*DO*), oxidation-reduction potential (*ORP*) can already be measured by sensors in the field and provide information about water quality (Faustine and Mvuma, 2014; Manjarrés *et al*., 2016). Water quality monitoring via WSs does not mean that it is necessary to collect as much data as possible. It provides the possibility of analysing such data in real-time and providing information about water quality. The analysis of a large volume of data can be developed through various software packages using statistical analysis techniques.

The application of different statistical techniques, such as regression analysis (RA), cluster analysis (CA), principal component analysis (PCA), and factor analysis (FA) helps to interprete complex data matrices to better understand water quality of water resources studied. The focus of the paper is to assess the relationship between each pair of water quality parameters over a large dataset generated by wireless sensors. The application of the regression analysis technique helps to interprete the relationship between one continuous variable (response variable) and one other variable (explanatory variable) (Helsel *et al*., 2020). The nature and magnitude of the relationship among various physicochemical parameters of stream water through regression analysis (RA) were explained by Bhat *et al*. (2014). The correlation test defining whether water quality parameters are correlated, in order to assess water quality of the Ganga River in Kanpur, India was carried out by Khatoon *et al*. (2013). Khatoon *et al*. (2013) emphasise that the correlation helps to identify parameters that can be measured frequently, so that the status of water, in terms of its quality, can be defined regularly. Statistical calculations, including regression analysis techniques, are used by many authors to interpret data related to the quality of water. These are mainly performed using Microsoft Office Excel, Statistica, Minitab 14, SPSS 14, 16, and 17, and PAST package programs (El-Korashey, 2009; Pejman *et al*., 2009; Bhat *et al*., 2014).

Nowadays, the WEKA open-source data mining software package for data analysis is used more frequently. The tool is widely used for data pre-processing, classification, regression, clustering, association rules, and visualisation (Salah *et al*., 2014). Many works highlight the potential of using the WEKA package for various data mining tasks on water quality classification. Bisht *et al*. (2018) developed a model which can be used to classify water quality of the Ganga River in India. It uses decision trees implemented in the WEKA tool, but over a dataset of only 900 records, and a traditional analysis method, i.e. in a laboratory. Gakii and Jepkoech (2019) present a classification model using also decision trees to predict whether drinking water is clean in different regions of Kenya. Muhammad *et al*. (2015) analyse and compare the performance of various classification models and algorithms through WEKA to identify significant features that contribute to the classification of water quality in the Kinta River in Perak Malaysia. Salah, Mocanu and Florea (2014) use classification decision trees (which describe data but not decisions) through the WEKA mining tool to assess water quality for the Tigris River within the Baghdad City. Sasikala (2017) compares several data mining tools, including WEKA, on their performance and results for the availability of water resources in different areas in Trichy, India. The study shows that WEKA takes less time to generate results, and it has high variation in settings in support of data analysis compared to other tools.

In our work presented in this paper, a rather larger dataset of 92,292 records gathered remotely through wireless sensors is analysed for correlation. Moreover, the case study involves observing water quality parameters, i.e. their correlation in the Sitnica River in Kosovo, relevant to local decision-makers when it comes to real-time water quality management.

MATERIALS AND METHODS

SITE SELECTION

The Sitnica River basin is located in the central part of Kosovo. The catchment area is about 2912 km^2, and the Sitnica River is 78 km long. The monitoring network covers five sampling sites controlled by the Hydro-Meteorological Institute of Kosovo – HMIK (Shqip: Institutit Hidrometeorologjik të Kosovës). For water quality monitoring, the HMIK uses methods that are mainly based on conventional grab samples and lab analyses. The water quality monitoring provided by the HMIK is characterised by low frequency (once a month) and small spacial deployment (only five measuring stations). Monitoring through a wireless sensors (WSs) installed in coordinates (42.70670319, 21.03843117) in the Sitnica River has enabled far more frequent and real-time quality monitoring and a large collection of data generated from the sensing system. The WS network system is implemented as part of the "InWaterSense: Intelligent Wireless Sensor Networks Monitoring Surface Water Quality" project, an EU funded project managed by European Union Office in Kosovo, and implemented by University of Prishtina (https://inwatersense.uni-pr.edu). Water quality monitoring via the WSs is conducted in the village of Plemetin (Fig. 1), near the capital city of Kosovo, Prishtina, as a monitoring point that has been selected given the site accessibility and security.

DATA PREPARATION

Wireless sensors (WSs) presented in this paper support real-time measurement of the following parameters: temperature (*T*) in °C, dissolved oxygen (*DO*) in mg·dm^{-3}, pH, and electrical conductivity (*EC*) in µS·cm^{-1}. The data collection frequency is configured every 10 min. Water quality monitoring through

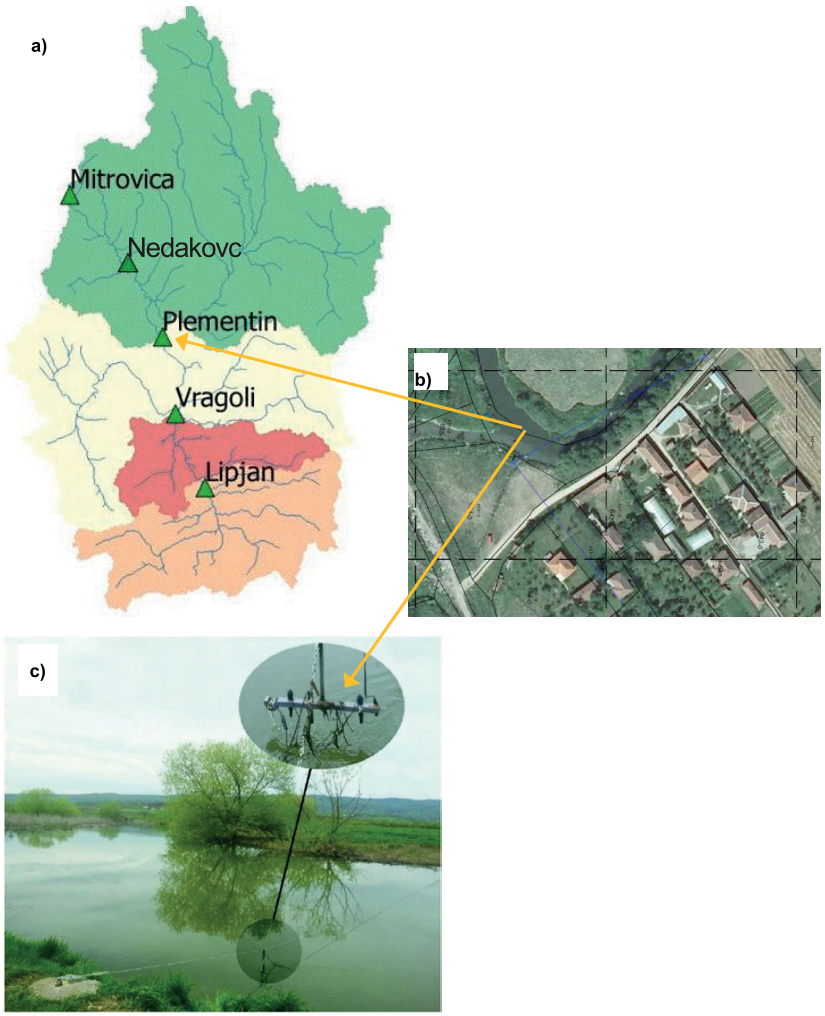

Fig. 1. Water quality monitoring site through wireless sensors: a) Sitnica River basin, b) monitoring site in the Sitnica River, c) wireless sensors deployed in the monitoring site; source: own elaboration based on Ahmedi *et al.* (2018)

wireless sensors lasted for eight months, from the beginning of May 2015 to the beginning of January 2016. This enabled to collect 92,292 water quality data (i.e. 23,073 per each of four parameters), including cases with missing values for certain parameters. For the correlation analysis, raw data are pre-processed to exclude out-of-range values. Out-of-range values are considered measurement ranges as defined by the sensor manufacturer. Additionally, missing values are removed, thus resulting in 18,500 out of 23,073 for each parameter. Pairs of water quality parameters are selected to assess the correlation. The regression analysis is applied to evaluate the correlation for the pairs as follows: electrical conductivity and temperature (*EC*&*T*), pH and temperature (pH&*T*), dissolved oxygen and temperature (*DO*&*T*), electrical conductivity and pH (*EC*&pH), electrical conductivity and dissolved oxygen (*EC*&*DO*), and dissolved oxygen and pH (*DO*&pH).

In this study, the calculation of correlation between each pair of parameters is performed using the WEKA 3.8.4 data mining software tool, as an open-source solution supporting the whole data mining life-cycle starting from data pre-processing, through classification, regression, clustering, or association rules, to visualisation. At the input, WEKA supports files in the CSV format, and hence CSV files with data are used to feed WEKA.

RESULTS AND DISCUSSION

The results of linear regression equations as well as potential correlations between each pair of parameters, temperature (*T*), dissolved oxygen (*DO*), pH, and electrical conductivity (*EC*), using the WEKA mining tool are shown below (Tab. 1).

Table 1. List of statistics related to linear regression for each pair of parameters

Pair of parameters	Regression equation	Correlation coefficient, *r*	Coefficient of determination, *R*
EC&*T*	12.2861*T* + 156.0011	0.717	0.51
pH&*T*	–0.0041*T* + 6.0208	–0.0019	0.00
DO&*T*	–5.0918*T* + 105.5266	0.7982	0.64
EC&pH	9.7442pH + 279.6085	0.1981	0.04
EC&*DO*	–1.2776*DO* + 373.6418	0.4942	0.25
DO&pH	5.5643pH + 0.1223	0.3093	0.10

Explanations: *EC* = electrical conductivity, *T* = temperature, pH = potential of hydrogen, *DO* = dissolved oxygen.
Source: own study.

The regression analysis carried out for water quality parameters shows different correlations between dependent and independent variables. The observed relationship seems to be insignificant for all parameters. Referring to the results presented in Table 1, the concentration of most dependent variables decreased as the independent variable increased, yielding a negative correlation.

The very weak negative correlation or completely uncorrelated variables are pH and temperature (pH&*T*). The regression equations (Tab. 1) show that dissolved oxygen has a significant negative correlation with temperature, whereas electrical conductivity has a significant positive correlation with temperature.

The correlation coefficients for dissolved oxygen and temperature, and for electrical conductivity and temperature are $r = 0.7982$ ($p < 0.01$) and $r = 0.717$ ($p < 0.01$), respectively. It can be inferred that temperature is a good predictor for dissolved oxygen and electrical conductivity in water. The coefficient of determination, which in the case of dissolved oxygen and temperature is $R > 0.64$ (as r^2), shows that for more than 64% of cases, the linear trend $-5.0918T + 105.5266$ gives "reasonable" results. More than 51% of cases with electrical conductivity as a dependent variable are explained by temperature. The observed relationship between electrical conductivity and dissolved oxygen explained by the regression equation is $-1.2776DO + 373.6418$ with the correlation coefficient of about 0.5. This shows that these two parameters have a moderate negative correlation. The same applies to dissolved oxygen and pH (*DO*&pH), while a weak relationship is shown to exist between electrical conductivity and pH (*EC*&pH). Since the value of the correlation coefficient between two variables *T*&*DO* is about 0.8. It indicates that a strong correlation between *T*&*DO* exists, as values of *T* increase, values of *DO* decrease. Based on the regression equation for *T*&*DO*, $-5.0918T + 105.5266$, it can be interpreted that the warmer the water, the less *DO* the water can hold, and the colder the water, the more *DO* it retains. The same correlation exists between *EC*&*T*, with a correlation coefficient of about 0.8. According to the regression equation of *EC*&*T*, $12.2861T + 156.0011$; it means that the warmer the water, the higher the conductivity.

In summary, a strong correlation identified to exist between certain pairs of parameters proved the importance of the regression equation for finding the value of one parameter if the value of another is known in the same water body. In the case represented in this paper, namely water quality parameters in the Sitnica River in Kosovo as measured by the wireless sensors, it has been shown that the WEKA software can be used to nearly real-time determination of strongly correlated parameters, i.e. loads of oxygen concentration and electrical conductivity can be estimated continuously once we know temperature values of the observed water body. In terms of the time spent by WEKA to run the analyses, the longest time taken for generating results of the linear regression between each pair of parameters was up to 0.03 s.

CONCLUSIONS

The statistical regression analysis is a highly useful technique in water quality analysis. While providing the analysis, WEKA seems to be very useful tool, especially in the case of large datasets generated by wireless water quality sensors. Thus, the linear regression between each pair of parameters is a useful step toward the management of water quality.

In this study, WEKA is used for the linear regression and six different regression equations are developed in total involving each pair of four water quality parameters, including temperature, dissolved oxygen, pH, and electrical conductivity. Among all candidate equations, the equations of regression for dissolved oxygen and temperature, as well as for electrical conductivity and temperature show that these are the best-correlated pairs of water quality parameters. It is because the ratio of the correlation coefficient is closer to 1 when compared to other pairs of parameters. The regression equations of these two pairs of parameters can be used to continuously estimate the concentration of dissolved oxygen and electrical conductivity based on temperature as an independent variable in the Sitnica River. This information may otherwise not be available by resource managers in Kosovo due to scarce infrastructure (e.g. sensors of certain water quality parameters are not in place). The regression equations presented in this study are site-specific and apply only to the Sitnica River.

It can be also emphasised by this study, that when dealing with a large dataset of water quality parameters as in this use case of continuous water quality monitoring using wireless sensors, the WEKA mining tool may be considered as an efficient time saving option for finding the correlation between parameters. Additionally, WEKA is easy-to-use and broadly available open-source regression analysis software. The use of such a regression analysis function using WEKA is particularly helpful to local decision-makers for nearly real-time management of surface water quality within the basin. It would be hard to otherwise manually oversee and interpret such data. In the future, further research is foreseen to find possible interesting patterns hidden within the data, yet using another approach, i.e. a clustering technique facilitated by WEKA.

FUNDING

This work has partially been supported by "InWaterSense: Intelligent Wireless Sensor Networks for Monitoring Surface Water Quality", an EU funded project (https://inwatersense.uni-pr.edu) managed by European Union Office in Kosovo, Implemented by the University of Prishtina.

REFERENCES

Ahmedi, F. *et al.* (2018) "InWaterSense: An intelligent wireless sensor network for monitoring surface water quality to a river in Kosovo," *International Journal of Agricultural and Environmental Information Systems*, 9(1), pp. 39–61. Available at: https://doi.org/10.4018/IJAEIS.2018010103.

Bhat, S.A. *et al.* (2014) "Statistical assessment of water quality parameters for pollution source identification in Sukhnag Stream: An inflow stream of Lake Wular (Ramsar Site), Kashmir Himalaya," *Journal of Ecosystems*, 2014, pp. 1–18. Available at: https://doi.org/10.1155/2014/898054.

Bisht, A.K. *et al.* (2018) "Development of an automated water quality classification model for the River Ganga," in *Communications in computer and information science*. Springer Science+Business

Media, pp. 190–198. Available at: https://doi.org/10.1007/978-981-10-8657-1_15.

DHI Gras Solution (2000) *Water quality monitoring from space: Baselines and up-to-date information*. [Online]. Available at: https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=-web&cd=&ved=2ahUKEwi-ud-E98yAAxW6JRAIHS-7CKAQF-noECBcQAQ&url=https%3A%2F%2Fwww.dhigroup.com%2F-%2Fmedia%2Fshared%2520content%2Fdhi%2Fflyers%2520and%2520pdf%2Fsolution%2520flyers%2Fwater%2520quality%2520monitoring%2520from%2520space%2520-%2520dhi%2520gras%2520solution.pdf&usg=AOvVaw3DM9Dg-qAfh1Bdp8HQVBODS&opi=89978449 (Accessed: September 14, 2022).

El-Korashey, R. (2009) "Using regression analysis to estimate water quality constituents in Bahr El Baqar drain," *Journal of Applied Sciences Research*, 5(8), pp. 1067–1076.

Faustine, A. and Mvuma, A.N. (2014) "Ubiquitous mobile sensing for water quality monitoring and reporting within Lake Victoria basin," *Wireless Sensor Network*, 06(12), pp. 257–264. Available at: https://doi.org/10.4236/wsn.2014.612025.

Gakii, C. and Jepkoech, J. (2019) "A classification model for water quality analysis using decision tree," *Journal of Chemical Information and Modeling*, 7(3), pp. 1–8.

GWA-DW (2009) *Water quality monitoring program design – A guideline for field sampling for surface water quality*. Perth: Government of Western Australia Department of Water. Available at: https://www.wa.gov.au/system/files/2023-05/water-quality-monitoring-program-design-a-guideline.pdf (Accessed: September 14, 2022).

Helsel, D.R. *et al.* (2020) "Statistical methods in water resources," *Techniques and methods*. Book 4, Chap. A3. U.S. Geological Survey Techniques and Methods. Available at: https://doi.org/10.3133/tm4a3.

Khatoon, N. *et al.* (2013) "Correlation study for the assessment of water quality and its parameters of Ganga River, Kanpur, Uttar Pradesh, India," *IOSR Journal of Applied Chemistry*, 5(3), pp. 80–90. Available at: https://doi.org/10.9790/5736-0538090.

Maasdam, R. (2000) *Exploratory data analysis in water quality monitoring systems*. MSc Thesis. University of Salford. Available at: https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=ed9150ff41e4a34b3b11b3b124552a689849cffc (Accessed: February 22, 2022).

Malik, H. and Szwilski, A.B. (2016) "Towards monitoring the water quality using hierarchal routing protocol for wireless sensor networks," *Procedia Computer Science*, 98, pp. 140–147. Available at: https://doi.org/10.1016/j.procs.2016.09.022.

Manjarrés, C.R. *et al.* (2016) "Chemical sensor network for pH monitoring," *Journal of Applied Research and Technology*, 14(1), pp. 1–8. Available at: https://doi.org/10.1016/j.jart.2016.01.003.

Muhammad, S.M. *et al.* (2015) "Classification model for water quality using machine learning techniques," *International Journal of Software Engineering and Its Applications*, 9(6), pp. 45–52.

O'Flynn, B. *et al.* (2007) "SmartCoast: A wireless sensor network for water quality monitoring," *32nd IEE Conference on Local Computer Networks*, pp. 815–816. Available at: https://doi.org/10.1109/LCN.2007.34.

O'Flynn, B. *et al.* (2010) "Experiences and recommendations in deploying a real-time, water quality monitoring system," *Measurement Science and Technology*, 21(12), 124004. Available at: https://doi.org/10.1088/0957-0233/21/12/124004.

Pejman, A.H. *et al.* (2009) "Evaluation of spatial and seasonal variations in surface water quality using multivariate statistical techniques," *International Journal of Environmental Science and Technology*, 6(3), pp. 467–476. Available at: https://doi.org/10.1007/BF03326086.

Pule, M., Yahya, A. and Chuma, J. (2017) "Wireless sensor networks: A survey on monitoring water quality," *Journal of Applied Research and Technology*, 15(6), pp. 562–570. Available at: https://doi.org/10.1016/j.jart.2017.07.004.

Salah, H.A., Mocanu, M. and Florea, A. (2014) "Analysis of data mining tools used for water resources management in Tigris River," *Advanced Management Science*, 3(2), pp. 1–10.

Sasikala, R. (2017) "A comparative analysis for smart water resource using data mining tools," *International Journal of Research – Granthaalayah*, 5(7(SE)), pp. 24–30. Available at: https://doi.org/10.29121/granthaalayah.v5.i7(se).2017.2039.

Sperling von, M. (2015) *Wastewater characteristics, treatment and disposal. Vol. 6: Sludge treatment and disposal*. London: IWA Publishing. Available at: https://doi.org/10.2166/9781780402086.

UN (2015) *Goal 6: Ensure access to water and sanitation for all*. [Online]. The Sustainable Development Goals. Available at: https://www.un.org/sustainabledevelopment/water-and-sanitation/ (Accessed: April 04, 2021).

Varol, M., Gökot, B. and Bekleyen, A. (2010) "Assesment of water pollution in the Tigris River in Diyarbakır, Turkey," *Water Practice & Technology*, 5(1). Available at: https://doi.org/10.2166/wpt.2010.021.

Ward, R., Loftis, J.C. and McBride, G.B. (1986) "The 'data-rich but information-poor' syndrome in water quality monitoring," *Environmental Management*, 10(3), pp. 291–297. Available at: https://doi.org/10.1007/bf01867251.