MethodsX 10 (2023) 102034

Contents lists available at ScienceDirect

MethodsX

journal homepage: www.elsevier.com/locate/mex

Method Article

An approach based on multivariate distribution and Gaussian copulas to predict groundwater quality using DNN models in a data scarce environment

Ayoub Nafii [a,b,*], Houda Lamane [a], Abdeslam Taleb [a], Ali El Bilali [a,b,*]

[a] Hassan II University of Casablanca, Faculty of sciences and techniques of Mohammedia, Morocco
[b] River Basin Agency of Bouregreg and Chaouia, 13000 Benslimane, Morocco

ARTICLE INFO

Method name:
An approach based on copulas to predict groundwater quality using DNN models with small data

Keywords:
Virtual sample generation
Deep neural network
Groundwater quality
Entropy Water quality Index

ABSTRACT

Machine Learning models have become a fruitful tool in water resources modelling. However, it requires a significant amount of datasets for training and validation, which poses challenges in the analysis of data scarce environments, particularly for poorly monitored basins. In such scenarios, using Virtual Sample Generation (VSG) method is valuable to overcome this challenge in developing ML models. The main aim of this manuscript is to introduce a novel VSG based on multivariate distribution and Gaussian Copula called MVD-VSG whereby appropriate virtual combinations of groundwater quality parameters can be generated to train Deep Neural Network (DNN) for predicting Entropy Weighted Water Quality Index (EWQI) of aquifers even with small datasets. The MVD-VSG is original and was validated for its initial application using sufficient observed datasets collected from two aquifers. The validation results showed that from only 20 original samples, the MVD-VSG provided enough accuracy to predict EWQI with an NSE of 0.87. However the companion publication of this Method paper is El Bilali et al. [1].

- Development of MVD-VSG to generate virtual combinations of groundwater parameters in data scarce environment.
- Training deep neural network to predict groundwater quality.
- Validation of the method with sufficient observed datasets and sensitivity analysis.

Specification table

Subject area:	Environmental science
More specific subject area:	Water Resources
Method Name:	An approach based on copulas to predict groundwater quality using DNN models with small data
Name and reference of original method:	A framework based on multivariate distribution-based virtual sample generation and DNN for predicting water quality with small data [1]
Availability Source:	Supplementary data

* Corresponding authors.
E-mail addresses: ayoubnafii@gmail.com (A. Nafii), Ali.elbilali-etu@etu.univh2c.ma (A. El Bilali).

https://doi.org/10.1016/j.mex.2023.102034
Available online 2 February 2023

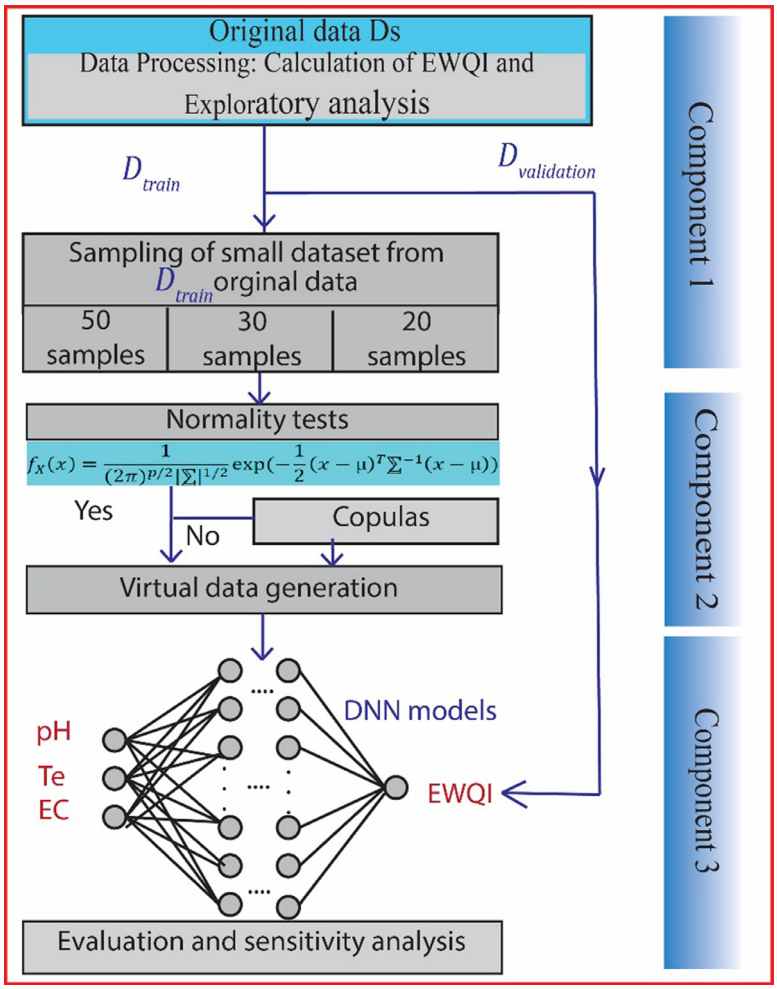

Fig. 1. Main components of the Multivariate Distribution-based Virtual Sample Generation Method. pH: Potential Hydrogen; Te: Temperature of the water; EC: Electrical Conductivity; EWQI: Entropy Weighted Water Quality Index; DNN: Deep Neural Network.

Method description

Introduction

Soft computing methods have become a powerful approach in groundwater modelling, as they can capture complex and nonlinear systems using available data compared to conceptual-based methods [2]. However, data availability is a keystone in developing and applying this approach. Especially for poorly monitored aquifers or when drilling new wells without sufficient observed datasets, the application of ML models is limited. The Virtual Sample Generation method (VSG) is valuable to overcome data shortage limitations, therefore, this can improve the accuracy of ML models even with small observed datasets.

Multivariate distribution - based virtual sample generation "MVD-VSG" method

This paper presented a novel method for generating virtual groundwater quality parameters to train DNN models to predict groundwater quality even with small datasets. It is based on multivariate distribution and is called MVD-VSG. However, the workflow of the developed method was built according to three components (Fig. 1): 1) Data processing and calculation of the Entropy Weighted Water Quality Index (EWQI), which is useful to reduce subjectivity in assessing groundwater quality [3–5]; 2) Virtual generation of groundwater quality samples using MVD-VSG and developing DNN models. DNN model is theoretically more accurate than traditional ML models as concluded in previous comparative studies [6,7]. Compared to existing VSG methods [8–10], the MVD-VSG allows the generation of groundwater quality samples respecting the inter-correlation between chemical and physical parameters and, consequently, the conservation of the physical information; 3) validation with observed groundwater samples, thereby the method was evaluated using sufficient observed datasets. These components are described in the following sub-section.

Component 1: Weighted entropy water quality index (EWQI)

The entropy was defined and applied for the first time by Shannon [11] in 1948 to thermodynamic sciences. It can evaluate an amount and a degree of pertinent information from disorderly and uncertain data pertaining to predicting the output of a probabilistic event. Its application was conducted in various hydrological studies, namely: drought indices, flood risk evaluation, and water quality assessment [3,12,13]. Hence, the EWQI was embedded in the method to minimize the subjectivity in assessing groundwater quality. However, to assess groundwater quality using n parameters j= {1, 2, 3…n} and m samples i= {1, 2, 3…m} using EWQI, we followed the steps below.

Step 1: Exploratory data analysis and cleaning process of raw data to select reliable datasets X_{ij}

$$X_{ij} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ x_{m1} & x_{m2} & & x_{mn} \end{bmatrix} \quad (1)$$

Step 2: Normalization process. The normalization process of datasets was carried out in this method according to the construction function as given by the Eq. (2) to reduce the dimensionality effects of groundwater quality parameters.

$$y_{(i,j)} = \frac{x_{max(i,j)} - x_{(i,j)}}{x_{max(i,j)} - x_{\min(j,j)}} \quad (2)$$

Hence,

$$Y_{ij} = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1n} \\ y_{21} & y_{22} & \cdots & y_{2n} \\ y_{m1} & y_{m2} & \cdots & y_{mn} \end{bmatrix} \quad (3)$$

Where Y_{ij} is a standard grade matrix and represents normalized datasets $y_{(i,j)\in[0,1]}$.

Step 3: Calculation of the entropy. Herein, the entropy was computed as following:

The ratio *Pij* of *j* index value in *i* sample was computed as follow:

$$P_{ij} = \frac{y_{ij}}{\sum_i^m y_{ij}} \quad (4)$$

Therefore, the information entropy e_j associated with parameter j is given by the Eq. (5).

$$e_j = -\frac{1}{\ln(m)} \sum_{i=1}^{m} P_{ij} * \ln(P_{ij}) \; where \; \lim_{P_{ij}\to 0} P_{ij} * \ln\left(P_{ij}\right) = 0 \quad (5)$$

Finally, the entropy weight ω_j associated with groundwater quality parameter j is computed by the following equation.

$$\omega_j = \frac{1 - e_j}{\sum_{i=1}^{m}(1 - e_j)} \quad (6)$$

Step 4: Calculation of EWQI

The rating quality scale q_j associated with groundwater quality parameter j is calculated by the following equation:

$$q_j = \frac{C_j}{L_j} * 100 \quad (7)$$

Where Cj is the measured groundwater quality parameter j and L_j is the limit value determined by the World Health Organization (WHO) of the parameter j for drinking purposes. When a parameter j is completely absent in the studied aquifer $Cj = 0$, then $q_j = 0$ meaning that there is no effect. The EWQI is calculated by the Eq. (8).

$$EWQI = \sum_{j=1}^{n} \omega_j * q_j \quad (8)$$

The developed method used physical parameters, such as the Electrical Conductivity (EC), Temperature (Te), and pH as feature variables to predict EWQI because they can be measured by automatic sensors in real-time. Therefore, it is valuable to optimize the process of the water quality assessment.

Component 2: Generating virtual datasets

As shown in the last sub-section, the EWQI is calculated using several physical and chemical parameters. More importantly, these parameters are inter-correlated, particularly the correlations that existed between the cations and anions associated with dissolved matter in groundwater due to various hydro-geochemical facies such as Na-Cl, Na-Mg-Ca-Cl, and Ca-Mg-HCO3-Cl [14]. Therefore, the generation of virtual datasets requires the conservation of the relationships that exist between the chemical parameters, so that the virtual generated data keeps the reel physical information that existed in the original dataset. The developed method MVD-VSG relies on multivariate distribution to generate appropriate combinations of virtual datasets, to overcome the data availability challenge in predicting groundwater quality using DNN models.

a-Multivariate normal distribution (MVN)

Let MVN of random vector variables with dimension m represented by $X=\{x_1, x_2, x_3 \ldots x_m\}$, ($X \sim N(\mu, \Sigma)$). It describes the distributions and the inter-correlations between random variables and is characterized by a covariance matrix $\Sigma = (\Sigma_{ij})_{m \times m}$ and a mean vector $\mu = \{\mu_1, \mu_2, \mu_3 \ldots \mu_m\}$. Also, MVN is a symmetric matrix (m x m). Thus, each random variable x_i has its normal distribution $N \sim (\mu_i, \Sigma_{ii})$. Indeed, the expression of the function density of the MVN is given by the Eq. (9):

$$f_X(x) = \frac{1}{(2\pi)^{\frac{p}{2}} |\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right) \; x \in \mathbb{R}^m \tag{9}$$

Therefore, through Cholesky decomposition [15], MVD-VSG can generate appropriate virtual combinations of groundwater quality parameters j, as it takes into account the inter-correlations between these parameters. Yet, it should be noted that if datasets follow an MVN distribution implies that all variables follow normal distributions, but the inverse is not necessarily true.

b-Copulas

The use of MVN to generate virtual combinations of groundwater quality parameters is suitable when the normality test of datasets is fairly acceptable. In dataset-scarce environments, the normality test of the MVN could be highly impacted either by dataset sizes or by the fact that the variables X follow different law distributions. Alternatively, the copulas are useful to generate appropriate virtual combinations when the normality is not verified [16,17]. Indeed, the copulas methods are multivariate cumulative distribution functions, as the distribution of the marginal probability of variables X is uniform [0,1] [18]. Consequently, it is valuable to separate the inter-correlation between random vectors from their marginal distribution.

Let $X \in \mathbb{R}^m$, the random vector variable considered. The expression of the joint distribution function of X can be written as follow:

$$H(x_1, x_2, \ldots., x_m) = C\left(F_1(x_1), F_2(x_2), \ldots \ldots, F_m(x_m)\right) \;\; X^T \in \mathbb{R}^m \tag{10}$$

C: $[0,1]^m \rightarrow [0,1]$ is the copula and F_i represents the ith marginal distribution function.

Hence, by turning the Eq. (10) around, any MVD can be projected to the unit square to recover the copula C.

$$C(u_1, u_2, \ldots., u_m) = H\left(F_1^{-1}(u_1), F_2^{-1}(u_2), \ldots \ldots, F_m^{-1}(u_m)\right) \tag{11}$$

Where F_i^{-1} is the inverse function of F_i.

However, the copulas can be classified into two categories, such as Archimedean and elliptical method families. Because of their simplicity and proprieties, Archimedean methods are powerful to simulate bivariate distributions [12,17,19]. Meanwhile, the MVD-VSG method adopted elliptical copulas, as we are studying several groundwater quality parameters and they can generate random variables with high dimensions. These copulas, however, were widely applied in hydrological sciences and showed their utility for generating synthetic datasets [20]. The projection of the MVN produces the Gaussian Copula with a function density given by the following equation [21]:

$$C_r(u) = \frac{1}{\sqrt{|r|}} \exp\left(-\frac{1}{2} F_i^{-1^T}\left(r^{-1} - I\right) F_i^{-1}\right) \tag{12}$$

Where I is the identity matrix, $r \in [-1, 1]^{m \times m}$ is the correlation matrix between variables with 1 in its diagonal.

Component 3: Validation and implementation processes

The reliability of the raw dataset was carried out through the calculation of the Charge Balance Error (CBE) of all samples [22]. The samples which have CBE of more than 5% in absolute value and/or those that correspond to outlier values were deleted. Finally, from 750 raw samples, 700 were retained for the case study. These samples were randomly divided into $D_{train} = 400$ samples for developing the MVD-VSG method and $D_{validation} = 300$ samples for the validation phase.

The Matlab software (R2021b) was used in which the Statistics and Machine Learning toolbox is called to implement the MVD-VSG method. Firstly, the logarithm transformation of the whole dataset was conducted to improve the distribution normality of the variables. Thereafter, the Mardia test [23] using the code developed by Trujillo-Ortiz (http://www.mathworks.com/matlabcentral/fileexchange/3519-mskekur) and Rayston test [24] were conducted for checking the multivariate normality of the original samples.

The MVD-VSG method is applied to generate 11 virtual datasets with sizes ranging from 500 to 10 000 samples from D_{train}, namely: 20, 30, 50, and 400 original samples (observed data) (Fig. 1). Then, a series of DNN models were trained using virtual datasets. Using only physical parameters as input features will improve the practical implication of the methodology by embedding the developed method in ML-based sensor technologies to now-cast groundwater quality even with small datasets.

The validation of the MVD-VSG method was carried out through the simulation of 300 measured samples using trained DNN models. The performances of DNN models for predicting EWQI were evaluated using statistical metric by comparing the simulated and observed EWQI values e.g. Root Mean Square Error, Nash-Sutcliff efficiency (NSE) [25]. Besides, the sensitivity analysis of the MVD-VSG method to the virtual datasets and to the original samples is valuable to generate optimal size of virtual datasets and the limits of the model performance that can be reached.

Reproducibility of the MVD-VSG method

The MVD-VSG method is valuable to predict groundwater quality using ML models even with small observed data. The results of the validation process using enough experimental data showed its robustness. Hence, the MVD-VSG is reproducible to overcome the data shortage in applying ML models, especially for either newly drilled wells or poorly monitored aquifers. Besides, a change in the baseline of the modelled system leads to the mandatory re-training of the models with new observed datasets (new discharge into the river as an example). In this scenario, the MVD-VSG is an alternative. Furthermore, ML and DNN cannot explain the process involved in aquifer modelling, as they are considered ''Black-Box'' models. Meanwhile, the physical-guided training of ML and DNN models by embedding the physical proprieties of the system to be modelled into the features during the training phase could be a promising alternative. Since it is difficult to get sufficient instances of the physical proprieties of the systems to be modelled to train DNN models, the MVD-VSG method is valuable to generate appropriate virtual combinations of the physical proprieties and keeps the physical information of the generated virtual datasets. Therefore, the MVD-VSG is not only a powerful method to broaden the application of DNN models with small data but is an armamentarium of a potential approach to develop explainable ML models.

Declaration of competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgment

This work is supported by the River Basin Agency of Bouregrge and Chaouia (ABHBC) by providing the material required for achieving this research. So, the authors thank the ABHBC teams for their help and assistance.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.mex.2023.102034.

Reference

[1] A. El Bilali, H. Lamane, A. Taleb, A. Nafii, A framework based on multivariate distribution-based virtual sample generation and DNN for predicting water quality with small data, J. Clean. Prod. 368 (2022) 133227, doi:10.1016/j.jclepro.2022.133227.
[2] A. El Bilali, A. Taleb, Y. Brouziyne, Comparing four machine learning model performances in forecasting the alluvial aquifer level in a semi-arid region, J. Afr. Earth Sci. 181 (2021) 104244, doi:10.1016/j.jafrearsci.2021.104244.
[3] R. Salman, M.R. Nikoo, S.A. Shojaeezadeh, P.H.B. Beiglou, M. Sadegh, J.F. Adamowski, N. Alamdari, A novel Bayesian maximum entropy-based approach for optimal design of water quality monitoring networks in rivers, J. Hydrol. (2021) 603, doi:10.1016/j.jhydrol.2021.126822.
[4] M. Hossain, P.K. Patra, Water pollution index – a new integrated approach to rank water quality, Ecol. Indic. (2020) 117, doi:10.1016/j.ecolind.2020.106668.
[5] N. Adimalla, H. Qian, P. Li, Entropy water quality index and probabilistic health risk assessment from geochemistry of groundwaters in hard rock terrain of Nanganur County, South India, Chemie Der Erde (2019), doi:10.1016/j.chemer.2019.125544.
[6] A. Korotcov, V. Tkachenko, D.P. Russo, S. Ekins, Comparison of deep learning with multiple machine learning methods and metrics using diverse drug discovery data sets, Mol. Pharm. 14 (2017) 4462–4475, doi:10.1021/acs.molpharmaceut.7b00578.
[7] K. Amasyali, N. El-Gohary, Machine learning for occupant-behavior-sensitive cooling energy consumption prediction in office buildings, Renew. Sustain. Energy Rev. 142 (2021) 110714, doi:10.1016/j.rser.2021.110714.
[8] R.Z. Xu, J.S. Cao, Y. Wu, S.N. Wang, J.Y. Luo, X. Chen, F. Fang, An integrated approach based on virtual data augmentation and deep neural networks modeling for VFA production prediction in anaerobic fermentation process, Water Res. 184 (2020) 116103, doi:10.1016/j.watres.2020.116103.
[9] T. Xu, G. Coco, M. Neale, A predictive model of recreational water quality based on adaptive synthetic sampling algorithms and machine learning, Water Res. 177 (2020) 115788, doi:10.1016/j.watres.2020.115788.
[10] A. MacAllister, A. Kohl, E. Winer, Using high-fidelity meta-models to improve performance of small dataset trained Bayesian networks, Expert Syst. Appl. 139 (2020) 112830, doi:10.1016/j.eswa.2019.112830.
[11] C.E. Shannon, A mathematical theory of communication, Bell Syst. Tech. J. 27 (1948) 379–423, doi:10.1002/j.1538-7305.1948.tb01338.x.
[12] J. Chang, Y. Li, Y. Wang, M. Yuan, Copula-based drought risk assessment combined with an integrated index in the Wei River Basin, China, J. Hydrol. 540 (2016) 824–834, doi:10.1016/j.jhydrol.2016.06.064.
[13] H. Al-Hinai, R. Abdalla, Mapping coastal flood susceptible areas using Shannon's entropy model: the case of Muscat governorate, Oman, ISPRS Int. J. Geo-Inf. 10 (2021), doi:10.3390/ijgi10040252.
[14] T. El Ghali, H. Marah, M. Qurtobi, F. Raibi, M. Bellarbi, N. Amenzou, B. El Mansouri, Geochemical and isotopic characterization of groundwater and identification of hydrogeochemical processes in the Berrechid aquifer of central Morocco, Carbonates Evaporites 35 (2020) 1–21.
[15] D. Dereniowski, M. Kubale, Cholesky factorization of matrices in parallel and ranking of graphs, Lect. Notes Comput. Sci. (Including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics). 3019 (2004) 985–992, doi:10.1007/978-3-540-24669-5_127.
[16] F. Alidoost, Z. Su, A. Stein, Evaluating the effects of climate extremes on crop yield, production and price using multivariate distributions: a new copula application, Weather Clim. Extrem. 26 (2019) 100227, doi:10.1016/j.wace.2019.100227.
[17] L. Wang, H. Yu, M. Yang, R. Yang, R. Gao, Y. Wang, A drought index: the standardized precipitation evapotranspiration runoff index, J. Hydrol. 571 (2019) 651–668, doi:10.1016/j.jhydrol.2019.02.023.
[18] A. Sklar, Fonctions de répartition à n dimensions et leurs marges (French), Publ. l'Institut Stat. Univ. Paris. 8 (1959) 229–231.
[19] X. Yang, Y.P. Li, G.H. Huang, Y.F. Li, Y.R. Liu, X. Zhou, Development of a multi-GCMs Bayesian copula method for assessing multivariate drought risk under climate change: a case study of the Aral Sea basin, Catena (2022) 212, doi:10.1016/j.catena.2022.106048.

[20] F. Tootoonchi, J.O. Haerter, O. Räty, T. Grabs, M. Sadegh, C. Teutschbein, Copulas for hydroclimatic applications – a practical note on common misconceptions and pitfalls, Hydrol. Earth Syst. Sci. Discuss. (2020) 1–31.
[21] P. Arbenz, Bayesian copulae distributions, with application to operational risk management-some comments, Methodol. Comput. Appl. Probab. 15 (2013) 105–108, doi:10.1007/s11009-011-9224-0.
[22] R.A. Freeze, J.A. Cherry, Groundwater, Prentice-Hall Inc, Englewood Cliffs, NJ, 1979.
[23] K.V. Mardia, Measures of multivariate skewness and kurtosis with applications, Biometrika 57 (1970) 519–530, doi:10.1093/biomet/57.3.519.
[24] P. Royston, Approximating the Shapiro-Wilk W-test for non-normality, Stat. Comput. 2 (1992) 117–119, doi:10.1007/BF01891203.
[25] J.E. Nash, J.V. Sutcliffe, River flow forecasting through conceptual models Part I—a discussion of principles, J. Hydrol. 10 (1970) 282–290.