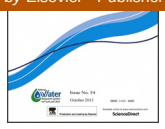
Research Center
المركز القومى لبحوث المياه

Water Science 27 (2013) 48–56

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



Identifying uncertainty of the mean of some water quality variables along water quality monitoring network of Bahr El Baqar drain

Hussein G. Karaman

Drainage Research Institute, National Research Center (NWRC), Ministry of Water Resources and Irrigation (MWRI), Egypt

Available online 1 March 2014

Abstract

Assigning objectives to the environmental monitoring network is the pillar of the design to these kinds of networks. Conflicting network objectives may affect the adequacy of the design in terms of sampling frequency and the spatial distribution of the monitoring stations which in turn affect the accuracy of the data and the information extracted. The first step in resolving this problem is to identify the uncertainty inherent in the network as the result of the vagueness of the design objective. Entropy has been utilized and adopted over the past decades to identify uncertainty in similar water data sets. Therefore it is used to identify the uncertainties inherent in the water quality monitoring network of Bahr El-Baqar drain located in the Eastern Delta. Toward investigating the applicability of the Entropy methodology, comprehensive analysis at the selected drain as well as their data sets is carried out. Furthermore, the uncertainty calculated by the entropy function will be presented by the means of the geographical information system to give the decision maker a global view to these uncertainties and to open the door to other researchers to find out innovative approaches to lower these uncertainties reaching optimal monitoring network in terms of the spatial distribution of the monitoring stations.

Keywords: Entropy theory; Water quality; Monitoring network; Uncertainty

1. Introduction

The ambiguity of the objective of any environmental monitoring network could be a threat to the accuracy of data measured and consequently on the information extracted from these data. Cavanagh et al. (1998) mentioned that defining clearly the monitoring objectives is the first and most crucial step in developing an experimental design. There are several water qualities monitoring objectives (i.e., assess compliance with standards, facilitate impact assessment studies, determine fate and transport of pollutants, etc.) where the information from each of these categories is obtained specifically for the needs of that particular category (Shaban, 2007). This situation exists in the National Water Quality Monitoring Network (NWQMN) of Egypt started in the early eighties of the last century. The initial objective of the network was to determine the status of water salinity in the drains and canals to satisfy the objectives of the national

E-mail address: Hussein.karaman@Gmail.com

Peer review under responsibility of National Water Research Center.

ELSEVIER Production and hosting by Elsevier

1110-4929

http://dx.doi.org/10.1016/j.wsj.2013.12.005

drainage water reuse project. In the nineteen's, the water quality issue became vital and the Egyptian Government represented by the National Water Research Center (NWRC) upgraded the network to monitor other extra water quality variables. The objective of the network was quantifying the variation in the drainage water quality in relation to the existing different pollution sources. Hence, the objective of the network had been changed from identifying the quality and quantity of drainage water reuse in agricultural and the potential of this activity to quantify the variation in the drainage water quality along the Nile Delta (NAWQAM, 2002). This conflict in objectives has raised doubts and uncertainties in the quality of the information extracted from the network. In addition to the uncertainty inherent in the water quality process itself (Sanders et al., 1983), the information extraction from the monitoring network become hazardous and the probability of making right decision will be under high risk (Karaman, 2007). Accordingly, identifying the uncertainties arose due to the difference in the monitoring objectives and the process itself should be under the scope of the decision maker and the water quality researchers. This has to be on line with the requirements of the adequacy of collected water quality data and the performance of the existing monitoring networks. (Harmancioglu et al., 1999) stated that, the evaluation of the monitoring water quality networks is triggered two main reasons. First an efficient information system is required to satisfy the needs of water quality management plans and to aid in the decision-making process. Second, this system has been realized under the constraints of limited financial resources, sampling, laboratory facilities, and manpower. Uncertainty in the water quality data is essential due to the difference between the real world (water quality in the environment) and the information we have about it (understanding of water quality conditions) which in scientific term called noise. Harmancioglu and Alpaslan (1997) identified the uncertainties raised from the monitoring practices in order to give the decision makers the capability to build right strategies and confident polices. To overcome the problems of uncertainty, several approaches have been used to evaluate and quantify these raised errors such as statistical and stochastic approaches. The main objective of this paper is to identify the uncertainties inherited in the data collected by using the Entropy function in the monitoring stations along one of the main drains located in the Eastern part of the Nile Delta in Egypt (Bahr El Baqar Drain). Also, the identified uncertainties will be demonstrated in spatially base by using the capabilities of the geographical information system (GIS).

2. Entropy theory

The entropy (or information) theory, developed by Shannon and Weaver (1949), has recently been applied in many different fields. This theory has also been applied in hydrology and water resources for measuring the information content of random variables and models, evaluating information transfer between hydrological processes, evaluating data acquisition systems, and designing water quality monitoring networks (Mogheir et al., 2004a). There are four basic information measures based on entropy, which are marginal, joint, and conditional entropies and trans-information (Karamouz et al., 2009). Shannon and Weaver (1949) were the first to define the marginal entropy, $H(x)$, of a discrete random variable x as:

$$H(x) = -\sum_{i=1}^{N} P(x_i) * \log P(x_i) \tag{1}$$

where, N represents the number of events x_i with probabilities $P(xi)$ ($i = 1,\ldots, N$). The total entropy of two independent random variables x and y is equal to the sum of their marginal entropies.

$$H(x, y) = H(x) + H(y) \tag{2}$$

The Entropy theory (information theory) has been used in several fields related to communications as (Shannon and Weaver, 1949) used entropy as a measure of uncertainty in the mind of someone receiving a message that contains noise (Karamouz et al., 2009). Later, in 1957, Jaynes made use of Shannon's entropy metric to formulate the maximum entropy principle that formed a basis for estimation and inference problems (Mogheir et al., 2004a,b). In 1973, Amorocho and Esplidora were the first to apply the entropy concept to hydrological modeling (Singh, 1997). Since then, there have been a great variety of entropy applications in hydrology and water resources management (e.g. Rajagopal et al., 1987; Singh and Rajagopal, 1987; Singh, 1998; Harmancioglu et al., 1999). Entropy theory also has been applied to assess and evaluate monitoring networks with respect to water quality (Harmancioglu et al., 1994; Ozkul et al., 2000), rainfall (Krastanovic and Singh, 1992) and groundwater (Bueso et al., 1999; Mogheir and Singh, 2002). Most of these applications involve applying Entropy theory to the evaluation, assessment and design of monitoring networks, and they used an analytical approach with a presumed knowledge of the probability distributions of the random variables

Table 1
Water quality variables definition.

Abbreviation	Variable name	Contamination type
BOD	Biological oxygen demand	Industrial and domestic
COD	Chemical oxygen demand	
DO	Dissolved oxygen	
NH_4	Ammonium	Agricultural
P	Phosphorus	Agricultural and domestic
NO_3	Nitrates	
TDS	Total dissolved salts	Agricultural
Fe	Iron	Heavy metals
Coliform	Coliform	Domestic

involved. Awadallah (2012) used the Entropy theory with the kriging approach to select the optimum location of the rainfall stations. He also recommended further research is needed to apply the methodology on a larger scale and to confirm that the proposed rainfall gauges add information to the spatial distribution of rainfall in the region. Furthermore, a comparison between entropy methods is needed to decide on the best suited method in selection of rainfall gauge locations.

2.1. Study zone and area description

The calculations of the Entropy values were applied on the Eastern Nile Delta of Egypt. The Eastern Delta – except for few catchments – drains its water into Lake Manzala, which in turn discharge freely into the Mediterranean Sea. A considerable area is drained by two main drainage systems: Bahr El Baqar and Bahr Hadus drainage systems. Bahr El Baqar drain was selected as a case study for the purpose of this paper. The choice of this drain was based mainly on: (1) its importance to the decision makers and the policy planners, (2) relatively high pollution levels in the drain; and (3) the availability of historical water quality data on the Drainage Research Institute (DRI) database. Bahr El Baqar drain is one of the most polluted drains in the Nile Delta as it collects agricultural, domestic and industrial wastewaters. In addition to that, it disposes its water to the Lake Manzala, negatively affecting its ecological system. Due to the variety of water types discharged in the drain, nine water quality variables have been used and tested using the Entropy function. These variables represent different pollution types where the biological oxygen demand (BOD), chemical oxygen demand and dissolved oxygen (DO) represent the contamination due to the industrial and domestic effluents and the Coliform represents the contamination due to the wastewater effluents. The water quality variables and its contamination types are presented in Table 1.

2.2. Bahr El Baqar drainage system

Bahr El Baqar originates at the confluence of two drains, namely Bilbeis drain and Qalyubeya drain as shown in Fig. 1. Bilbeis drain receives its water originally through a lifting pump lifted drainage water of urban areas (industrial and domestic sewage water) of greater Cairo. Then, it is fed gravitationally by agricultural drainage throughout its main course. It is important to mention that Qalyubeya drain is completely fed by gravity. Both (Bilbeis and Qalyubeya) drains discharge approximately equal quantities of drainage water annually, but the main quantity of the Qalyubeya drain discharge is pumped into El Wadi El Sharky irrigation canal by Wadi P.S. (EB03). The remaining water flows together with Bilbeis drain to continue thereafter as Bahr El Baqar drain. The first reach of Bahr El Baqar drain is fed by gravity, further downstream the pumping stations Saada (EB06) and Bahr Baqar (EB10) lift water into the drain. From Bahr El Baqar bridge until the outfall at Lake Manzala withdrawals and additions take place diffusely without control as illustrated in Fig. 2. The analysis is carried out on the monitoring stations distributed along the catchment of the drain. The total number of stations is15; however 5 stations do not have data covering the whole duration (1997–2010). This enforces the analysis to be limited to the other 10 stations having data covering the entire study duration (1997–2010). Table 2 shows data available at each monitoring stations.

Fig. 1. Layout of Bahr El Baqar drain and its tributaries.

2.3. Methodology

There are several methods to identify the uncertainty associated with the water quality data such as the stochastic modeling (Time series analysis), statistical approaches and many other methods. In this research the entropy information will be used to identify the uncertainty in the water quality data as it is a robust measure and easy to calculate. The main hypothesis in this study is the ability to estimate the amount of information contained in the selected stations of the

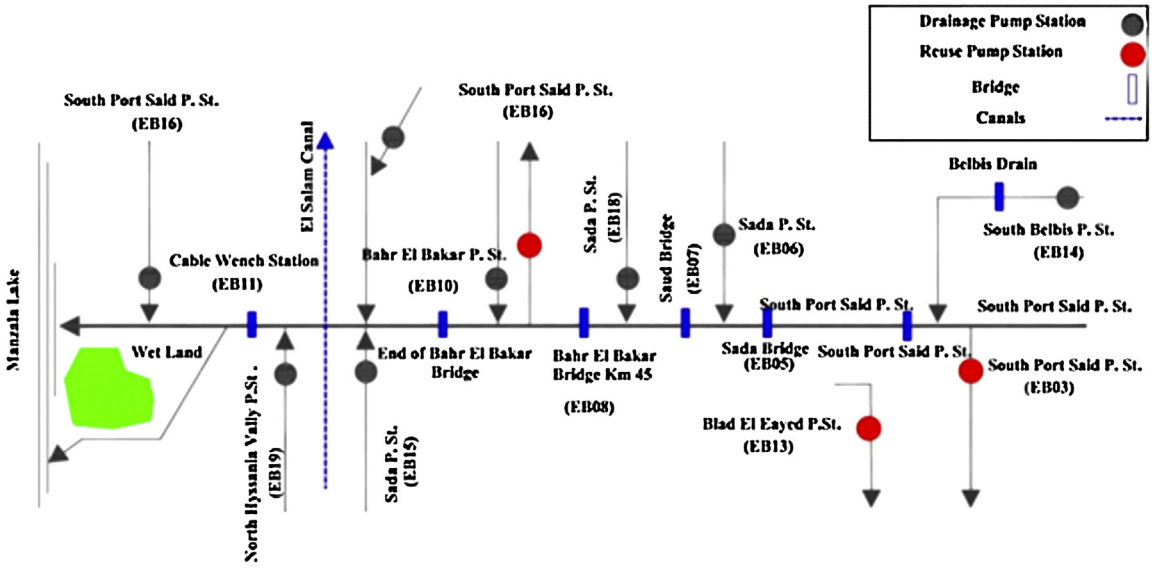


Fig. 2. Bahr El Baqar drain schematic diagram.

Table 2
Monitoring stations description and data duration.

Code	Description	Data duration	Analysis status
EB03	Wadi pump station	1997–2002	Not used
EB04	Wadi railway bridge	1997–2010	Used
EB05	Saada bridge	1997–2010	Used
EB06	Saada pump station	1997–2010	Used
EB07	Saud bridge	1997–2010	Used
EB08	Bahr El Baqar bridge	1997–2010	Used
EB10	Bahr El Baqar pump station	1997–2010	Used
EB14	South Belbies pump station	1997–2010	Used
EB15	Bahr El Baqar outfall	1997–2010	Used
EB31	Bahr El Baqar drain	1997–2002	Not used
EB36	Bahr El Baqar drain	1997–2002	Not used
EB38	Bahr El Baqar drain	1997–2010	Used
EB40	Bahr El Baqar drain	1997–2002	Not used
EB43	Bahr El Baqar drain	1997–2002	Not used
EB47	Bahr El Baqar drain	1997–2010	Used

monitoring network of Bahr El Baqar drainage system in terms of information entropy for the selected water quality variables. The higher the values of the entropy the lower the information extracted from these stations and consequently affecting the reliability of the decisions built on this information. The calculation of the entropy information was based on a spatial basis to give a holistic picture to the uncertainty distribution along the whole drain. The capabilities of the geographical information system were used to present the results of the entropy information. The analysis starts with the calculation of the entropy information values for the selected water quality variables at each monitoring station along the selected drain where the values of the Entropy have been calculated by the Entropy package that implements various estimators of the Shannon Entropy where the maximum likelihood method was the core of the analysis and it was implemented under the R statistic platform (Hausser and Strimmer, 2009). Since there is no threshold value to the entropy information in the literature to judge the uncertainty in the selected water quality variables objectively, a classification tree based on the capabilities of the cluster analysis (K-Mean Algorithm) has been used to categorize the stations into groups based on the similarities between the values of the calculated entropy as shown in Fig. 3. Fig. 4 shows the methodology of applying the K-Mean Algorithm on the entropy values (adapted from Karaman, 2007). The results of the calculated similarity distances of the constructed clusters have been used in the process of the entropy classification. It is assumed that the locations in the cluster which has maximum mean entropy values will be dealt as

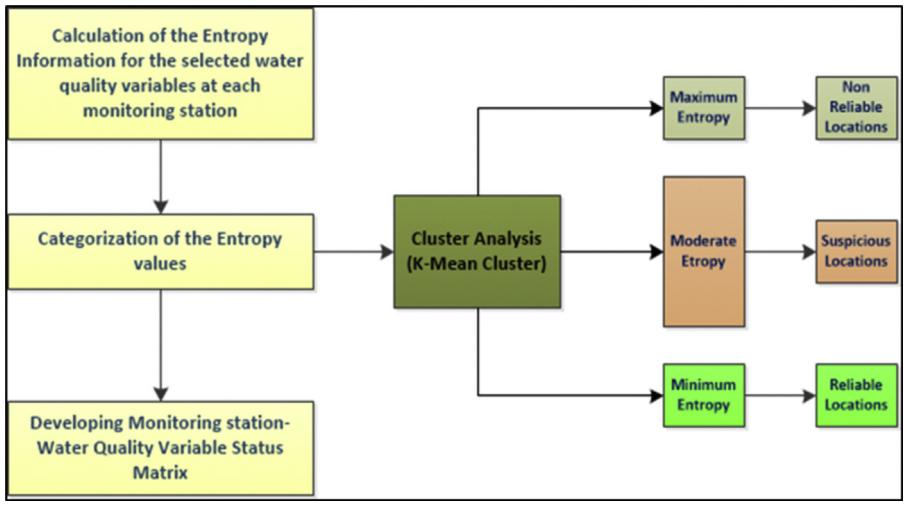


Fig. 3. Proposed methodology.

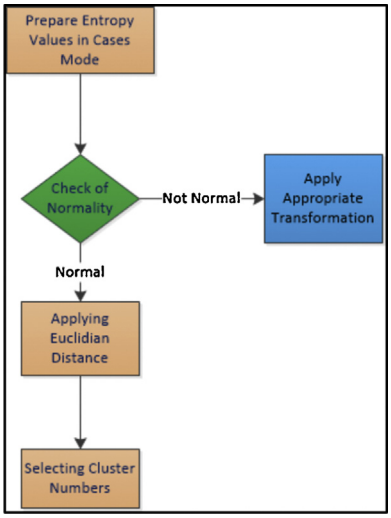


Fig. 4. K-Means Algorithm application.

non-reliable locations (the data at these locations are considered of less confidence) while the locations in the cluster which has minimum entropy values will be dealt as a reliable locations (the data at these locations are considered of greater confidence). Finally, the locations in the cluster which has entropy values laid among the minimum and the maximum values will be dealt as doubtful locations and for decision making safety considerations, it will be judged as the locations that need to be checked. Based on the results of the entropy interpretation, a monitoring stations-water quality variables status matrix will be introduced to the decision maker showing the variability status of the selected water quality variables for each monitoring station. Finally the variability status of each variable at each monitoring station will be presented spatially by the capabilities of the Geographical Information System (GIS) to have a holistic picture to the variability status along the whole drain system. Fig. 3 shows the steps of the proposed methodology to fulfill the objective of the study. This variability is in some sense a measure of the uncertainty (i.e. more sampling for example may be required to have more confidence in the station information).

3. Results and discussions

Table 3 shows the results of the application of the entropy function for each monitoring station to the selected water quality variables. Since the entropy information is measured in (Nats), the comparison of the entropy values for the selected variables in all monitoring stations is consistent. Table 4 shows the results of the normality test applied on the entropy values for the selected variables where the results show that all variables are normally distributed.

Table 3
Results of entropy for the selected water quality variables.

Code	BOD	COD	DO	NO_3	NH_4	P	TDS	Fe	Coliform
EB04	3.618	4.584	4.685	4.309	4.549	4.905	5.010	4.763	2.142
EB05	4.711	4.663	4.455	4.242	4.518	4.951	5.013	4.858	1.343
EB06	4.586	4.453	4.925	4.112	4.499	4.484	5.014	4.785	0.867
EB07	4.574	4.550	4.583	4.241	3.608	4.937	5.013	4.795	4.101
EB08	4.632	4.568	4.655	2.264	4.531	4.956	4.989	4.833	3.150
EB10	2.670	4.558	4.938	4.420	4.550	4.884	4.971	4.791	2.503
EB14	4.610	4.561	4.500	4.274	5.501	4.302	4.990	4.810	2.102
EB15	3.578	4.503	4.958	2.425	4.522	3.838	4.999	4.789	1.399
EB38	4.031	4.630	4.682	4.144	4.472	4.239	4.977	4.746	3.557
EB47	4.709	4.658	4.580	4.239	4.456	4.943	5.010	4.693	4.148

Table 4
Normality tests results.

	Kolmogorov–Smirnov			Shapiro–Wilk		
	Statistic	*df*	Sig.	Statistic	*df*	Sig.
BOD	.199	10	.200*	.900	10	.220
COD	.164	10	.200*	.948	10	.646
DO	.224	10	.168	.889	10	.166
NO3	.217	10	.200*	.915	10	.317
NH4	.148	10	.200*	.966	10	.849
P	.327	10	.523	.642	10	.624
TDS	.262	10	.061	.861	10	.078
Fe	.189	10	.200*	.961	10	.802
Coliform	.272	10	.085	.831	10	.064

* High significance, greater than 0.05.

Table 5 shows the classification of the stations based on the cluster analysis applied on the entropy values for each water quality variable. It was assumed that the monitoring station included in the cluster which have the highest mean entropy value will be unreliable and the information extracted from it need to be checked in terms of field and laboratory steps while the monitoring stations included in the cluster which have lowest mean entropy value will be reliable and the information extracted from it is to some extent confident. For the cluster which has mean entropy value in between the maximum and the minimum, it was assumed that the stations included in this cluster need also to be checked and the information extracted from it is questionable and the dependence process on this information will be subject to the preference of the decision maker. The table indicates the number of clusters and the values of the similarity distance of each cluster based on the entropy values. It is obvious from the table that cluster No. 2 represents the highest entropy values and cluster No. 3 represents the lowest entropy values while cluster No. 1 represents the moderate entropy values. From the three clusters, the monitoring stations with high uncertainty with low confidence, monitoring stations with low uncertainty with high confidence and finally the monitoring stations that have moderate uncertainty and might be relied on in the decision making process according to the decision maker preference can be concluded as shown in Table 6.

Based on the classification of the entropy of the monitoring stations. Fig. 5 shows the spatial distribution of the monitoring station status along the system of Bahr El Baqar drain based on the entropy classification.

Table 5
Final cluster centers and entropy classification.

	Cluster		
	1	2	3
BOD	3.6400	5.6458	2.5820
COD	4.5595	4.6088	4.4780
DO	4.7190	7.6067	2.9415
NO_3	3.3470	6.2398	4.4685
NH_4	4.2255	4.5223	1.5105
P	3.0930	4.9218	0.6610
TDS	4.0805	5.0020	1.0065
Fe	4.2005	4.7813	1.7870
Coliform	1.3025	8.0735	0.1330
Mean center	3.6853	5.7114	2.1742
	Moderate entropy	High entropy	Low entropy

Table 6
The components of each cluster.

Cluster No. 1	Cluster No. 2	Cluster No. 3
EB10	EB04	EB06
EB14	EB05	EB15
	EB07	
	EB08	
	EB38	
	EB47	
Moderate entropy	High entropy	Low entropy

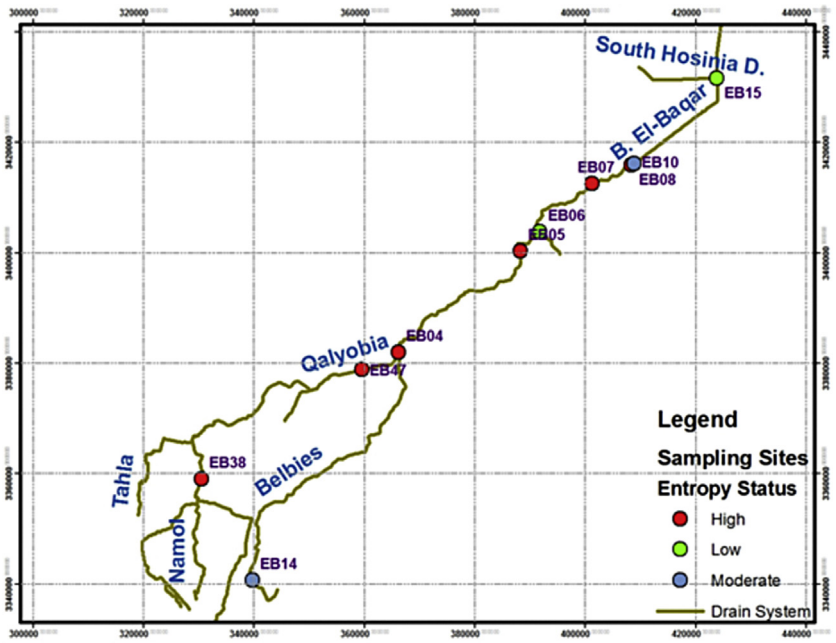


Fig. 5. Spatial distribution of the stations status along Bahr El Baqar drain system.

4. Conclusion

Identifying the uncertainty in the data sets of any monitoring network can be the main objective of any data analyst and the decision makers as well. The worth of identifying the uncertainty raised up due to its importance to build a decision making tree stand on a reliable data and hence taking actions with high confident. Many researchers studied the uncertainty in the monitoring network data especially the water quality data to fulfill this truth and they used many tools to pinpoint these uncertainties such as stochastic models, statistical approaches and many other methods. Accordingly, the main objective of this study was to identify the uncertainty in the water quality monitoring data for selected water quality variables in different water quality monitoring stations along the Nile Delta in Egypt. The Entropy information has been used in estimating the uncertainty in the water quality variables along the monitoring stations of Bahr El Baker drain due to its robustness and calculation easiness. Also, the cluster analysis technique (K-Means Algorithm) has been used to classify the monitoring stations entropy to have an objective judgment about the status of the entropy then the action needed to be taken from the decision maker. The results indicate that the entropy values for the water quality variables in the selected monitoring stations can be classified into three main clusters. The first cluster show the monitoring stations that have moderate entropy (include 2 monitoring stations) while the second cluster designate the monitoring stations that have the high entropy values (include 6 monitoring stations) and finally the third cluster that indicate the monitoring stations that have low entropy values (include 2 monitoring stations). According to the

results of entropy information, it can be concluded that most of the entropy values for most of the monitoring stations along the drain system for the selected water quality variables have moderately high entropy values and lay among the maximum and minimum values calculated. Based on that the analysis process for the water samples at these locations may need check and review and finally the locations which have highest entropy values should have a complete check in the water sampling process in the field and the physical and chemical analysis procedures in the laboratory as well. Finally, the Entropy concept could be used as a diagnostic procedure to the reliability of the information extracted from the collection process of data from water quality monitoring networks.

References

Amorocho, J., Esplidora, B., 1973. Entropy in the assessment of uncertainty of hydrologic systems and models. Water Resources Research 9, 1522–1551.

Awadallah, A.G., 2012. Selecting optimum locations of rainfall stations using kriging and entropy. International Journal of Civil & Environmental Engineering IJCEE-IJENS 12 (01).

Bueso, M.C., Angulo, J.M., Cruz-Sanjulian, J., Carcia-Arostegui, J.L., 1999. Optimal spatial sampling design in a multivariate framework. Mathematical Geology 31 (5), 507–525.

Cavanagh, N., Nordin, R.N., Pommen, L.W., Swain, L.G., 1998. Guidelines for Designing and Implementing a Water Quality Monitoring Program in British Columbia. British Columbia Ministry of Environment, Lands and Parks, Available from: http://www.llbc.leg.bc.ca/Public/PubDocs/bcdocs/323987/design_guidelnes.pdf

Harmancioglu, N.B., Alpaslan, M.N., 1997. Redesign of water quality monitoring networks. In: Refsgaard, J.C., Karalis, E.A. (Eds.), Operational Water Management. A.A. Balkema, Rotterdam, pp. 57–64.

Harmancioglu, N.B., Alpaslan, N., Singh, V.P., 1994. Assessment of the entropy principle as applied to water monitoring network design. In: Hipel, K.W., Mcleod, A.I., Panu, U.S., Singh, V.P. (Eds.), Stochastic and Statistical Methods in Hydrology and Environmental Engineering, vol. 3. Kluwer, Dordrecht, pp. 135–148.

Harmancioglu, N.B., Fistikoglu, O., Ozkul, S.D., Singh, V.P., Alpaslan, M.N., 1999. Water Quality Monitoring Network Design. Kluwer, Boston, pp. 299.

Hausser, J., Strimmer, K., 2009. Entropy inference and the James–Stein estimator, with application to nonlinear gene association networks. Journal of Machine Learning Research 10, 1469–1484.

Jaynes, E.T., 1957. Information theory and statistical mechanics I. Physics Revision 106, 620–650.

Karaman, H.G., (Ph.D. thesis) 2007. Riskand Uncertainty Assessment Applied in Water Quality Management: A Probabilistic Decision Support System. Faculty of Engineering, Cairo University.

Karamouz, M., Nokhandan, A., Kerachian, R., Maksimovic, C., 2009. Design of on-line river water quality monitoring systems using the entropy theory: a case study. Environmental Monitoring and Assessment 155, 63–81.

Krastanovic, P.F., Singh, V.P., 1992. Evaluation of rainfall networks using entropy II. Water Resources Management 6, 295–314.

Mogheir, Y., Singh, V.P., 2002. Application of information theory to groundwater quality monitoring networks. Water Resources Management 16 (1), 37–49.

Mogheir, Y., De Lima, J.L.M.P., Singh, V.P., 2004a. Characterizing the spatial variability of groundwater quality using the entropy theory: I. Synthetic data. Hydrological Processes 18, 2165–2179.

Mogheir, Y., De Lima, J.L.M.P., Singh, V.P., 2004b. Characterizing the spatial variability of groundwater quality using the entropy theory: II. Case study from Gaza Strip. Hydrological Processes 18, 2579–2590.

NAWQAM (National Water Quality and Availability Management), 2002. Water Quality Status for Surface and Ground Water Baseline Report October 1999–September 2001.

Ozkul, S., Harmancioglu, N.B., Singh, V.P., 2000. Entropy-based assessment of water quality monitoring networks. Journal of Hydrologic Engineering, American Society of Civil Engineers 5 (1), 90–100.

R Development Core Team, 2011. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, Available from: http://www.R-project.org

Rajagopal, A.K., Teitler, S., Singh, V.P., 1987. Some new perspectives on maximum entropy techniques in water resources research. In: Singh, V.P. (Ed.), Hydrologic Frequency Modelling. Reidel, Dordrecht, pp. 247–366.

Sanders, T.G., Ward, R.C., Loftis, J.C., Steele, T.D., Adrian, D.D., Yevjevich, V., 1983. Design of Networks for Monitoring Water Quality. Water Resources Publications, Littleton, CO, 328 pp.

Shaban, M., (A Thesis Submitted for the Fulfillment of the Requirements for The Degree of Doctorate in Natural Sciences) 2007. Spatial and Temporal Consolidation of Drainage Water Quality Monitoring Networks: A Case Study From The El-Salam Canal Project in Egypt. Universität Lüneburg, Germany.

Shannon, C.E., Weaver, W., 1949. The Mathematical Theory of Communication. University of Illinois Press, Urbana, IL.

Singh, V.P., 1997. The use of entropy in hydrology and water resources. Hydrological Processes 11, 587–626.

Singh, V.P., 1998. Entropy-Based Parameter Estimation in Hydrology. Kluwer, Boston.

Singh, V.P., Rajagopal, A.K., 1987. Some Recent Advances in Application of the Principle of Maximum Entropy (POME). International Association of Hydrological Sciences Publication, pp. 353–364.