

A spatiotemporal Bayesian maximum entropy-based methodology for dealing with sparse data in revising groundwater quality monitoring networks: the Tehran region experience

Zahra Alizadeh¹ · Najmeh Mahjouri¹

Received: 9 March 2017 / Accepted: 14 June 2017 / Published online: 26 June 2017
© Springer-Verlag GmbH Germany 2017

Abstract Data inadequacy is a common problem in designing or updating groundwater monitoring systems. The developed methodologies for the optimal design of groundwater monitoring systems usually assume that there is a complete set of data obtained from existing monitoring wells and provide a revised configuration for the system by analyzing the current data. These methodologies are not usually applicable when the current groundwater quantity and quality data are highly sparse. In this paper, a new simulation–optimization approach based on Bayesian maximum entropy theory (BME) is proposed for revising spatial and temporal monitoring frequencies in a sparsely monitored aquifer. The BME is used to simulate the spatial and spatiotemporal variations of groundwater indicators, incorporating the space/time uncertainties due to insufficient data. Comparing the obtained estimations with observations, the best BME model was selected to be linked with an optimization model. The main goal of optimization was to find out the spatial and temporal sampling characteristics of the monitoring stations using the concepts of Entropy theory and a groundwater vulnerability index. The results show the BME estimations are less biased and more accurate than Ordinary Kriging in both spatial and spatiotemporal analysis. The improvements in the BME estimates are mostly related to incorporating hard (accurate) and soft (uncertain) data in the estimation process. The applicability and efficiency of the

proposed methodology have been evaluated by applying it to the Tehran aquifer in Iran which is suffering from high groundwater table fluctuations and nitrate pollution. Based on the results, in addition to the existing monitoring wells, seven new monitoring stations have been proposed. Few stations which potentially can be removed or combined with other stations have been identified and a monthly sampling frequency has been suggested.

Keywords Bayesian maximum entropy (BME) · Geostatistics · Kriging · Updating groundwater monitoring network · Optimization

Introduction

A monitoring network can provide data and information to achieve or improve understanding of the state of groundwater quantity and quality and its changes over time, which provides data and information required for aquifer management along with operating or updating an existing monitoring system.

Geostatistical approaches, especially the Kriging method, are among the most widely used approaches to interpolate spatiotemporal data required for optimization of groundwater monitoring networks. Kriging is mostly used for regionalizing or interpolating variables in different points of an area where no measurement exist. Several interesting applications of Kriging in the area of evaluation and optimization of groundwater monitoring systems can be found in the literature. For example, Dhar (2013) developed a methodology based on the Kriging as an external model for spatial estimation of piezometric head and the concentration of water quality indicators for optimal design of groundwater monitoring network. To

✉ Najmeh Mahjouri
mahjouri@kntu.ac.ir

Zahra Alizadeh
alizadeh.zahra92@gmail.com

¹ Faculty of Civil Engineering, K. N. Toosi University of Technology, Tehran, Iran

determine the number and location of monitoring stations, an optimization model with objectives of minimizing the estimation error of the values of piezometric head and water quality indicators was developed. Bhat et al. (2015) applied a geostatistical method to determine the optimum number and locations of monitoring wells which provide more useful groundwater data compared to an existing monitoring network. The redesigned network reduced the mean prediction standard error compared to the former network.

Several other successful applications of Kriging method have shown the efficiency of this method in optimal design or evaluation of groundwater monitoring networks (Theodossiou and Latinopoulos 2006; Triki et al. 2012; Datta and Singh 2014; Ran et al. 2015), especially when not dealing with highly sparse data.

Varouchakis and Hristopulos (2013) compared the performance of some deterministic interpolation methods, such as inverse distance weight (IDW) and minimum curvature (MC) with stochastic methods of Ordinary Kriging (OK) and Universal Kriging (UK) and showed better performance of the stochastic methods comparing to the deterministic methods especially when dealing with uncertainties.

To incorporate the uncertainties resulting from sparse monitoring data or errors inherent in the models, one can either use additional monitoring wells and more frequent samplings, which can be budgetary inefficient, or a more suitable method for interpolating scattered data to map the groundwater quantity and quality variability. The Bayesian maximum entropy (BME) of modern geostatistics combines various types of information for more accurate estimation of groundwater quantity and quality variables at desired locations and times. These estimations can be expressed with some degree of uncertainty reflecting the uncertainty inherent in the underlying information (LoBueglio et al. 2007).

The Bayesian maximum entropy, as a non-linear geostatistical approach, balances two requirements. The first requirement incorporates the prior information and knowledge related to the spatial variability of the estimated variables which involves the maximization of an entropy function. The second requirement which leads to a posterior probability with minimum uncertainty involves the maximization of a Bayesian function (Christakos 1990).

Coulliette et al. (2009) integrated a hydrologic-driven mean trend model in a BME framework to obtain informative space/time maps of fecal contamination. Money et al. (2009) used BME for integrating monitored and predicted water quality data to produce maps of estimated concentration along a river basin. They showed that by adding soft data, as secondary information in the BME structure, the estimation error decreased by about 30%. Yu

and Chu (2010) used BME for estimating and analyzing changes of groundwater level using monthly spatiotemporal piezometric heads from 66 wells.

Bayat et al. (2012) modeled spatial and spatiotemporal variations of annual precipitation with and without incorporating elevation variations using the BME and Ordinary Kriging (OK) methods. They showed that more detailed and reliable results were achieved using the BME estimation. In another research, Bayat et al. (2014) used OK and BME spatiotemporal analysis for producing meteorological drought occurrence probability maps and illustrated the superiority of BME over OK in their work.

Studying previous works shows that the BME concept has not been widely used in water resources management problems. In this paper, a BME spatiotemporal simulation–optimization model is developed for revising and updating groundwater monitoring networks. The uncertainties resulted from sparse data are incorporated through the use of interval information which is referred to as soft data in the BME. Finally, an optimal set of number and locations of the monitoring wells is proposed through the use of the concepts of marginal entropy, transient information and a vulnerability index. Details of the proposed methodology will be discussed in the following section.

Methodology

A framework of the proposed methodology is given in Fig. 1. The structure of the presented framework is discussed in the following sections.

Selecting the water quantity and quality indicators

Usually, different pollution sources are responsible for degrading groundwater quality and increasing the concentration of many water quality variables (Mahab Ghods Consulting Engineers 2008). It is not cost-effective to choose several water quality variables for designing or updating a groundwater monitoring system (Asadollahifard 2015). Therefore, it seems reasonable to consider only few water quality indicators which can serve as representatives of water quality condition of the aquifer.

Preparing data and information

After collecting the required data and information from the existing monitoring stations, they should be prepared to be used in models. Usually, it is necessary to normalize probability density function of data. To do this, different transform functions such as Box-Cox, logarithm and square root transformations can be utilized. Also, terms of trend and seasonality are eliminated from time series of data.

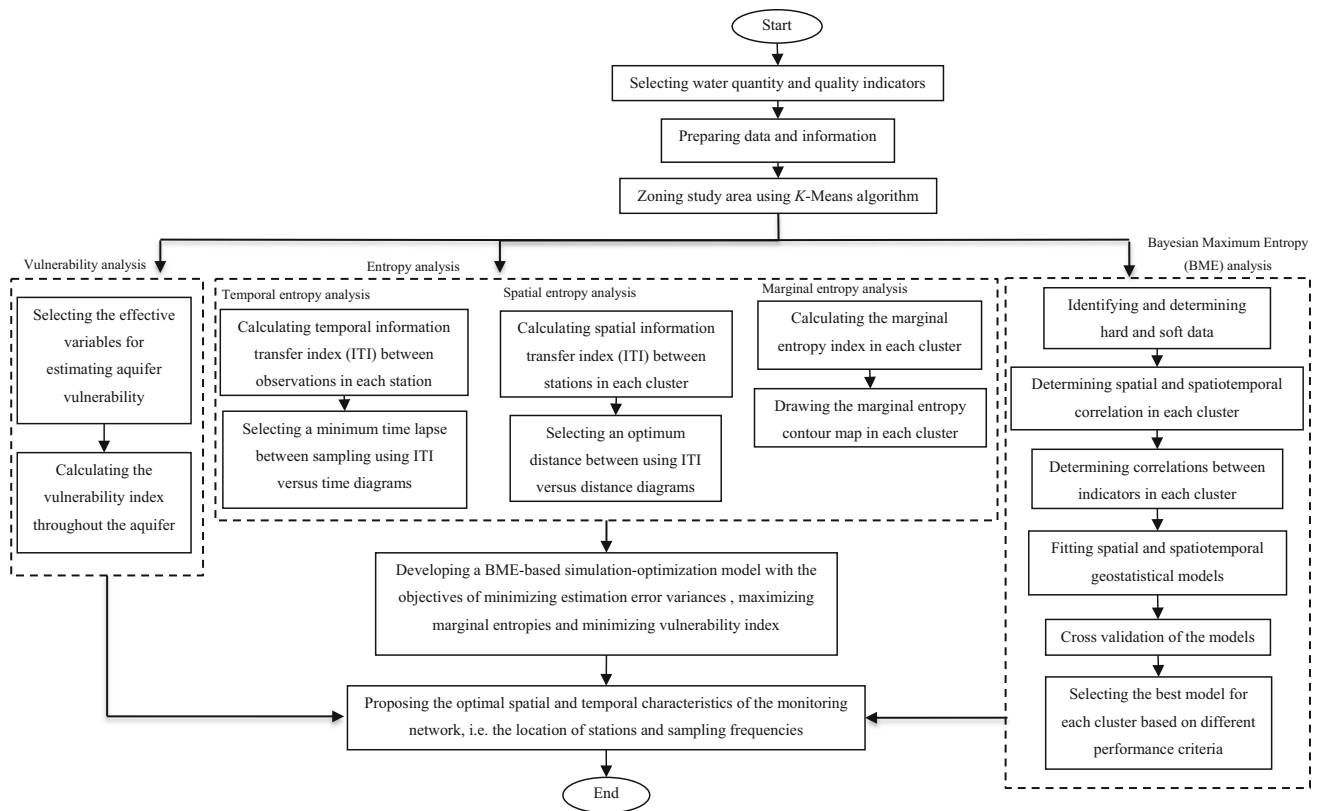


Fig. 1 A framework of the proposed methodology

Data clustering using *K*-means algorithm

The value of groundwater quantity and quality indicators can highly vary throughout the aquifer due to the spatial distribution of pollution sources and aquifer heterogeneity. Therefore, to improve the estimation accuracy of groundwater indicators, it is suggested to categorize the existing stations into clusters. In this paper, existing monitoring wells are clustered based on their locations and values of the observed water quantity and quality data using the *k*-means clustering method (Zalik 2008). In the *k*-means clustering, *n* observations are segmented into *k* clusters in a way that each observation goes to the cluster with the mean to which it is closest. Details of this method can be found in Zalik (2008).

Analysis of groundwater quantity and quality variations using the Bayesian maximum entropy (BME)

Geostatistics is a specialized branch of statistical analysis which concerns spatial or spatiotemporal correlations among data in a two- or three-dimensional coordinate space. In this paper, the geostatistical method of BME, which incorporates the uncertainties resulting from sparse

monitoring data, is used for estimation of the values of water quantity and quality indicators throughout the aquifer. Different knowledge bases and shapes of physical knowledge are considered and combined in the BME (Christakos et al. 2002). In the BME, two requirements are balanced: first, high prior information, through maximization of an entropy function and second, high posterior probability about the estimated data, which results in minimum uncertainty attached, through maximization of a Bayes function (Christakos 1990). One of the major advantages of the BME over most of the other geo-statistic-based methods is that it addresses uncertainties by considering two types of data, namely soft and hard data. These types of data will be discussed in the following section.

Selecting hard and soft data

To incorporate the uncertainties in the existing groundwater quantity and quality data, the data are divided into two categories of soft and hard. Hard data represent accurate measurements or values calculated using numerical simulations. They are mainly considered accurate with negligible measurement errors. Soft data, representing uncertainties of estimation, usually contain qualitative

statements and/or incomplete and uncertain observations which can be expressed in the form of interval values, probability statements and empirical charts (Christakos et al. 2002). For example, if soft data are represented in interval numbers, with the lower and higher bounds of data, a uniform distribution is fitted to them. For probabilistic presentation, Gaussian or Normal distribution can usually be a useful choice (Kotulski and Szczepinski 2010), since this distribution is a good representative of the probability distribution of many types of data. In this paper, to classify hard and soft data, all observations in any station are considered as hard data, while observations with gaps or insufficient measurements are addressed as soft data. In bivariate estimations, the soft data were defined in interval form and their lower and upper bounds were selected based on the recorded data.

Determining spatial and spatiotemporal correlations

The geostatistical methods are based on relations between observations in time or space. Measurements made at different locations mostly are spatially dependent. For example, observations from nearby locations may be closer in value compared to those from locations farther apart. This fact is also applied to temporal data. In the geostatistical analysis, spatial and spatiotemporal dependence among monitoring stations can be taken into account.

In addition to spatial and spatiotemporal correlation within observations of the same variable, correlations usually exist between different groundwater variables. To increase the accuracy of the estimations, in addition to correlations within observations, one can take advantage of dependence between different groundwater quantity and quality variables through multivariate analysis. In a multivariate geostatistical interpolation, a secondary or auxiliary variable can be used to improve the estimation accuracy of the main or primary variable. The secondary variable usually has less variability and a relatively high correlation with the main variable. This can sometimes improve estimation accuracy of a less densely sampled primary variable.

Fitting spatial and spatiotemporal univariate and bivariate models

The analysis of spatial and spatiotemporal variations of groundwater quantity and quality indicators is done using the BME. Hard data are used to obtain variogram/covariance models. These models illustrate the variation of correlations between data and distance. The optimum structure of covariance models is used for estimating the values of water quantity and quality indicators in the aquifer.

Spatiotemporal covariance (C_{st}) and variogram (γ_{st}) are defined as Eqs. (1) and (2) (De Cesare et al. 2001):

$$C_{st}(h) = \text{Cov}(\vec{Z}(\vec{s} + h_s, t + h_t), Z(\vec{s}, t)) \quad (1)$$

$$\gamma_{st}(h) = \frac{\text{Var}(Z(\vec{s} + h_s, t + h_t), Z(\vec{s}, t))}{2} \quad (2)$$

where h_s and h_t are spatial and temporal lags, respectively. In this paper, Z represents groundwater quantity and quality indicators, $\vec{s} = (s_1, s_2)$ and t represent spatial (two-dimensional) and temporal coordinates, respectively. The general mathematical structure of covariance and variogram models has been described by De Cesare et al. (2001) as Eq. (3):

$$C_{st}(h_s, h_t) = C_s(h_s) \times C_t(h_t) \quad (3)$$

where C_s and C_t are spatial and temporal covariances, correspondingly. To evaluate the estimated variograms/covariances, leave-one-out cross-validation analysis is used. The statistics of the coefficient of determination (R^2) and Nash–Sutcliffe efficiency (NSE) are used to evaluate the estimated variograms (James et al. 2013).

Entropy analysis

Entropy is a measure of uncertainty in the information content. Monitoring stations with higher information content would generally provide more valuable data. Therefore, when selecting an optimum set of stations among a number of candidates, stations with higher information content may be given a higher priority over stations with lower information content (Yang and Burn 1994).

In this paper, the concept of marginal entropy is used to measure the information content of observations throughout the aquifer (Mahjouri and Kerachian 2011). Also, the concept of information transient index (ITI) is used to measure mutual information or information transferred among the stations. Shannon and Weaver (1949) defined the marginal entropy of a discrete random variable x as Eq. (4):

$$E(I(x)) = H(x) = \sum_{i=1}^N p(x_i) \log p(x_i) \quad (4)$$

where, N represents the number of events x_i with probabilities $p(x_i)$ ($i = 1, \dots, N$). Marginal entropy contour maps can be used to evaluate the existing groundwater monitoring network and selecting the optimum monitoring locations for the revised network.

Transinformation ($T(X, Y)$) which measures the redundant or mutual information between X and Y , can be calculated as follows (Mogheir et al. 2004a, b):

$$T(X, Y) = \sum_{i=1}^n \sum_{j=1}^m p(x_i, y_j) \ln \left[\frac{p(x_i, y_j)}{p(x_i)p(y_j)} \right] \tag{5}$$

where $p(x_i)$ and $p(y_j)$ are probabilities of occurring X and Y , respectively. Also, $p(x_i, y_j)$ is the probability of occurring random variables X and Y both at the same time. The standardized information transferred from one monitoring station to another is called Information transfer index (ITI) which is computed as Eq. (6):

$$ITI = \frac{T(X, Y)}{H(X, Y)} \tag{6}$$

In Eq. (6), $H(X, Y)$ is total entropy of two independent random variables X and Y :

$$H(X, Y) = H(X) + H(Y) \tag{7}$$

In this paper, to find out the optimum time laps between samplings, temporal information transfer index (ITI) between observations in each monitoring station is calculated. In the temporal entropy analysis, the transferred information between observations in different time steps are computed and a graph of the calculated temporal ITIs versus time laps is obtained.

Assessment of the aquifer vulnerability

Infiltration of domestic wastewater into groundwater, which increases the nitrate concentration may cause the allowed standards for water quality variables to be exceeded. To determine the spatial variability of the potential of aquifer for being polluted, a vulnerability index can be used. The vulnerability index is calculated based on the topographical and climatological factors including groundwater depth and slope, groundwater flow parameters such as hydraulic conductivity and rainfall depth and distribution. This index will be used in the optimization model for taking account of areas with higher risk of contamination when choosing monitoring locations. The vulnerability index is computed using four weighted variables of groundwater depth, rainfall depth, aquifer hydraulic conductivity and slope and is calculated as follows:

$$D_i = \sum_{j=1}^4 (w_j \times Q_j) \tag{8}$$

where D_i the value of the vulnerability index; w_j the j th variable weight; Q_j the j th variable value.

Developing a simulation–optimization model based on the BME and entropy analyses

A simulation–optimization model is developed to identify the optimal locations of groundwater monitoring wells

among a number of potential locations. The objective is to minimize the sum of estimation error variances corresponding to the values of groundwater quality indicators at potential locations, which are calculated using the BME, the values of marginal entropy, and the amount of transferred information, the values of vulnerability index calculated at stations. The objective function is defined as follows:

$$OF = \text{Min} \left(\sum_{j=1}^n \left[\sum_{i=1}^m w_j^{\text{var}} \cdot (\text{EEV}_{ij} + \sum_{j=1}^m w_j^{\text{ent}} \cdot (\text{Ent}_{ij}) + (w^{\text{vul}} \cdot \text{Vul}_i) + 0.5 \sum_{k=1}^n |r_{ik} - r_{\text{opt}}| / (r_{\text{max}} - r_{\text{opt}}) \right] \right) / n \tag{9}$$

where n number of candidate locations; m number of groundwater quality indicators; EEV estimation error variance; Ent marginal entropy values; Vul vulnerability index values; r_{ik} distance between two candidate locations; r_{opt} optimum distance between two stations; r_{max} maximum distance among candidate locations; w^{var} weight of estimation error variance; w^{ent} weight of marginal entropy weight; r_{opt} denotes the maximum required distance between the stations at which ITI values become negligible. Also, Eq. (9) illustrates a minimum required distance between stations at which ITI values are considerable.

The result of the optimization model is a trade-off curve between values of the objective function and the number of stations. Based on the available budget and the desired accuracy, the number of monitoring stations along with their corresponding optimum layout can be chosen by the decision maker. Usually, after a certain number of potential monitoring stations, no significant improvement of the objective functions is observed. This point can potentially be selected as an upper bound for the number of required stations, for which the best arrangement of the monitoring wells are to be found.

The study area

The proposed methodology is applied for the optimal redesign of the groundwater quality monitoring system in the Tehran aquifer, Iran. Tehran aquifer is bounded by the Kan River in the West and the Sorkhehesar River in the East. About one billion cubic meters of water is annually provided for domestic consumption of 12 million people in the Tehran metropolitan area. More than 60% of consumed water in this city returns to the Tehran aquifer via traditional absorption wells. Figure 2 shows a Google Earth image of the Tehran region. Tehran aquifer is mainly recharged by inflow at the boundaries, precipitation, local

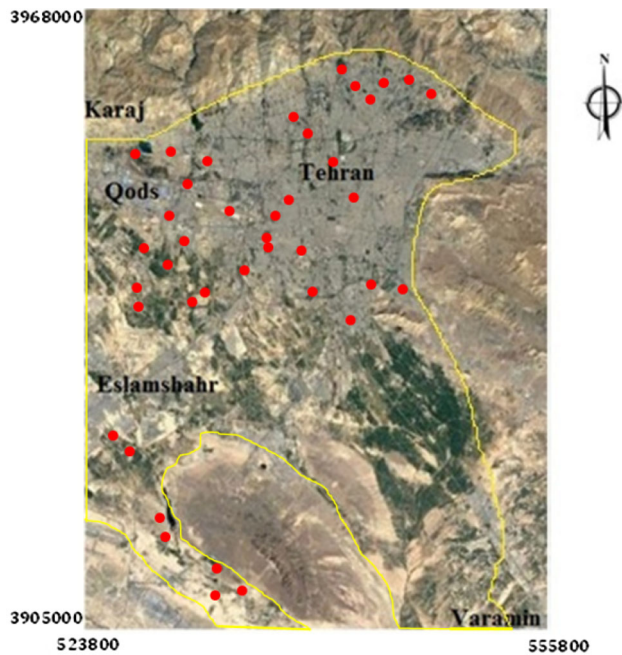


Fig. 2 Boundaries of the study area and location of existing monitoring wells Coordinates are in UTM (Universal Transverse Mercator)

rivers and return flows from domestic, industrial and agricultural sectors. The main characteristics of the Tehran aquifer are presented in Table 1.

Wastewater disposal in Tehran is carried out through more than three million absorption wells, which are often 15–20 m deep. The use of absorption wells has caused groundwater pollution and a significant rise of the water table in the southern part of Tehran (Bazargan-Lari et al. 2009; Kerachian et al. 2010). According to the existing groundwater quality data, some water quality variables

such as total dissolved solids, nitrate, and coliform bacteria are violating the standards (Masoumi and Kerachian 2008; Rafipour-Langeroudi et al. 2014).

Results and discussion

In this paper, groundwater quality indicators are selected considering the locations of the pollution sources and concentration of the water quality variables. The main pollution source in the study area is municipal wastewater which is discharged into the groundwater through absorption wells. The concentration of water quality variables such as nitrate (NO_3) and total dissolved solids (TDS) violate water quality standards in many regions of the aquifer. Therefore, these two water quality variables along with groundwater depth are considered as the main indicators.

The observed data are normalized and prepared for using in the BME simulation model. In order to increase the estimation accuracy, the *k*-means method is used to spatially cluster data into two groups based on the observations of the water quality indicators. The resulted clusters are mostly in accordance with the spatial variations of the observed data (Fig. 3); that is, the higher pollution concentrations mostly occur in the southern part of the aquifer and pollution density in the north of the aquifer is generally lower.

For evaluating the performance of spatial and spatiotemporal models, leave-one-out cross-validation procedure is used. The performance criteria calculated for the spatial and spatiotemporal Kriging and BME models are shown in Tables 2 and 3. The best covariance model

Table 1 Main characteristics of the Tehran aquifer

Properties	Units	Value
UTM coordinates		
X	Meters	5,23,800–5,55,800
Y	Meters	3,905,000–3,968,000
Period of observations	Year	2002–2012
Interval of monthly observed NO_3	mg/L	6.4–152
Interval of monthly observed total dissolved solids (TDS)	mg/L	420–1380
Range of variations of monthly observed piezometric head	Meters	10–147
Average/standard deviation (NO_3 observations)		
Cluster 1	mg/L	15.8/8.5
Cluster 2	mg/L	26.5/19.1
Average/standard deviation (TDS observations)		
Cluster 1	mg/L	1311.5/201.1
Cluster 2	mg/L	594.8/154.6
Average/standard deviation (piezometric head observations)		
Cluster 1	Meters	21.3/8.1
Cluster 2	Meters	69.7/21.5

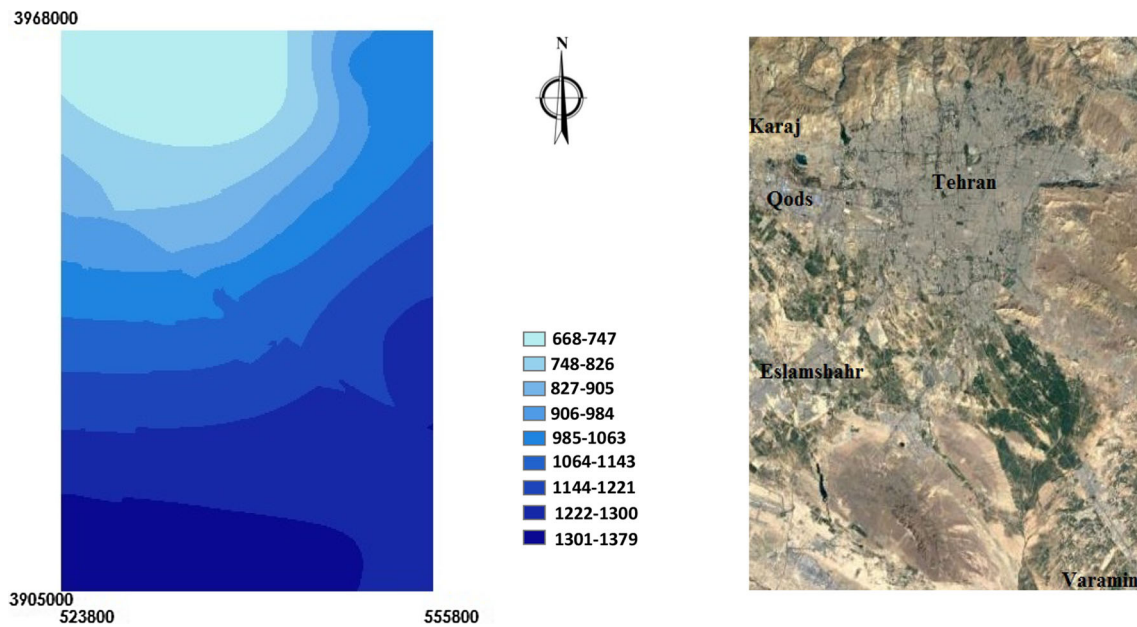


Fig. 3 Spatial variations of the average concentration of TDS (mg/L) (2002–2012)

Table 2 Performance criteria calculated for the spatial Kriging and BME univariate and bivariate models

Variables	Method	Cluster number	The best fitted curve	Cross-validation performance criteria	
				R^2 ^a	NSE ^b
TDS (univariate)	Kriging	1	Spherical	0.61	0.59
		2	Gaussian	0.74	0.68
	BME	1	Spherical	0.9	0.81
		2	Gaussian	0.9	0.81
Depth (univariate)	Kriging	1	Spherical	0.81	0.76
		2	Gaussian	0.8	0.73
	BME	1	Spherical	0.82	0.77
		2	Gaussian	0.86	0.85
NO ₃ (univariate)	Kriging	1	Spherical	0.56	0.49
		2	Spherical	0.43	0.38
	BME	1	Spherical	0.57	0.51
		2	Spherical	0.44	0.39
TDS (bivariate)	Kriging	1	Spherical	0.78	0.71
		2	Gaussian	0.83	0.77
	BME	1	Spherical	0.98	0.9
		2	Gaussian	0.95	0.83

^a Coefficient of determination (Cameron and Windmeijer 1997)

^b Nash–Sutcliffe efficiency (Nash and Sutcliffe 1970)

corresponding to each indicator in every cluster is selected based on the values of the performance criteria. The spatial variation of the nitrate concentration in the study area is very high due to numerous wastewater absorption wells. This problem has affected the fitted variograms, and the performance criteria in fitting variograms for estimating nitrate concentration are not very high.

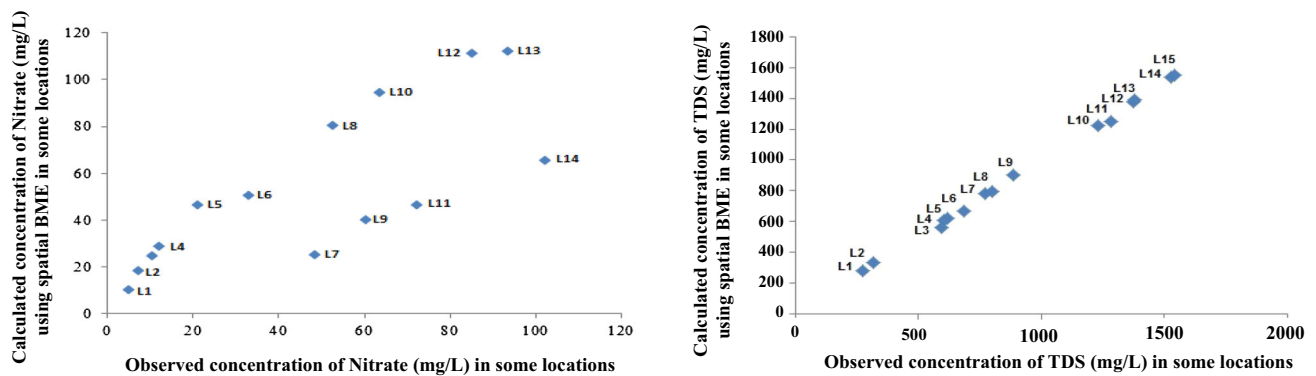
According to Tables 2 and 3, the spatiotemporal models have not performed as suitably as spatial models which can be due to temporal data insufficiency. Also, estimation error variance for BME results is less than that of the OK in both spatial and spatiotemporal analyses. Thus, the spatial BME and estimation error variances calculated for the simulation results are used in the simulation–optimization

Table 3 Performance criteria calculated for the spatiotemporal Kriging and BME models

Variables	Method	Cluster	The best fitted surface	Cross-validation performance criteria	
				R^2 ^a	NSE ^b
TDS (univariate)	Kriging	1	Spherical–Spherical	<0.5	<0.5
		2	Spherical–Spherical	<0.5	<0.5
	BME	1	Spherical–Spherical	<0.5	<0.5
		2	Spherical–Spherical	<0.5	<0.5
Depth (univariate)	Kriging	1	Spherical–Spherical	0.81	0.78
		2	Spherical–Exponential	0.91	0.85
	BME	1	Spherical–Spherical	0.81	0.79
		2	Spherical–Exponential	0.92	0.88
NO ₃ (univariate)	Kriging	1	Exponential–Spherical	<0.5	<0.5
		2	Exponential–Exponential	<0.5	<0.5
	BME	1	Exponential–Spherical	<0.5	<0.5
		2	Exponential–Exponential	<0.5	<0.5

^a Coefficient of determination (Cameron and Windmeijer 1997)

^b Nash–Sutcliffe efficiency (Nash and Sutcliffe 1970)

**Fig. 4** Observed and calculated concentration of nitrate and TDS using the spatial BME in some representative locations

model. As an example, Fig. 4 shows the variation of the observed values with the calculated ones for nitrate and TDS in some representative locations. These stations are selected as representatives in a way that they spatially cover the whole study area. Also, the maps of estimated nitrate and TDS and corresponding maps of variances of estimation error obtained using the best BME model are presented in Figs. 5 and 6, respectively.

The values of marginal entropy are calculated throughout the aquifer considering the existing information from the 43 groundwater quality monitoring stations and 60 water level monitoring stations in the Tehran aquifer. For example, Fig. 7a, b and c illustrates the marginal entropy maps for nitrate, TDS and depth of ground water, respectively. As seen in these figures, the values of the marginal entropy in some regions are considerably higher than others. This can illustrate more important information in these areas which also shows higher uncertainty resulted from high temporal variability.

The region with the highest marginal entropy has presumably the highest temporal variations in water quantity and quality, and stations in such region have the greatest potential for providing more information (Memarzadeh et al. 2013). Figure 8a shows the values of ITI versus distance between the stations for the nitrate (NO₃) variable. The values of ITI versus distance between the stations for TDS are represented in Fig. 8b. As shown in these figures, the transferred information between the stations is not considerable and no meaningful relationship can be seen between the values of ITI and distance for water quality variables. Also, the ITI values for groundwater depth have larger values compared to those of nitrate and TDS (Fig. 8c). This can be due to higher correlations among the data of groundwater depth in the monitoring wells. It can be inferred that the current stations are not excessive, though, in further analysis, their locations may be improved.

The sampling interval of water quality variables in the existing monitoring system is sometimes up to several

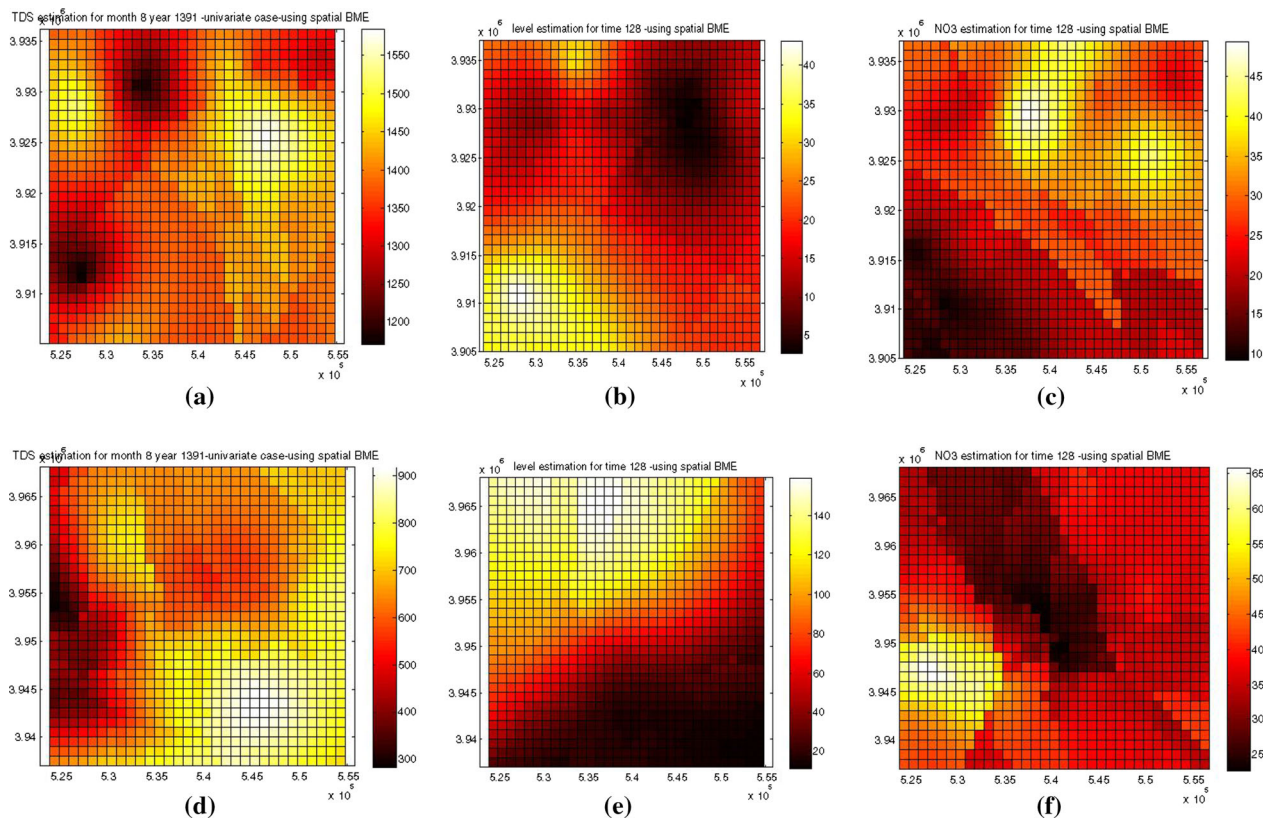


Fig. 5 Maps of variations of estimated nitrate and TDS using spatial univariate BME model, **a** TDS-cluster 1, **b** depth-cluster 1, **c** NO₃-cluster 1, **d** TDS-cluster 2, **e** depth-cluster 2, **f** NO₃-cluster 2

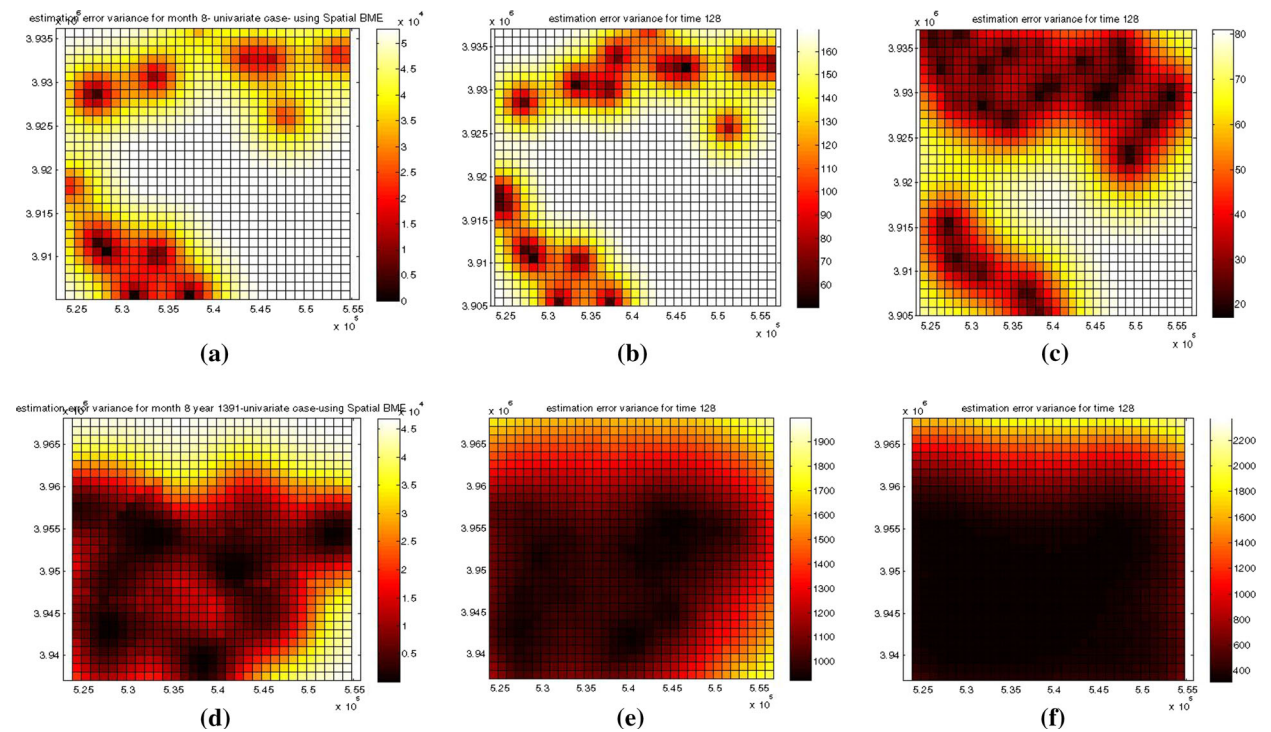


Fig. 6 Maps of variances of estimation error using spatial univariate BME model, **a** TDS-cluster 1, **b** depth-cluster 1, **c** NO₃-cluster 1, **d** TDS-cluster 2, **e** depth-cluster 2, **f** NO₃-cluster 2

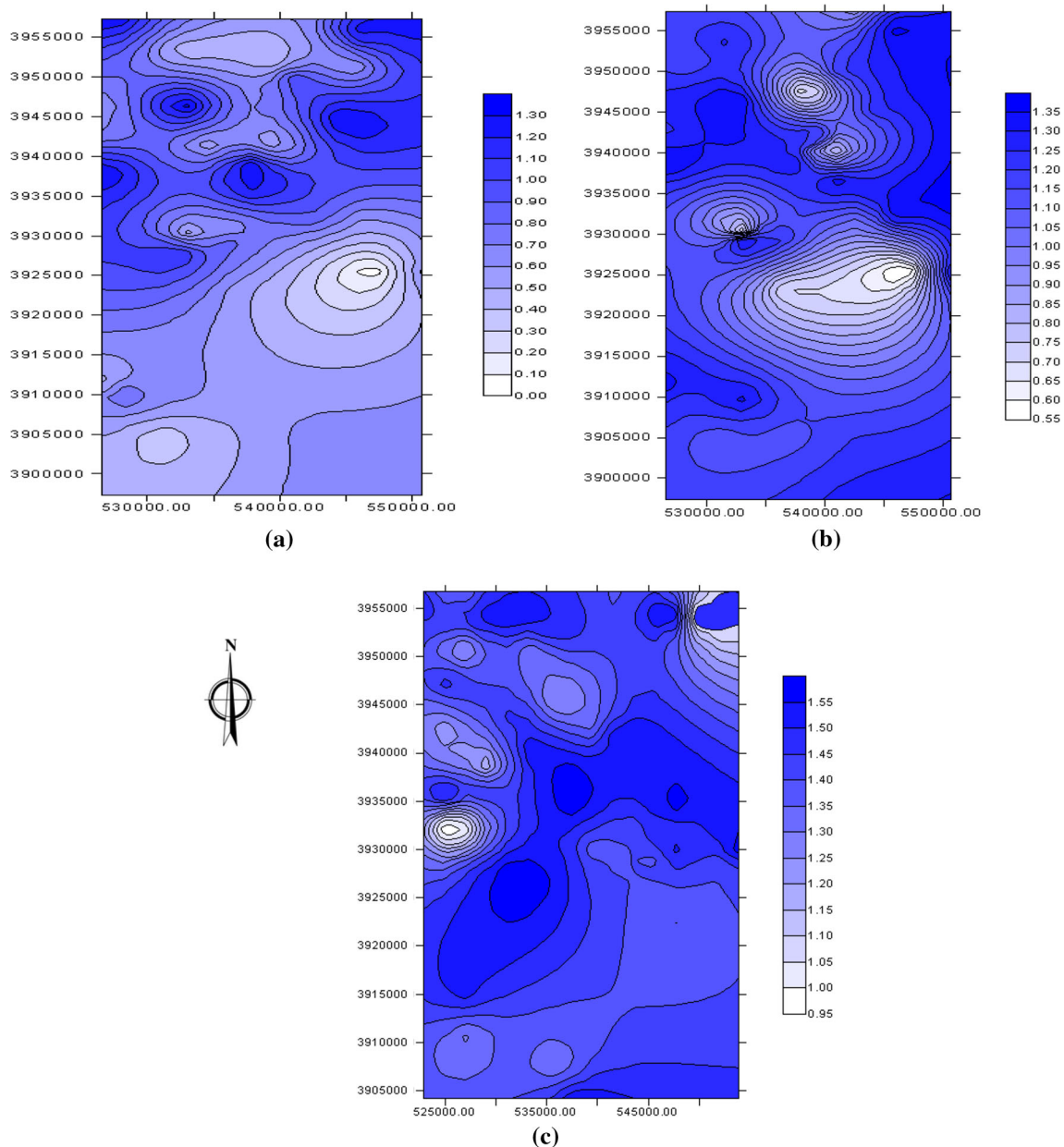


Fig. 7 The spatial distribution of marginal entropy in the study area for **a** NO₃, **b** TDS and **c** depth

months. Using the time series of depth data, the values of the temporal ITI for 1–4 time lags are obtained. Figure 9, for instance, shows the variations of the temporal ITIs versus monthly time steps for some representative stations. This figure illustrates no meaningful pattern in ITI variation with time lags and no correlation between the ITIs and the number of time lags. As seen in this figure, even for 1 month time lag, ITI values are not significant. It can be inferred that temporal sampling interval of 1 month results in very low redundant information in the data.

The contour maps of standardized variables of groundwater depth, rainfall and the vulnerability index are

displayed in Fig. 10. Average values of groundwater depth observed in the monitoring wells from 2002–2012 are applied for groundwater depth mapping. Also, the rainfall mapping is done by using average values of rainfall obtained from synoptic stations.

To find out the optimum locations of the monitoring wells, the selected BME model is linked with an optimization model with the objectives of minimizing the total marginal entropy of the system, the variances of the estimation error and the vulnerability of the system.

The particle swarm optimization (PSO) algorithm (Parsopoulos and Vrahatis 2002) is used to solve the

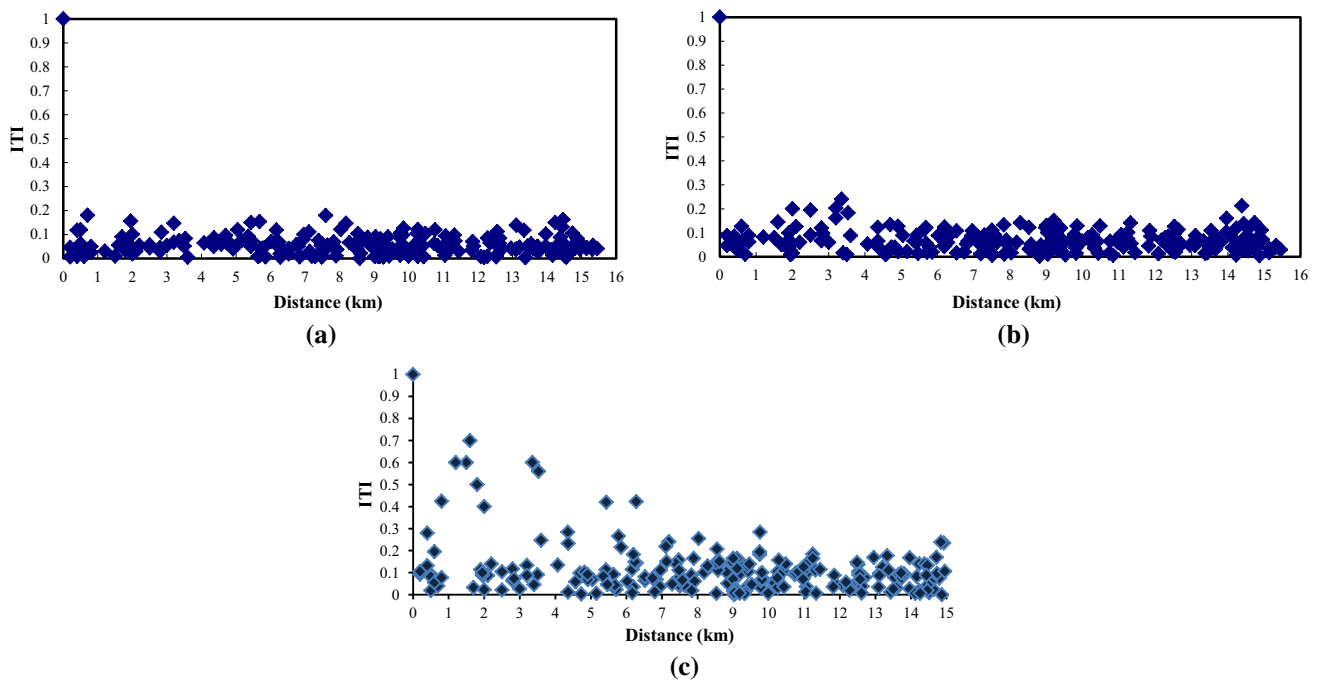


Fig. 8 Variation of information transient index (ITI) versus distance for the indicators **a** NO₃, **b** TDS and **c** groundwater depth

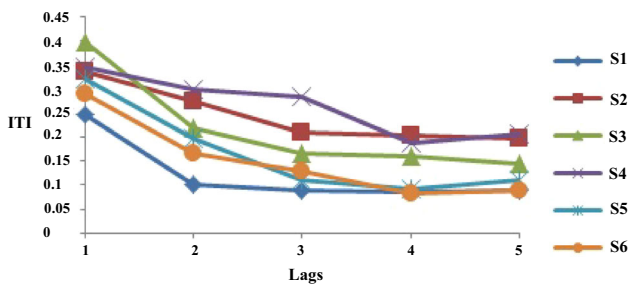


Fig. 9 Variation of temporal ITIs versus monthly lags in some representative stations

optimization problem. The simulation–optimization model is executed considering a number of potential locations for monitoring station. As a result, a trade-off curve between the number of monitoring stations and the corresponding values of the objective function based on the selected locations is obtained. Figure 11 shows the resulting objective function in each cluster. Obviously, as the number of potential stations increases, the value of the objective function increases. The optimum configuration of monitoring wells (the best locations of stations corresponding to a certain number of monitoring stations) can be chosen based on the values for the objective function and available budget. The objective function does not significantly vary after a certain number of stations (25, here). Also, increasing the number of stations mostly leads to an increase in the operational costs; therefore, the number of required stations has been chosen to be 25.

The spatial distribution of the proposed and the current monitoring stations for each cluster is depicted in Fig. 12a, b. According to these figures, some of the current monitoring

stations are in the immediate vicinity of the proposed stations obtained from the simulation–optimization model. Therefore, for efficiency purposes, these proposed stations are replaced by the monitoring wells already in the current monitoring network.

Also, since a number of current monitoring wells are very closely located to each other, to reduce monitoring cost, stations with more transient data with the nearby stations can be omitted. The updated monitoring network in each cluster is displayed in Fig. 13. The final monitoring network in the Tehran aquifer can be seen in Fig. 14.

Summary and conclusions

In this paper, a new approach was proposed for revising and updating monitoring locations and sampling frequencies of an existing groundwater monitoring system which is highly suffering from data inadequacy. The variations of groundwater depth and quality indicators were simulated using the BME model. This BME was especially used to deal with sparse data and incorporate uncertainties caused by insufficient information. The influence of soft data on spatial and spatiotemporal estimations has been evaluated by comparing the results with those of the Ordinary

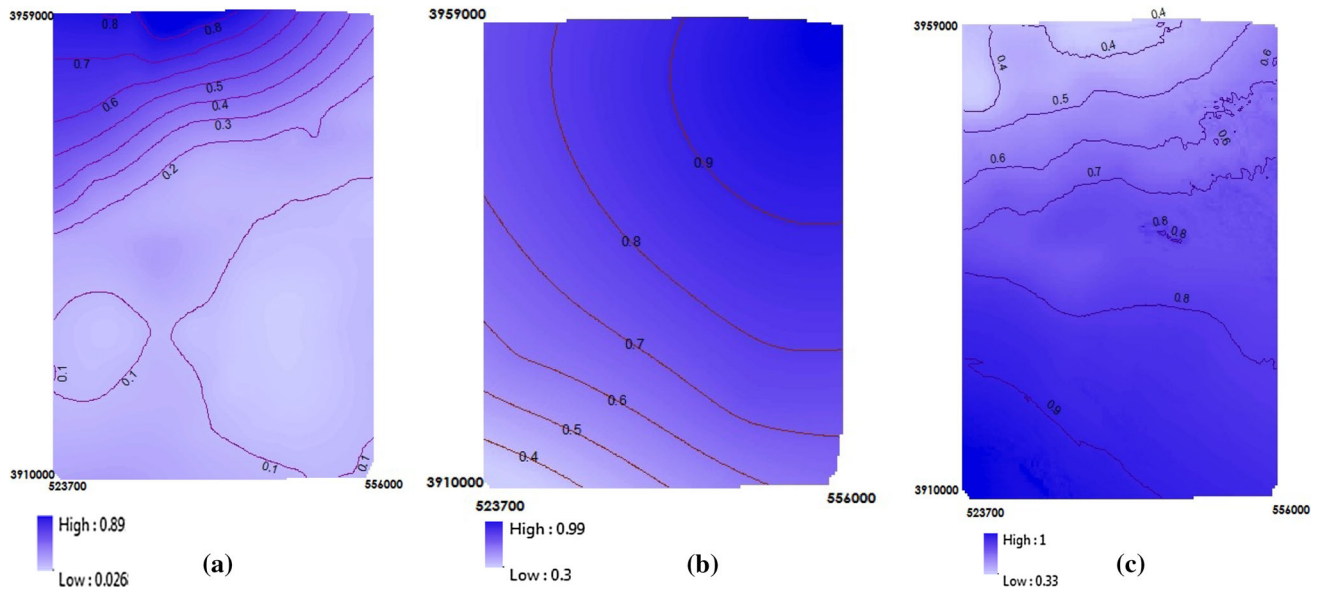
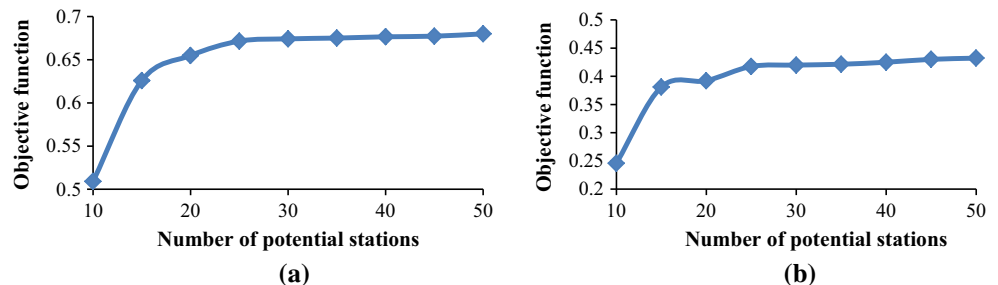


Fig. 10 Contour maps of standardized average **a** groundwater depth, **b** rainfall and **c** vulnerability index (2002–2012)

Fig. 11 The objective function versus the number of potential stations, **a** cluster 1 and **b** cluster 2



Kriging. The cross-validation analysis confirms that incorporating the concept of soft data results in better estimations in terms of reliability and accuracy. The results also showed that using the BME for estimating the concentration of the groundwater quality indicators in the Tehran aquifer, the average variance of estimation error can be reduced by more than 15% in both spatial and spatiotemporal analyses, compared to the results of the OK.

The best BME model was selected to be linked with an optimization model. The optimization model mainly aimed at finding the spatial and temporal sampling characteristics of the monitoring stations using both concepts of Entropy theory and a groundwater vulnerability index. The results of the temporal entropy analysis showed that there is almost no temporal redundant information in the data. In marginal entropy analysis, the areas of higher spatial and temporal variability, which generally needed stricter monitoring, got a higher number of monitoring stations and sampling frequencies. Also, according to information transient index, to reduce the spatial data redundancy, the

set of stations with minimum common information were chosen.

Based on the results of the optimization model, in addition to the existing monitoring wells, seven new monitoring stations have been proposed. Few stations which potentially can be removed or combined with other stations were identified. As expected, more stations were suggested in areas with more spatial heterogeneity, in terms of groundwater quality, such as areas close to high wastewater discharges. By proposing new stations and sampling frequencies, the variance of estimation error can decrease by 20%, compared to the existing monitoring network. On the whole, the results showed that the combined BME-optimization model for revising the existing groundwater monitoring network had an appropriate performance and BME gave mostly reliable and comparatively precise results, when dealing with sparse data.

In future works, to better incorporate the spatiotemporal variations of groundwater level and quality, the results of the proposed BME-based methodology can be combined

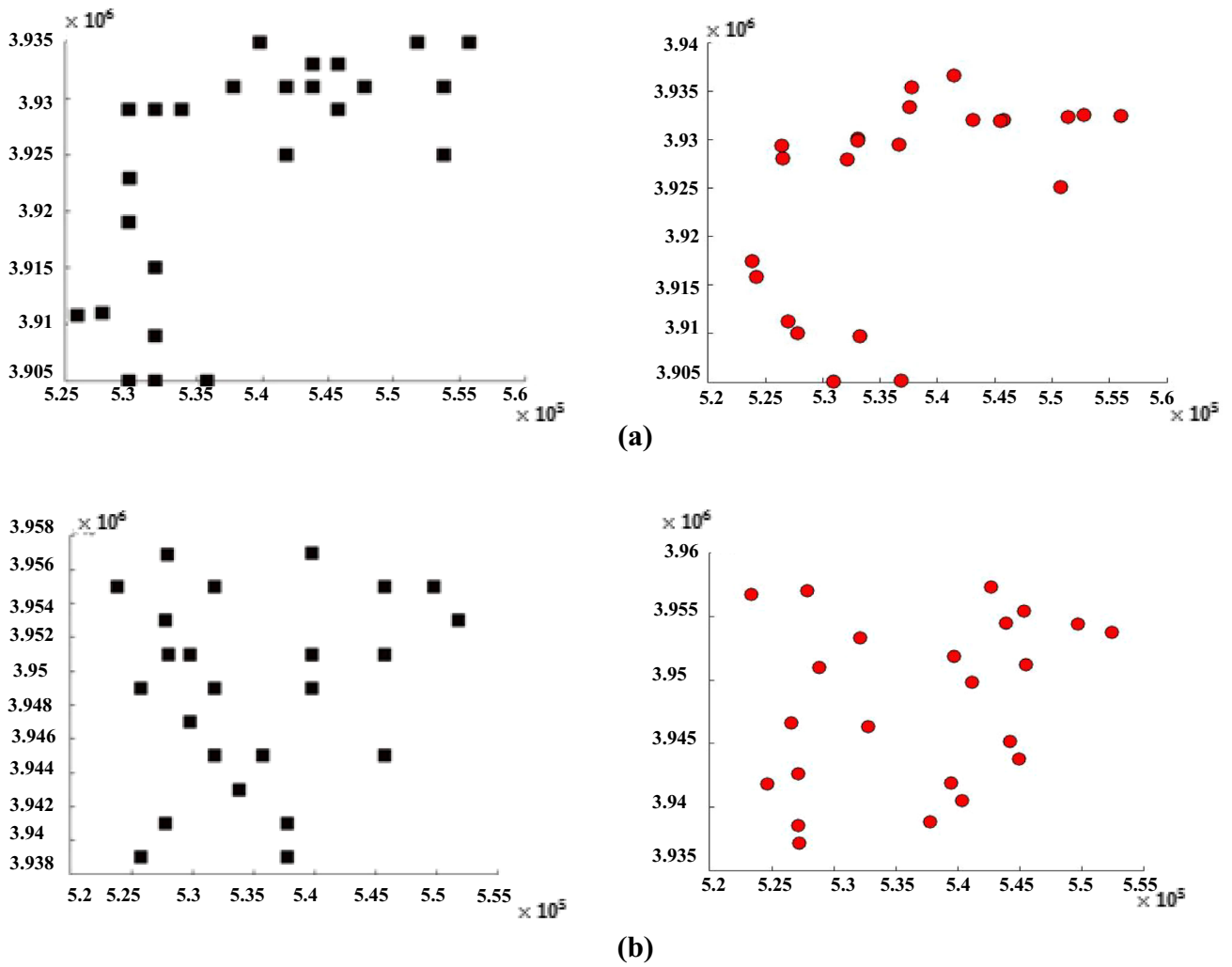


Fig. 12 Spatial distribution of the proposed and current monitoring stations for a cluster 1 and b cluster 2

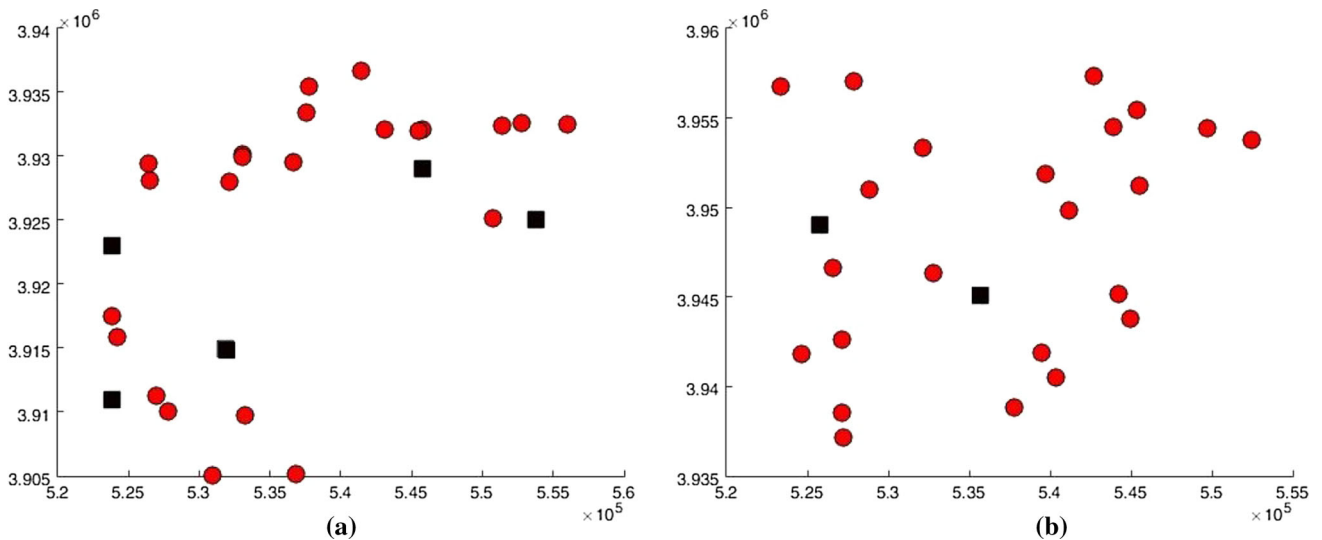


Fig. 13 The updated monitoring network (after omitting very close stations), a cluster 1 and b cluster 2

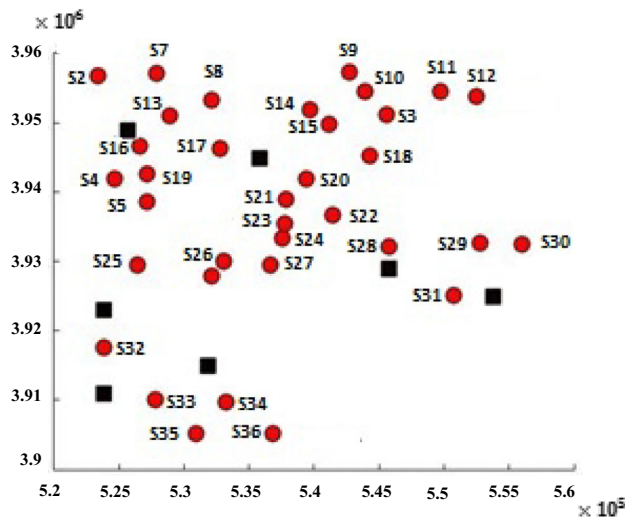


Fig. 14 The final monitoring network of the Tehran aquifer

with results of a numerical groundwater quantity and quality simulation model by using a data fusion technique. The combined data can be used in the proposed optimization model for redesigning monitoring networks.

Acknowledgements This study was financially supported by a grant from the Tehran Regional Water Company and Water Resources Management Company of Iran Ministry of Energy.

References

- Asadollahifard G (2015) Water quality management-assessment and interpretation. Springer, Berlin
- Bayat B, Zahraie B, Taghavi F, Nasser M (2012) Evaluation of spatial and spatiotemporal estimation methods in simulation of precipitation variability patterns. *Theor Appl Climatol* 113:429–444
- Bayat B, Nasser M, Naser G (2014) Improving Bayesian maximum entropy and Ordinary Kriging methods for estimating precipitations in a large watershed: a new cluster-based approach. *Can J Earth Sci* 51:43–55
- Bazargan-Lari MR, Kerachian R, Mansoori A (2009) A conflict-resolution model for the conjunctive use of surface and groundwater resources that considers water-quality issues: a case study. *Environ Manag* 43:470
- Bhat S, Motz LH, Pathak C, Kuebler L (2015) Geostatistics-based groundwater-level monitoring network design and its application to the Upper Floridan Aquifer, USA. *Environ Monit Assess* 187:4183
- Cameron AC, Windmeijer FA (1997) An R-squared measure of goodness of fit for some common nonlinear regression models. *J Econom* 77(2):329–342
- Christakos G (1990) A Bayesian maximum-entropy view to the spatial estimation problem. *Math Geol* 22:763–777
- Christakos G, Bogaert P, Serre ML (2002) *Temporal GIS*. Springer, New York
- Coulliette AD, Money E, Serre ML, Noble R (2009) Space/time analysis of fecal pollution and rainfall in an eastern North Carolina estuary. *Environ Sci Technol* 43:3728–3735
- Datta B, Singh D (2014) Optimal groundwater monitoring network design for pollution Plume estimation with active sources. *Int J GEOMATE* 6:864–869
- De Cesare L, Myers D, Posa D (2001) Estimating and modeling space-time correlation structures. *Stat Prob Lett* 51:9–14
- Dhar A (2013) Geostatistics-based design of regional groundwater monitoring framework. *J Hydraul Eng* 19:80–87
- James G, Witten D, Hastie T, Tibshirani R (2013) *An introduction to statistical learning*. Springer, Berlin
- Kerachian R, Fallahnia M, Bazargan-Lari MR, Mansoori A, Sedghi H (2010) A fuzzy game theoretic approach for groundwater resources management: application of Rubinstein Bargaining theory. *Resour Conserv Recycl* 54(10):673–682
- Kotulski ZA, Szczepinski W (2010) *Error analysis with application in engineering*. Springer, Berlin
- LoBuglio JN, Characklis GW, Serre ML (2007) Cost-effective water quality assessment through the integration of monitoring data and modeling results. *Water Resour Res* 43:W03435
- Mahab Ghods Consulting Engineers (2008) *Optimal water quantity and quality management in Tehran–Shahriar plain*. Technical Report
- Mahjouri N, Kerachian R (2011) Revising river water quality monitoring networks using discrete entropy theory: the Jajrood River experience. *Environ Monit Assess* 175(1–4):291–302
- Masoumi F, Kerachian R (2008) Assessment of the groundwater salinity monitoring network of the Tehran region: application of the discrete entropy theory. *Water Sci Technol* 58(4):765–771
- Memarzadeh M, Mahjouri N, Kerachian R (2013) Evaluating sampling locations in river water quality monitoring networks: application of dynamic factor analysis and discrete entropy theory. *Environ Earth Sci* 70(6):2577–2585
- Mogheir Y, de Lima JLMP, Singh VP (2004a) Characterizing the spatial variability of groundwater quality using the entropy theory: I Synthetic data. *J Hydrol Process* 18:2165–2179
- Mogheir Y, de Lima JLMP, Singh VP (2004b) Characterizing the spatial variability of groundwater quality using the entropy theory: II. Case study from Gaza Strip. *J Hydrol Process* 18:2579–2590
- Money E, Carter G, Serre ML (2009) Modern space/time geostatistics using river distances: data integration of turbidity and *E. coli* measurements to assess fecal contamination along the Raritan River in New Jersey. *Environ Sci Technol* 43(10):3736–3742
- Nash JE, Sutcliffe JV (1970) River flow forecasting through conceptual models part I—A discussion of principles. *J Hydrol* 10(3):282–290
- Parsopoulos EK, Vrahatis MN (2002) Particle swarm optimization method in multiobjective problems. In: *Proceedings of the 2002 ACM symposium on applied computing*, pp 603–607
- Rafipour-Langeroudi M, Kerachian R, Bazargan-Lari MR (2014) Developing operating rules for conjunctive use of surface and groundwater considering the water quality issues. *KSCE J Civ Eng* 18(2):454–461
- Ran Y, Li X, Ge Y, Lu X, Lian Y (2015) Optimal selection of groundwater-level monitoring sites in the Zhangye Basin, Northwest China. *J Hydrol* 525:209–215
- Shannon CE, Weaver W (1949) *A mathematical theory of communication*. University of Illinois Press, Urbana
- Theodossiou N, Latinopoulos P (2006) Evaluation and optimization of groundwater observation networks using the Kriging methodology. *Environ Model Softw* 21:991–1000
- Triki I, Zairi M, Dhia HB (2012) A geostatistical approach for groundwater head monitoring, network optimisation: case of the Sfax superficial aquifer (Tunisia). *Water Environ J* 27:362–372
- Varouchakis EA, Hristopulos DT (2013) Comparison of stochastic and deterministic methods for mapping groundwater level spatial

- variability in sparsely monitored basins. *Environ Monit Assess* 185:1–19
- Yang Y, Burn DH (1994) An entropy approach to data collection network design. *J Hydrol* 157:307–324
- Yu HL, Chu HJ (2010) Understanding space–time patterns of groundwater system by empirical orthogonal functions: a case study in the Choshui River Alluvial Fan Taiwan. *J Hydrol* 381:239–247
- Zalik KR (2008) An efficient k-means clustering algorithm. *Pattern Recogn Lett* 29:1385–1391