BAYESIAN METHODS FOR ASSESSING WATER QUALITY

Khalil Shihab[1] and Nida Al-Chalabi[2]

[1]College of Engineering & Science, Victoria University, Australia
Khalil.shihab@gmail.com
[2]Department of Computer Science, SQU, Oman
nida@squ.edu.om

ABSTRACT

This work presents the development of Bayesian techniques for the assessment of groundwater quality. Its primary aim is to develop a predictive model and a computer system to assess and predict the impact of pollutants on the water column. The process of the analysis begins by postulating a model in light of all available knowledge taken from relevant phenomenon. The previous knowledge as represented by the prior distribution of the model parameters is then combined with the new data through Bayes' theorem to yield the current knowledge represented by the posterior distribution of model parameters. This process of updating information about the unknown model parameters is then repeated in a sequential manner as more and more new information becomes available.

KEYWORDS

Bayesian Belief Networks, Water Quality Assessment, Data Mining

1. INTRODUCTION

Water is an essential requirement for irrigated agriculture, domestic uses, including drinking, cooking and sanitation. Declining surface and groundwater quality is regarded as the most serious and persistent issue and has become as a global issue effecting the people and the ecosystem. Anthropogenic sources of pollution such as agriculture, industry, and municipal waste, contribute to the degradation of groundwater quality, which may limit the use of these resources and lead to health-risk consequences. For these reasons, the need for intensive groundwater resources management has become more urgent.

In this work, we studied the Salalah area of Oman because the groundwater has been an important natural resource and the only available water source other than the seasonal rainfall.

Groundwater quality and pollution are determined and measured by comparing physical, chemical, biological, microbiological, and radiological quantities and parameters to a set of standards and criteria. A criterion is basically a scientific quantity upon which a judgment can be based [1]. In this work, however, we considered only the chemical parameters: total dissolved solids (TDS), electrical conductivity (EC) and water pH.

David C. Wyld et al. (Eds) : CCSIT, SIPP, AISC, PDCTA, NLP - 2014
pp. 397–407, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.4234

2. UNCERTAINTY ANALYSIS

The Ministry of Water Resources (MWR) maintains data on the concentration of the harmful substances in the groundwater at Taqah monitoring sites, which are located to the south of the Sultanate of Oman, in the Salalah plain [2, 3]. We observed that good quality data were obtained from several monitoring wells in this region. Because of the lack of monitoring wells in certain areas in that region, we filled in the missing measurements with data obtained from Oman Mining Company (OMCO) and Ministry of Environmental and Regional Municipalities (MRME) [4].
Data for water quality assessment are normally collected from various monitoring wells and then analyzed in environmental laboratories in order to measure the concentration of a number of water quality constituents. We realized that the methods used by these laboratories do not emphasize accuracy. There is a lack of awareness among both laboratory and validation personnel regarding the possibility of false positives in environmental data. In order to overcome this problem and to have representative data, we, therefore, used the following modified Bayesian model to that developed by Banerjee, Planting and Ramirez [6], to preprocessing the datasets used for the development of the Bayesian Networks.

2.1. Bayesian Models

The formulation of the model is as follows:

Let **S** denote a particular hazardous constituent of interest. Since the concentration of the substance may vary from well to another, it is necessary to consider each well separately. Let x_t= (x_{t1}, x_{t2}, x_{t3}, x_{tm}) be the vector of **m** measurements of the concentration of **S** in **m** distinct water samples from a given well at a given sampling occasion where (**m**>=1) and (t=1, 2, . . .). Each measurement consists of the true concentration of **S** plus an error.

Let X_t be the true concentration of **S** in the groundwater at sampling occasion t. If we assume that the true concentration X_t is unknown and is a random variable, the model evaluates the posterior distribution of X_t given the sample measurements x_t at sampling occasion t.

Using the normality assumption and given $X_t = x_t$ and δ^2, the concentration measurements in x_t represent a random sample of size m for random distribution with mean x_t and variance δ^2.

We assume that the parameters x_t and δ^2 of the normal distribution are random variables with certain prior probability distribution. Therefore, the model for prior distribution of X_t and δ^2 can be presented as follows:

For t =1, 2… and given δ^2 the conditional distribution of X_t at sampling occasion t is a normal distribution with mean μ_{t-1} and variance $\delta^2_{t-1}\ \delta^2$. The marginal distribution of δ^2 is an inverted gamma distribution with parameter β_{t-1} and ν_{t-1}.

This model uses the following prior distribution, which represents the concentration measurements before the first sampling.

The pdf of the prior distribution of X_0 is:

$$f_0(x_0) = \left\{ 1 + \frac{1}{2v_0} \left[\frac{x_0 - \mu_0}{\sigma_0 \sqrt{\beta_0 / v_0}} \right]^2 \right\}^{-(2v_0+1)/2} \tag{2.1}$$

which is the pdf of the student's t-distribution with $2v_0$ degrees of freedom, location parameters μ_0 and variance $\delta_0^2\beta_0/v_0$.

Now suppose that the observations are available on the concentration of **S**, given the sample X_t the posterior marginal distribution of X_t is a student's t-distribution with $2v_t$ degree of freedom, location parameters μ_t and variance $\delta^t \beta_t/v_t$ where the pdf has the form:

$$f_t(x_t / x) = \left\{ 1 + \frac{1}{2v_t} \left[\frac{x_t - \mu_t}{\sigma_t \sqrt{\beta_t / v_t}} \right]^2 \right\}^{-(2v_t+1)/2} \tag{2.2}$$

where:

$$\beta_t = \beta_{t-1} + \sum_{j=1}^{m} (x_{tj} - \bar{x})/2 + m(\mu_{t-1} - \bar{x}_t)/\left[2(1 + m\sigma_{t-1}^2)\right]$$

$$v_t = v_{t-1} + m/2$$

$$\mu_t = (\mu_{t-1} + m\bar{x}_t\sigma_{t-1}^2)/(1 + m\sigma_{t-1}^2) \tag{2.3}$$

$$\sigma_t^2 = \sigma_{t-1}^2 / (1 + m\sigma_{t-1}^2)$$

$$\bar{x}_t = \sum_{j=1}^{m} x_{tj} / m$$

It is obvious from the equation of μ_t the sequential nature of this posterior distribution. Therefore, in order to present the true unknown concentration of the substance **S** in the well under consideration, it is frequently more convenient to put a range (or interval) which contains most of the posterior probability. Such intervals are called highest posterior density (HPD) intervals. Thus for a given probability content of (1-α), 0< α<1, a 100(1- α) percent HPD interval for X_t, is given by:

$$\mu_t \pm t_{2v_t}(\alpha/2)\sigma_t \sqrt{\beta_t / v_t} \tag{2.4}$$

when $t_{2vt}(\alpha/2)$ is the 100(1- α/2) percentile of the student's t-distribution with $2v_t$ degree of freedom.

2.2. Bayesian Algorithm

In brief, the monitoring algorithm, which is based on the Bayesian model, is as follows:

(1) Fix a value of α (0< α <1) based on the desired confidence level. In this case, we chose α to be 0.01.

(2) Since we do not have enough data to work with, we used the same parameters of the prior distribution used in the model of Banerjee, Plantinga and Ramirez. These parameters are :
β_0= 0.0073 , v_0=2.336 , μ_0= 9.53 , δ_0^2 =3056.34

(3) At each sampling occasion t , (t= 1,2,...), compute the parameters β_t , ν_t , μ_t and δ_t of the posterior distribution X_t given the set of observations in $\mathbf{x_t}$ on the concentration of S available from a given well in a given site using (2.3). Compute LHPD and UHPD using these parameter estimates and (2.4).

(4) Plot μ_t, LHPD, and UHPD that are obtained in step 3 above against sampling occasion t.

(5) For the next sampling occasion, update the values of the parameters β_t, ν_t, μ_t and δ_t using (2.3) and the datasets just obtained. Recomputed LHPD, and UHPD using the updated parameter values in (2.4) and repeat step 4 above.

Some of these datasets needed to be scaled down using the following normalization technique:

$$x = \frac{\bar{x} - \mu}{\sigma}$$, where $$\bar{x} = \sum_i^n x_i \Big/ n$$, and $$\sigma = \sqrt{\frac{\sum_i^n x_i^2 - n\bar{x}^2}{n-1}}$$

2.3. Implementation

The pre-processing system is implemented on PC platform using Visual Basic programming language.

Table 1 presents the concentration data for TDS (Total Dissolved Solids) for Well 001/577 in the Taqah area. In particular, the table shows the true concentration data for TDS produced by our pre-processing system.

Table 1. Concentration Data of TDS for Well001/577 in the Salalah plain, where OC stands for Observed Concentration and ETC stands for Expected True Concentration.

Te	OC	LHPD	ETC	UHPD
84	1.147	0.85	1.15	1.45
85	1.106	1	1.13	1.26
86	1.938	1.12	1.4	1.68
87	2.237	1.33	1.61	1.88
88	3.857	1.6	2.06	2.52
89	3.834	1.91	2.35	2.79
90	3.957	2.18	2.58	2.98
91	3.761	2.38	2.73	3.08
92	4.3	2.58	2.9	3.23
93	3.958	2.72	3.01	3.3
94	1	2.54	2.83	3.11
95	3.714	2.64	2.9	3.16
96	3.65	2.73	2.96	3.19
97	3.381	2.78	2.99	3.2
98	3.396	2.83	3.02	3.2
99	3.477	2.87	3.04	3.22
00	3.498	2.91	3.07	3.23
01	3.23	2.93	3.08	3.23
02	3.243	2.95	3.09	3.22
03	3.267	2.97	3.1	3.22
04	3.297	2.99	3.11	3.22

3. BAYESIAN NETWORKS

After the pre-processing stage, we constructed a Bayesian Network (BN) by using the Hugin system. We then used this BN as an initial building network for the construction of two Dynamic Bayesian Networks in order to predict the impact of pollution on groundwater quality.

3.1. Dynamic Bayesian Networks (DBNs)

DBNs extend Bayesian Networks from static domains to dynamic domains [7, 8]. This is achieved by introducing relevant temporal dependencies between the representations of the static network at different times.

The main characteristic of DBNs is as follows:

Let $\mathbf{X}_t$ be the state of the system at time t, and assume that

(1) The process is Markovian, i.e.,

$$P(\mathbf{X_t/X_0, X_1, . . ., X_{t-1}})= P(\mathbf{X_t/X_{t-1}})$$

(2) The process is stationary or time-invariant, i.e.,

$$P(\mathbf{X_t/X_{t-1}}) \text{ is the same for every t.}$$

Therefore, we just need $P(\mathbf{X_0})$, which is a static Bayesian network (BN), and $P(\mathbf{X_t/X_{t-1}})$, which is a network fragment, where the variables in $\mathbf{X_{t-1}}$ have no parents, in order to have a Dynamic Bayesian Network (DBN).

3.2. Bayesian Networks Development

Among more than twenty wells in the Taqah area, we selected only four wells for this study. Those four wells have had, to the greatest extent, complete data measurements and provide sufficient information for the assessment of the groundwater quality for this area.

The electrical conductivity (EC) of the water has been used as a measure for the salinity hazard of the groundwater used for irrigation in the Salalah plain. The total dissolved solid (TDS) limit is 600 mg/L, which is the objective of the current plan of the MWR. TDS contains several dissolved solids but 90% of its concentration is made up of six constituents. These are: sodium Na, magnesium Mg, calcium Ca, chloride Cl, bicarbonate HCO_3 and sulfate SO_4. We, therefore, considered only these elements in the calculation of TDS.

We also used the following relationship between TDS and EC.

TDS = A * EC; where A is a constant with value between 0.65 and 0.77.

Both TDS and EC can affect water acidity or water pH. Solute chemical constituents are variable in high concentration at lower pH (higher acidity). On the other hand, acidity allows migration of hydrogen ions (H+), which is an indication of conductivity. Therefore, our work concentrated on the following relations.

$$TDS \rightarrow EC, EC \rightarrow pH, TDS \rightarrow pH$$

Reaching to these relations we used two learning approaches to construct and parameterize a simple static BN that have three nodes, each node represents a groundwater quality constituent

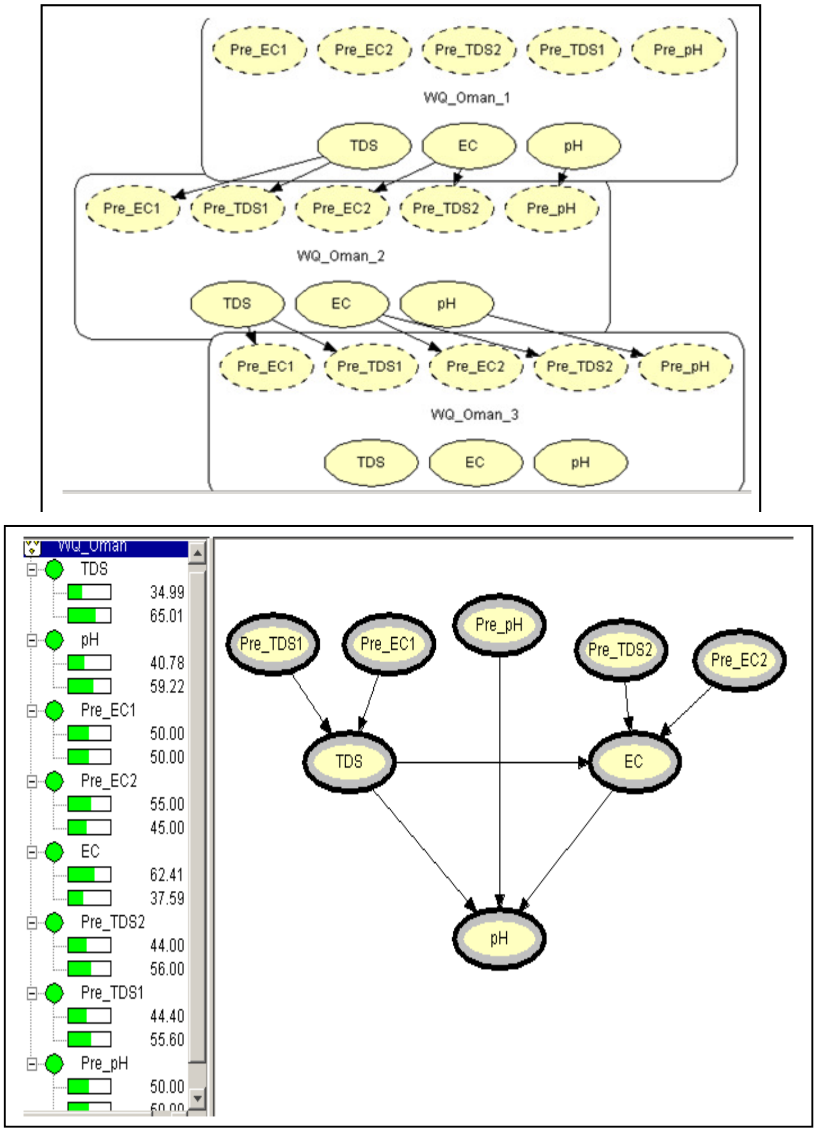
(TDS, EC or pH). Learning basically consists of two different components: 1) learning the network structure, 2) learning the conditional probability distributions.

For the first component, we used the Hugin system that supports structure and parameter learning in Bayesian networks. We also developed a program written in C++ to generate the conditional probabilities for TDS, EC and pH using Table 2 as input.

Once the static BN model (static model) for each monitoring well was built, parameterized and tested, we used these models as initial building networks in the construction of OOBNs. Figure 1 models the time slices for each well characterizing the temporal nature of identical model structures, where the initial building network, see Figure 2, describes a generic time-sliced network.

Table 2. TDS, EC, and pH data for the well Well 001/577.

Yr	**TDS mg/L**	**EC µS/cm**	**pH**
84	542.7	548	7.85
85	525.5	548	7.8
86	565.4	579	7.75
87	604.2	588	7.57
88	541.8	601	7.43
89	565.9	625	7.34
90	558.6	638	7.32
91	640.4	798	7.27
92	754.5	739	7.24
93	798.7	758	7.28
94	746.4	799	7.29
95	615.8	514	7.3
96	737.5	619	7.28
97	753.6	869	7.19
98	935.6	558	7.15
99	1174	855	7.15
0	1021	796	7.06
1	1067	855	6.98
2	1223	844	6.94
3	1055	881	6.9



WQ_Oman
TDS
34.99
65.01
pH
40.78
59.22
Pre_EC1
50.00
50.00
Pre_EC2
55.00
45.00
EC
62.41
37.59
Pre_TDS2
44.00
56.00
Pre_TDS1
44.40
55.60
Pre_pH
50.00
50.00

Figure 2. The initial building block representing one time-sliced network

4. USING CLASSICAL TIME SERIES FOR THE ASSESSMENT OF GROUNDWATER QUALITY

The purpose of this section is to apply the classical time series analysis to groundwater quality data and to compare the results with that obtained by the application of Dynamic Bayesian Networks (DBNs). The continuous and regular monitoring data of electrical conductivity (EC), total dissolved solid (TDS), pH measured by the Ministry of Water Resources (MWR) were also used here for the time series analysis.

Time series analyses of water supply wells with respect to the concentration of chemical constituents are presented in Figures 3-8.

Total dissolved solids (TDS) are a measure of the dissolved minerals in water and also a measure of drinking water quality. There is a secondary drinking water standard of 500 milligrams per liter (mg/L) TDS; water exceeding this level tastes salty. Groundwater with TDS levels greater than 1500 mg/L is considered too saline to be a good source of drinking water. Figure 3 shows the concentration of TDS for the well Well001/577 for a period of twenty one years.

The fluctuation of the concentration of the chloride (Cl), sodium (Na), and calcium (Ca) with respect to time is shown in Figure 5. The values were averaged during the initial analysis as there were no significant differences among the monthly data. Chloride values above 250 mg/l give a slight salty taste to water which is objectionable by many people.

Relationships between TDS, EC and pH are examined using multiple regression analysis, see Figure 5. Multiple regression analysis is used to explain as much variation observed in the response variable as possible, while minimizing unexplained variation from "noise". The results of this analysis are used to produce the moving average chart, Figure 7, and the linear regression chart, Figure 8. We used Excel Business Tools, Microsoft Excel, and Matlab for producing these and other charts.

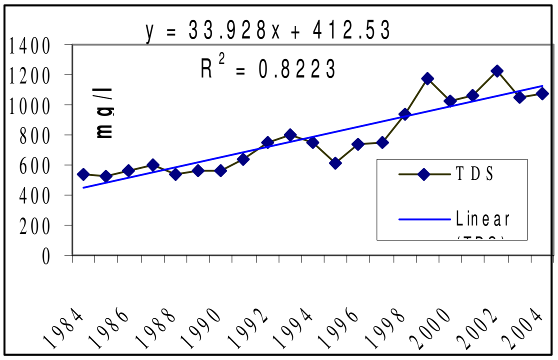


Figure 3. Fluctuation of TDS concentration for the well Well001/577

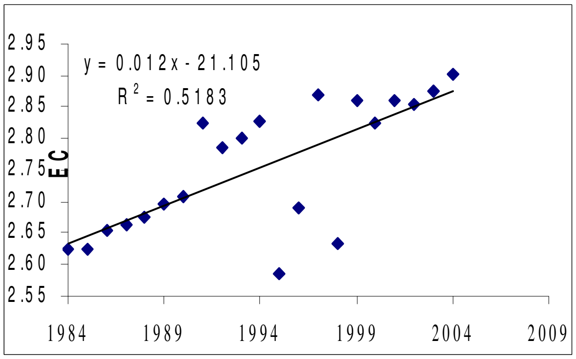


Figure 4. EC concentration is poorly represented for the well Well001/577

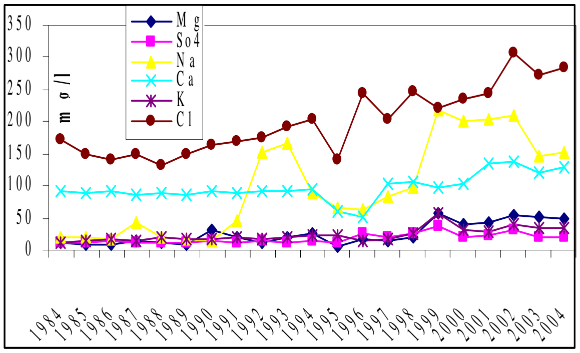
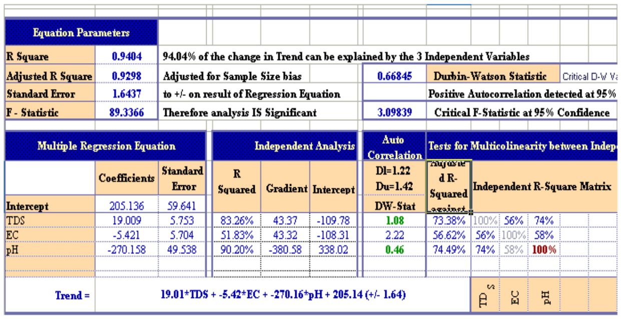


Figure 5. Fluctuation of the concentration of the major chemical constituents for Well001/577 for a period of 21 years



Equation Parameters					
R Square	0.9404	94.04% of the change in Trend can be explained by the 3 Independent Variables			
Adjusted R Square	0.9298	Adjusted for Sample Size bias	0.66845	Durbin-Watson Statistic	Critical D-W Va
Standard Error	1.6437	to +/- on result of Regression Equation		Positive Autocorrelation detected at 95%	
F - Statistic	89.3366	Therefore analysis IS Significant	3.09839	Critical F-Statistic at 95% Confidence	

Multiple Regression Equation			Independent Analysis			Auto Correlation	Tests for Multicolinearity between Indep			
	Coefficients	Standard Error	R Squared	Gradient	Intercept	Dl=1.22 Du=1.42 DW-Stat	Adjusted R-Squared against	Independent R-Square Matrix		
Intercept	205.136	59.641								
TDS	19.009	5.753	83.26%	43.37	-109.78	1.08	73.38%	100%	56%	74%
EC	-5.421	5.704	51.83%	43.32	-108.31	2.22	56.62%	56%	100%	58%
pH	-270.158	49.538	90.20%	-380.58	338.02	0.46	74.49%	74%	58%	100%
Trend =	19.01*TDS + -5.42*EC + -270.16*pH + 205.14 (+/- 1.64)							TDS	EC	pH

Figure 6. Excel templates for financial analysis and business productivity from Excel Business Tools

As is shown in Figure 5 that the trend is as follows:

$$TrendWQ=19.01*TDS - 5.42*EC -270.16*pH + 205.14$$

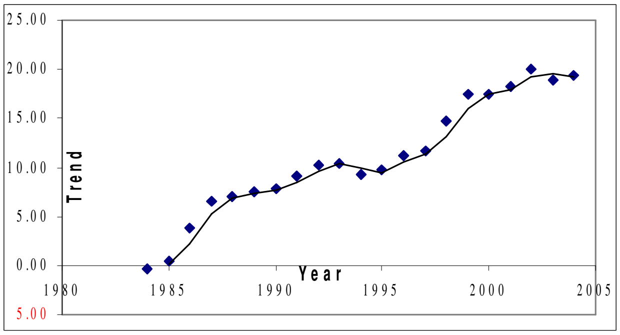


Figure 7. Moving average chart of 2-year period for groundwater quality trend

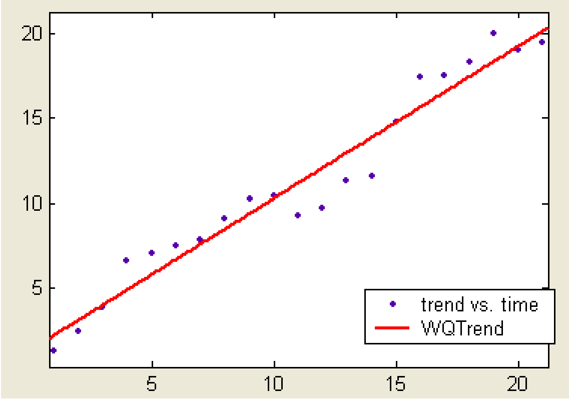


Figure 8. A curve fitting chart showing groundwater quality trend over time

Figure 7 shows the groundwater quality trend over time (linear regression). The trend has the following properties:

Linear model Poly1:

$$f(x) = p1*x + p2$$

Coefficients (with 95% confidence bounds):

p1 = 0.8954 (0.7962, 0.9947)
p2 = 1.332 (0.08589, 2.579)

Goodness of fit:

SSE: 32.91
R-square: 0.9494
Adjusted R-square: 0.9467
RMSE: 1.316

Although the classical time series models are used here to assess the presence and strength of temporal patterns of groundwater quality. These models are based on the assumption of stationary (i.e. time invariant). They have been widely used in many domains such as financial data and weather forecasting. Yet these models do not readily adapt to domains with dynamically changing model characteristics, as is the case with groundwater quality assessment. In addition to the above mentioned assumption, the classical models are restricted in their ability to represent the general probabilistic dependency among the domain variables and they fail to incorporate prior knowledge.

The observed groundwater quality data are irregularly spaced and not predetermined as in the case with ordinary time series. This may cause the traditional time series techniques to be ineffective (Prediction: what is the predicted value for one period a head). It is evident that the time series casts doubts on the positive or negative effects of any chemical constituent on the groundwater quality for the long run, and is thus not as clear and reliable as in the case of using Dynamic Bayesian Techniques. While some groundwater quality constituents, such as chloride and TDS, show an increasing trend, the other constituents, such as pH, Mg, and SO4 do not demonstrate obvious trends. Therefore, we can draw a reliable conclusion on the cause of the increasing trend of the groundwater quality and we cannot investigate the effect of the increasing or decreasing other constituents, such as pH and EC. In addition to this ignorance of the cause-effect relationships, classical time series models assume the linearity in the relationships among variables and normality of their probability distributions.

5. CONCLUSION AND FURTHER WORK

This work presents the assessment of groundwater quality. Bayesian methods have been investigated and shown to offer considerable potential for use in groundwater quality prediction. These methods are based on reasoning under conditions of uncertainty. This work is the first step towards having a comprehensive network that contains the other variables that are considered by the researchers significant for the assessment of groundwater quality in the Salalah plain in particular.

Also we showed that the classical time series models do not readily adapt to domains with dynamically changing model characteristics, as is the case with groundwater quality assessment. This is mainly because these models are restricted in their ability to represent the general probabilistic dependency among the domain variables and they fail to incorporate prior knowledge.

REFERENCES

[1] Wu-Seng, L. 1993. Water Quality Modeling, CRC Press, Inc.

[2] Dames and Moore. 1992. Investigation of The Quality of Groundwater Abstracted from the Salalah Plain: Dhofar Municipality, Final Report.

[3] Ministry of Water Resources (MWR), Sultanate of Oman. 2004. Law on the Protection of Water Resources, promulgated by Decree of the Sultan No. 29 of 2004, and its implementing regulations (Regulations for the organization of wells and aflaj, and Regulations for the use of water desalination units on wells), (in Arabic).

[4] Shihab, K. and Al-Chalabi, N. 2004. Treatments of Water Quality Using Bayesian Reasoning, Lecture Notes in Computer Science, 3029, 728–738.

[5] Shihab, K and Nida Al-Chalabi, 2007. Dynamic Modeling of Groundwater Quality Using Bayesian Techniques, Journal of the American Water Resources Association (JAWRA), Blackwell Publishing (Online Blackwell Synergy), Vol. 43, No. 3, pp. 664-674.

[6] Banerjee A. K. et al. 1985. TR no. 773, Monitoring groundwater quality, Department of Statistics, University of Wisconsin.

[7] HUGIN Expert Brochure. 2005. HUGIN Expert A/S, P. O.Box 8201 DK-9220, Aalborg, Denmark, (http://www.hugin.com).

[8] Kjaerulff, U. 1995. dHugin: A computational system for dynamic time-sliced Bayesian Networks, International Journal of Forecasting, 11, 89-111.

[9] Shihab, K. 2008. Analysis of Water Chemical Contaminants: A Comparative Study, Applied Artificial Intelligence (AAI), Vol 22, No. 4, pp. 352-376.