Discovery of Temporal Variation of Arsenic in a Historical Blackfoot Disease Territory by Time Series Analysis

Jan-Yee Lee, Ting-Nien Wu*
Department of Environmental Engineering, Kun Shan University, Tainan 71003, Taiwan, R.O.C.
jylee@mail.ksu.edu.tw, wutn@mail.ksu.edu.tw

Abstract

Time series analysis is useful tool for extracting interesting pattern from ordered sequence of observations. The Chianan Blackfoot disease region was selected as study area, and the monitoring data of arsenic in groundwater during the period of 2003 and 2008 was subjected to time series analysis. This study attempted to discover the temporal trend of arsenic level in groundwater by applying the tool of time series analysis. ARMA and ARIMA, the common time series modelling methods, were employed to interpret the information beneath the monitoring data of groundwater quality. Through further verification, the selected ARMA(1,1) model fits the data set well over the other three models. The result showed that this developed numerical model can effectively interpret and forecast the arsenic level in groundwater from area affected by salinization and high arsenic level in Chianan Plain based on the known information.

Keywords: groundwater management, data mining, time series analysis, arsenic, water quality

1. Introduction

There are many monitoring stations for investigating environmental quality in Taiwan. The established environmental database including air quality and water quality produces a huge amount of spatial and temporal data. These environmental data directly reflects environmental quality through index calculation, but the implicated knowledge behind the data is still under mining. The typical data mining tools embrace exploratory data analysis, linked map and graph display, multivariate statistical methods, linear and logistic regression, classification, decision tree and regression tree, association rule detection, and neural networks [1]. This study is concentrated on knowledge discovery of groundwater management through mining monitoring data. Groundwater quality could be influenced by many factors including climate, topography, aquifer lithology, surface water recharge, saline water intrusion, human activities, etc. Either naturally occurring processes or human activities may have a significant impact on the quality of subsurface waters and further limit its use as water supply.

As a general rule, the mineral contents found in groundwater samples are closely related to dissolution processes of geological formation in the studied area. Our previous works have successfully utilized multivariate statistical tools to identify the sources of groundwater contamination on the south region of Taiwan, and the familiar factors affecting groundwater quality consisted of salinization, organic pollution, mineralization, and agricultural nutrient release [2, 3]. In Chianan Plain groundwater subregion, there exists a historical Blackfoot disease region. The abundance of arsenic in drinking water may result in Blackfoot disease due to a long-term uptake. Another previous study utilizing principal component analysis and cluster analysis has spatially allocated the Chianan Blackfoot disease region based on the mapping of monitoring wells [5].

In the Chianan Blackfoot disease region, groundwater possesses an extremely high level of arsenic and a strong magnitude of salinity revealing the associated impact of salinization and arsenic dissolution in the aquifer [5]. Whether the Chianan Blackfoot disease domain chronologically expands or shrinks is still unknown. It has drawn our interest to examine the linkage of arsenic level in groundwater with time. In this paper, time series analysis was employed as data mining tool. Several approaches of time series analysis were investigated to ascertain the trend of temporal variations on arsenic level in groundwater; including autoregressive process (AR), moving average process (MA), autoregressive moving average process (ARMA), and autoregressive integrated moving average process (ARIMA) models. This study provided a time-series
showcase to implement knowledge discovery from mining groundwater monitoring data, and also expected to interpret the nature of arsenic dissolution in the Chianan Blackfoot disease territory.

2. Study Area

The study area belongs to Chianan Plain groundwater subregion locating on the south of Taiwan. Chianan Plain groundwater subregion is a shallow aquifer consisting of 3 to 5 layered geological structures that are mainly the mixed layers of clay, silt and silty sand. This type of geological structures may hamper percolation and limit groundwater recharge to the deeper layers. Annual groundwater recharge is estimated only 120 million tons in study area, while the withdrawal of groundwater resource sums up to 351 million tons each year, mainly for the demand of agricultural irrigation. In the Chianan Plain aquifer, the transmissivity and storativity are 0.014 m²/min and 3×10⁻⁴ respectively [6]. As shown in Figure 1, the Chianan Blackfoot disease region rests on the northwest seashore of the Chianan Plain. In this region, groundwater flows in the direction of east to west and finally discharges to Taiwan Strait.

3. Groundwater Monitoring Data

In Chianan Plain groundwater subregion, eighty-four monitoring wells were established for routine ground-water monitoring. As shown in Fig 1, the locations of analyzed monitoring wells are scattering in the west portion of the study area. The quarterly monitored data was recorded in groundwater monitoring database since 2002. Groundwater quality data including pH, electrical conductivity (EC), hardness, total dissolved solid (TDS), total organic carbon (TOC), ammonia, nitrate, chloride, sulfate, Fe, Mn, As, Na, K, Ca, Mg, Cd, Cr, Cu, and Pb. According to our previous study, the domain of the Chianan Blackfoot disease region was spatially allocated as Figure 1. Only the monitoring wells around the Chianan Blackfoot disease region were subjected to time series analysis. The input data was extracted from historical monitoring data of arsenic concentration in groundwater during the time period of 2003 and 2008. There were only 6 years of monitoring data available, so a quarter was selected as the interval to extend the time span of time series analysis.

4. Method

Figure 2 showed several steps of data processing including data pre-processing, time series analysis, information interpretation, and knowledge discovery.
4.1 Time Series Analysis

Time series analysis has been widely used in a variety of fields such as well-known areas as economic forecasting, signal processing and communications. The practical application cases of time series analysis include neurophysiology to astrophysics. Time series analysis is known as a powerful tool for summarizing ordered sequences of observations with various degrees of correlations over time by employing the concepts of stochastic process for dynamic systems. Typically, time series analysis is based on the assumption that the data consist of systemic pattern and random error. Through extracting information from autocorrelation functions (ACF) and partial autocorrelation functions (PACF), it can filter out noise in order to make the pattern more salient. The steps for analyzing and modelling time series include constructing the graph, data transformation, examining the ACF and PACF, identifying proper orders of differencing and moving average, and model diagnosis.

In order to determine the relationship between arsenic and groundwater sampling time, we mainly focused on ARMA and ARIMA models. In this study, all the analysis was computed by the SPSS-12.0 software.

![Figure 3. Variations of arsenic level (ppm) in groundwater during the period of 2004 and 2008.](image)

4.2 Data Preprocessing

The first step in time series analysis is to perform some preliminary processing to make the raw data suitable for further analysis. The plot of the data is given as Figure 3. It indicated that the series is stationary in the mean but may not be stationary in variance. Hence, the power transformation was applied to meet the requirement of variance stabilization. The result suggested that the natural logarithmic transformation is appropriate for data preprocessing.

4.3 ARMA and ARIMA Models [7]

4.3.1 Autoregressive Moving Average Process Model (ARMA)

The process \( \{Z_t\} \) is said to be an ARMA model if it can be expressed as

\[
\Phi_p(B)Z_t = \theta_q(B)a_t,
\]

where,

\[
\Phi_p(B) = 1 - \Phi_1 B - \cdots - \Phi_p B^p,
\]

\[
\theta_q(B) = 1 - \theta_1 B - \cdots - \theta_q B^q
\]

(1)

(2)

(3)

The model is usually referred to ARMA(p,q) where p, d, and q are integers greater than or equal to zero and refer to the order of the autoregressive, and moving average parts of the model.

4.3.2 Autoregressive Integrated Moving Average Process Model (ARIMA)

On the other hand, if the process \( \{Z_t\} \) can be expressed as the following form, then the model is said to be an ARIMA model.

\[
\Phi_p(B)(1-B)^dZ_t = \theta_0 + \theta_q(B)a_t,
\]

where,

\[
\Phi_p(B) = 1 - \Phi_1 B - \cdots - \Phi_p B^p
\]

\[
\theta_0(B) = 1 - \theta_1 B - \cdots - \theta_q B^q
\]

(4)

It is a generalization of ARMA model. The model is generally referred to as an ARIMA(p,d,q) model where p, d, and q are integers greater than or equal to zero and refer to the order of the autoregressive, integrated, and moving average parts of the model respectively.

5. Result and Discussion

As mentioned, time series analysis is based on extracting information from ACF and PACF that filter out the trend, interesting patterns, rules, or models for the data set. For model diagnostic checking, if the residual estimation error for a particular model is white noise within a specified tolerance level, the fitted model is in goodness of fit.
The residual ACF, PACF, and Box-Ljung test are commonly applied for diagnostic checking. In time series analysis, different models might describe the data adequately. The Akaike’s information criterion (AIC) is generally served as a criterion for assessing the quality of model fitting. It is based on residual log-likelihood function for model comparison. Being the model selection filter, the sole AIC value has no meaning unless applying for comparison with a series of related models. Subsequently, universal guidance of AIC value for evaluating model performance does not exist. As a usual rule, the smaller AIC and simpler model has better fitness for a given data set.

Some scientific works have pointed out that salinization is associated with a high arsenic content of groundwater in Chianan Plain [4, 5]. In this study, the data matrix is extracted from a total of 908 historical groundwater samples from 39 monitoring wells that were affected by salinization and geological arsenic release. The results of the ARMA and ARIMA showed that time series analysis can briefly reduce the complexity of the original data set.

After natural logarithmic transformation, the AIC value decreased tremendously from over 10,000 to about 650. This result successfully shows the stabilization of the noisy variation. It also indicates that the fitted models improve the accuracy for managing the relationship between arsenic concentration and groundwater sampling time. The obtained AIC for four fitted ARMA and ARIMA models were shown in Table 1. From the result of test for significance of parameters, the square terms for both autoregressive and moving average are not suggested to be included in interpreting model. Based on AIC, the minimum AIC occurs in ARMA model with p=1 and q=1. It is also the simplest one among all fitted models. Hence, it is the one should be selected to interpret the analyzed data.

<table>
<thead>
<tr>
<th>Models</th>
<th>AIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARMA(1,1)</td>
<td>640</td>
</tr>
<tr>
<td>ARMA(2,1)</td>
<td>642</td>
</tr>
<tr>
<td>ARMA(1,2)</td>
<td>642</td>
</tr>
<tr>
<td>ARIMA(1,1,1)</td>
<td>666</td>
</tr>
</tbody>
</table>

The estimates of parameters are all statistically significant. The estimation results were shown in Table 2. We obtained the interpreting model as following:

\[ A_s + 2.042 = 0.935(A_{s_{t-1}} + 2.042) + a_t - 0.381a_{t-1}, \]

where \( A_s \) represents the concentration of arsenic at time \( t \), and \( a_t \) represents the white noise at time \( t \). In other words, the given the previous arsenic information \( A_{s_{t-1}} \), the arsenic at time \( t \) is subject to stochastic influence by time \( t-1 \).

The above fitted interpreting model provides a possible solution to predict the upcoming arsenic level based on the prior arsenic monitoring. It could be served as alarm or baseline for the temporal dynamics of arsenic groundwater monitoring in the region affected by salinization and geological arsenic release. It is also found that the contribution of
salinization to the total variance is more remarkable. This finding is coincided with the results of other research works with the tools of factor analysis, cluster analysis and principal component analysis [4, 5].

### Table 2. Estimation Results of ARMA(1,1) model.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>St. Error</th>
<th>t-Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>-2.042</td>
<td>0.107</td>
<td>-19.031</td>
</tr>
<tr>
<td>AR1</td>
<td>0.935</td>
<td>0.014</td>
<td>68.250</td>
</tr>
<tr>
<td>MA1</td>
<td>0.381</td>
<td>0.036</td>
<td>10.660</td>
</tr>
</tbody>
</table>

6. Conclusions

High level of arsenic in drinking water leading to Blackfoot disease has been proven by the long-term retrospective medical research. In the Chianan Blackfoot Disease region, the high arsenic level in groundwater resulted from arsenic dissolution of geological formation. By using time series analysis, the numerical models were developed for interpreting and forecasting the arsenic level in groundwater from the region affected by salinization in Chianan Plain. The fitted model explained the overall outcome of the underlying processes that influence the trend of the arsenic level in salinized groundwater. In other words, it can help discovering the existing relationship between the arsenic level and time-domain of groundwater in salinized region. This study has tentatively found out the temporal trend of the arsenic level in groundwater that can provide the forecasting information for the ease and improvement of groundwater management.

7. Acknowledgement

The authors would like to thank Mr. Chiu-Sheng Su for his support on data collection.

8. References