05JEI00047
1726-2135/1684-8799
© 2005 ISEIS
www.iseis.org/jei

Journal of Environmental Informatics 5 (2) 61-72 (2005)

Water Quality Management Using GIS Data Mining

F. Karimipour[1], M. R. Delavar[1*] and M. Kinaie[2]

[1]Department of Surveying and Geomatic Engineering, Engineering Faculty,
University of Tehran, Tehran, Iran
[2]Informatics and Statistics Center, University of Tehran, Tehran, Iran

ABSTRACT. Nowadays scientists, managers and decision makers have faced with ever increasing production of digital geospatial data acquired at various geometric, thematic and temporal characteristics. Geospatial information systems (GISs) have been widely considered to handle such a diverse range of geospatial data. One of the important issues in geospatial data management is to explore the relationships and future trends of the data, which is possible through geospatial data mining and knowledge discovery. Geospatial data mining, its need and analyses have been investigated in this paper. In addition, applications of geospatial data mining in environmental data management and especially in water quality management have been introduced. Finally, regarding the abundance of industrial centers in Western and Eastern Azerbaijan Provinces in North-West of Iran and their effects on water quality in this region, correlation between industrial pollutions and water quality indicators through geospatial data mining has been modeled as a case study. The results have clearly identified the relationship between number and location of industrial pollutions and water quality indicators to be used in environmental protection and land use planning.

Keywords: Association analysis, environmental protection, geospatial data mining, GIS, water quality management

1. Introduction

Parallel to ever increasing uses of geospatial information technology (GIT), its computing environment, scope, coverage and volume of geospatial data are growing fast. A number of agencies in public and private companies involve in acquisition, processing and display of geospatial data. In addition, spatial data acquisition systems are evolving from speed and accuracy point of view. For example, remote sensing systems and global positioning systems (GPS) are used to capture huge amount of geospatial data.

Increasing spatial data sharing and interoperability throughout the world have resulted availability of large amount of data to be used in development of spatial data infrastructures (SDI) in a rapid and unprecedented pace.

In spite of production of such a huge amount of spatial and thematic data, proper use and management of the data are important issues to achieve sustainable development. It has been clear that conventional approaches of statistical analyses of geospatial data in analogue form and in small amount cannot further be efficiently implemented for digital data and in large volume, which are being produced nowadays. Therefore, conventional analysis approaches cannot be implemented to explore the hidden relationship between and among spatial data and their future trends, which are quite important functionalities in optimum geospatial data management. Such functionalities can be efficiently employed using geospatial data mining and knowledge discovery.

One of the important applications of GIS is environmental data management. GIS can be used to provide scientists and managers with a range of scenarios for spatial distribution of the data and predict future trends of the data to avoid possible environmental crisis. Geospatial data mining can be used to assess hidden relationships of the crisis and environmental pollutions, sources, causes and amount of pollutions to take necessary measures for environmental protection.

In 1972 Clean Water Act established a National Pollution Discharge Elimination System (NPDES), which requires an easily revoke permit for any industry, municipality or other entity dumping wastes in surface water (Cunningham and Saigo, 1999).

Most of those efforts have been aimed at point sources, especially to build or upgrade thousands of municipality sewage treatment plants.

Attempts were made by Indian geographers to delineate ‘Physical complexes’ on the principle of relationship among the natural elements. One study consists of 16 macro and 58 microphysical complexes based on regional grouping of administrative districts (Singh, 1995). Such systems are considered as starting point in integrating the dimensions of environment with spatial monitoring and forecasting. This system had many problems in building a unified database due to increasing volume of source information.

In 1996, the U.S. Environmental Protection Agency (EPA)

[*] Corresponding author: mdelavar@ut.ac.ir

announced that toxic, chemical sewage or other pollutants contaminated about 16000 segments of surface water in U.S. and its territories. However, the method to compare this situation with past pollution levels and with the other countries, to make an efficient decision was not specified (Cunningham and Saigo, 1999).

Coordination of the above-mentioned samples can be done in a GIS environment. Especially, when the volume of data is increased, geospatial data mining is an appropriate candidate to extract more information from such a data warehouse.

North-West of Iran is an industrial region where agriculture and animal husbandry are well established and there are huge amount of underground chemical resources, too. Therefore food, leather, chemical and mineral industries and population are concentrated there. Since this region is rainy and many rivers are originated from there, proper management of water quality in this region is very essential for Iranian Department of Environment.

This paper is concentrated on water quality management using geospatial data mining in a GIS environment. Especially, dependency analysis which is one of the spatial data mining analyses is developed to find hidden relationships between concentrations of the above-mentioned industries, population and water quality parameters to assist the strategic planning to establish new industrial centers in this region.

Section 2 describes data mining. In this section concept of data mining, its stages and analyses are introduced. In Section 3, spatial data mining and its individual characteristics are mentioned. Section 4 contains reasons for water quality management and past efforts to use GIS in this direction. Section 5 concentrates on the case study and explains used spatial and attribute data. In this section, dependency analysis is used to extract hidden relationships between industrial and population concentration and water pollutant parameters. In Section 6 the validity of achieved results are verified. Finally, Section 7 contains conclusions and open opportunities for future works.

2. Data Mining

Every organization collects its required data to solve some specific problems. Indeed, these data are collected for specific purposes. When the existing data are hoarded (so called data warehouse), some extra information can be extracted form this huge amount of data. This process is known as data mining. In this section, fundamental aspects of data mining, its requirements, processes and analysis approaches are elaborated.

2.1. Data Mining Concepts

Data mining is an approach for information extraction from huge amount of data stored in a database (Miller and Han, 2001). The concept of data mining is illustrated in Figure 1. Recent trends in information technology (IT) and its growing application areas in addition to increase of available databases, along with the data mining are being used to extract and interpret information available in the databases, and explore the necessary information and their relationships to produce useful information/knowledge for decision making.

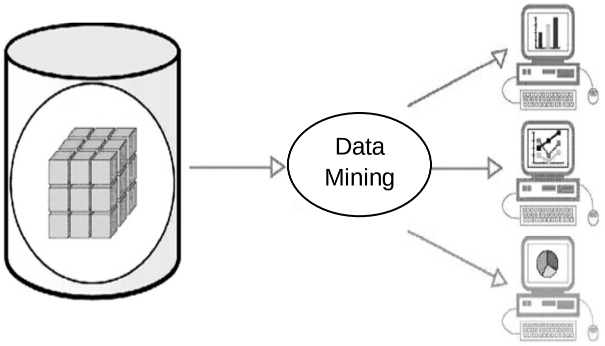


Figure 1. Data mining concept.

In other words, data mining can be considered as an approach to determine the *valid, novel, useful* and ultimately *understandable* data patterns in a large database (Miller and Han, 2001).

Valid means that the extracted patterns can be applied for new dataset as well and are not explicitly useful for the data based on which the relationship are derived. *Novel* means that the patterns are non-trivial and unexpected. *Useful* means that the extracted patterns can be used for future activities, because the data mining is used as a means of decision support systems. Finally, *understandable* means that discovered relationships should be simple and interpretable.

It has to be pointed out that data mining analyses are valid while having huge amount of data; otherwise, the achieved results may not be extended further. One solution to this problem is to consider *data warehousing (DW)* concepts (Jerk et al., 2000). DW can be regarded as the infrastructure for data mining and is a repository that integrates existing data in one or several databases in a corporate database. A DW usually exists to support strategic and scientific decision-making based on integrated and shared information. DWs are also used to save legacy data for library and other porous (Jerk et al., 2000).

2.2. Data Mining Stages

Data mining analyses follow several stages including data cleaning, data selection, data reduction, information extraction, interpretation and reporting (Miller and Han, 2001) as shown in Figure 2.

Since most data are provided through observations and measurements, they must be checked against errors and distortions. Data cleaning covers cleaning of data in addition to removing noises and duplications.

Data selection refers to selection of some fields and re-

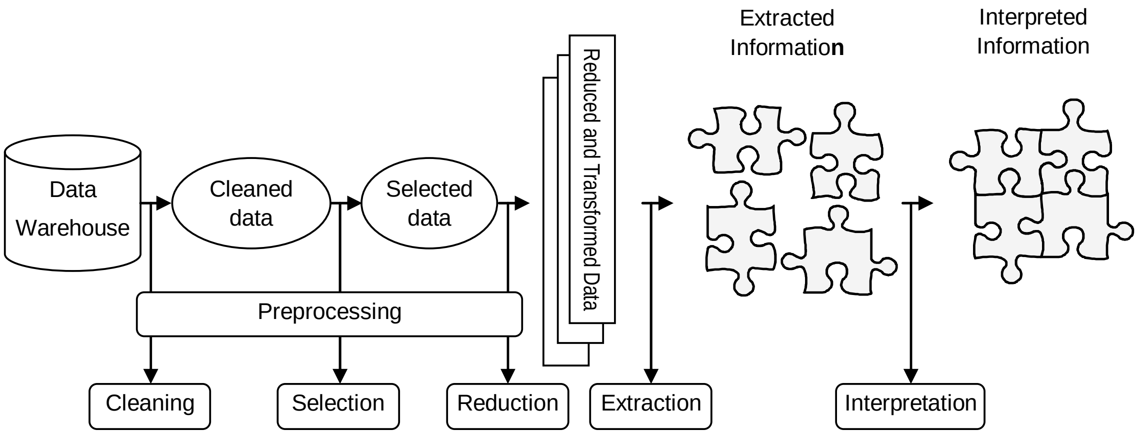


Figure 2. Data mining stages.

cords in a database in order to be used in the analysis operations as well as maintaining the data integrity to provide the proper results.

Data reduction can be considered as a reselection of data. In this stage, volumes of the selected data are reduced from attribute and records point of view. Furthermore, some records are properly merged or transformed.

At information extraction stage, the data are investigated to extract patterns and relationships, therefore, this is the main stage of a data mining process.

Finally, at interpretation stage, the discovered patterns are evaluated and integrated using the essential parameters in a data mining process as pointed out in Section 2.1.

2.3. Analyses Used in Data Mining

There are several analyses used in data mining which are mainly classified as clustering, classification, dependency analysis and trend prediction. They are elaborated as follows (Miller and Han, 2001):

2.3.1. Clustering

Clustering means specifying some implicit classes in which the selected data are classified. In this case, the number of classes is not known beforehand and existing data in terms of specific characteristics are divided to some classes having similar characteristics. Some statistical approaches exist for clustering. These approaches recursively divide the data to some classes with less variances form a certain tolerance. In addition, the average of the elements of each class must be far from that of others regarding a predefined tolerance.

However, statistical approaches for clustering face some computational problems while having huge amount of data. This problem can be solved using Artificial Neural Networks (ANNs) that have ability to cluster huge amount of data efficiently.

2.3.2. Classification

Classification is an analysis in which some rules to distribute features in classes with pre-determined conditions exist.

Similar to clustering, classification can be done using statistical approaches and AANs, too. In this case, different ANNs such as unsupervised networks (e.g. Kohonen) and supervised ones (e.g. back propagation) have been extended.

One of the main advantages of ANNs is their ability to nonlinear classification, the case may be frequently occurs in most applications.

2.3.3. Dependency Analysis

Dependency analysis involves finding rules to predict the value of some attributes based on the value of other attributes (Ester et al., 1997, cited in Miller and Han, 2001). Dependency analysis is used to determine the correlation between different measures (Figure 3). Regression analysis (Tobler, 1994), decision trees and ANNs are among the dependency analysis approaches.

2.3.4. Trend Prediction

Trend prediction is another analysis used in data mining. A line or curve is fitted to the data to formulate the change over time in order to predict the future trend. Furthermore, using sequential pattern extraction, possible existing temporal correlation in data can be extracted. Trend prediction is a data mining analysis most used for time dependent data. Trend prediction may be done using regression analysis and ANNs. In addition, time series analysis is a quite appropriate approach in this area.

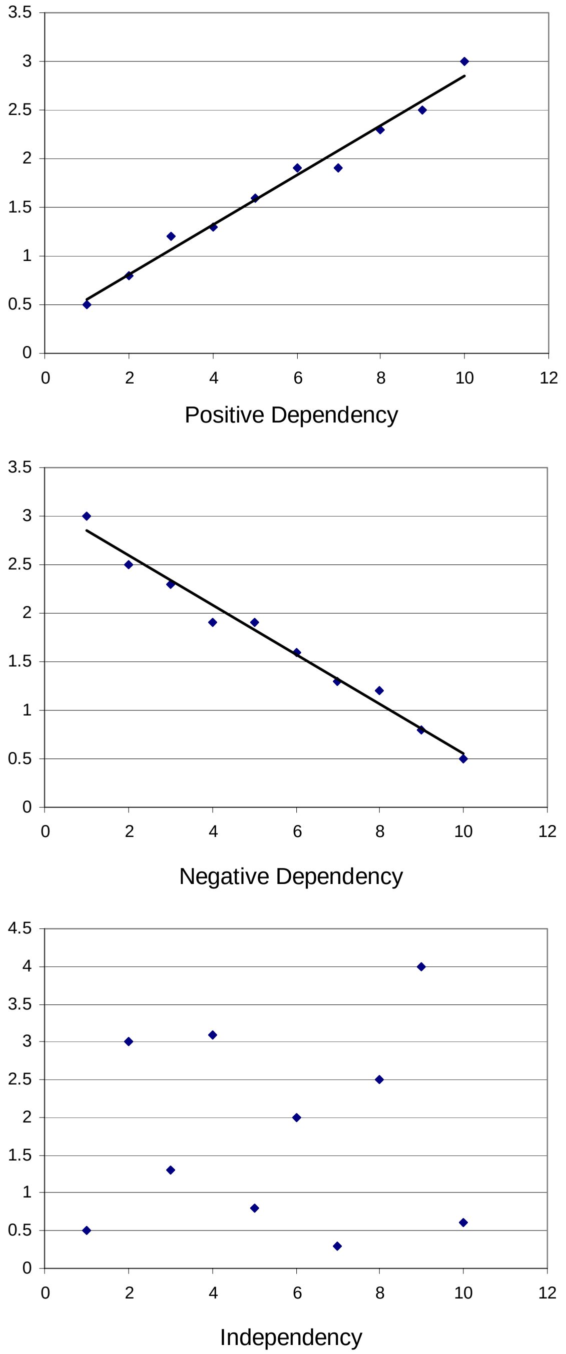


Figure 3. Different dependency options.

3. Geospatial Data Mining

Geospatial data are data with some important positional components. Data in a non-spatial database are not associated to a certain position. However, in a spatial database there is a reference framework where geospatial data are associated with it through their coordinates. Experiences show that about eighty percent of existing data have some geospatial components. Therefore, extension of the data mining concept to geospatial data seems quite essential. This section reviews data mining form a GIS perspective to determine the specific characteristics of geospatial data mining.

3.1. Geospatial Data Mining Characteristics

There are a number of characteristics to distinguish geospatial data from others. The specific characteristics of geospatial data necessitate more complicated analyses to be implemented in geospatial data mining. The mentioned characteristics of spatial data are geospatial measurement framework, spatial dependency and complex objects (Miller and Han, 2001). They are described below.

3.1.1. Geospatial Measurement Framework

Data used in data mining processes, are usually multi-dimensional which are mainly independent. However, spatial data in addition to multi-dimensionality, are interrelated in three spatial and may be one temporal dimensions which form a framework for other dimensions.

3.1.2. Spatial Dependency

Geospatial characteristics measured usually present a spatial dependency indicating the existence of some relationships among some characteristics at specific locations. Spatial dependency causes the extracted information in one location may not be applicable elsewhere. The mentioned locations are usually proximal in Euclidian space. However, direction, connectivity and other geospatial attributes (e.g. terrain and land cover) can also affect spatial dependency (Miller and Han, 2000).

3.1.3. Complex Objects

Non-spatial data can be considered as a point in a database. However, geospatial data cannot. Geospatial data may have size, shape and orientation that are important for most spatial analyses and incorporating of such properties in data mining analyses needs more complex strategies.

3.2. Analyses Used in Geospatial Data Mining

Although the analyses used in geospatial data mining seem similar in nature to those of data mining, they can be much more complex from conceptual and implementation point of view. These analyses include spatial clustering, spatial classification, spatial dependency and spatial trend prediction (Miller and Han, 2001). In these analyses at least one locational component has to be incorporated as a main parameter. Spatial dependency analysis as an example of spatial data mining is implemented in this research.

Spatial dependency analysis involves determination of some rules to estimate the value of one or more data using the value of other data while one or some of their main components are spatial data. Such a relationship can be between a spatial and an attribute component or between two spatial components.

In the reminder of the paper, importance of water quality management and some past efforts in this area using GIS are investigated. In addition, a case study in GIS data mining applied on water quality management is introduced.

4. The Importance of Water Resource Management

Water is one of the most important requirements in daily life which contains a major parts of the earth’s hydrosphere. Scientific investigations regarding lack of enough water resources, increase of pollution in water resources in major parts of the world and increase of man’s destructive activities affecting water resources are going to make a disaster in the near future. Implementation of proper and practical policies to evaluate the water resources through integrated exploitation, management and planning is vital. This is to ensure the availability of enough qualified water for the whole planet in addition to maintain the hydrologic, biologic and chemical ecosystem in a proper way.

Considering the vital role of water resources in national, regional and global ecosystems and a number of socio-economic disputes currently exist especially in the Middle East which seems to be intensified in the near future, assessment of existing situation and evaluation of the potentials of available water resources are vital steps for proper water resource management.

Water crisis in international domain has been considered especially in recent years. The existence of a number of environmental effects such as increasing the earth temperature has changed the spatio-temporal distribution of rain. This will lead to desertification, flooding, etc. (Cunningham and Saigo, 1999). The relationship between development and environment, population increase and the importance of maintaining food security are among the important challenges in water demand and supply management. Therefore, water supply and quality protection have been considered in Agenda 21 in Rio Conference in 1992 (www.un.org).

The importance of water quality management forced a number of countries to investigate ways for pollution control. In this direction, GISs are among the most useful approaches. Recently, many efforts have been undertaken to use GISs for water quality assessment and management in different scales such as streams, rivers, lakes, seas and oceans. For example “The Alabama Watershed Demonstration” project links land use patterns and water quality through GIS (Flynn, 1999). Also some successful efforts about satellite and GIS tools to assess lake quality have been reported in University of Minnesota (Brezonik et al., 2002).

In some previous researches predefined indicators and relationships were considered and GISs were used to manage these situations. However, in some cases no certain relationships between parameters and their extraction have been directly reported. In such situations, some statistical analyses can be used which lead us to geospatial data mining.

A number of applications of geospatial data mining in water quality management have been recorded in the literature to explore spatio-temporal dependencies between water quality measures, their sources and the trend prediction (Miller and Han, 2001). However, in this paper a static situation is considered. In the next section, dependency analysis is used to extract hidden relationships between water pollutant sources and water quality parameters.

5. Case Study

As a case study undertaken in this research, geospatial data mining for water quality management in Western and Eastern Azerbaijan Provinces in North-West of Iran situated between 44.04° to 48.50° W Longitude and 35.59° to 39.71° N Latitude is considered. In this region, agriculture and animal husbandry are well established and there are huge amount of underground chemical resources. Therefore food, leather, chemical and mineral industries as well as population are concentrated there. Since this region is rainy and many rivers are originated from there, proper management of its water quality is very essential for Iranian Department of Environment. In this section dependency analysis as one of the data mining analyses is used to extract hidden relationships and patterns between the above-mentioned potential source of pollutants and water quality parameters of existing rivers.

The spatial data related to some rivers in the study region have been collected from 1:25000 maps produced by Iranian Department of Environment. Major rivers of the study area that have stations for measurement of water quality parameters have been extracted. In addition, the data related to some cities, industrial centers including food, chemical and leather industries and mines have been acquired using 1:25000 topographic maps produced by Iran's National Cartographic Center (NNC). The acquired data are shown in Figure 4.

In addition to major data available on quality parameters of industrial waste water, few data were simulated from neighboring and similar industries where the actual data were unavailable. Since the volume of waste water and kind of industry has a relatively direct relation, the simulated data can be considered as adequate approximations and they have not distorted the results. Attribute data are shown in Figure 5.

For each river, number of existing distinct industrial centers and populations within a 20 kilometers radius around water quality measurement stations have been achieved through a buffer analysis and associated to the tuple of that river in the table shown in Figure 5 as a new column.

A number of attribute data have been selected and parameters of dependency between each two of them for first, second and third degree polynomials and exponential curve calculated. Among them the best fitted curve has been specified thorough a goodness of fitness test. If at least one of the fitted curves was significant, then the existence of dependency between its two related parameters is reported. The above procedure is described bellow for one of the mentioned cases:

To check the existence of dependency between popula-

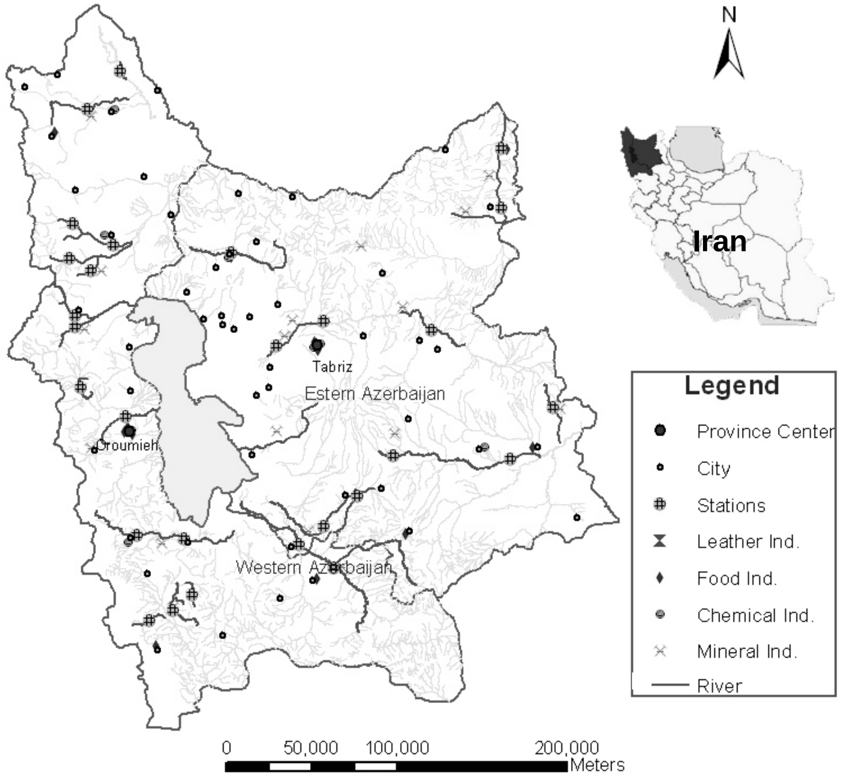


Figure 4. The study area.

tion concentration and *DO* (Dissolve Oxygen) (the case shows in Figure 6.i), first to third degree polynomials and an exponential curve have been fitted to population concentration in the selected buffer (20 km from quality parameter measurement stations) and their respected *DO*. The calculated *DO* ($\hat{DO}$) regarding these fitted curves for each case and each river have been calculated. The calculated and observed value for *DO* of each river and sum of the squares of the residuals (*J*) has been calculated for each case using Equation (1) are shown in Table 1:

$$J = \sum_{i=1}^{n} (DO_i - \hat{DO}_i)^2 \qquad (1)$$

These parameters can be used for goodness of fitness test (Ford and Zanelli, 1985). The best fitted curve is selected where *J* is minimum. On the other hand, it can be proved that *J* has a Chi-square distribution with *n-p* degrees of freedom where *n* and *p* are number of observations and unknowns, respectively (Ford and Zanelli, 1985). Therefore its significant region at a confidence level *α* is (2):

$$J < \chi^2_{\alpha,n-p} \qquad (2)$$

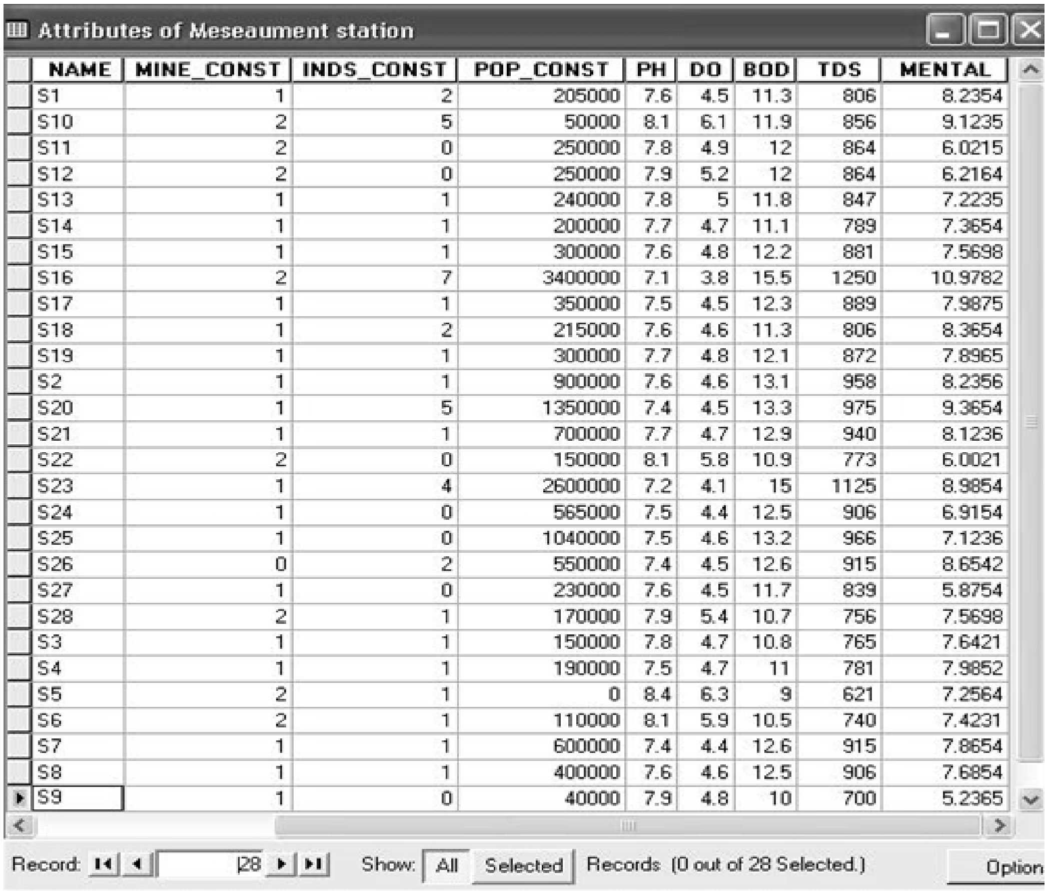
where $\chi^2_{\alpha,n-p}$ is the value of the Chi-square distribution at significance level *α* and *n-p* degree of freedom (*df*). In this case study, number of observations is 28 and number of unknowns depend on type of curve and confidence level based on user needs. The value of Chi-square for confidence level 99.5% for each case have been obtained and compared with the calculated *J* to test its significance (Table 1).

Finally the second degree polynomial ($DO = 2\times10^{-13} p - 10^{-6} p + 5.2792$) where the calculated *J* is minimum (*J* = 5.424) and the $J_{\min} < \chi^2_{0.5,25} = 10.520$ has been selected for the best fit as shown in Figure 6.i.

This process has been done for each case and the results for all of the checked dependencies are shown in Figure 6 and elaborated in Section 6.

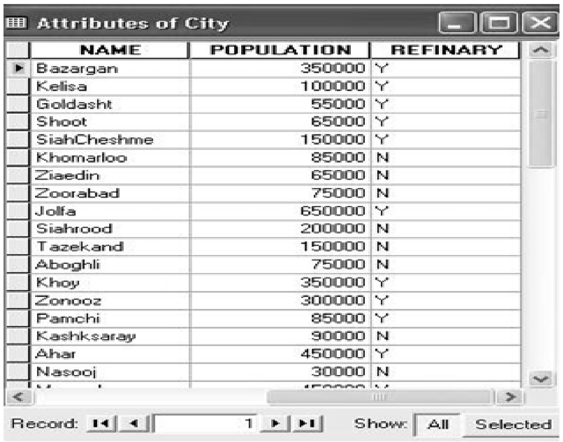
6. Discussions

The validity of achieved results have been verified and described further. Figures 6.a, 6.b and 6.c represent the existence of positive dependencies between inclusion of industrial centers and *TDS* (Total Dissolve Solid), *BOD* (Biological Oxygen Demand) and density of heavy metals, respectively. However, Figure 6.d represents the existence of a negative dependency between inclusions of industrial centers and *DO*



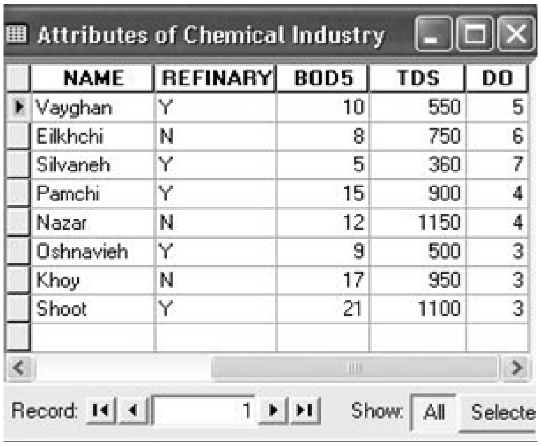
NAME	MINE_CONST	INDS_CONST	POP_CONST	PH	DO	BOD	TDS	MENTAL
S1	1	2	205000	7.6	4.5	11.3	806	8.2354
S10	2	5	50000	8.1	6.1	11.9	856	9.1235
S11	2	0	250000	7.8	4.9	12	864	6.0215
S12	2	0	250000	7.9	5.2	12	864	6.2164
S13	1	1	240000	7.8	5	11.8	847	7.2235
S14	1	1	200000	7.7	4.7	11.1	789	7.3654
S15	1	1	300000	7.6	4.8	12.2	881	7.5698
S16	2	7	3400000	7.1	3.8	15.5	1250	10.9782
S17	1	1	350000	7.5	4.5	12.3	889	7.9875
S18	1	2	215000	7.6	4.6	11.3	806	8.3654
S19	1	1	300000	7.7	4.8	12.1	872	7.8965
S2	1	1	900000	7.6	4.6	13.1	958	8.2356
S20	1	5	1350000	7.4	4.5	13.3	975	9.3654
S21	1	1	700000	7.7	4.7	12.9	940	8.1236
S22	2	0	150000	8.1	5.8	10.9	773	6.0021
S23	1	4	2600000	7.2	4.1	15	1125	8.9854
S24	1	0	565000	7.5	4.4	12.5	906	6.9154
S25	1	0	1040000	7.5	4.6	13.2	966	7.1236
S26	0	2	550000	7.4	4.5	12.6	915	8.6542
S27	1	0	230000	7.6	4.5	11.7	839	5.8754
S28	2	1	170000	7.9	5.4	10.7	756	7.5698
S3	1	1	150000	7.8	4.7	10.8	765	7.6421
S4	1	1	190000	7.5	4.7	11	781	7.9852
S5	2	1	0	8.4	6.3	9	621	7.2564
S6	2	1	110000	8.1	5.9	10.5	740	7.4231
S7	1	1	600000	7.4	4.4	12.6	915	7.8654
S8	1	1	400000	7.6	4.6	12.5	906	7.6854
S9	1	0	40000	7.9	4.8	10	700	5.2365

(a) Attribute data of water quality measurement stations.



NAME	POPULATION	REFINARY
Bazargan	350000	Y
Kelisa	100000	Y
Goldasht	55000	Y
Shoot	65000	Y
SiahCheshme	150000	Y
Khomarloo	85000	N
Ziaedin	65000	N
Zoorabad	75000	N
Jolfa	650000	Y
Siahrood	200000	N
Tazekand	150000	N
Aboghli	75000	N
Khoy	350000	Y
Zonooz	300000	Y
Pamchi	85000	Y
Kashksaray	90000	N
Ahar	450000	Y
Nasooj	30000	N

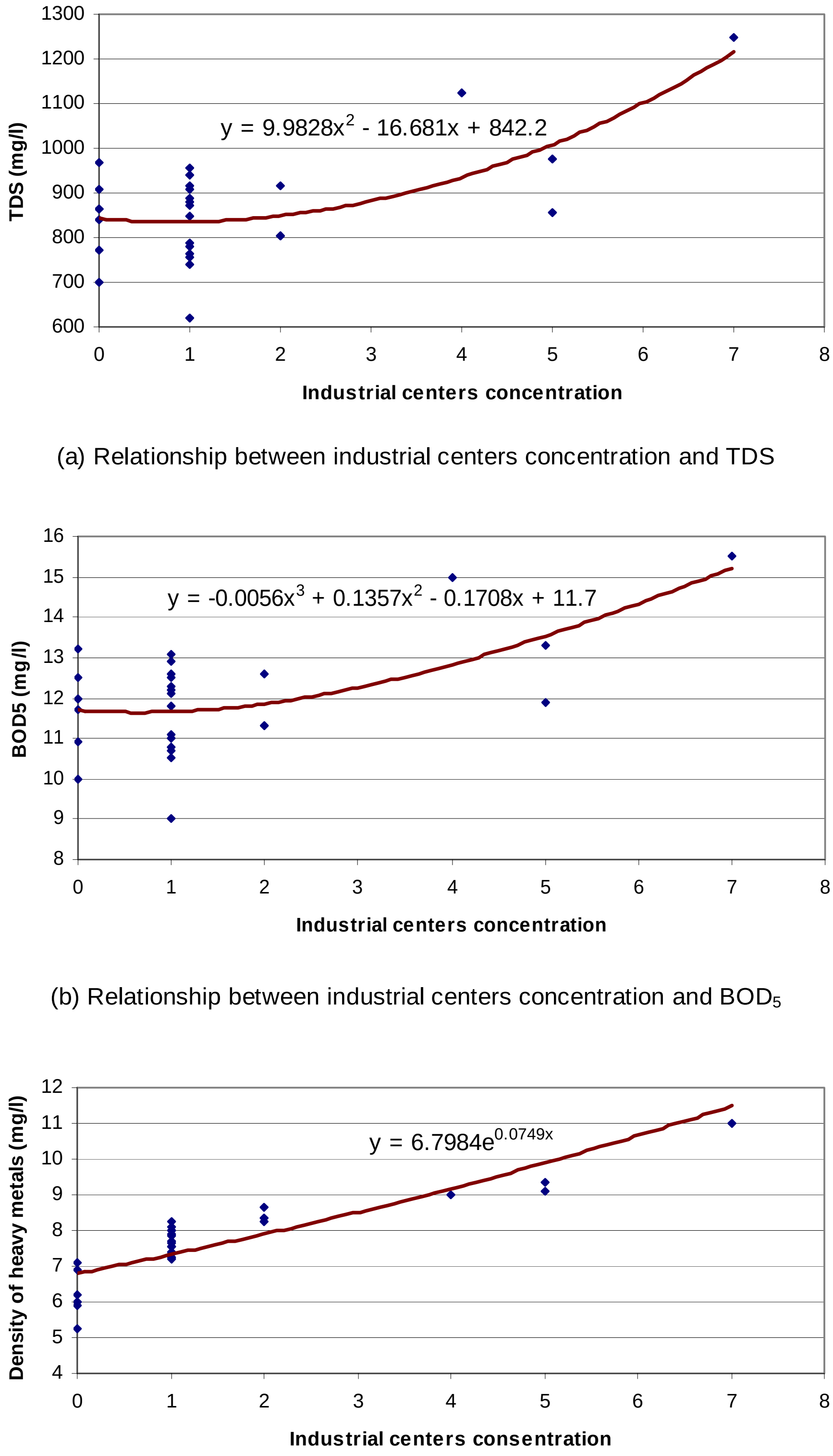
(b) Attribute data of cities



NAME	REFINARY	BOD5	TDS	DO
Vayghan	Y	10	550	5
Eilkhchi	N	8	750	6
Silvaneh	Y	5	360	7
Pamchi	Y	15	900	4
Nazar	N	12	1150	4
Oshnavieh	Y	9	500	3
Khoy	N	17	950	3
Shoot	Y	21	1100	3

(c) Attribute data of industrial centers
(e.g. chemical centers)

Figure 5. The attribute data used.

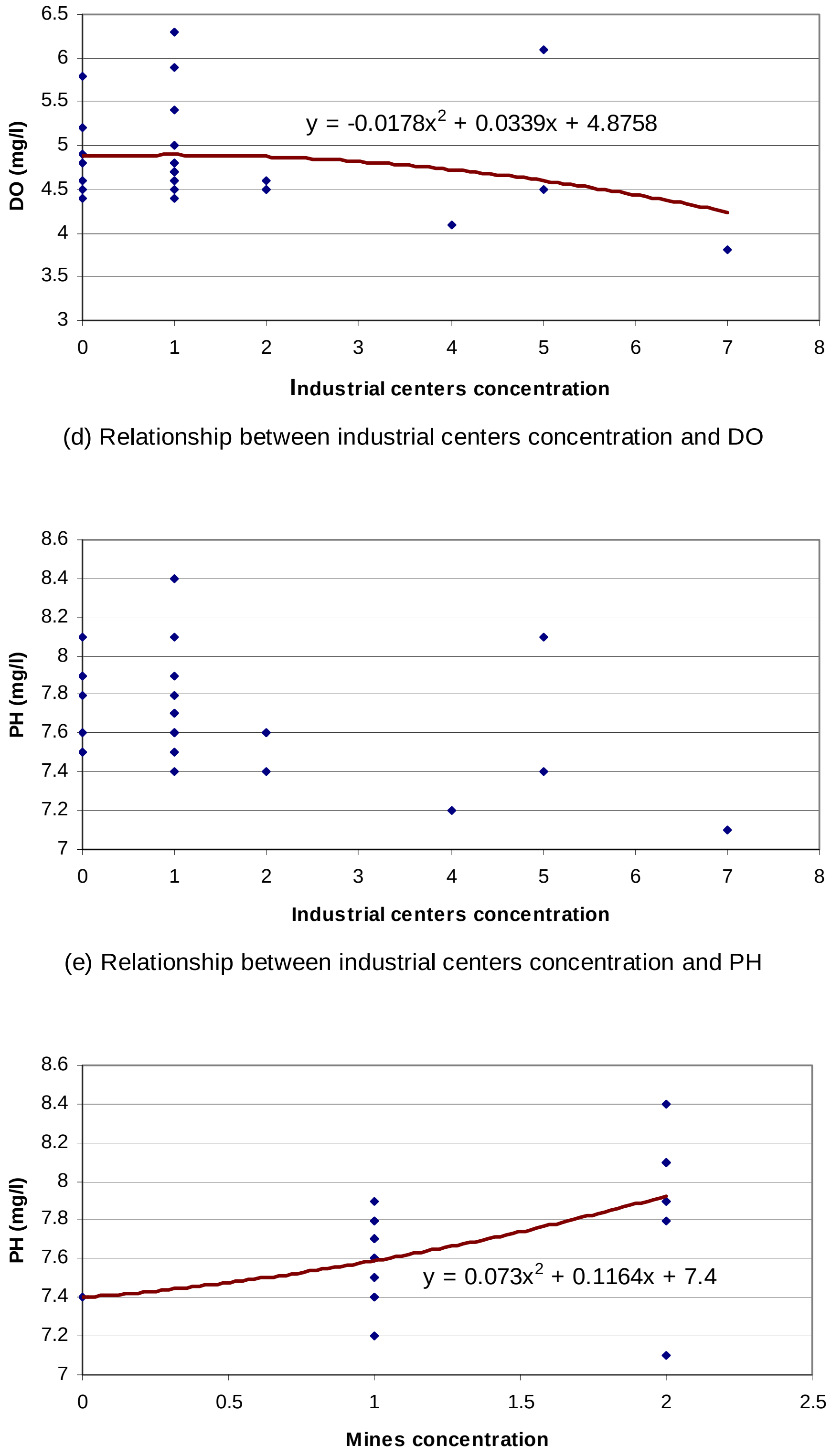


(a) Relationship between industrial centers concentration and TDS

(b) Relationship between industrial centers concentration and BOD_5

(c) Relationship between industrial centers concentration and Mental Densification

Figure 6. The results of dependency analyses.

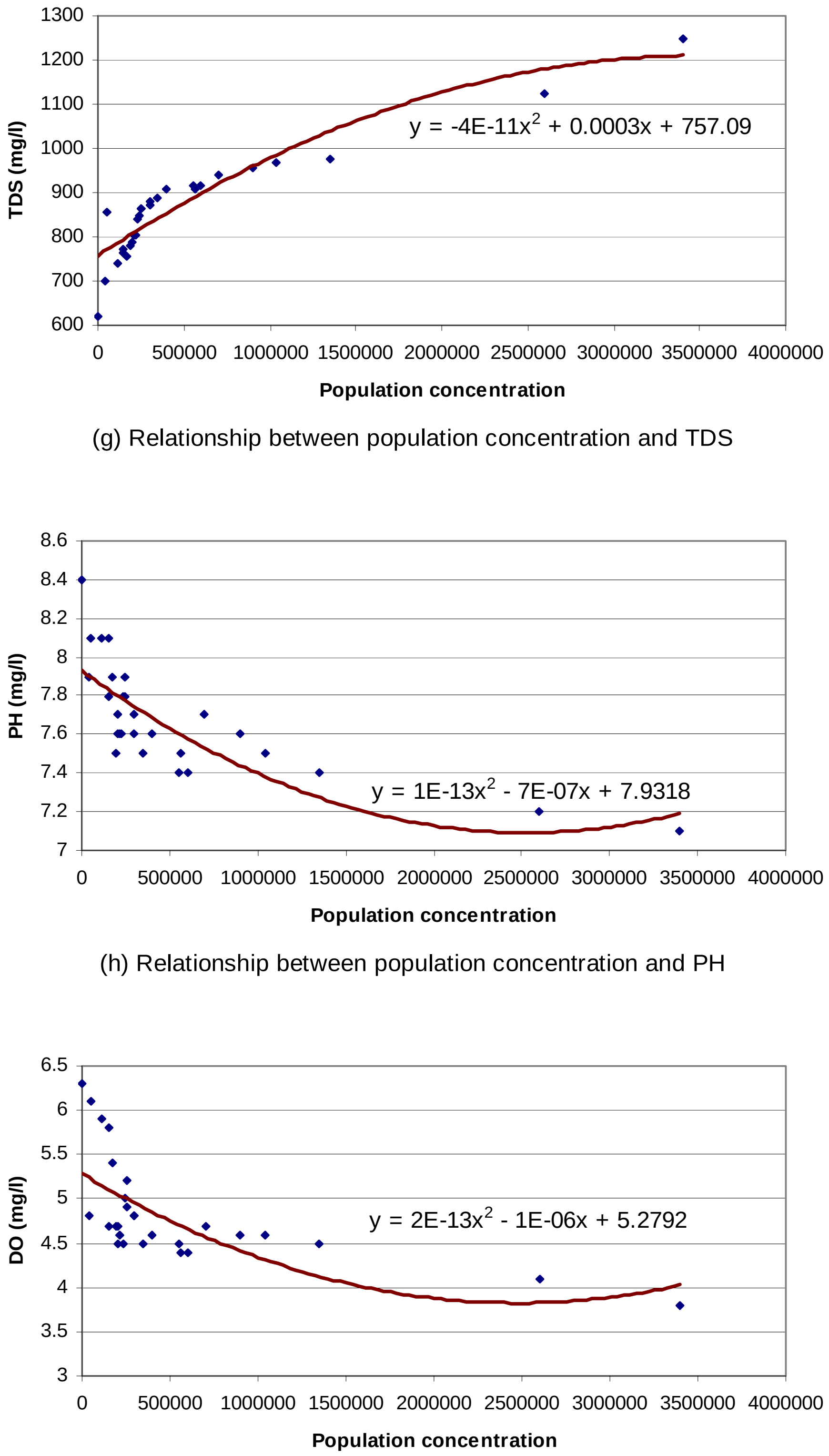


(d) Relationship between industrial centers concentration and DO

(e) Relationship between industrial centers concentration and PH

(f) Relationship between Mines concentration and PH

Figure 6. The results of dependency analyses (Cont.).



(g) Relationship between population concentration and TDS

(h) Relationship between population concentration and PH

(i) Relationship between population concentration and DO

Figure 6. The results of dependency analyses (Cont.).

Table 1. Calculations of Function of Dependency, Goodness of Fitness Test and its Significant for DO Given Its Relation to Population Concentration

Population Con.	*DO*	First Degree Polynomial		Second Degree Polynomial		Third Degree Polynomial		Exponential	
		$\hat{DO}$	$(DO-\hat{DO})^2$	$\hat{DO}$	$(DO-\hat{DO})^2$	$\hat{DO}$	$(DO-\hat{DO})^2$	$\hat{DO}$	$(DO-\hat{DO})^2$
205000	4.5	5.000	0.250	5.083	0.339	4.974	0.224	4.961	0.213
900000	4.6	4.722	0.015	4.541	0.003	4.209	0.153	4.628	0.001
150000	4.7	5.022	0.104	5.134	0.188	5.101	0.161	4.988	0.083
190000	4.7	5.006	0.094	5.096	0.157	5.007	0.094	4.968	0.072
0	6.3	5.082	1.483	5.279	1.042	5.507	0.629	5.064	1.528
110000	5.9	5.038	0.743	5.172	0.531	5.201	0.489	5.008	0.795
600000	4.4	4.842	0.196	4.751	0.123	4.362	0.001	4.769	0.136
400000	4.6	4.922	0.104	4.911	0.097	4.608	0.000	4.865	0.070
40000	4.8	5.066	0.071	5.240	0.193	5.390	0.349	5.044	0.059
50000	6.1	5.062	1.077	5.230	0.757	5.362	0.544	5.039	1.127
250000	4.9	4.982	0.007	5.042	0.020	4.878	0.001	4.939	0.002
250000	5.2	4.982	0.047	5.042	0.025	4.878	0.104	4.939	0.068
240000	5	4.986	0.000	5.051	0.003	4.898	0.010	4.944	0.003
200000	4.7	5.002	0.091	5.087	0.150	4.985	0.081	4.964	0.069
300000	4.8	4.962	0.026	4.997	0.039	4.779	0.000	4.914	0.013
3400000	3.8	3.722	0.006	4.191	0.153	6.636	8.043	3.604	0.038
350000	4.5	4.942	0.196	4.954	0.206	4.689	0.036	4.890	0.152
215000	4.6	4.996	0.157	5.073	0.224	4.952	0.124	4.956	0.127
300000	4.8	4.962	0.026	4.997	0.039	4.779	0.000	4.914	0.013
1350000	4.5	4.542	0.002	4.294	0.043	4.364	0.018	4.424	0.006
700000	4.7	4.802	0.010	4.677	0.001	4.284	0.173	4.721	0.000
150000	5.8	5.022	0.605	5.134	0.444	5.101	0.488	4.988	0.659
2600000	4.1	4.042	0.003	4.031	0.005	5.954	3.439	3.904	0.038
565000	4.4	4.856	0.208	4.778	0.143	4.397	0.000	4.786	0.149
1040000	4.6	4.666	0.004	4.456	0.021	4.213	0.150	4.564	0.001
550000	4.5	4.862	0.131	4.790	0.084	4.412	0.008	4.793	0.086
230000	4.5	4.990	0.240	5.060	0.313	4.919	0.176	4.949	0.201
170000	5.4	5.014	0.149	5.115	0.081	5.054	0.120	4.978	0.178
J		6.046		5.424		15.616		5.887	
$\chi^2_{0.5\%,df}$		11.160		10.520		9.88		11.160	
df		26		25		24		26	

(Dissolved Oxygen). Experiments confirm these results, too. Waste water of industries in the study area has significant pollutants of these parameters and therefore concentration of industrial centers increases these types of pollutants in the rivers. It is notable that the amount of *DO* and water pollution have a reverse relation, that is decrease of *DO* increases the water pollution.

Figure 6.e illustrates no dependency between inclusions of industrial centers and *PH* value of water. Industrial centers of this region do not have *PH* pollutant and amount of *PH* of their waste water does not have a significant trend and therefore concentration of industrial centers does not affect *PH* of the rivers.

On the other hand, Figure 6.f represents existence of a positive dependency between the amount of mine concentration and *PH* value of water. Mine of this region are alkali and therefore concentration of them increases the amount of *PH* of the rivers.

Figure 6.g illustrates the existence of a positive dependency between the amount of population concentration and *TDS*. Finally, Figures 6.h and 6.i represent the existence of negative dependencies between the amounts of population concentration, *PH* and *DO* values of water, respectively. Both increase of *TDS* and decrease of *DO* imply the pollution of water. There is no information about these pollution parameters of cities in the study area, however, these dependencies may be as a result of existence of such pollutions in the waste water of the cities and therefore concentration of the populations increases the pollution of the rivers.

Some environmental parameters, which have not been considered at this stage of the research, could affect the results. For example, the high value of *BOD* in a river could be due to the existence of alga which increases the value of oxygen through oxygen production representing a high value of *BOD* in a river in an unexpected way.

7. Conclusions and Future Works

This paper provided a brief review on geospatial data mining, the need and required analyses to address the importance of water quality management. Some applications of data mining technique have been tested.

Due to the importance of water quality management and increase volume of its databases, the use of new methods for database management and information extraction is unavoidable. However, considering the importance of spatial component in water quality management, the use of GIS and geospatial data mining for information extraction is essential.

The achieved results regarding the spatial distribution of water quality parameters can be used for industrial site selection in master planning stage. In this case any proposed industrial center will be checked against its associated water quality parameters and the construction permission will be assigned provided the quality parameters do not exceed the accepted threshold.

Although, lack of qualified and complete dataset required was one of the major problems faced in this research, the results proved the suitability of geospatial data mining for water quality management.

In addition, regarding the statistical nature of geospatial data mining analyses, existence of more qualified and complete datasets in the database, could lead to better results.

As discussed before, all industrial centers in the study area within a 20 km buffer zone selected on the basis of trial and error are considered. To improve the buffer zone selected, water body of each river can be determined and the industrial centers located within the zone considered.

In this paper, static data with no temporal components were used. However, with the evolution of temporal GISs recording spatio-temporal water resource data, some results can be provided to better evaluate the existing situation and predict the future trend.

Acknowledgments. The authors wish to thank Iranian Department of Environment (DOE) due to very useful helps in delivering the required data. Especially, The comments given by Mrs. Nadia Roostaei (Eng.) from DOE on initial draft of this paper is highly appreciated.

References

Brezonik, P.L., Kloiber, S.M., Olmanson, L.G. and Bauer, M.E. (2002). *Satellite and GIS Tools to Assess Lake Quality*, Water Resources Center, University of Minnesota.

Cunningham, W.P. and Saigo, B.W. (1999). *Environmental Science a Global Concern (Fifth Edition)*, McGraw-Hill.

Ester, M., Kriegel, H.P. and Sander, G. (1997). *Spatial Data Mining: A Database Approach*, Springer, Berlin.

Flynn, K.M. (1999). *Water Quality and Geographic Information Systems: The Alabama Watershed Demonstration Project*, Auburn University.

Ford, G.E. and Zanelli, C.I. (1985). Analysis and quantification of errors in the geometric correction of sattellite images, in L.K. Fenstermarker (Ed.), *Remote Sensing Thematic Accuracy Assessment: A Compendium*, American Society for Photogrammetry and Remote Sensing.

http://www.un.org/esa/sustdev/documents/docs.htm.

Jerk, M., Lenzerini, M., Vassiliou, Y. and Vassiliadis, P. (2000). *Fundamentals of Data Warehouses*, Springer, Berlin.

Miller, H.J. and Han, J. (2000). Discovering geographic knowledge in rich environment, *SIGKKD Exploration*, pp. 105-108.

Miller, H.J. and Han, J. (2001). *Geographic Data Mining and Knowledge Discovery*, Taylor and Francis.

Roddick, J.F. and Spiliopoulou, M. (1999). A bibliography of temporal, spatial and spatio-temporal data mining research, *SIG Explorations*, pp. 34-38.

Singh, R.B. (1995). *Global Environmental Change Perspective of Remote Sensing and Geographical Information System*, A.A. Balkema, Rotterdam.

Tobler, W.R. (1994). Bidimentional regression, *Geographic Analysis*, pp. 187-212.

Web site: http://www.un.org/esa/sustdev/documents/docs.htm (accessed may 2005).