International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-2, December 2019

Preprocessing the Groundwater Quality data by LSR and QDR techniques

Arunkumar Rajamani, Velmurugan Thambusamy

Abstract: ***Data mining is the process of identifying patterns and their relationships to solve problems through data analysis. Data mining is utilized to haul out working information from a colossal dataset of any crude information. Environmental mining is one of the wide areas to find impact on environment. Data mining encourages the usage of essential strategies and finds noteworthy information from gigantic measure of environmental information. Data preprocessing techniques are very essential in data mining, which uses various techniques to convert the raw data into a meaningful data to further research work. In this research work, Logical Similarity Replacement (LSR) and Quantity based Discrepancy Replacement (QDR) algorithms are proposed to ascertain the quality of groundwater. The numerical information are preprocessed by the statistical techniques Mean, Median methods and non-numeric information are preprocessed by the proposed LSR and QDR methods to satisfy the fragmented and conflicting information in the dataset. The conflicting and the missing information are corrected by the picked strategies for preprocessing. In the wake of applying these preprocessing systems connected in the dataset, the nature of the informational index is improved.***

Keywords: Preprocessing, Statistical Techniques, Logical Similarity Replacement, Quantity based Discrepancy Replacement, Groundwater Quality.

I. INTRODUCTION

Data and Information encase a noteworthy job on human exercises. Data mining is the way toward breaking down gigantic volume of information from various perspectives. Technically, data mining is the activity of detecting patterns among huge relational databases. Techniques of data mining pertinent to all domains are based on the need of the applications. Data mining has an extremely important role in different fields of human life and it is generally utilized in a several zones, for example, money related information investigation, media transmission industry, natural information examination, and so on. In data mining, environmental mining is one of the fundamental fields in our everyday life. Environmental technologies are used for pollution abatement, waste management, energy, water and material conservation, and for improving technological efficiency of production [1].

Revised Manuscript Received on December 30, 2019.
***** Correspondence Author
Arunkumar Rajamani*, Ph.D Research Scholar, Department of Computer Science, D.G. Vaishnav College, Chennai, India.
Velmurugan Thambusamy, Associate Professor In The PG And Research Department of Computer Science and Applications, D. G. Vaishnav College, Chennai, India.

Groundwater is an essential resource of drinking water for people around the world, particularly in rural areas. It is accessible wherever under the world's surface not in a solitary wide spread aquifer but rather in a great many nearby aquifer frameworks and segments that have comparable characters. Substitution and recuperation of groundwater has an exceptional

essentialness in parched and semi-dry locales. Because of disparity in monsoonal precipitation, lacking surface waters and over drafting of groundwater assets happens. Water pollution not only affects water quality but also threats human health, economic development, and social prosperity [2]. Quality of groundwater depends on the quality of recharged water, atmospheric precipitation, inland surface water, and on sub-surface geochemical processes [3]. The groundwater quality dataset was gotten from the Government of Tamil Nadu Water Resource Department, Public Works Department, Taramani, Chennai. The dataset have 5 years information from 2012 to 2016. Six areas groundwater quality information were gathered for this exploration work. The names of the areas are Thiruvallur, Kancheepuram, Cuddalore, Thiruvannamalai, Vellore and Villupuram. In this exploration work, MATLAB (rendition R2015a) programming is utilized for preprocessing to dissect the groundwater quality.

Raw data may mislead the target prediction in research work. Data preprocessing has several techniques, which changes over the crude information into sensible information to empower the forthcoming examination works. It analyzes the raw data and segregates noisy data, missing values and inconsistent data in the dataset. Preprocessing utilizes appropriate systems to correct these sorts of information into important profitable information. At first missing qualities, inadequate information and boisterous information are distinguished and supplanted by information cleaning forms, at that point incorporate the cleaned information with the dataset. The data transformation technique standardizes the dataset, which merges the dataset utilizing accumulation, smoothing and speculation forms. At long last excess qualities and insignificant information are diminished and invalid information is expelled by discretization process in the dataset. The dataset is dissecting the groundwater quality in six regions just in Tamil Nadu. It contains missing qualities and conflicting information as it were. LSR, QDR strategies and Statistical methods are utilized to clean the dataset and regularize to additionally examine work.

Organization of the paper is portrayed as pursues. Section II examines the survey of writing about groundwater quality procedures and preprocessing systems, which are utilized in ground water quality.

Section III investigates materials and strategies that are utilized in preprocessing. The results and dialog are depicted in section IV. Finally, the research work is concluded in section V.

II. REVIEW OF LITERATURE

Groundwater is the most fundamental one to each living thing and it is a standout amongst the most imperative inexhaustible assets everywhere throughout the world. Individuals principally rely upon the groundwater for water system, local and modern purposes each day. N. S. Magesh et al. proposed GIS mapping technique to determine the groundwater quality in Dindigul district. The researchers found out presence of major ions (HCO_3, Cl, FSO_4, Ca, Mg, Na and K) and physicochemical parameters (pH, EC, TDS and TH) are primarily controlled by rock–water interaction and residence time of groundwater [4]. K. Dhanasekar and P. Partheeban were using American Public Health Association (APHA) standard methods and procedures to determine the water quality in Chennai. They analyzed the physico-chemical parameters in ground water [5]. Kazem Nosrati et al. suggested a multivariate statistical technique for assess the groundwater quality, which was combination of Cluster Analysis (CA), Factor Analysis (FA), Principal Component analysis (PCA) and DA. The researchers analyzed various groundwater parameters in this research work [6].

The physico-chemical parameters levels in groundwater in Periyakallapadi village, Thiruvannamali district assessed by B. Sundararaman and K. Muthuramu. The researchers used standard methods to calculate quality of the groundwater. They described the quality of the groundwater was inappropriate for drinking and agricultural use [7]. S. Krishna Kumar et al. proposed fuzzy logic method to carry out the quality of the groundwater in south Chennai coastal area, Tamil Nadu. They classified the groundwater quality parameters into 5 major groups, which are 'Excellent water', 'Good water', 'Poor water', 'Very poor water' and 'Water Unsuitable for drinking purposes' based on the Water Quality Index (WQI) [8]. Boyacioglu described WQI improves understanding of water quality issues by integrating complex data and generating a score that describes water quality status and evaluates water quality trends [9]. S. Krishna Kumar et al. evaluated the groundwater and drinking water geochemical attributes in Anna Nagar, Chennai. The researchers used World Health Organization (WHO) and Bureau of Indian Standards (BIS) water quality standards to find out the quality of the groundwater.

They illustrated majority of the water samples were excellent, good and suitable for drinking [10]. Bing Zhang et al. proposed fuzzy membership analysis multivariate statistical technique to evaluate the surface water and groundwater quality in Songnen plain, Northeast China. That multivariate statistical technique was combination of Hierarchical Cluster Analysis (HCA) and PCA, which were used for cluster and assess the following water quality parameters Na, HCO_3, NO_3, Fe, Mn and EC in their research [11]. Sudhir Dahiya et al. proposed fuzzy synthetic evaluation model for analyze groundwater quality in southern Haryana. They analyzed the physico-chemical quality in the groundwater for drinking purpose and categorized the groundwater into 'desirable', 'acceptable' and 'not acceptable' based on the quality [12]. Kunwar P. Singh et al. analyzed fluride level in soil and groundwater using chemometric techniques such as PCA, DA and Partial Least Squares (PLS) [13]. Factor analysis techniques used by Chen-Wuing Liu et al. for assess groundwater quality. They described the quality of groundwater affected ,due to sea water salinization and arsenic pollution [14]. Ting-Nien Wu and Chiu-Sheng Su used statistical techniques such as average and standard deviation to replace the half of the missing values in their dataset [15].

M. Kanevskia et al. approached ANN techniques to removed unwanted non-linear spatial structures in environmental data mining. They used Machine learning algorithms to improve the image processing [16]. Quality of groundwater could analyze by A. Saberi Nasr et al. in Yazd Province, Iran. The researchers proposed mamdani fuzzy inference system to analyze the groundwater into acceptable and non-acceptable group [17]. Keunje Yoo et al. analyzed hydrological parameters and groundwater pollution with the use of decision tree based technique. They normalized the data using min-max normalization method, which is a simplest statistical procedure [18]. Harald Genther and Manfred Glesner implemented fuzzy maximum likelihood estimation (FMLE) method to preprocess the raw data. They achieved good classification results by using fuzzy classification system [19]. Groundwater level was predicted using several data driven techniques by Bagher Shirmohammadi et al. They used adaptive neuro-fuzzy inference system (ANFIS) model for forecast the groundwater level. Fuzzy inference system improved the system identification models [20].

III. MATERIALS AND METHODS

Data preprocessing is the principal venture in information mining process. It is the underlying stage to change the raw data into reasonable data. The precision and productivity of information can be expanded by information preprocessing procedures. Preprocessing systems are utilized to achieve simplicity of mining and upgrade the information portrayal.

A. Description of Dataset

The dataset contains Tiruvallur, Kancheepuram, Cuddalore, Thiruvannamalai, Vellore and Villupuram districts groundwater quality data, and these districts are located in northern parts of Tamil Nadu state. The groundwater is a standout amongst the most essential wellsprings of residential and farming use in these areas. In pre monsoon and post monsoon periods of consistently the groundwater quality information are gathered by the Water Resource Department and the nature of the information is dissected. In this research work, the dataset enclosed pre monsoon and post monsoon information in the time of 2012 to 2015 however 2016 has just pre monsoon information. The dataset has 24 parameters and these parameters are characterized into two subgroups, which are numeric and non-numeric.

The quantities of numeric and non-numeric parameters are 17 and 7 respectively. The numeric parameters are utilized to decide the nature of the groundwater quality in those areas. These parameter names are Total Dissolved Solids (TDS), NO_2+NO_3, Calcium (Ca), Magnesium (Mg), Sodium (Na), Potassium (K), Chloride (Cl), Sulphate (SO_4), Carbonate (CO_3), Bicarbonate (HCO_3), Fluoride (F), pH_GEN, Electrical Conductivity_GEN (EC_GEN), Total Hardness, Sodium Adsorption Ratio (SAR), Residual Sodium Carbonate (RSC) and Na%. The non-numeric parameters are utilized to recognize the area of the groundwater quality information. Well Number, District, Taluk, Village, Latitude, Longitude, and Date of Collection are the non-numeric parameters in the dataset.

Table- I. Invalid data in the dataset

Type of Data / Type of Invalid Value	Missing Values	Inconsistent Values	Total Invalid Values
Non-numeric	170	69	239
Numeric	843	4701	5544

In the dataset, 1013 missing values and 4770 inconsistent values are distinguished. Absolutely 5783 invalid information are preprocessed and changed into important information to this research work. The total missing and inconsistent values in the dataset are represented in Table 1. In view of the Table 1 information the complete number of invalid qualities in the dataset is represented in Fig 1.

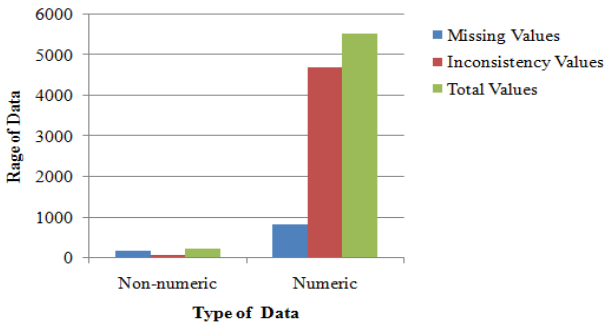

Fig 1. Representation of invalid data

The dataset thoroughly contains 3746 records, which are utilized to break down the nature of the groundwater in six regions. The example contribution of the groundwater quality dataset is outlined in Table 2. This sample input contains missing values and inconsistent esteems and these qualities are changed over into significant profitable information by proposed LSR, QDR strategies and measurable methods.

B. Methodology

In Data mining, preprocessing is the initial move towards the mining procedure. The preprocessing underpins a few techniques that empower to deliver the legitimate information and to improve the exactness of the information. For getting the high precision in predication, preprocessing is a conspicuous procedure in research work. Information preprocessing is finished utilizing the accompanying strides to complete the all around shaped information to further research work. The stream of information preprocessing is represented in Fig 2.

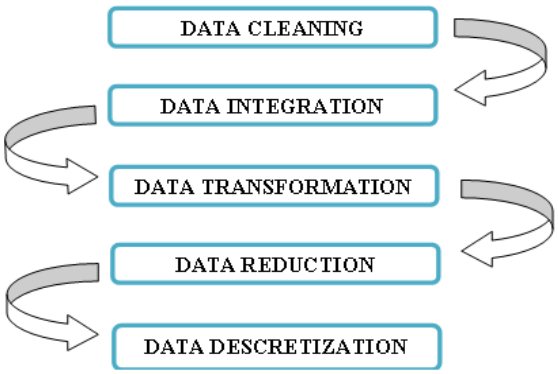

Fig 2. Flow of data preprocessing

In this research work, the data cleaning process replaces the missing information and the loud information in the dataset. The cleaned information is then fused with the dataset by data integration. From that point forward, the dataset is combined into appropriate structure for the mining procedure.

3.2.1 Statistical Methods

Statistical methods are used to analyze the information and to apply it in various types of factual issues. Some of the statistical methods are mean, median, standard deviation and so forth. In this research work, mean and median strategies are utilized to satisfy the fragmented numerical information in the dataset.

(a) Mean: Mean is one the most widely recognized and effective numerical strategy to fill the missing qualities in the dataset, which finds the average estimation of a lot of information. As such, mean is the whole of every datum in a set partitioned by number of information in a set. On the off chance that x_1, x_2….x_n be a lot of 'n' values for numeric property x, at that point the mean esteem is

Table 2. Sample input of groundwater

5	13057A		Tiruvallur	Tirupalaivanam	13°2'409"	80°14'58"	7/3/2015	2548	8	120	116.64	688	23	1134	10	0	841.8	0.23	7.8	4610	780	10.71993	0	64.91436

Parameters / S.No.	1	2	3	4
Well no	13090	23040	A23033	HP31572
District	Thiruvallur	Thiruvannamalai	Vellore	Cuddalore
Taluk	Ambatthur	Chengam	Arakkonam	Cuddalore
Village	Velappanchavadi	Pachal	Sendamangalm	Periakattuppalayam
Latitude	13°03'38"	12°17'38"	12°59'13"	11°51'05"
Longitude	80°03'30"	78°57'26"	79°40'39"	79°47'30"
Date of Collection	1/11/2013	1/6/2012	1/6/2016	1/2/2013
TDS		1330		1374
NO_2+NO_3		42		15
Ca				128
Mg				-65.61
Na				389
K				86
Cl				355
SO_4				235
CO_3				-12
HCO_3				0
F				0.31
pH_GEN		8		7.6
EC_GEN		2010		2220
HAR_Tota				50
SAR				24.06229
RSC		0		0
Na%				84.14371

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n} \quad (1)$$

Where,

$\bar{x}$ - Mean

x_1 - The first value

x_2 - The second value

x_n - The last value

n - Total number of values for attribute *x*

$\sum_{i-1}^{n} x_i$ - Summation of 'n' values for attribute *x*

(b) Median: Middle is another preeminent strategy to finish the missing numeric qualities. It very well may be any an incentive inside two middlemost estimation of among all qualities in a lot of information. The data are grouped in intervals according to data values and that the frequency (i.e., number of data values) of each interval, the interval that contains the median frequency be the median interval [21].

$$Median = l + \frac{h}{f}\left(\frac{N}{2} - c\right) \quad (2)$$

Where,

l - Lower class boundary of the median class

h - Size of the median class interval

f - Frequency corresponding to the median class

N - Total number of observation i.e. sum of the frequencies

c - Cumulative frequency preceding median class.

3.2.2 Semantic Analysis Techniques

Semantic analysis is a strategy for relating syntactic arrangements from phrases, conditions, sentences and sections to the dimension of composing all in all, to their language-autonomous implications. It gives the connection between individual words.

The lack of semantic information may give series drawbacks, when implementing data mining algorithms with high cardinality and scattered data collected in particular fields. Semantic calculations or techniques will be utilized in Pattern extraction, Anomaly deletion, Similarity computation (Similarity Matching) and Classification. The proposed LSR and QDR methods are utilized to supplant the invalid information in the non-numerical qualities. LSR and QDR algorithms are supplanted the missing values and conflicting values in the dataset separately. These proposed calculations substitutes the most related values in the dataset. In this exploration work, the non-numeric missing values in the dataset can be practiced by LSR algorithm.

Logical Similarity Replacement Algorithm: The LSR algorithm is used to fulfill the missing non-numeric information in the groundwater quality dataset. The non-numeric missing values in the characteristic are supplanted by comparable values in the similar trait with the utilization of remarkable property coordinating in the dataset.

Step 1: Select one or more than one unique value attributes in the dataset for similarity matching.

Step 2: Check whether the attribute is numeric or non-numeric in the dataset.

Step 3: If the attribute is non-numeric, then check whether the attribute value in the row is empty or not.

Step 4: If the missing value is identified, then store the unique value of the corresponding missing value row into new factor.

Step 5: compare the new factor value and the relating unique value parameter in the dataset. If the values are matched, check whether the required value of the non-numeric attribute is empty or not.

Step 6: If the required value is NOT NULL, then replace the missing value by required value in that corresponding attribute. The missing value is fulfilled by 'similarity value'.

Step 7: Repeat step 4 until the parameter has no more missing values.

Step 8: Repeat step 3 until all the non-numeric attributes are handled in the dataset.

The LSR algorithm identifies the non-numeric missing value in the dataset. The one of a kind columns value in the dataset used to discover the missing values in non-numeric traits. At last supplant the missing value by required value in the relating quality.

Quantity based Discrepancy Replacement Algorithm: The QDR algorithm is used to rectify the inconsistent data in the dataset. In the dataset, District, Taluk and Village attribute values are comprises some wrong esteem in light of incorrectly spelled. It may lead to produce an inaccurate result in research work. The QDR strategy for non-numeric conflicting information is represented to beneath.

Step 1: Check whether the attribute is numeric or non-numeric in the dataset.

Step 2: If the attribute is non-numeric, then find out the number of distinct values and the count of each distinct values in the attribute.

Step 3: Find the inconsistent data in the list of distinct values.

Step 4: Supplant the conflicting information by utilizing most astounding include of that information in the characteristic.

Step 5: Repeat step 2 until the attribute has no more inconsistent values.

Step 6: Repeat step 1 until all the non-numeric properties are handled in the dataset.

The above algorithm proficiently supplanted the invalid information into sensible information in the dataset. The QDR approach supplanted the base number of conflicting information by most extreme number of predictable incentive in a similar quality.

IV. RESULTS AND DISCUSSION

In this research work, the groundwater quality dataset was preprocessed utilizing proposed methods and statistical techniques. Two diverse proposed calculations (LSR and QDR) utilized for non-numerical characteristics and the statistical techniques (mean and median) utilized for numerical qualities in this preprocessing research work. This exploration work, the Table 2 which is test information contains the missing qualities and conflicting information. These qualities will create the not exact outcome in the examination work for further research work. Table 3 and Table 4 speak to the all out number of missing qualities and inconsistent data respectively in the table by characteristic shrewd. After the pre processed work, the ground water quality data contains the suitable data to further research work. The dataset have 1013 missing values (blank values) and 4770 inconsistent values (zeros and invalid information). These values are amended and rectified into reliable incentive by information mining strategies. Table 5 portrays the sample output of the groundwater quality dataset. The LSR and QDR methods are very useful to pre processed the data set.

A. Methods for Missing Values

In data cleaning process, missing values are extremely basic unavoidable issue in immense datasets. Different strategies are accessible to process the missing information in datasets and stay away from troubles brought about by it. Disregarding the missing qualities now and again may misdirect the expectation of research work. The groundwater quality dataset encases 20 missing worth parameters among 24 parameters. The dataset contained 1013 missing values and these values are likewise isolated into 843 numeric missing values and 170 non-numeric missing values. The non-numeric missing values are identified by LSR method. The LSR approach for non-numeric properties dependent on the Latitude and Longitude parameters, which dissects each missing value in the attribute and process data and as well as find out the relationship within the data. The LSR technique has two unique attributes which are latitude and longitude attributes. This technique replaces the missing values in that property by the similar characteristic value, which matches the latitude and longitude estimation of the missing worth. Numeric missing values are satisfied by mean technique, which computes the average value of each numeric property and fills the missing values in the suitable characteristics.

Table 3: Results of finished missing qualities

S.No.	Type of Value / Attributes	Missing Values
1	Well No	NIL
2	District	NIL
3	Taluk	170
4	Village	NIL
5	Latitude	NIL
6	Longitude	NIL
7	Date of collection	NIL
8	TDS	49
9	NO_2+NO_3	50
10	Ca	49
11	Mg	49
12	Na	50
13	K	50
14	Cl	51
15	SO_4	53
16	CO_3	50
17	HCO_3	50
18	F	53
19	pH_GEN	48
20	EC_GEN	49
21	HAR_Total	49
22	SAR	49
23	RSC	45
24	Na%	49
	Total	**1013**

B. Methods for Inconsistent Values

Nearness of deficient data, entering an erroneous catch or absence of consideration may make conflicting information. Inconsistent information may happen notwithstanding when distinctive information sources are incorporated into a solitary one. With the conflicting information mayn't create the best possible expectation. So as to preprocess the inconsistency esteems in groundwater quality dataset by utilizing QDR method. The dataset contained 3 non numeric parameters with conflicting qualities, which are District, Taluk and Villages. These attributes discrepancy values are rectified by proposed QDR algorithm. The dataset contains two classifications of numeric conflicting data, which are zeros (0) and negative values.

The statistical data mining procedures convert the inconsistency information into reliable significant information in the dataset. In the research work, Mean and Median statistical techniques are used to get the reliable data for further research work. The Mean technique is utilized to displace the estimations of zero into significant information. Negative qualities are another sort of inconsistency values in the dataset. These values are replaced by the technique for Median. The dataset have two negative value parameters, which are Mg and CO_3. These values are supplanted by the estimation of 42.5250. The conflicting estimation of each characteristic is depicted in Table 4. The replaced inconsistent values are represented in Fig 4.

Table 4: Results of supplanted conflicting qualities

S. No.	Type of Value / Attributes	Inconsistent Values
1	Well No	NIL
2	District	1
3	Taluk	9
4	Village	59
5	Latitude	NIL
6	Longitude	NIL
7	Date of collection	NIL
8	TDS	NIL
9	NO_2+NO_3	16
10	Ca	NIL
11	Mg	6
12	Na	NIL
13	K	4
14	Cl	NIL
15	SO_4	1
16	CO_3	1783
17	HCO_3	2
18	F	NIL
19	pH_GEN	NIL
20	EC_GEN	NIL
21	HAR_Total	NIL
22	SAR	1
23	RSC	2887
24	Na%	1
	Total	**4770**

The groundwater quality dataset consists of missing values and conflicting data. These values are rectified into meaningful data by proposed LSR, QDR methods and statistical Mean, Median techniques. In the preprocessing work, among 24 attributes 20 attribute values are redressed and 1013 missing values and 4770 conflicting values are replaced. In the dataset, the Taluk is one and just non-numeric missing value parameter, which contains 170 deficient information. The missing data is satisfied by the LSR method. Six characteristics (Well No, District, Village, Latitude and Longitude) have no missing values among 24 properties.

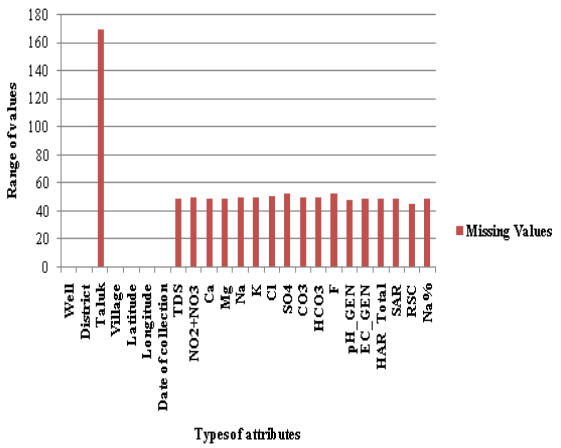

Fig. 3. Demonstrate the finished missing values

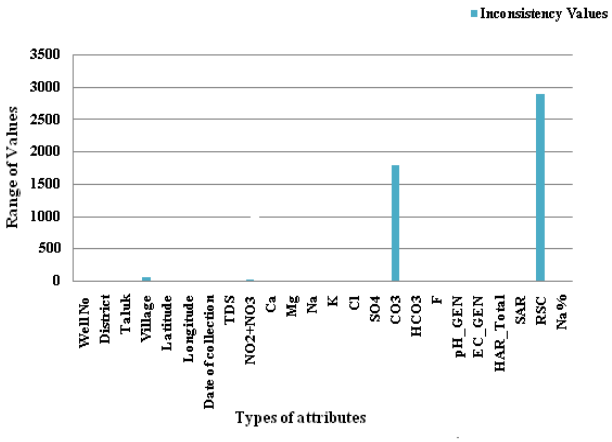

Fig. 4. Illustrate the completed inconsistency values

The QDR method settled 69 non-numeric conflicting values.

For example in Taluk attribute one of the inconsistent data ‘Ambatthur’ has been changed into ‘Ambattur’.The dataset contained zero values in 8 numerical attributes and negative values in two attributes. These values are replaced by median method. For example the zero values in attribute K are replaced by the value 12.5960. Table 5 speaks to the example yield of groundwater quality, which demonstrates satisfied missing values and conflicting values. Finally 5783 values are changed into significant information in the preprocessing work.

Table 5. Sample output of ground water quality

Parameters /S.No.	1	2	3	4	5
Well no	13090	23040	A23033	HP31572	13057A
District	Thiruvallur	Thiruvannamalai	Vellore	Cuddalore	Tiruvallur
Taluk	Ambattur	Chengam	Arakkonam	Cuddalore	Ponneri
Village	Velappanchavadi	Pachal	Sendamangalam	Periakattuppalayam	Tirupalaivanam
Latitude	13°03'38"	12°17'38"	12°59'13"	11°51'05"	13°24'09"
Longitude	80°03'30"	78°57'26"	79°40'39"	79°47'30"	80°14'58"
Data of Collection	1/11/2013	1/6/2012	1/6/2016	1/2/2013	7/3/2015
TDS	822.84	1330	822.84	1374	2548
NO_2+NO_3	11.38	42	11.38	15	8
Ca	61.87	61.87	61.87	128	120
Mg	51.72	51.72	51.75	42.53	116.64
Na	163.61	163.61	163.61	389	688
K	12.61	12.60	12.61	86	23
Cl	260.68	260.68	260.68	355	1134
SO_4	71.40	71.38	71.40	235	10
CO_3	14.96	13.50	21.133	16.25	18.42
HCO_3	266.81	266.66	266.81	266.66	841.8
F	0.56	0.56	0.56	0.31	0.23
pH_GEN	8.14	8	8.145	7.6	7.8
EC_GEN	1473.22	2010	1473.22	2220	4610
HAR_Tota	367.51	367.51	367.51	50	780
SAR	3.73	3.73	3.73	24.06229	10.71993
RSC	0.59	0.49	1.01	0.56	0.82
Na%	44.01	44.01	44.025	84.14371	64.91436

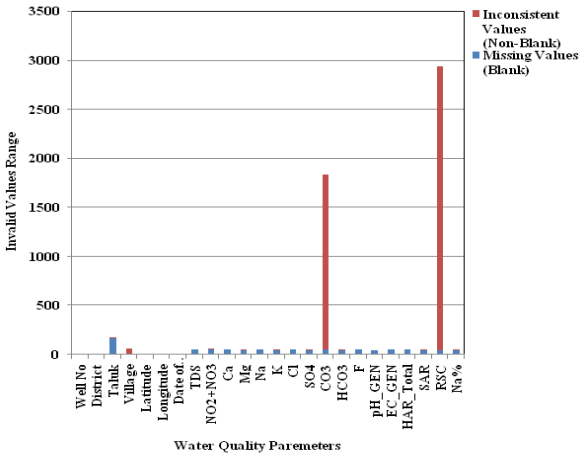

Fig. 5. Representation of completed invalid data

After the preprocessing, the missing values and inconsistent values in the groundwater quality dataset is achieved with important significant information. The replaced values are represented in Figure 5.

V. CONCLUSION

Groundwater quality is essential for everyday existence of people. The parameters of the groundwater are utilized to analyze the nature of the water. In this research work, proposed LSR technique is connected to correct the trait estimations of District, Taluk, and Villages in the dataset. This technique satisfies the missing values in the informational index productively and consequently facilitates further handling. Likewise, another proposed QDR method finds the disparity of information in these parameters. The non-numeric information is splendidly supplanted by the proposed technique. The Statistical techniques are replaced the missing numeric values and conflicting values in the dataset. All the 24 parameters are preprocessed viably by the proposed and statistical techniques. In this research, proposed methods and statistical methods are effectively supplanted the sporadic qualities in the dataset with an abnormal state of attractive with high level of satisfactory. The finished preprocessed dataset has been utilized for further research work. In future, it is distinguished that the nature of information by considering different parameters of the informational collection.

REFERENCES

1. Paul Shrivastava, “Environmental Technologies and Competitive Advantage", Strategic Management Journal , Vol . 16, 2016, pp. 183–200.
2. Mimoza Milovanovic, “Water quality assessment and determination of pollution sources along the Axios/Vardar River, Southeastern Europe”, *Desalination*, Vol. 213, 2007, pp. 159–173.
3. M. Vasanthavigar, K. Srinivasamoorthy, K. Vijayaragavan, R. Rajiv Gandhi, S. Chidambaram, P. Anandhan, R. Manivannan and S. Vasudevan, “Application of water quality index for groundwater quality assessment : Thirumanimuttar sub-basin, Tamilnadu, India”, Environ. Monit Assess, 2010, Vol . 171, pp. 595–609.
4. N.S. Magesh, S. Krishnakumar, N.Chandrasekar and John Prince Soundranayagam, “Groundwater quality assessment using WQI and GIS techniques, Dindigul district, Tamil Nadu, India”, Vol. 6, 2013, pp. 4179–4189.
5. K. Dhanasekar and P. Partheeban, “Water quality index for groundwater in Chennai, Tamilnadu, India", Vol 33(2), EM International, 2014, pp. 327-335.
6. Kazem Nosrati and Miet Van Den Eeckhaut, “Assessment of groundwater quality using multivariate statistical techniques in Hashtgerd Plain, Iran”, Environ. Earth Sci., Vol. 65(1), Nov 2015, pp. 331-.
7. B. Sundararaman and K. Muthuramu, “Assessment of Ground Water Quality in Tiruvannamalai District-Random Study in Periyakallapadi Village”, Int. Journal of Environ. and Nat. Res., Vol. 9(2), Mar 2018, pp. 1–5.
8. S. Krishna Kumar, R. Bharani, N. S. Magesh, Prince S. Godson and N. Chandrasekar, “Hydrogeochemistry and groundwater quality appraisal of part of south Chennai coastal aquifers, Tamil Nadu, India using WQI and fuzzy logic method”, Appl. Water Sci., Vol. 4, 2014, pp. 341–350.
9. Hulya Boyacioglu, “Development of a water quality index based on a European classification scheme”, Vol. 33(1), 2007, pp. 101–106.
10. S. Krishna Kumar, A. Logeshkumaran, N. S. Magesh, Prince S. Godson and N. Chandrasekar, “Hydro-geochemistry and application of water quality index (WQI) for groundwater quality assessment, Anna Nagar, part of Chennai City, Tamil Nadu, India”, Appl. Water Sci., Vol. 5, 2015, pp. 335–343.
11. Bing Zhang, Xianfang Song, Yinghua Zhang, Dongmei Han and Changyuan Tang, “Hydrochemical characteristics and water quality assessment of surface water and groundwater in Songnen plain, Northeast China”, A Journal of the Inter. Water Association, Vol. 46(8), May 2012, pp. 2737-2748.
12. Sudhir Dahiya, Bupinder Singh, Shalini Gaur, V. K. Garg and H. S. Kushwaha, “Analysis of groundwater quality using fuzzy synthetic evaluation”, Journal of Hazardous Materials, Vol. 147, 2007, pp. 938–946.
13. Kunwar P. Singh, Amrita Malik, Vinod. K. Singh, Dinesh Mohan and Sarita Sinha, “Chemometric analysis of groundwater quality data of alluvial aquifer of Gangetic plain, North India”, Analytica Chimica Acta, Vol. 550, 2005, pp. 82–91.
14. C. Liu, K. Lin, and Y. Kuo, “Application of factor analysis in the assessment of groundwater quality in a blackfoot disease area in Taiwan,” vol. 313, no. 2, pp. 77–89, 2003.
15. Ting-Nien Wu, Chiu-Sheng Su, “Application of Principal Component Analysis and Clustering to Spatial Allocation of Groundwater Contamination”, 5th Int. National Conf. on Fuzzy Systems and Knowledge Discovery, 2008, pp. 236–240.
16. M. Kanevski, R. Parkin, A. Pozdnukhov, V. Timonin, M.Maignan, B.Yatsalo and S. Canu, “Environmental data mining and modelling based on machine learning algorithms and geostatistics”, 1st Int. Congress on Environ. Modelling and Software, Jun 2002, pp. 414–419.
17. A. Saberi Nasr, M. Rezaei and M. Dashti Barmaki, “Analysis of Groundwater Quality using Mamdani Fuzzy Inference System Analysis of Groundwater Quality using Mamdani Fuzzy Inference System (MFIS) in Yazd Province, Iran”, Int. Journal of Comp. Appl. Vol. 59(7), Dec 2012, pp. 0975-8887.
18. Keunje Yoo, Sudheer Kumar Shukla, Jae Joon Ahn, Kyungjoo Oh and Joonhong Park, “Decision tree-based data mining and rule induction for identifying hydrogeological parameters that influence groundwater pollution sensitivity”, Journal of Clean. Prod., vol. 122, 2016, pp. 277–286.
19. Harald Genther and Manfred Glesner, “Advanced data preprocessing using fuzzy clustering techniques”, Fuzzy Sets and Systems, Vol. 85, 1997, pp. 155–164.
20. Bagher Shirmohammadi, Mehdi Vafakhah, Vahid Moosavi and Alireza Moghaddamnia “Application of Several Data-Driven Techniques for Predicting Groundwater Level”, Water Resoource Manage, Vol. 27, 2013, pp. 419–432.
21. Jiawei Han, Micheline Kamber and Jian Pei, "Data Mining Concepts and Techniques", Morgan Kaufmann pulication, 3rd Edition, 2011, ISBN: 978-93-80931-91-3.

AUTHORS PROFILE

Dr. T. Velmurugan is working as an Associate Professor in the PG and Research Department of Computer Science and Applications, D. G. Vaishnav College, Chennai-600106, India. Also, he is the Head of the Department of Computer Science and Applications (UG). He holds a Ph.D. degree in Computer Science from the University of Madras. He elected as a Senate Member, University of Madras.

He has published more than 90 articles indexed in SCOPUS and SCI such as Applied Soft Computing, Journal of Computer Science and etc. He was an invited speaker of the 10th Int‟l Conference on Computational Intelligence and Software Engineering (CiSE 2018) held from January 5-7, 2018 in Bangkok, Thailand. He has guided more than 300 M.Phil. Research Scholars in the field of Computer Science. He guided 7 Ph.D. scholars and currently guiding 10Ph.D. scholars in the same field. He served as a nominated Senate Member in Middle East University, Dubai, UAE for a period of three years. He is a member in Board of studies for many autonomous institutions and Universities like Periyar University, Salem, India. He hosted a lot of programmes in Doordharsan television about recent topics in Information Technology field. He arranged and acted as an Organizing Secretary of International Conference on Computing and Intelligence Systems (ICCIS 2015). In addition, he was a resource person for various national workshops entitled “Scientific Research Article Writing and Journal Publications”. He is an editorial board member of many International Journals. He is also a reviewer in many peer reviewed journals like Elsevier, Springer, IOS Press Journals etc. Further, he is a visiting faculty for M.Phil. Courses for various universities throughout India.

R. Arunkumar is a full time Ph.D research scholar in the department of computer science, D.G. Vaishnav College, Chennai-600106, India. He is research scholar under the guidance of Dr. T. Velmurugan in University of Madras, Chennai-600005. He participated in International/National conferences and published one journal. Also, he took a “National Level Python Workshop” in Sridevi Arts and Science College, Ponneri-601204 with the cooperation of ResProLabs private Ltd. Participated in various International/National Level workshops and seminars.