J. Inst. Eng. India Ser. A
https://doi.org/10.1007/s40030-024-00796-3

CASE STUDY

Identifying the Sources of Contaminants in Groundwater Within the Alluvial Region Between two Perennial Rivers of Central Gujarat

Mukesh A. Modi[1] · N. J. Shrimali[1]

Received: 19 July 2023 / Accepted: 9 March 2024

Abstract This research is mainly focused on identifying the sources of groundwater contaminants present in the alluvial region between Mahi and Narmada rivers of central Gujarat using multivariate statistical analysis. Furthermore, Ground Truth Study (GTS) will be conducted to determine the source of these contaminants and ascertain whether they type from anthropogenic (caused by human activities) or geogenic sources. In the study area, shallow depth of groundwater varies from 9.54–97.45 feet below ground level (fbgl). Using Principal Component Analysis (PCA), Factor Score Analysis and Hierarchical Cluster Analysis (HCA), 50 aquifer openwells and shallow tubewells have been analyzed using PCA, Factor Score Analysis, and Hierarchical Cluster Analysis (HCA). The PCA revealed 3 significant components explaining 77.15% of total variance in which PC-1 (43.04%) described high positive loadings of (TDS, NO_3^-, Cl^-, SO_4^{2-}, Mg^{2+}) parameters. The PC-2 (16.11%) and PC-3 (18%) showed high positive loadings of (F^-, ALK) and (pH, TH, Ca^{2+}) parameters. The application of Hierarchical Cluster Analysis (HCA) revealed the presence of distinct clusters influenced by either anthropogenic or geogenic sources. Notably, Clusters 3 and 4 exhibited substantial anthropogenic influence, emphasizing the urgency for immediate remediation measures. Cluster 2 demonstrated a combination of geogenic and anthropogenic sources, underscoring the importance of continuous monitoring in this area. In contrast, Cluster 1 represented relatively non-polluted regions, serving as valuable reference points. The findings emphasize the necessity for targeted mitigation strategies to safeguard groundwater resources, particularly in highly contaminated Cluster 4. Additionally, the inclusion of a Ground Truth Study (GTS) further enhanced the accuracy and reliability of the findings, providing corroborating evidence to support the identification and characterization of contamination sources within these clusters.

Keywords Groundwater · Contamination · Source identification · Factor score · Multivariate statistical analysis · Ground truth study

Introduction

The quality of the water instead of the amount is a constraint in arid areas of the world. Hydrogeological sciences study and practical applications have focused on aquifer pollution since it could prevent using groundwater for uses such as residential and drinking [14], Fetter and Fetter [12]. Once the groundwater quality has been affected from undesired contaminant sources, it becomes a very difficult task to bring it back within its permissible standards. Over the past few decades, the multivariate statistical analysis approach has been widely used to address groundwater contamination source identification. Consequently, most researchers have relied on the results without adequate verification (Prasanna M. et al. [33], Nosrati and Eeckhaut [31], Kanchan and Ghosh [15], Machiwal and Jha [30], Loganathan and Ahmed [1].

The principal component analysis (PCA), factor score analysis (FSA) and hierarchical cluster analysis (HCA) are some of these statistical methods used extensively with GIS environment to identify pockets of anthropogenic sources

✉ Mukesh A. Modi
researchofwater@gmail.com

N. J. Shrimali
narendra.shrimali-ced@msubaroda.ac.in

[1] Civil Engg. Department, Faculty of Technology and Engg., M.S. University of Baroda, Gujarat, India

Published online: 12 April 2024

falling within one's area of study [37], Kanchan and Ghosh [15], Machiwal and Jha [30]. Standard textbooks on groundwater cover a variety of common tools and strategies for better understanding of groundwater chemistry with statistics and graphs (Freeze and Cherry, [13, 26],Sara and Gibbons, [35]. Various multivariate statistical and machine-learning techniques are used in recent research works for groundwater contamination (Chan and Huang, [2], [34], if the investigation is restricted to finding sources of anthropogenic or geogenic pollution.

The statistical methods that solve problems consisting of numerous input variables have found to be a useful tool in the assessment of spatiotemporal dissimilarities and simplification of groundwater physico-chemical datasets. As a result of such approaches, it is also possible to allocate groundwater pollution sources (natural or anthropogenic) and to conceptualize a supervisory system that is more efficient at meeting the needs of water resources [24], Guler and Thyne [19]. For example, by using HCA, PCA, and geochemical modeling approaches, Demirel and Guler [6] found human intervention influencing the existing geochemistry for the coastal aquifer of Mediterranean region, Mersin-Erdemli basin (Turkey).

PCA and HCA were used by Cloutier et al. [3] to determine the geochemical mechanisms governing the physicochemical nature of groundwater in the Paleozoic Basses-Laurentides (Canada). The PCs 1 and 2 were identified as "salinity" and "hardness" of groundwater based on loadings. In the humid Manukan Island's aquifer in Malaysia, Lin et al. [29] assessed temporal inconsistency and variables influencing geochemistry of shallow groundwater considering analysis of variance, PCA, HCA, and geostatistical approaches. To distinguish between the effects of natural and human causes toward the groundwater quality in South Korea, Kim et al. [27] used cluster analysis modeling.

Multivariate statistical analyzes, such as factor analysis (FA) and cluster analysis (CA), serve as invaluable tools for managing extensive datasets and numerous parameters. In the research conducted by Simeonov [36], these analytical methods proved instrumental in clarifying the underlying variable structure, facilitating the derivation of simplified groups. Yang [39] applied multivariate statistical analyzes to investigate the spatio-temporal patterns of water pollution in the Dianchi Lake basin. Guler [20] commenced a comprehensive examination of clustering techniques, including principal component analysis, in specific regions of Southwestern USA, delving into the advantages and disadvantages of each method. The utilization of appropriate statistical techniques emerges as a crucial approach for achieving meaningful generalizations in scientific inquiries.

Despite extensive research on identifying anthropogenic and natural contamination sources, none of the previous studies have adequately addressed the crucial aspect of determining the root causes of anthropogenic contamination. The lack of understanding regarding the fundamental factors driving human-caused contamination is a matter of utmost significance for decision makers. This research surpasses conventional methods by incorporating field conditions and attempting the primary causes of contamination. Its principal goal is not only to pinpoint contamination sources, but to precisely determine their specific origins through a robust Ground Truth Study (GTS). Through extensive data collection, the study seeks to establish the precise factors responsible for the disturbingly elevated levels of contamination observed. Geogenic contaminant sources pose tough challenges to prevention, this study emphasizes that anthropogenic sources can be effectively mitigated through rigorous monitoring. Furthermore, the GTS required to verify whether the identified sources were of geogenic origin or the result of human activities.

Study Area Description

The alluvial region between Mahi and Narmada rivers, which are flowing through the central part of Gujarat state, covers six blocks of Vadodara, three blocks of Panchmahal and three blocks of Bharuch districts. A total of 2750 km^2 are covered by the alluvial region, which is located between 72.51° and 73.64° east longitude and 21.78°–22.83° northern latitude (Fig. 1). A sub-humid climate prevails in the alluvial region of Gujarat, which is situated between the high rainfall areas of south Gujarat and the dry plains of north Gujarat. Approximately, 850 mm of rain is received annually in this semi-arid climate. From the middle of June through the middle of September, the southwest monsoon creates a humid environment.

The alluvial region of Vadodara taluka has observed a steady development, accommodating a range of industries such as oil refineries, petrochemical plants, fertilizer factories and heavy water projects. Additionally, Gujarat Industrial Development Corporation (GIDC) has played a crucial role in establishing and managing significant industries involved in the production of engineering and mechanical components, rubber-plastic goods, non-metallic mineral products and metal products ([7], [8], [9] DIPS-2016). Groundwater can be found in both the confined and unconfined zones in present research region. The unconfined aquifers are formed by weathered zones, shallow depth jointed and fractured rocks, and saturated zones of unconsolidated shallow alluvium. Interflow zones of basalts, inter-trapping beds, deep-seated fracture zones, shear zones in basalts, granites and gneisses, as well as multi-layered aquifers under impermeable clay horizons in the production of alluvium, generate semi-restricted to limited conditions ([16], [17], [18] CGWB-2014).

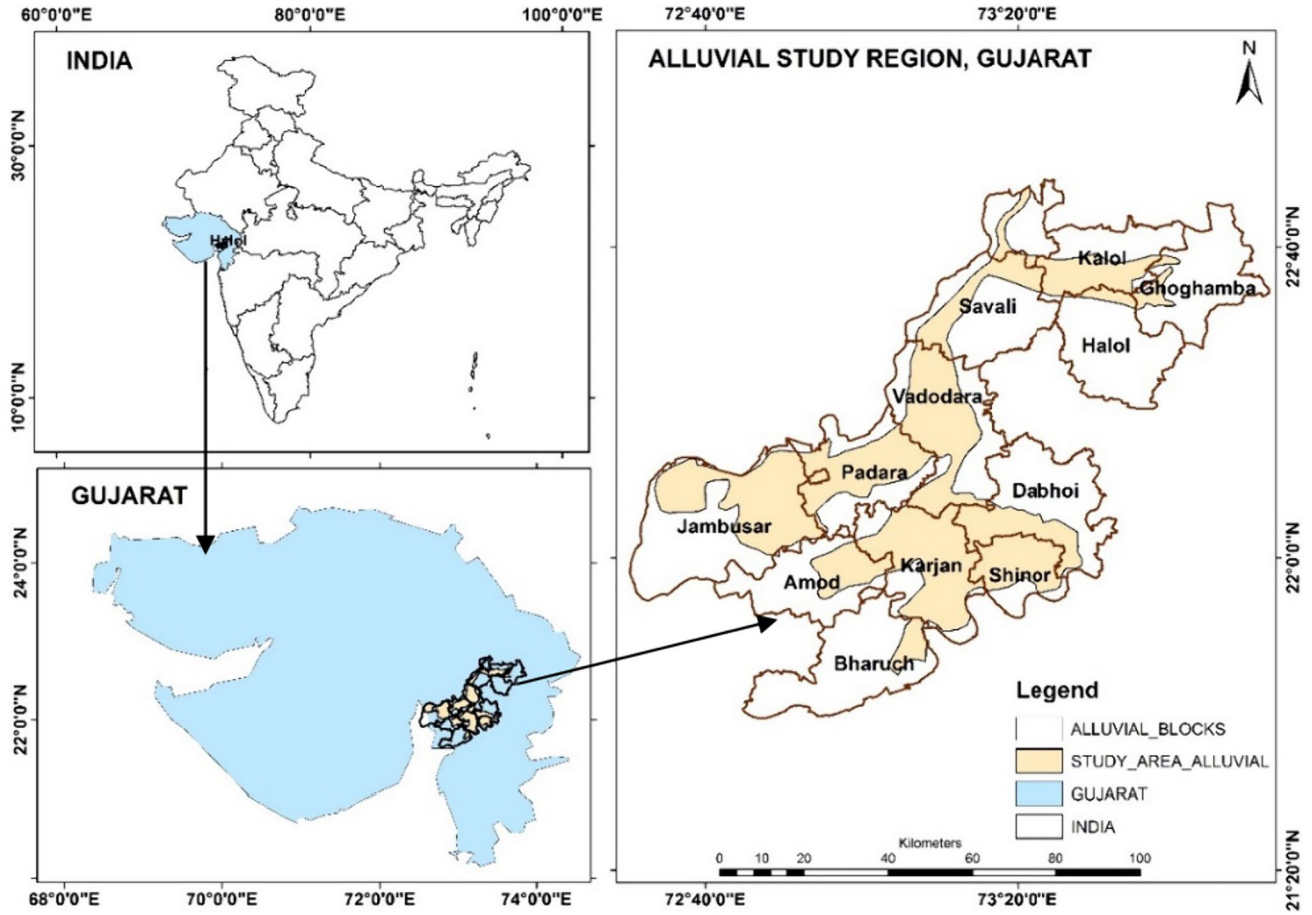

Fig. 1 Alluvial region within Mahi and Narmada Rivers of Gujarat

Materials and Methods

Figure 2 illustrates the step-wise methodology employed for identifying groundwater contaminant sources. Initially, 50 no. of wells groundwater quality data obtained from the MOJS-DDWAS portal of May 2018 were subjected to multivariate statistical analysis. Common groundwater quality parameters (Potential of Hydrogen (pH), Total Dissolved Solid (TDS), Nitrate (NO_3^-), Fluoride (F^-), Chloride (Cl^-), sulfate (SO_4^{2-}), Calcium (Ca^{2+}), Magnesium (Mg^{2+}), Total Hardness (TH) and Alkalinity (ALK)) were available at each well location and imported into SPSS software. The parameter values were then transformed into *Z*-scores. Next, a principal component analysis (PCA) was conducted on these *Z*-score values, resulting in the extraction of principal components (PCs) $1-n$ and the formation of factor scores (FS) for all well locations. Further, the FS values were input into a GIS (Geographic Information System) environment to obtain their spatial distribution. The IDW (Inverse Distance Weightage) interpolation method was employed to generate the spatial distribution of FS values. A composite map considering all PCs was also obtained using overlay analysis in GIS. The study conducted a Hierarchical Cluster Analysis (HCA) on *Z*-score values derived from groundwater quality parameters at various well locations, leading to the creation of a dendrogram clarifying the distribution of wells within distinct clusters. To verify the outcomes of both Principal Component Analysis (PCA) and Hierarchical Cluster Analysis (HCA), a Ground Truth Study (GTS) was undertaken. This study aims to ascertain whether the identified sources of contamination are anthropogenic, geogenic, or a combination of both. The GTS involves the collection of data on site characteristics and the sampling of groundwater quality for a comprehensive verification process.

Multivariate Statistical Analysis

To simplify groundwater quality data with a high number of geochemical characteristics for a better knowledge of the local hydrogeology, multivariate statistical analysis is performed. Such method provides strong bases for the problem of groundwater contamination being influenced by either geogenic or anthropogenic activities when combined with spatial and temporal distribution. This

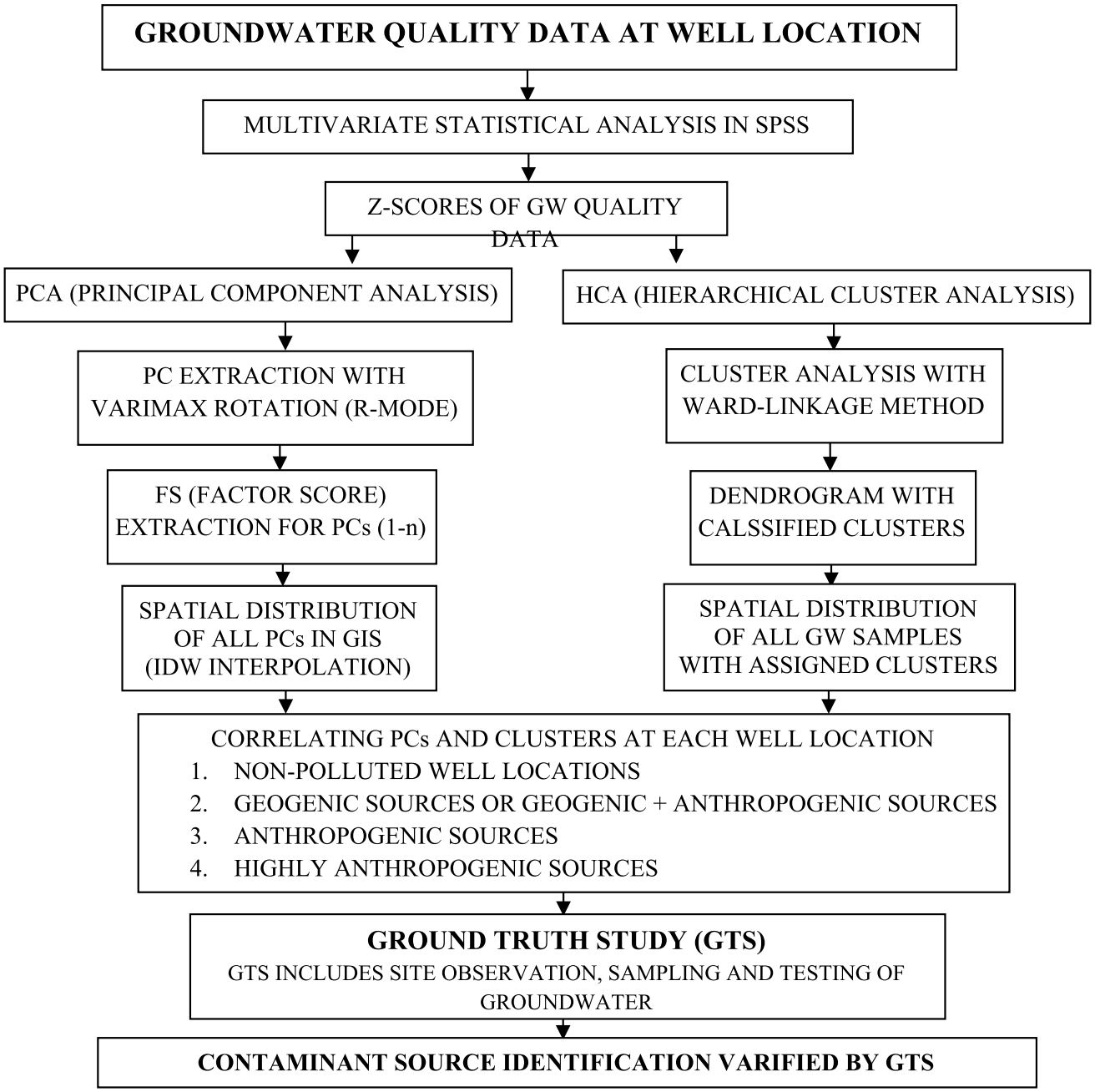

Fig. 2 Methodology-flow chart of contaminant source identification

technique focuses on pinpointing the factors that contribute to groundwater system governance and highlights the importance of employing tools to address challenges associated with groundwater contamination. Principal Components (PCs) and Factor Scores (FS) were computed using IBM SPSS software for the analysis.

Principal Component Analysis and Factor Score

A vital statistical approach for examining the chemistry of groundwater is principal component analysis (PCA) [10]. By determining correlated concentrations linked to particular pollution sources or processes, the aim is to reduce

multidimensional water quality datasets to a smaller collection of interpretable "principal components" (PCs). The two phases of PCA are PC extraction and data standardization. PCA was performed on the groundwater quality data (pH, TDS, NO_3^-, F^-, Cl^-, Mg^{2+}, TH, ALK, SO_4^{2-}, Ca^{2+}) in this study during the 2018 pre-monsoon season using IBM SPSS Statistics. The data on the observed groundwater quality were transformed using a *z*-scale before PCA. To begin PCA, the data correlation matrix is restructured to better understand the structure of the underlying system. PCs are linear combinations of the initial variables and are new variables produced from the original dataset. Starting with the correlation matrix, eigenvalues and eigenvectors are extracted while less important ones are removed [5]. Then, the eigenvectors are converted into PCs. The variation is best explained by the first PC, which accounts for the higher part of the variance. The Kaiser Normalization Criterion [25] is used to determine the number of retained PCs. PCs that may be reliably understood and have eigenvalues larger than 1 are accepted for further study [21]. The equation for calculating the *Z*-score for an individual data point *X* in a variable with mean μ and standard deviation σ is given by equation no (1):

$$Z = \frac{(X - \mu)}{\sigma} \tag{1}$$

Where *Z* is the *Z*-score, *X* is the raw score, μ is the mean, and σ is the standard deviation.

Factor analysis is a multivariate statistical technique employed to separate relationships among variables. The primary objective of this analysis is to diminish the dimensionality of the data while retaining a set of interrelated variables without sacrificing pertinent information (Farnham, [11]. Helena [22] previously utilized this statistical approach to interpret intricate, interrelated processes governing general water chemistry. In the current study, factor analysis was applied to extract factors using the "Kaiser Criterion," where eigenvalues exceeding unity (1) were considered [4]. The screen test, involving a descending order of eigenvalues relative to factors, was employed. A break in the scree plot indicated the number of factors to be taken into account. To ensure maximum variability, "varimax rotation" was implemented. This rotation method allowed for a more efficient analysis of the interrelationships among variables, as it grouped the numerous variables into a reduced set. The factor score for an observation on a particular principal component is computed using the coefficients obtained during the PCA. Let *Fi* be the factor score for the ith observation on a given principal component, and $X_1, X_2, \ldots, X_p$ be the standardized variables (variables with Z-scores). The equation is:

$$Fi = \sum_{J=1}^{P} (aij.Zj) \tag{2}$$

Here a_{ij} is the loading of the *j*th variable on the *i*th principal component, and Z_j is the *Z*-score of the *j*th variable for the *i*th observation.

Hierarchical Cluster Analysis (HCA)

Numerous groundwater sample locations were grouped using the Hierarchical Cluster Analysis (HCA) method based on qualitative similarities in the data. This separation aids in the identification of contamination sources by highlighting groups of wells with similar physicochemical characteristics. Hierarchical Cluster Analysis (HCA) was utilized to classify groundwater sample sites into groups or classes that had similarities in terms of groundwater quality but were distinct from one another. HCA is an unsupervised method that shows a dataset's underlying behavior without making assumptions. In order to identify the origins of groundwater pollution, it attempts to categorize components based on their similarities. Cluster Analysis can be classified into two major categories: hierarchical and non-hierarchical. Hierarchical Cluster Analysis (HCA) involves clustering similar pairs sequentially and combining them in progressively larger clusters. Clustering results are presented in the form of a dendrogram, which provides a visual overview of the process. The *Z*-score dataset was used in this study for HCA, which minimizes sums of squares by evaluating distances between clusters [20], Kruskal and Landwehr, [28], Otto, [32], [38].

Results and Discussion

The data on groundwater levels were collected from India-WRIS portal for the year 2018-pre monsoon season. The shallow depth of groundwater varies from 9.54 to 97.45 fbgl in the alluvial region (Fig. 3). The area shown with blue color highlights shallow depth in comparison of the area shown with red color.

Principal Component Analysis (PCA) examining intercorrelations among various groundwater quality parameters provides valuable insights into their relationships. The intercorrelation results (Table 1 and 3) demonstrate significant associations between several parameters. For example, the correlation between nitrate (NO_3^-) and total dissolved solids (TDS) is good correlation at 0.70, suggesting a notable relationship between these two variables. Similarly, chloride (Cl^-) exhibits a strong correlation of 0.83 with TDS, indicating a strong connection between chloride levels and overall dissolved solids.

Moreover, sulfate (SO_4^{2-}) shows a good correlation of 0.76 with TDS, further emphasizing its relationship with the dissolved solids in the groundwater. The strong correlation between Mg and TH can be scientifically explained

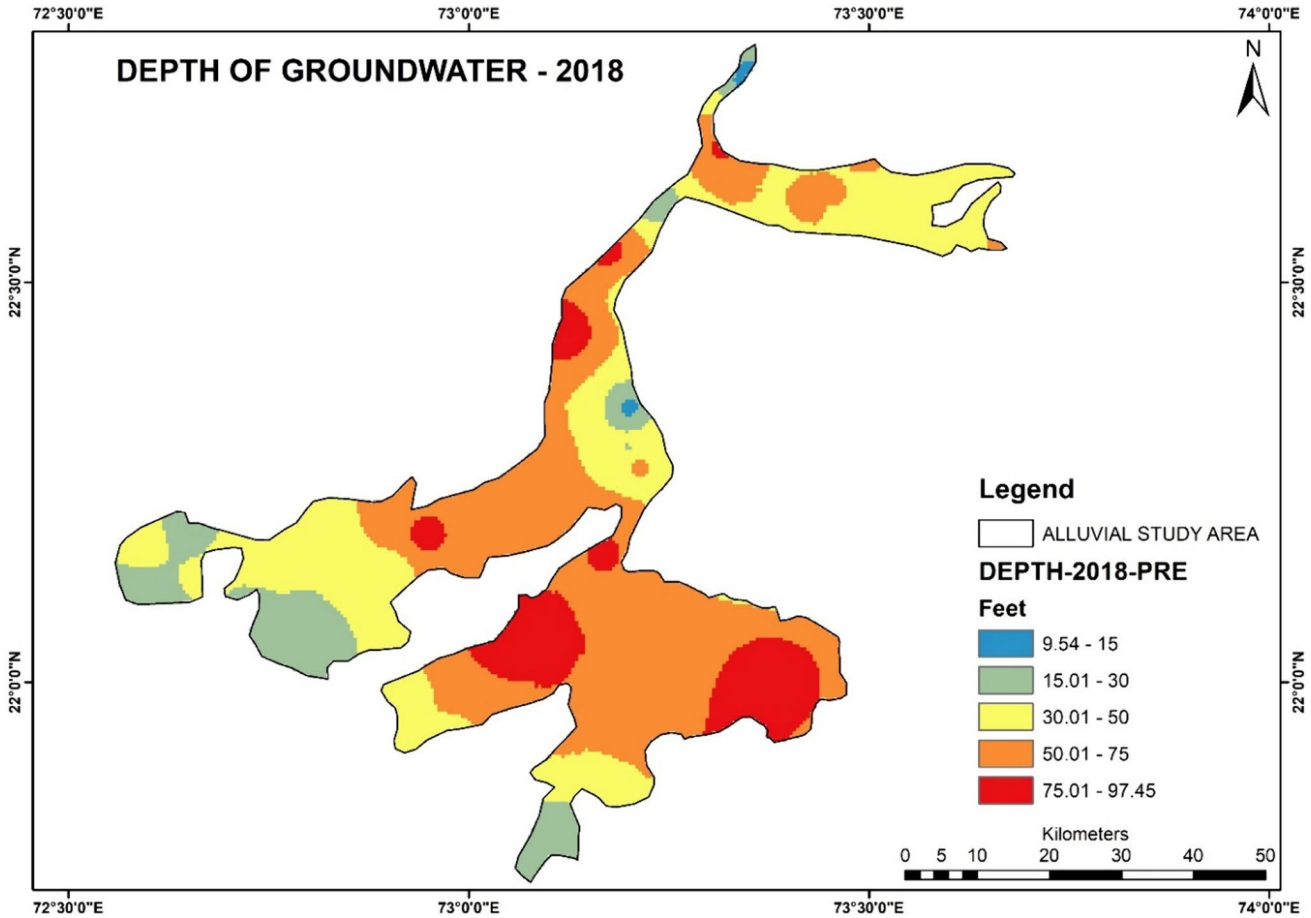

Fig. 3 Shallow depth of groundwater, 2018-pre monsoon

Table 1 Groundwater quality parameters-spearman correlation

Parameters	pH	TDS	NO_3^-	F^-	Cl^-	SO_4^{2-}	Ca^{2+}	Mg^{2+}	ALK	TH
pH	1.00									
TDS	−0.04	1.00								
NO_3^-	−0.24	**0.70**	1.00							
F^-	0.19	0.35	0.15	1.00						
Cl^-	0.14	**0.83**	**0.50**	0.20	1.00					
SO_4^{2-}	−0.31	**0.76**	**0.52**	0.17	**0.52**	1.00				
Ca^{2+}	−0.05	0.41	**0.55**	−0.18	0.41	0.30	1.00			
Mg^{2+}	−0.32	**0.78**	**0.81**	0.09	**0.60**	**0.72**	**0.63**	1.00		
ALK	0.06	**0.51**	0.19	**0.64**	0.24	0.42	−0.23	0.20	1.00	
TH	−0.25	**0.70**	**0.80**	−0.04	**0.59**	**0.59**	**0.83**	**0.94**	0.03	1.00

The bold values indicate good as well as high correlation

by their interdependence in groundwater chemistry. Magnesium (Mg^{2+}) is a common cation found in groundwater, often associated with the dissolution of magnesium-bearing minerals in the aquifer. Total Hardness (TH) is a measure of the combined concentrations of calcium and magnesium ions in water. Given their common source and shared geological origins, Mg^{2+} and TH exhibit a high positive correlation at 0.94. The dissolution of minerals containing both magnesium (Mg^{2+}) and calcium (Ca^{2+}) contributes to the simultaneous increase in their concentrations in groundwater, thereby resulting in a strong correlation.

Table 2 Principal component analysis (MOJS-DDWAS-2018-Pre)

Total variance explained									
PC	Initial eigenvalues			Extraction sums squared loadings			Rotation sums squared loadings		
	Total	Variance%	Cumulative%	Total	Variance%	Cumulative%	Total	% Variance	Cumulative %
1	4.304	43.04	43.04	4.304	43.04	43.04	4.963	49.634	49.634
2	1.611	16.11	59.15	1.611	16.11	59.15	1.383	13.830	78.691
3	1.8	18	77.15	1.8	18	77.15	1.523	15.226	64.860
4	0.827	8.27	85.42						
5	0.512	5.12	90.54						
6	0.468	4.68	95.22						
7	0.348	3.48	98.7						
8	0.063	0.63	99.33						
9	0.06	0.6	100						
10	0.00002	0.0002	100						

Principal Components and Factor Scores (PCA & FS)

Table 2 shows the initial eigenvalues signifying the variances attributed to factors influencing groundwater quality. Table 2 displays the results of the statistical analysis, incorporating 10 principal components, corresponding to the inclusion of 10 water quality parameters in the analysis. Specifically, the consideration was given to the first three Principal Components, as their eigenvalues surpassed the threshold of 1. The Extraction Sums of Squared Loadings, based on eigenvalues greater than 1, indicate that these factors collectively account for a larger proportion of the total variation in the data compared to individual parameters. Conversely, factors with eigenvalues less than 1 explain a smaller proportion of the total variation and are not considered further in the Rotation Sums of Squared Loadings, resulting in a reduction in the number of variables.

Three principal components with varimax rotation having eigenvalues greater than 1 which explained 77.15% of cumulative variance (Table 2) were extracted in SPSS environment. The factor loadings for PC-1 (TDS, NO_3^-, Cl^-, SO_4^{2-}, Mg^{2+}) showed 43.04% of total variance from the dataset. The PC-2 highlighted 16.11% of variance with high positive loadings on (F^-, ALK) parameters. The PC-3 observed 18% variability containing (pH, Ca^{2+}, TH) parameters. The factor loadings for PCs are described in Table 3. After the PC extraction, factor scores at each well location were used for spatial distribution of individual PCs as well as a composite PC in GIS environment considering IDW interpolation method (Ghosh and Kanchan, [15], Loganathan and Ahmed [1]. These maps highlighted the areas contaminated from anthropogenic sources having FS > 2 as shown in the Figs. 4, 5, 6, 7.

The PC-1 map (Fig. 4), associated with high positive loadings of TDS, NO_3^-, Cl^-, SO_4^{2-}, and Mg^{2+} groundwater quality parameters which can be attributed as anthropogenic factor occurring from sources such as extensive use of fertilizers in agriculture, wastes from animal husbandry and improper arrangement of septic tanks and soak pits. The agricultural activities of rural area, petro-chemicals based industries and shallow depth of groundwater (around 35 fbgl) are the prominent causes for high factor scores of PC-1 in the northern parts of the alluvial region. The wells 24 and 50 showing higher factor scores are located near a small river stream and an industrial pocket respectively. The dumping of ill-treated industrial waste water into small streams that eventually percolates toward groundwater is another significant factor which explains high loadings of Mg and TDS parameters. The gradual decrease in factors

Table 3 Factor loadings for PCs after Varimax Rotation

Rotated component matrix			
GWQ parameters	Component		
	PC-1	PC-2	PC-3
TDS	0.919		
NO_3^-	0.783		
Cl^-	0.705		
SO_4^{2-}	0.943		
Mg^{2+}	0.954		
TH			0.511
ALK		0.818	
F^-		0.793	
pH			0.350
Ca^{2+}			0.939

Extraction Method: Principal Component Analysis, Rotation Method: Varimax with Kaiser Normalization. Rotation converged in 4 iterations

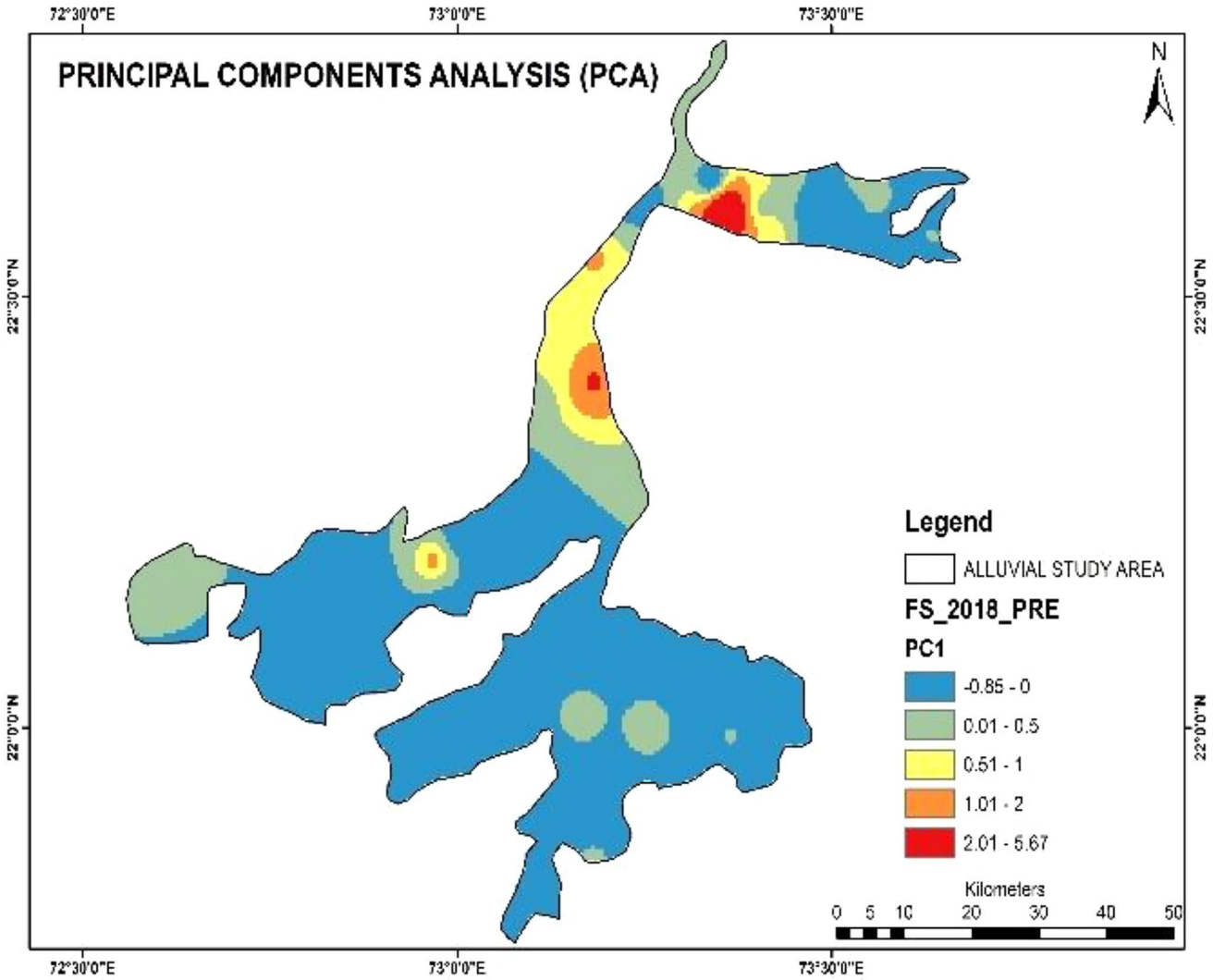

Fig. 4 PC1-2018-pre (TDS, NO_3^-, Cl^-, SO_4^{2-}, Mg^{2+})

scores is in alignment with the movement of groundwater from North-East to South-West direction.

In Fig. 5, the PC-2 map shows significant positive loadings for the F^- and ALK groundwater quality parameters. Higher factor scores are observed in the northern and central parts of the alluvial region, where the groundwater depth is relatively deeper (approximately 70 fbgl). The elevated factor scores in these areas can be attributed to several factors, including excessive groundwater exploitation, extensive use of phosphate-based fertilizers and the discharge of industrial wastes. Specifically, in the northern zone, well 26 and in the central zone, wells 42 and 43 are situated near small river stretches, while well 44 is located in close proximity to small-scale industries within the central zone.

The PC-3 map (Fig. 6) shows high positive loadings of both TH, Ca^{2+} and pH parameters observed mostly in northern parts of the alluvial region where the depth of groundwater is comparatively shallow (around 40 fbgl). The higher factor scores from PC-2 also pointed toward the anthropogenic category of contamination source such as an industrial area that causes mixing of waste water into small drains which infiltrates into the groundwater. The well 8 located in the northern zone is falling in such an industrial pocket where majority of the medium scale materials manufacturing units exist.

The composite map (Fig. 7) obtained from overlay analysis in GIS environment helped in identification of the contamination sources. The majority of the contaminated pockets were identified in the northern (wells 6, 8, 15 and 24) and central parts (43 and 44) of the alluvial region showing factor score greater than 2 being considered as the anthropogenic contamination sources. The conventional agriculture practices, rapid growth of minerals,

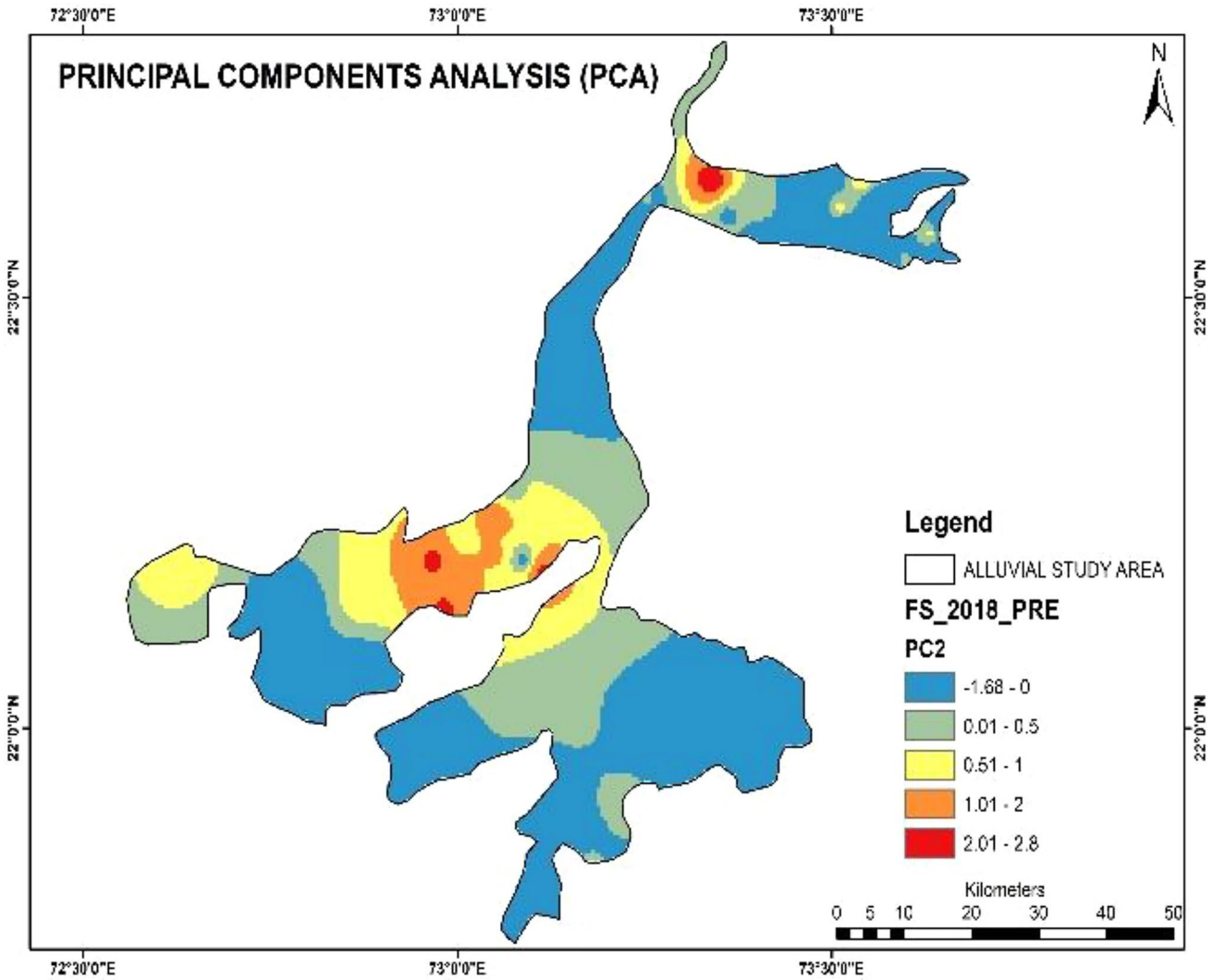

Fig. 5 PC2-2018-pre (F^-, ALK)

disposal of industrial effluents as well as shallow depth of groundwater are the key anthropogenic sources responsible for contamination. The overlay analysis also highlighted a few pockets in the central alluvial region having factor scores greater than 1 which required continuous monitoring. The non polluted wells were majorly observed in the southern and western parts of the alluvial region with comparatively high depth of groundwater and the negative factor scores toward perennial Mahi and Narmada rivers. Table 4 displays the potential origin of contamination that requires additional verification through a Ground Truth Study (GTS).

Hierarchical Cluster Analysis (HCA)

The Fig. 8 below displays the cluster distribution of well sites as determined by HCA. In addition to being used in PCA, the *Z*-scores from the groundwater quality dataset were also used in hierarchical cluster analysis (Machiwal and Jha, [30]. Squared Euclidean distances were used as a similarity metric in the analysis along with the ward's linkage approach (Loganathan and Ahmed, [1]. The dataset was divided into 4 significant clusters using the hierarchical cluster analysis. Understanding the distribution of well locations within each cluster (Table 5) and the average values of each quality parameter (Table 6) was made simpler with the aid of the Dendrogram generated from the HCA and displayed in Fig. 9.

Dendrogram is created by iteratively merging or splitting clusters in a hierarchical manner. The vertical lines in the dendrogram represent the individual well number, and the horizontal lines indicate the merging or splitting of clusters at different levels of similarity. The height at which two

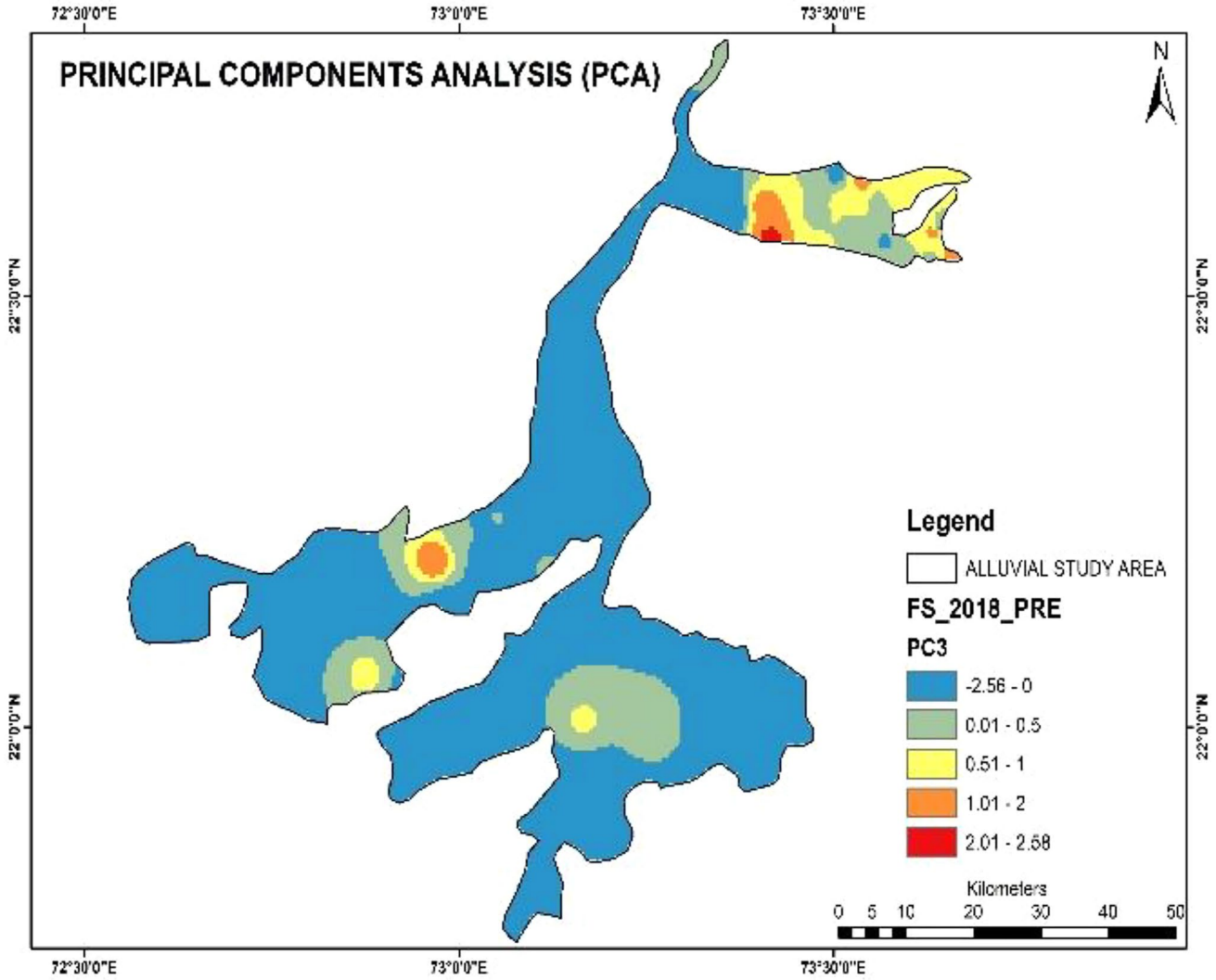

Fig. 6 PC3-2018-pre (pH, TH, Ca^{2+})

branches merge in the dendrogram corresponds to the level of similarity or distance at which the clustering occurs.

The cluster 1 majorly located in southern parts showed 64% of the well samples with moderate mean values of pH (7.83), Ca^{2+} (73 mg/L) and TH (418 mg/L) groundwater quality parameters. The Fig. 8 highlighted cluster 2 containing 18% of well samples in the central and western parts with high ALK (644 mg/L) and F^{-} (1.19 mg/L) parameters. The Cluster 3, which included 12% of well locations with high mean values of NO_3^{-}(98 mg/L), Mg^{2+} (122 mg/L) and TH (912 mg/L) explaining high anthropogenic sources in shallow aquifers of alluvial region. The well number 24 came under cluster-4 with the highest amount of NO_3^{-} (284 mg/L), TDS (4120 mg/L), SO_4^{2-} (473 mg/L), Mg^{2+} (424 mg/L) and TH (1810 mg/L) pointing toward severe contamination from anthropogenic sources.

After conducting Principal Components Analysis (PCA), Factor Score Analysis (FSA) and Hierarchical Cluster Analysis (HCA) using secondary data of 50 well locations, the wells were classified into three Principal Components (PCs) and four Clusters.

In case, the NO_3^{-} concentration more than 45 mg/l observed in groundwater at particular well, then there are various possible sources either rural sanitation or industrial effluents or usage of excess fertilizers in farms. In this situation, Ground Truth Study (GTS) confirm Specific sources along with specific course of action. A GTS was afterward conducted exclusively for those wells where at least one or more physico-chemical parameters exceeded the permissible limits. While the analyzes of PCA, FSA and HCA provided insights into the classification and

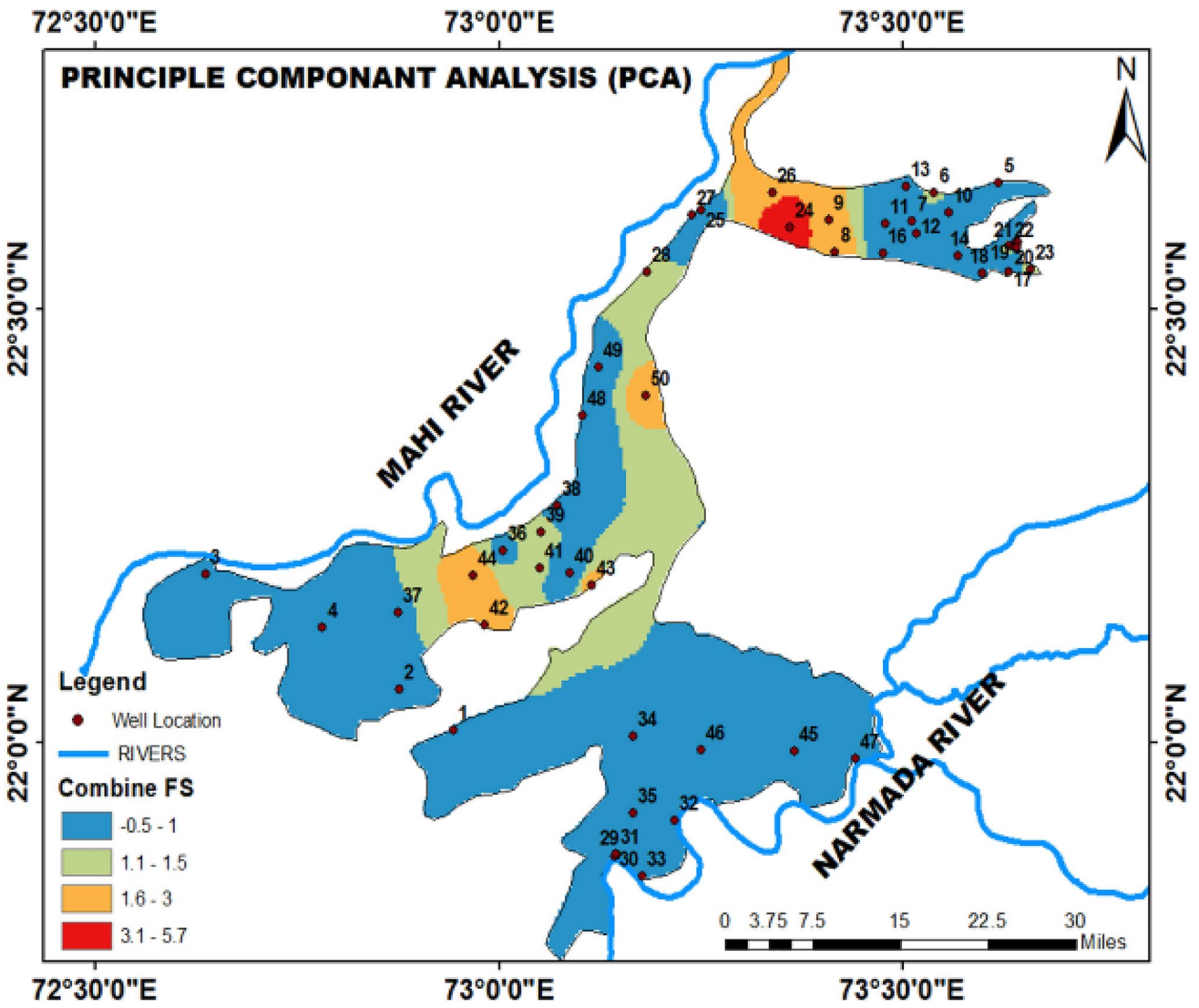

Fig. 7 Composite PC (2018-pre)

grouping of well locations, they were unable to pinpoint the actual causes of contamination, be it anthropogenic, geogenic or a combination of both. However, through the incorporation of location-specific characteristics, integral to the GTS, we could discern and elucidate the precise causes of contamination.

Table 4 Source identification by PCA

Sr. No	Factor score range	Remarks
1	< 1	Non polluted
2	1–1.5	Geogenic + anthropogenic or geogenic
3	1.5–2	Anthropogenic
4	> 2.0	Highly anthropogenic

Ground Truth Study (GTS)

After identification of the contamination sources based on secondary data, it is at most necessary to verify present status of contaminants in groundwater by Ground Truth Study. GTS includes conducting field investigations, sampling of groundwater and analyzing the collected samples to gather precise and reliable information about the contamination sources and their characteristics. The well mentioned in Table 7 was subjected to testing, which confirmed that the dominant parameter for each well exceeded its permissible limit as per IS 10500: 2012 [23].

Based on the analysis of FS (Factor Score) and HCA (Hierarchical Cluster Analysis), total 4 clusters have been identified. Out of the 50 well locations studied, only 15 wells exhibited factor scores indicating a significant presence of

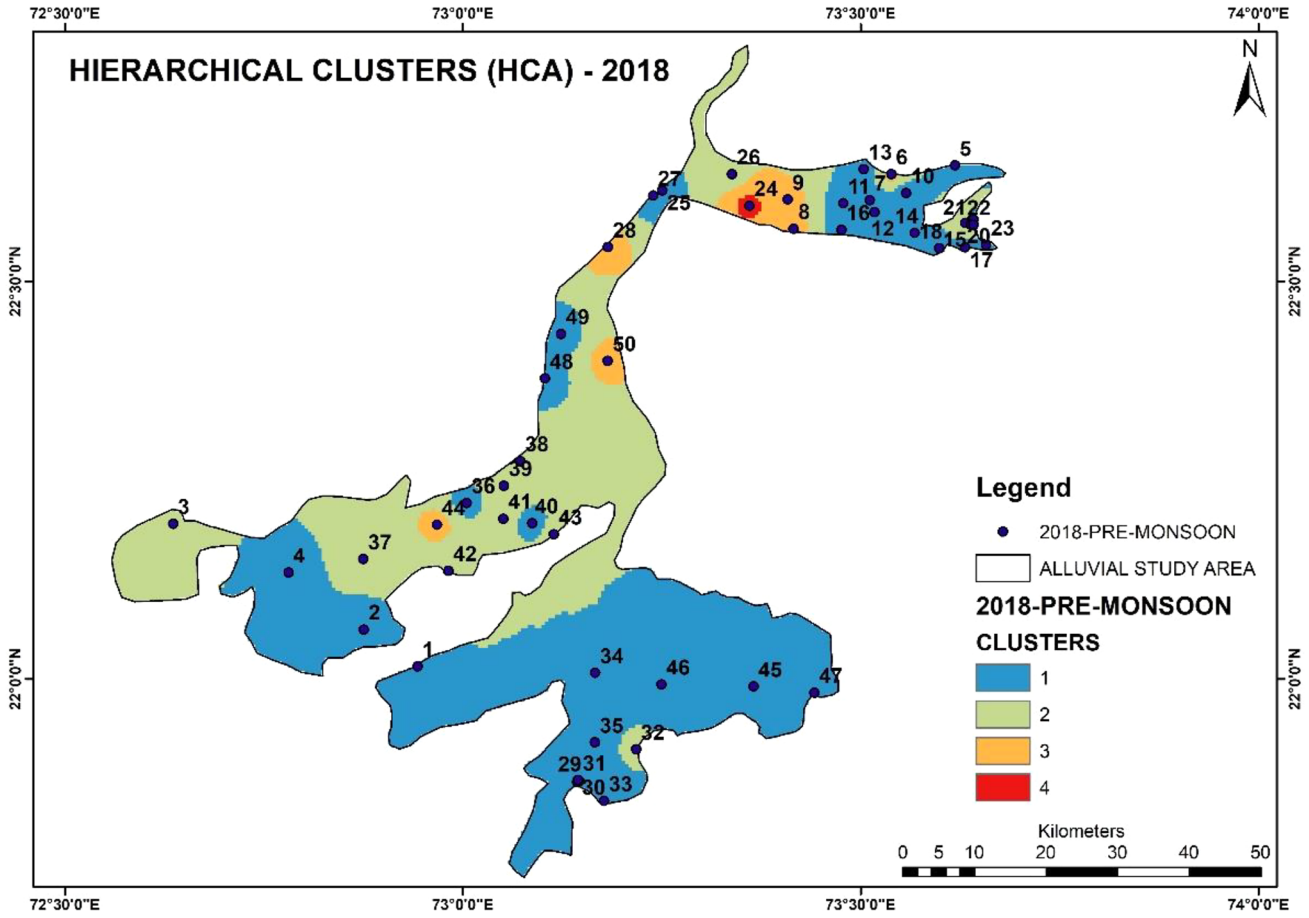

Fig. 8 Spatial distribution of Clusters from HCA

Table 5 Cluster classification

Clusters	Well locations
Cluster-1	1, 2, 4, 5, 7, 10, 11, 12, 13, 14, 16, 17, 18, 19, 20, 22, 23, 25, 27, 29, 30, 31, 33, 34, 35, 36, 40, 45, 46, 47, 48, 49
Cluster-2	3, 26, 32, 37, 38, 39, 41, 42, 43
Cluster-3	6, 8, 9, 15, 21, 28, 44, 50
Cluster-4	24

anthropogenic activities as well as geogenic in their vicinity. Moreover, a Ground Truth Study (GTS) was conducted to verify actual cause of type of contamination source.

Conclusion

Present study area includes high industrial zone of central Gujarat, where due considerations of hazardous activities of disposal of harmful effluents, high usage of fertilizer for agriculture, animal husbandry activities in near Vadodara district are responsible for groundwater contamination. This research stands out from conventional methods by integrating field conditions and delving into the root causes of contamination. This research is not limited to merely identifying contamination sources but it includes precisely determine the actual causes of sources of contamination through Ground Truth Study for decision makers to prescribe appropriate groundwater management strategy.

Table 6 Average of GWQ parameters for each cluster

Cluster	pH	TDS	NO_3^-	F^-	Cl^-	SO_4^{2-}	Ca^{2+}	Mg^{2+}	TH	ALK
1	7.83	859	39	0.63	191	29	73	57	418	381
2	7.88	1076	19	1.19	208	29	30	36	226	644
3	7.70	1700	98	0.87	362	78	161	122	912	543
4	7.00	4120	284	0.54	848	473	24	424	1810	602

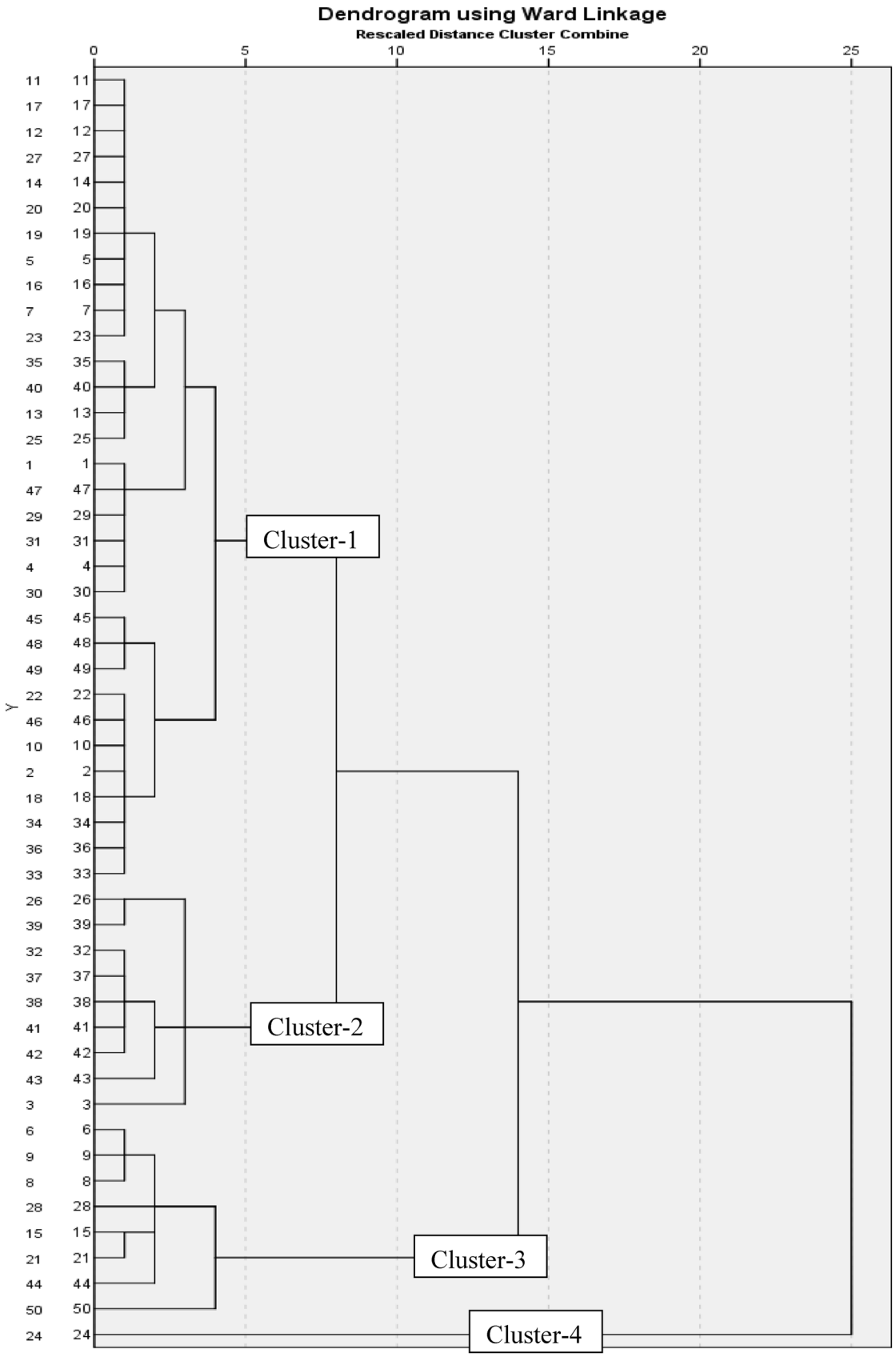

Fig. 9 Dendrogram for 2018-Pre-Monsoon data (MOJS-DDWAS)

Table 7 Actual cause of type of contamination source by GTS

Cluster No.	Well No.	Well Name	pH (6.5–8.5)	TDS (500–2000)	NO_3^- (45)	F^- (1–1.5)	Cl^- (250–1000)	SO_4^{2-} (200–400)	Ca^{2+} (75–200)	Mg^{2+} (30–100)	TH (200–600)	ALK (200–600)	Dominant Parameter	PCA–FS	Type of contaminant source	Actual cause of type of contamination source verified by GTS
1	23	Savapura, Ghoghamba, Panchmahal	7.82	1102	83	0.7	128	32	154	92	768	228	TH (768)	PC3, 1.55	Anthropogenic	Excess use of fertilizer (Urea, DAP) in the agricultural activity
2	26	Dipapura, Savli, Vadodara	8.01	1020	34	2.15	112	45	14	24	136	704	ALK (704)	PC2, 2.81	Highly Anthropogenic	Rural sewage disposal to effluent stream
	39	Dabhasa, Padra, Vadodara	7.98	886	45	1.98	176	38	61	60	404	446	F (1.98)	PC2, 1.53	Anthropogenic	Pharmaceuticals effluent disposal to nearby land
	41	Pipli, Padra, Vadodara	7.74	1102	31	0.88	172	12	16	15	104	700	ALK (700)	PC2, 1.1	Geogenic + Anthropogenic	Rural sewage disposal to village pond and alkaline soil
	42	Sadra, Padra, Vadodara	8.02	1403	23	1.18	272	13	20	24	152	825	ALK (825)	PC2, 2.28	Highly Anthropogenic	Rural sewage disposal to village pond and animal husbandry
	43	Shihor, Padra, Vadodara	8.41	1618	9	0.82	568	36	57	57	380	772	ALK (772)	PC2, 2.15	Highly Anthropogenic	Rural sewage disposal to village pond
3	6	Alali, Kalol, Panchmahal	7.8	1354	129	1.46	224	35	166	99	828	568	TH (828)	PC3, 1.6	Anthropogenic	Excess use of manure and Urea for crops and animal husbandry by farmers
	8	Bakrol, Kalol, Panchmahal	7.66	1809	104	0.78	232	112	239	144	1196	428	TH (1196)	PC3, 2.6	Highly Anthropogenic	Metal industrial effluent along with agricultural activity

Table 7 (continued)

Cluster No.	Well No.	Well Name	pH (6.5–8.5)	TDS (500–2000)	NO_3^- (45)	F^- (1–1.5)	Cl^- (250–1000)	SO_4^{2-} (200–400)	Ca^{2+} (75–200)	Mg^{2+} (30–100)	TH (200–600)	ALK (200–600)	Dominant Parameter	PCA–FS	Type of contaminant source	Actual cause of type of contamination source verified by GTS
3	9	Bedhiya, Kalol, Panchmahal	7.88	1144	132	0.62	76	41	166	99	832	628	TH (832)	PC3, 1.53	Anthropogenic	Improper rural sewage disposal and excessive use of urea and DAP fertilizers
	15	Govindpuri, Halol, Panchmahal	7.9	1478	28	1.3	448	31	144	86	720	444	TH (720)	PC3, 1.41	Geogenic	Minerals deposit from effluent stream (Karad River)
	21	Ranipura (Dam), Ghoghamba, Panchmahal	7.78	1778	29	0.72	508	41	184	110	920	556	TH (920)	PC3, 1.83	Anthropogenic	Agricultural and domestic waste disposal to Karad River
	28	Wankaner, Savli, Vadodara	7.3	1310	168	0.34	135	40	100	140	828	487	NO_3 (168)	PC3, 1.09	Geogenic + Anthropogenic	Rural sewage disposal to village pond
	44	Vishrampura, Padra, Vadodara	7.98	2236	102	1.36	440	114	182	169	1160	828	ALK (828)	PC3, 2.27	Highly Anthropogenic	Excessive use of fertilizers and pesticides. Improper rural sewage disposal to land
	50	Sisva, Vadodara	7.31	2494	95	0.38	832	210	109	131	812	404	NO_3 (95)	PC1, 2.14	Highly Anthropogenic	Effluent of petro chemical industrial waste along with excess use of fertilizer

Table 7 (continued)

Cluster No.	Well No.	Well Name	pH (6.5–8.5)	TDS (500–2000)	NO_3^- (45)	F^- (1–1.5)	Cl^- (250–1000)	SO_4^{2-} (200–400)	Ca^{2+} (75–200)	Mg^{2+} (30–100)	TH (200–600)	ALK (200–600)	Dominant Parameter	PCA–FS	Type of contaminant source	Actual cause of type of contamination source verified by GTS
4	24	Dhantej, Savli, Vadodara	7	4120	284	0.54	848	473	24	424	1810	602	NO_3 (284) TDS (4120)	PC1, 5.69	Severely Anthropogenic	Excess use of fertilizer (DAP, Urea), along with animal husbandry. Disposal of domestic waste nearby village pond

*All the unit of water quality parameters are in mg/l except for pH. Permissible limits of water quality parameters have been mentioned in title row as per IS 10500:2012

1. The utilization of multivariate statistical analysis methods, including Principal Component Analysis (PCA), Factor Analysis (FSA) and Hierarchical Cluster Analysis (HCA), on a substantial groundwater quality dataset has proven effective in reducing data dimensionality, interpreting variable structures, and identifying inherent clusters. Following this, the integration of Ground Truth Study (GTS) enhances the results by precisely locating contamination sources within the alluvial region.
2. In this study, the initial step involves narrowing down the contaminated zones and identifying a specific number of wells exhibiting high and severe contamination. Subsequently, the GTS tool has been strategically employed to verify specific sources of contamination in those prioritized areas.
3. The PC-1 (TDS, NO_3^-, Cl^-, SO_4^{2-}, Mg^{2+}), PC-2 (F^-, ALK) and PC-3 (pH, TH, Ca^{2+}) all three components pointed toward the same pockets i.e., northern and central parts which have been highly contaminated with anthropogenic sources thus a composite PC map prepared in GIS environment using overlay analysis better visualized the highly contaminated zone.
4. Based on individual PC maps, the two main perennial rivers Mahi and Narmada flowing in the western and eastern parts respectively of the alluvial region and shallow depth of groundwater toward these rivers is influencing the gradual decrease in factor scores (Fig. 7).
5. The well locations showing factor scores higher than 1 came under cluster 3 and 4 indicating these clusters to be affected from anthropogenic sources. The cluster 2 also contained wells with positive factor scores greater than 1 which pointed toward continuous monitoring of such clusters which are being affected with geogenic and anthropogenic sources combined.
6. A large number of wells fell under cluster 1 having negative factor scores being considered as non-polluted were observed in the southern and western parts of the alluvial region.
7. Cluster 2 primarily encompasses the central region of the study area, with the majority of wells exhibiting a shared characteristic of high alkalinity. Consequently, PC-2 assumes prominence within this cluster, and the majority of wells exhibit factor scores exceeding 1. Notably, GTS observations in this specific area suggest that the contamination source for Cluster 2 and PC-2 is linked to significant improper solid waste disposal activities.
8. Cluster 3 predominantly comprises wells located in the upper north zone of the study area. The analysis of Hierarchical Cluster Analysis (HCA) and Principal Component Analysis (PCA) indicates that the

primary component influencing cluster 3 is PC3 and PC1, which includes variables such as Total Dissolved Solids (TDS) and Total Hardness (TH). A distinct characteristic shared by all wells in Cluster 3 is a high level of Total Hardness, which can be attributed to their proximity to industrial areas, as verified by the Ground Truth Study (GTS). Consequently, a significant number of these wells exhibit high levels of Nitrate concentration.

9. Well number 24 within Cluster 4 revealed a highly contaminated area in the northern part. Ground Truth Study (GTS) verification confirmed that the primary factors contributing to this contamination are the excessive use of Urea, coupled with practices associated with animal husbandry and the improper disposal of domestic waste near a pond.
10. The recommendations of groundwater management can be evaluated with MCDM and numerical simulation as a future scope of this research.

Funding No institutional funding is involved.

Data Availability The collection of physicochemical data obtained from the esteemed Ministry of Jal Shakti portal (Source:https://ejalshakti.gov.in/IMISReports/Reports/WaterQuality/rpt_WQM_GPwiseTesting_S.aspx?Rep=0&RP=Y), alongside groundwater depth data collected from the India—WRIS (Water Resources Information System) (Source: https://indiawris.gov.in/wris/#/groundWater). The collected secondary data specifically focuses on shallow wells within various blocks in Bharuch, Panchmahal, and Vadodara District. Additionally, these selected wells have been strategically distributed across the entire study area. These data are valuable asset for this research.

Declarations

Conflict of interest The authors do not have any conflicting interests.

References

1. A.J. Ahamed, K. Loganathan, S. Ananthakrishnan, Ahmed and J.K.C. Ashraf, Evaluation of graphical and multivariate statistical methods for classification and evaluation of groundwater in Alathur block, Perambalur district, India". International water, air & soil conservation society, Kuala lumpur, Malaysia (2017)
2. C.W. Chan, G.H. Huang, Artificial intelligence for management and control of pollution minimization and mitigation processes. Eng. Appl. Artif. Intell. **16**(2), 75–90 (2003)
3. V. Cloutier, R. Lefebvre, R. Therrien, M.M. Savard, Multivariate statistical analysis of geochemical data as indicative of the hydrogeochemical evolution of groundwater in a sedimentary rock aquifer system. J. Hydrol. **353**(3–4), 294–313 (2008)
4. J.C. Davis, *Statistics and data analysis in geology*, 2nd edn. (Wiley, New York, 1986)
5. J.C. Davis, *Statistics and data analysis in geology* (Wiley, Singapore, 2002), pp.526–540
6. Z. Demirel, C. Guler, Hydro geochemical evolution of groundwater in a Mediterranean coastal aquifer, Mersin-Erdemli basin (Turkey). Enviro. Geol. **49**, 477–487 (2006)
7. District Industrial Potential Survey Report of Bharuch District, (2016–2017), A Report of MSME – Development Institute, Ministry of Micro, Small & Medium Enterprises, Govt. of India
8. District Industrial Potential Survey Report of Panchmahal District, (2016–2017), A Report of MSME – Development Institute, Ministry of Micro, Small & Medium Enterprises, Govt. of India
9. District Industrial Potential Survey Report of Vadodara District, (2016–2017), A Report of MSME – Development Institute, Ministry of Micro, Small & Medium Enterprises, Govt. of India
10. G.H. Dunteman, *Principal component analysis* (Sage, California, 1989)
11. I.M. Farnham, A.K. Singh, K.J. Stetzenbach, K.H. Johannesson, Treatment of nondetects in multivariate analysis of groundwater geochemistry data. Chemometr. Intell. Lab. Syst. **60**, 265–281 (2002)
12. C.W. Fetter, C. Fetter, *Contaminant Hydrogeology*, vol. 500 (Prentice Hall, New Jersey, 1999)
13. R.A. Freeze, J.A. Cherry, *Groundwater* (Prentice-Hall Inc, Englewood, 1979)
14. L.W. Gelhar, Stochastic Subsurface Hydrology (1993)
15. T. Ghosh, R. Kanchan, Geoenvironmental appraisal of groundwater quality in Bengal alluvial tract, India: a geochemical and statistical approach. Environ. Earth Sci. **72**, 2475–2488 (2014). https://doi.org/10.1007/s12665-014-3155-3
16. Groundwater Brochure, Bharuch District, A Report of Central Ground Water Board, West Central Region, Ahmedabad, (March 2014), Ministry of Water Resources, Government of India
17. Groundwater Brochure, Panchmahal District, A Report of Central Ground Water Board, West Central Region, Ahmedabad, (March 2014), Ministry of Water Resources, Government of India
18. Groundwater Brochure, Vadodara District, A Report of Central Ground Water Board, West Central Region, Ahmedabad, (March 2014), Ministry of Water Resources, Government of India
19. C. Guler, G.D. Thyne, Hydrologic and geologic factors controlling surface and groundwater chemistry in Indian Wells Owens Valley area, southeastern California, USA. J. Hydrol. **285**, 177–198 (2004)
20. C. Guler, G.D. Thyne, J.E. McCray, A.K. Turner, Evaluation of graphical and multivariate statistical methods for classification of water chemistry data. Hydrogeol. J. **10**, 455–474 (2002)
21. H.H. Harman, *Modern factor analysis* (University of Chicago Press, Chicago, 1960)
22. B. Helena, R. Pardo, M. Vega, E. Barrado, J.M. Fernandez, L. Fernandez, Temporal evolution of groundwater composition in an alluvial aquifer (Pisuerga River, Spain) by principal component analysis. Water Res. **34**(3), 807–816 (2000)
23. Indian Standards Specifications for drinking water IS: 10500–2012. http://cgwb.gov.in/Documents/WQ-standards.pdf
24. C.H. Jeong, Effect of land use and urbanization on hydrochemistry and contamination of groundwater from Taejon area. Korea. J. Hydrol. **253**, 194–210 (2001)
25. H.F. Kaiser, The varimax criterion for analytic rotation in factor analysis. Psychometrika **23**, 187–200 (1958). https://doi.org/10.1007/BF02289233
26. K.R. Karanth, *Groundwater assessment: development and management* (Tata McGraw-Hill Publishing Company Limited, New Delhi, 1987), p.720
27. K.-H. Kim, S.-T. Yun, S.-S. Park, Y. Joo, T.-S. Kim, Model-based clustering of hydrochemical data to demarcate natural versus

human impacts on bedrock groundwater quality in rural areas, South Korea. J. Hydrol. **519**, 626–636 (2014)
28. J.B. Kruskal, J.M. Landwehr, Icicle plots: better displays for hierarchical clustering. Am. Stat. **37**(2), 162–168 (1983)
29. C.Y. Lin, M.H. Abdullah, S.M. Praveena, A.H.B. Yahaya, B. Musta, Delineation of temporal variability and governing factors influencing the spatial variability of shallow groundwater chemistry in a tropical sedimentary island. J. Hydrol. **432–433**, 26–42 (2012)
30. D. Machiwal, M.K. Jha, Identifying sources of groundwater contamination in a hard-rock aquifer system using multivariate statistical analyzes and GIS-based geostatistical modeling techniques. J. Hydrol. **4**, 80–110 (2015). https://doi.org/10.1016/j.ejrh.2014.11.005
31. K. Nosrati, M. Van Den Eeckhaut, Assessment of groundwater quality using multivariate statistical techniques in Hashtgerd Plain. Iran. Environ. Earth Sci **65**, 331–344 (2012)
32. M. Otto, Multivariate methods, in *Analytical chemistry*. ed. by R. Kellner, J.M. Mermet, M. Otto, H.M. Widmer (Wiley-VCH, Weinheim, 1998), p.916
33. M.V. Prasanna, S. Chidambaram, K. Srinivasamoorthy, Statistical analysis of the hydrogeochemical evolution of groundwater in hard and sedimentary aquifers system of Gadilam River basin, South India. J. King Saud Univ. (Sci) **22**, 133–145 (2010). https://doi.org/10.1016/j.jksus.2010.04.001
34. A. Rasekh, K. Brumbelow, Machine learning approach for contamination source identification in water distribution systems. In: World Environmental and Water Resources Congress. Palm Springs, CA (2012)
35. M.N. Sara, R. Gibbons, Organization and analysis of water quality data, in *Practical handbook of ground-water monitoring*. ed. by D.M. Nielsen (Lewis Publishers, Michigan, 1991), pp.541–588
36. V. Simeonov, J.A. Stratis, C. Samara, G. Zachariadis, D. Voutsa, A. Anthemidis, M. Sofoniou, Th. Kouimtzis, Assessment of the surface water quality in Northern Greece. Water Res. **37**, 4119–4124 (2003)
37. C. Steube, S. Richter, C. Griebler, First attempts toward an integrative concept for the ecological assessment of groundwater ecosystems. Hydrogeol. J. **17**(1), 23–35 (2009)
38. J.H. Ward, Hierarchical grouping to optimize an objective function. J. Am. Stat. Assoc. **58**(301), 236–244 (1963)
39. Y.H. Yang, F. Zhou, H.C. Guo, H. Sheng, H. Liu, X. Dao, C.G. He, Analysis of spatial and temporal water pollution patterns in Lake Dianchi using multivariate statistical methods. Environ. Monit. Assess. **170**, 407–416 (2010)