Desalination and Water Treatment
www.deswater.com
doi: 10.5004/dwt.2020.25432

187 (2020) 311–320
May

Appraisal of groundwater alarming zones in district Okara-Pakistan using dimension reduction and Kriging procedures

Nadia Idrees[a,b,*], Shahid Kamal[c], Maqsood Ahmad[b,d]

[a]*Department of Mathematics and Statistics, University of Agriculture, Faisalabad 38040, Pakistan, email: nadia.stat.uaf@gmail.com (N. Idrees)*
[b]*College of Statistical and Actuarial Sciences, University of the Punjab, Lahore, Pakistan, email: maqsoodzain5@gmail.com (M. Ahmad)*
[c]*Government College University Faisalabad, Pakistan, email: kamal_shahid@hotmail.com (S. Kamal)*
[d]*Department of Statistics, University of Okara, Okara, Pakistan*

Received 18 April 2019; Accepted 4 January 2020

ABSTRACT

In this paper, our objective is to identify alarming areas of contaminated water in district Okara through dimension reduction techniques and spatial analysis of the water quality parameters (WQPs). For this purpose, water samples from 217 locations of district Okara, Pakistan are considered and their quality was evaluated by 14 international standard WQPs. Initially, we drew summary statistics, compared the WQPs with permissible limits of world health organization and assessed their distribution. To identify the most significant WQPs, total dissolved solids (TDS) was regressed upon other 13 WQPs. Through several dimension reduction techniques, results showed that Bicarbonate, Alkalinity, Sodium, Sulfate, Magnesium, and Chloride are important WQPs having a significant effect on TDS. General correlation matrix and distance-based correlation matrix (cross-variogram) highlighted the same positively correlated variables. To estimate the unobserved spatial locations, ordinary kriging, and cokriging techniques are applied and predictive results are presented by contour plots. Maps showed that the locations falling between 30.6 –30.8 latitude and 73.5 –73.7 longitude are alarming with respect to water quality measures. It is, therefore, recommended that the inhabitants of these vicinities must avoid drinking water without purification and the government should install water purification plants to save the lives of these residents.

Keywords: Water quality; Spatial mapping; Predictive analysis; Geostatistics; Desalination

1. Introduction

Appraisal of the groundwater quality is very important as it is highly associated with public health and many chronic diseases [1]. It is also imperative for industrial, agricultural, and domestic necessities [2]. Extensive usage of groundwater affects the aquifers, thus more than one billion people do not have clean drinking water. Moreover, unsuitable elimination of waste ingredients, industrial leftovers in water, food scum, and unnecessary usage of agrochemicals in agriculture contaminate the groundwater severely [3]. Since two billion people use groundwater for drinking around the world [4]; therefore, groundwater care is of vital importance to governments and other organizations, as its purification is closely related to socioeconomic growth.

In Pakistan, people get drinking water from hand pump, tube well, house water resources, and injector pumps [5]. Ever since slight consideration has been given to increase the drinking water quality; consequently, water obtained from these sources is mostly polluted [6]. Supply of water is generally unbalanced and water-borne infections such as typhoid, hepatitis, diarrhea, and stomach infection are very common in Pakistan [7]. Water sanitary condition in urban areas is also unsatisfactory [8]. In Pakistan, about 40% deaths and

* Corresponding author.

30% diseases are due to polluted and contaminated drinking water. In the meantime, each fifth resident of Pakistan is suffering from such infections; and about 0.1 million deaths occur each year [7]. There is a quick need to search those water quality parameters (WQPs) that are the real basis of this severe contamination of groundwater [9].

In hydrological researches, water samples are collected and analyzed for different WQPs and compared with the permissible limits. Generally, much more water samples are collected as compared to the samples that are statistically analyzed [1]. The reason behind this gap is the challenges faced during the analysis. Moreover, the analysis of groundwater quality has gained much importance in order to assess the factors making groundwater contaminated.

Statistical methods like principal component and cluster analysis have been considered [10] to predict the distribution of WQPs due to several chemical and physical parameters of groundwater [11,12]. These techniques lead towards effective learning of water quality and help in identifying those factors that influence the quality of water and provide consistent solutions to improve the quality of groundwater [4]. Usually, general statistical methods are applied to see some descriptive measures. Since it is not possible to collect samples from all locations; therefore, the prediction is used to assess those locations which are unmeasured. In this way, the spatial autocorrelation based techniques are of great concern. Non-normal data is often challenging for the researchers and it is also puzzling to estimate the missing observations.

In geostatistical analysis, generally, the hypotheses are: (1) Spatial process is stationary, (2) the neighboring observations are spatially auto-correlated. When data set is collected from large number of locations, it becomes tedious to manage such situations. To cope with such conditions, advanced statistical techniques are used that are able to deal the big n problems [13]. Geostatistics suggests different methods for modeling and prediction of the spatially varying groundwater parameters. Mainly, geostatistics comprises the variogram modeling to evaluate the correlation, cross validation, Kriging, and spatial mapping [14]. Kriging has been widely adopted to appraise spatial variability of groundwater quality parameters [15].

In this study, our main objective is to explore several predictive modeling techniques available for variable selection and to assess their predictability. These methods, however, reduce the dimensionality of the data. To deal with the physical and chemical variables in groundwater, most of the research work was based on some multivariate techniques. The selected and most effecting WQPs have been then spatially mapped by cokriging geostatistical method to observe the level of concentration at unobserved locations in sampled regions.

2. Materials and methods

2.1. Area under study

In this study, water samples from 217 locations of district Okara-Pakistan were collected from Pakistan Council of Research in Water Resources (PCRWR), Lahore. Okara is a district of Punjab, Pakistan that falls in the Sahiwal Division. It became a district in 1982 and lies at 30 48′29′′ north latitude, 73 26′45′′ east longitude. The Multan Road joins the district Okara with Lahore 110 km away and Faisalabad is 100 km far from Okara through Ravi River. According to the census of 2017, its population is 3,039,139. Okara cantonment is a beautiful Cant of Pakistan. Okara has a desert climate with 296 mm average precipitation and 24.5°C average temperature. Fig. 1 shows the location map of 217 samples of WQPs, collected from district Okara, Pakistan. The 14 WQPs that were measured from the groundwater of district Okara includes total dissolved solids (TDS), turbidity, calcium, magnesium, hardness, bicarbonate, alkalinity, potassium, chloride, sodium, iron, nitrate-N, sulfate, and fluoride.

2.2. Regression methods

Regression methods are the general approaches for data-driven modeling. A multiple linear regression (MLR) model is estimated by using ordinary least squares (OLS) method when there is one response variable and more than one predictors (say p). The usual MLR model with p-predictors is:

$$Z = X\beta + \varepsilon \tag{1}$$

where $X_{n\times p}$ is the design matrix of p-predictors, $\beta_{p\times 1}$ is the vector of unknown parameters and $\varepsilon_{n\times 1}$ is the vector of independently and identically distributed normal noise with mean 0 and covariance $\sigma^2 I_n$, and $Z_{n\times 1}$ is the random response vector. The least squares method minimizes the sum of squared deviations of the observed and estimated response, that is:

$$\hat{\beta}_{\text{ols}} = \arg\min_{\beta}\left\{\sum_{i=1}^{n}\left(z_{(i)} - \hat{z}_{(i)}\right)^2\right\} \tag{2}$$

where $z_{(i)}$ and $\hat{z}_{(i)}$ are the i^{th} observed and estimated responses. If the assumptions of MLR model fulfilled, than various inferences are possible to draw from the fitted model. However, the assumptions of MLR may not hold in practice particularly with several predictors or when the predictors are highly correlated. In such cases, OLS estimates of MLR model may not be unique and results in large prediction errors. In such situations, different alternative approaches that are able to deal with high levels of collinearity have to be considered. The detail of such approaches is briefly described in the next section considered in this study.

2.3. Dimension reduction techniques

A considerable number of regression methods for dimension reduction are available for modeling, the relationship between response variable and multiple predictors. However, these methods have their own assumptions regarding the nature and distribution of the variables. According to the algorithmic and calculating similarities, we grouped them in two classes: regularized regression methods and latent variable regression methods. In addition, their predictive performance depends on the characteristic under study as well as degree of sparsity and collinearity among predictors. A model is sparse, if it contains few important predictors from several candidate predictors. If most of the variables

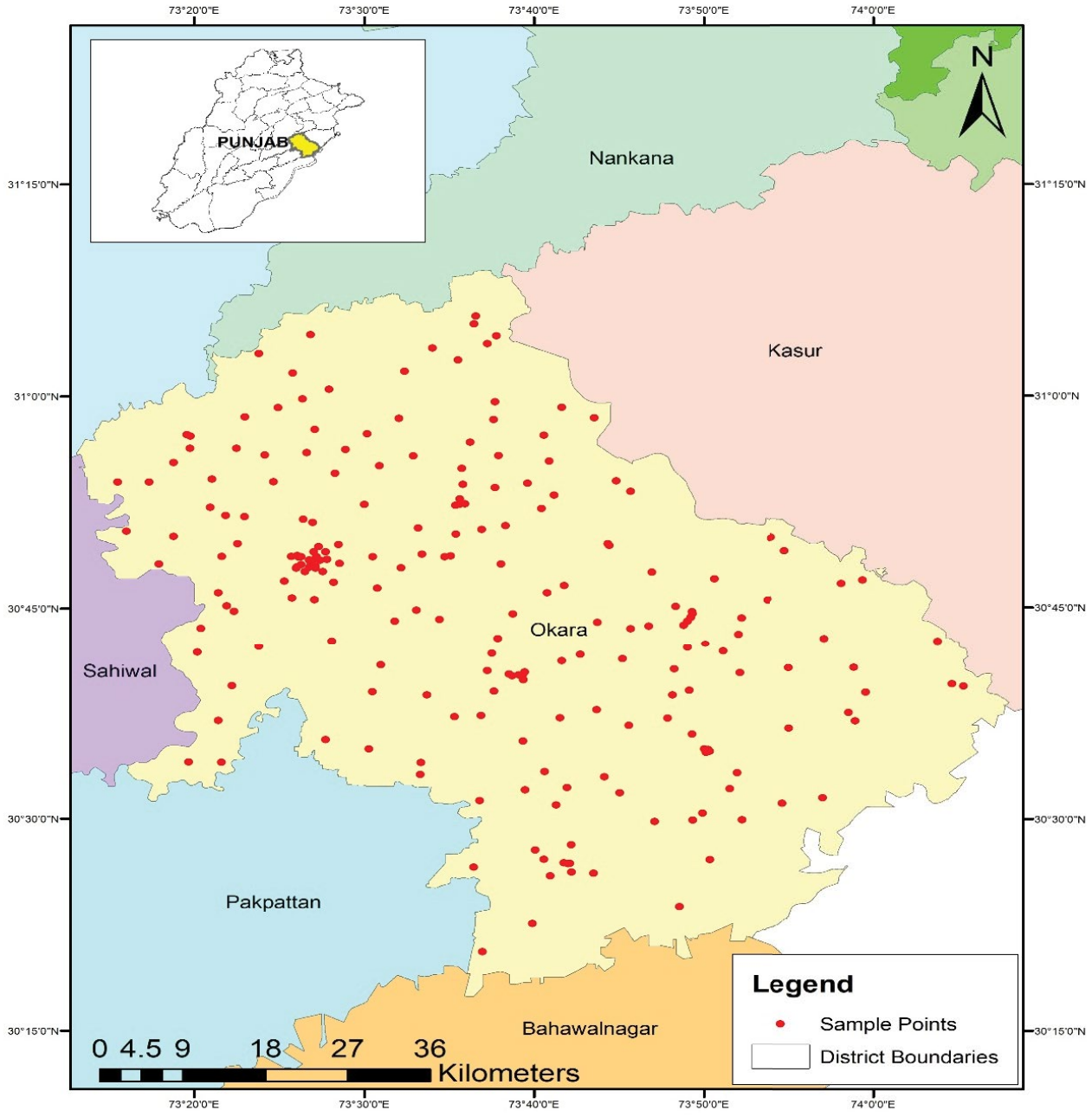

Fig. 1. Spatial distribution of 217 water samples taken from district Okara-Pakistan.

significantly affect the response, then it is not a sparsity scenario. On the other hand, the existence of correlation among the candidate predictors results in the problem of collinearity. Both (sparsity and collinearity) situations create problems in the development/estimation of the regression model. In case of sparse structure, a good predictive technique estimates the model by including only relevant/important predictors while leave aside the irrelevant predictors.

2.3.1. Regularized (penalized) regression methods

In order to stabilize the OLS estimates in the presence of collinearity, the class of regularized (penalized) regression methods had been proposed in the literature by the addition of a penalty term to Eq. (2). The penalty term incorporated to impose some restrictions on the magnitude of the regression coefficients. Hence, it shrinks regression coefficients towards zero, reduce the model complexity and increase the prediction accuracy. Three methods from this group of penalized regression were considered. These are the least absolute shrinkage and selection operator (lasso) [16], adaptive lasso (AL) [17] and elastic net (EN) [18].

The first regularized method comprised in this comparative study was proposed by Tibshirani [16] and is called "least absolute shrinkage and selection operator" (lasso). This method impose penalty to the sum of the absolute values of

the beta coefficients. Mathematically, regression coefficients are estimated by the method of lasso as:

$$\hat{\beta}_{\text{lasso}} = \arg\min_{\beta}\left\{\sum_{i=1}^{n}\left(z_{(i)} - \hat{z}_{(i)}\right)^2 + \lambda\sum_{j=1}^{p}\left|\beta_j\right|\right\} \quad (3)$$

The first part is in the right side of Eq. (3) is the usual least squares condition and second part is the regularized sum of the absolute values of beta coefficients. Furthermore, a decision has to make about the regularized parameter (λ), and 10-fold cross validation was adopted in this study. Due to restriction on the regression coefficients, some of the coefficients turn out to be exactly zero, and hence fitted model consists of few regressors as compared to the total. The second penalized regression method included in this study is EN, which is a generalization of lasso and introduced by Zou and Hastie [17]. This method performs better than lasso in situations where the sample size is less than the predictors or if there is a grouping effect present and if predictors are highly correlated. Elastic net estimator is defined as:

$$\hat{\beta}_{\text{EN}} = \arg\min_{\beta}\left\{\sum_{i=1}^{n}\left(z_{(i)} - \hat{z}_{(i)}\right)^2 + \lambda_1\sum_{j=1}^{p}\left|\beta_j\right| + \lambda_2\sum_{j=1}^{p}\left(\beta_j\right)^2\right\} \quad (4)$$

where λ_1 and λ_2 are regularized parameters usually specified by user or by cross validation. The penalty term of Elastic net is the combination of L_1 and L_2 norm.

The last method included in this comparative study from this class is the AL, proposed by Zou [18]. This method is also a generalization of lasso method that suggests the inclusion of adaptive weights in the penalty term. The AL estimator is:

$$\hat{\beta}_{\text{AL}} = \arg\min_{\beta}\left\{\sum_{i=1}^{n}\left(z_{(i)} - \hat{z}_{(i)}\right)^2 + \lambda\sum_{j=1}^{p} w_j\left|\beta_j\right|\right\} \quad (5)$$

where λ is the regularization parameter selected by 10-fold cross validation, $\hat{w} = 1\Big/\left|\hat{\beta}\right|^{\gamma}$ are the adaptive weights with $\gamma > 0$, $\hat{\beta}$ is any consistent initial estimator of β, for example it may be $\hat{\beta}_{\text{ols}}$. The use of weights in estimating the beta coefficients reduces the bias of Lasso estimates.

2.3.2. Latent variables regression methods

The regularized regression are based on the assumption that comparatively some predictors effect more to the response variable, and hence the fitted model have less number of predictors than the total predictors under the study. This methodology reduces the complexity of model and hence increases the prediction accuracy. In contrast, the latent variable methods have been established on the assumption that only a small number of linear combinations of original predictors rule the observed variation. Consequently, these techniques are appropriate for high collinearity problems as all the predictors are considered while estimating the regression model. Two approaches from this group were considered: principal component regression (PCR) [19,20] that extends the PC analysis to regression approaches and partial least squares regression (PLSR) [21].

Both the methods PCR and PLSR are used in case of several predictors or if there is strong collinearity among the predictors. Both methods construct new predictors, known as latent variables, as linear combinations of the original predictors. The latent variables are constructed in both methods in different way. PCR creates latent variables to explain the variability in the predictors only. While, PLSR does take into account the covariance between predictors and response variable in addition to the variability of predictors, and hence leads to models that fit the response variable with less number of latent variables and therefore results in better prediction.

PCR searches for a small number of linear combinations of the predictors which contains the maximum variability. Those linear combinations (called the factors/components) are uncorrelated and only a small number of those components are required to explain the over-all data variation. The PCR attempts to find the decomposition of X and is described as $X = RU' + A$, where U is the matrix of loadings, R is the score matrix, and A is the residual matrix.

PLSR has been used in many fields where predictive linear modeling with large number of predictors is necessary. This predictive model is obtained by extracting a set of orthogonal factors from the predictors. PLSR attempts to find the linear decomposition of X and Y such that $X = RU' + A$, and $Z = SV' + B$, where R and S are the score matrices of X and Z respectively, U and V are loading matrices and A and B are the matrices of residuals.

2.4. Geostatistical cokriging

To deal with the multivariate spatially auto-correlated data, geostatistical cokriging is an efficient prediction technique [4,10,22]. The best linear unbiased estimate of Z value at any unmonitored location s_0, is mathematically expressed as:

$$Z(s_0) = \sum_{i=1}^{n}\lambda_i Z(s_i) + \sum_{k=1}^{m}\omega_k Y(s_k) \quad (6)$$

where $Z(S_i)$ are the observed values of primary variable Z taken at S_i and $Y(S_k)$ are the observed values of secondary variable Y taken at S_k, where $k = 1, 2,\ldots, m$. Here λ_i and ω_k are weights associated with primary and secondary variables. For unbiased cokriging estimator, the sum of weights are $\sum_{i=1}^{n}\lambda_i = 1$ and $\sum_{i=1}^{m}\omega_k = 0$. For a detailed methodology on geostatistical cokriging see, Subyani et al. [22].

2.5. Comparison framework and performance measures

The predictive performance of different regression methods described in previous section was evaluated through a sound comparison framework. This comparison strategy was based on the simulations and cross validation. The assessment procedure starts by defining the number of simulations. In each simulation run, the complete data set was randomly divided into training and test set (70% of total observations for training and the remaining 30% for test set). The training data set was used to fit the models, and these

built-in models were used for predicting the test set observations. The performance evaluation measures (described below) were calculated in the test set for different regression methods in the current simulation and saved. The distribution of the error performance measure characterizes the predictive performance of a method (method having small measure is better in terms of prediction accuracy). As the data set was divided randomly into training and test set, so, different data sets were obtained in each simulation and different values of a performance evaluation measure were computed for a regression method. In addition, the fitted models and the value of regularized parameters for each regression method might be varied as a result of different random splits. However, this variation provides useful information about the regression approach and data. The model that was most repeatedly fitted during the random splits was considered as final model. For all the regression methods, the median of error prediction measures over 100 simulations was used for comparison purpose.

An important concern during the model estimation from training data is the selection of regularized parameter. Various approaches like Mallow's Cp, leave one out cross validation, *k*-fold cross validation, and generalized cross validation are used for this purpose. In this article, *k*-fold (k = 10) cross validation methodology was adopted as this approach has been frequently and successfully adopted in diverse statistical approaches. In addition, it is directly associated to the predictive performance.

2.6. Measures for performance evaluation

To assess the performance of regression methods, three performance evaluation measures were used: mean absolute error (MAE), root mean square prediction error (RMSPE), and relative prediction error (RPE). These performance measures are defined as:

$$\mathrm{MAE} = \frac{1}{n^*}\sum_{i=1}^{n^*}\left|Z_i - \hat{Z}_i\right| \tag{7}$$

$$\mathrm{RMSPE} = \sqrt{\frac{1}{n^*}\sum_{i=1}^{n^*}\left(Z_i - \hat{Z}_i\right)^2} \tag{8}$$

$$\mathrm{RPE} = \frac{\sum_{i=1}^{n^*}\left(Z_i - \hat{Z}_i\right)^2}{\sum_{i=1}^{n^*}\left(Z_i - \bar{Z}_p\right)^2} \tag{9}$$

where Z_i denotes the i^{th} observed response in test data, $\hat{Z}_i$ is the predicted value of response in test data obtained by using the model fitted from the training data, $\hat{Z}_p$ is the arithmetic mean of predicted values and the sum is over all observations in the test data. The mean response of training data set may be used as the predicted response in case where no variables were selected from training data.

3. Results

In this section, we analyzed the 14 WQPs from 217 locations, observed from the groundwater of district Okara. Initially, we analyze the data of WQPs for some summary statistics, see Table 1. It is evident that all WQPs have maximum values significantly larger than the allowable limits described by World Health Organization (WHO) [23]. The very high coefficient of variation (CV) for all the WQPs reflects the high variation in the groundwater of district Okara.

Generally, for fitting a MLR model, a basic assumption is about the normality of error term and hence of response. Another assumption is about the internal relationship of the predictors, that is, predictors must be independent to each other (no multicollinearity). However, it is not a prerequisite to assume normality of error term to perform regression. A normality assumption about error term is required to make inferences about the regression coefficients after fitting the regression model. Therefore, before establishing any statistical models, the distribution of response variable and collinearities among water parameters have to be evaluated.

Table 1
Descriptive statistical measures of 217 water quality parameters in the groundwater of Okara district

Variable	Min.	Max.	Mean	CV	Skewness	Kurtosis	WHO limits
TDS, mg/L	166.00	8,366.0	1,191.50	108.33	3.36	13.22	≤1,000
Turbidity, NTU	0.10	84.0	3.93	190.46	6.95	64.49	≤5
Calcium, mg/L	14.00	426.0	92.49	55.60	3.05	16.31	≤75
Magnesium, mg/L	6.00	168.0	35.95	64.65	2.01	6.28	≤150
Hardness, mg/L	65.00	1,510.0	379.30	49.54	2.34	9.85	≤60
Bicarbonate, mg/L	120.00	1,350.0	395.30	42.31	1.40	4.98	Not defined
Alkalinity, m mol l	2.40	27.0	7.92	42.14	1.39	5.02	≤200
Chloride, mg/L	7.00	2,695.0	196.50	185.51	4.48	23.28	≤250
Potassium, mg/L	1.00	64.0	9.10	97.19	2.40	8.15	≤12
Sodium, mg/L	15.00	2,900.0	327.90	149.60	3.24	11.76	≤200
Sulfate, mg/L	15.00	3,125.0	427.00	130.83	2.96	9.49	≤250
Iron, mg/L	0.02	1.7	0.22	81.56	4.66	28.15	≤0.3
Nitrate-N, mg/L	0.10	15.6	2.10	97.97	2.80	11.20	≤50
Fluoride, mg/L	0.04	66.0	1.20	376.20	13.92	200.74	≤1.5

For testing the normality, Anderson-Darling normality test was performed which shows the non-normality of TDS (A^2 = 24.82, p-value < 0.005). However, after applying log transformation, the response variable transformed to approximately a normal distribution. In Fig. 2, histogram (left panel) is also showing the positively skewed behavior and the spatial distribution of TDS (mg/L) developed by linear interpolation is shown (right panel) where the area shaded by all other colors except blue color is alarming.

Since, one of the objectives is to assess the most significant WQPs causing the contamination in drinking water; therefore initially we constructed correlation matrix (shown in Fig. 3). Highest positive correlation among WQPs is shown by red color. It is evident that TDS have highest positive correlation with sodium (r = 0.98) and chloride (r = 0.95), sulfate exhibited maximum correlation with TDS (r = 0.97) and with sodium (r = 0.95), chloride showed maximum correlation with sulfate (r = 0.88), and with sodium (r = 0.94) similarly, hardness have maximum association with calcium (r = 0.88) and with magnesium (r = 0.78). Turbidity and iron showed negative correlation with many other WQPs. A MLR model estimated by using OLS was also used to assess if there is any multicollinearity effect present in the data. The variance inflation factor (VIF) was calculated to assess the presence of multicollinearity. The VIF obtained for 8 out of 13 predictors was very large (>10) and thus, there was a strong evidence of the presence of multicollinearity. Because of multicollinearity, the combine effect of all the water parameters on TDS was highly significant (F-ratio = 84.52, p-value = 0.000) while they were non-significant individually. Examination of correlation matrix (Fig. 3) also confirms the presence of collinearity. So, a MLR model estimated by OLS was not appropriate to model the water parameters. Hence, the regression techniques based on regularized methods and latent variables regression methods were used to establish the relationship between TDS and 13 WQPs. Table 2 presents the median of error prediction measures over 100 simulations for all the regression methods.

In terms of RMSPE, MAE, and RPE, AL outperformed other methods. Elastic net and lasso also showed better performance in reducing the prediction error. Results of PCR and PLSR revealed that first five components explain 84% and 72% of the total variation. First component of both the PCR and PLSR accounts for 42% of the total variation that is associated with the high loadings of magnesium, bicarbonate, alkalinity, chloride, sodium, and sulfate (>0.30).

Different numbers of variables were selected by different regression methods. Table 3 presents the variables that were consistently selected in the 100 (simulation runs) random splits, but three variables (Magnesium, Alkalinity, and Sulfate) were frequently selected by all the five methods. In addition, three other variables (Bicarbonate, Chloride, and Sodium) were also identified as significant by Elastic net, PLSR, and PCR. Dendrogram based on complete linkage methods (cluster analysis) for the assessment of most correlated factors also identify these six variables in the first cluster (see Fig. 4). Table 4 shows the estimates of regression coefficients estimated by different regression methods in one simulation. The analysis suggested that TDS have more dependence on these six variables and it is essential to think through the spatial distribution of these variables.

3.1. Predictive maps

In this section, cokriging technique was used for the spatial prediction of the selected variables. Initially, we estimated the three parameters (Sill = σ^2, Range = ϕ, Nugget = τ^2) of the spherical variogram model using the command "eyefit" and later on confirmed the parameters estimation using OLS method.

Later on, we visualize cross-variogram using gstat package [24] of R statistical software [25] in order to show the correlation structure with respect to distance. Six variables (selected by PLSR, PCR, and Elastic net) have been processed in cross-variogram by considering the TDS as primary variable while other five as secondary variables. As the results

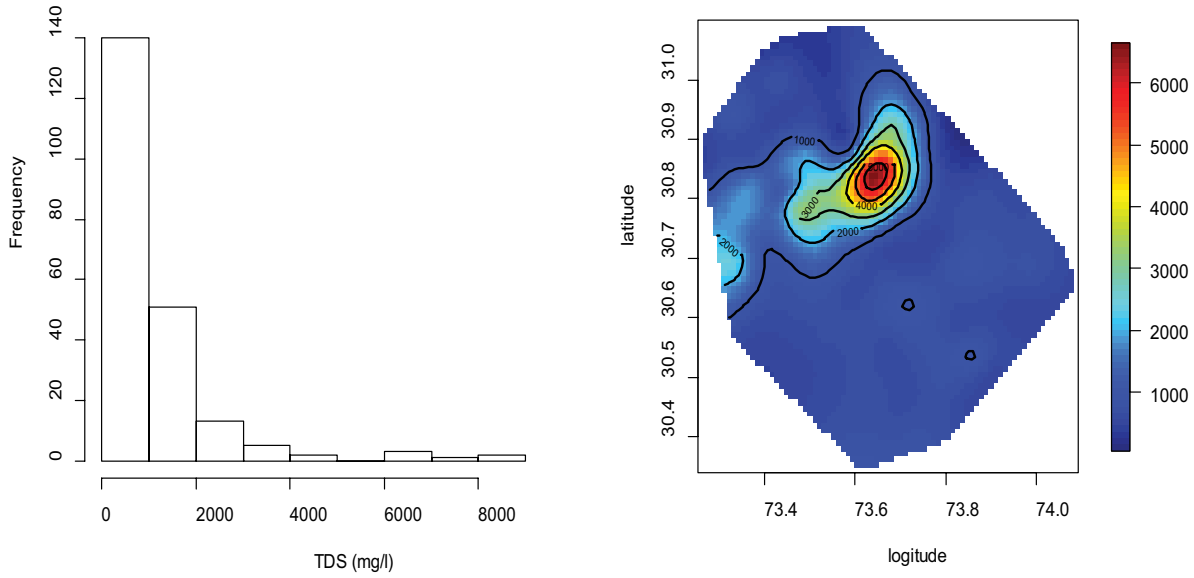

Fig. 2. Histogram (left panel) and graphical description of the spatial distribution of TDS in the groundwater of district Okara.

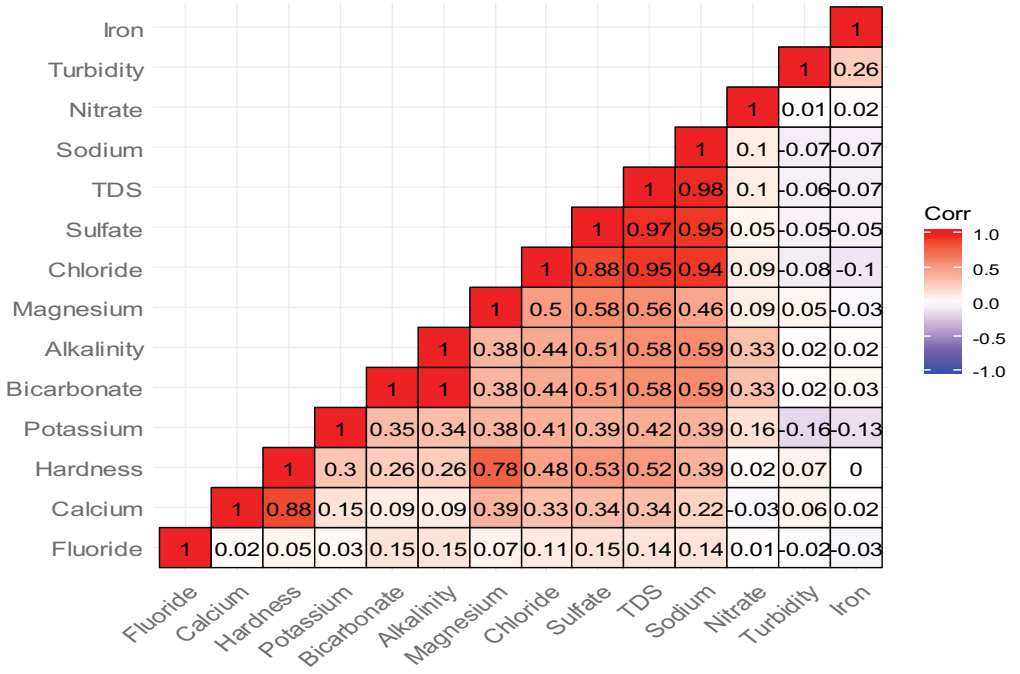

Fig. 3. Graphical display of a correlation matrix among all water quality parameters.

Table 2
Median of different performance measures from regression methods

Regression methods	RMSPE	MAE	RPE	R^2
PLS	0.478	0.356	0.174	71.9
PCR	0.499	0.394	0.195	83.8
LASSO	0.467	0.364	0.173	88.9
Adaptive lasso	0.454	0.335	0.163	88.9
Elastic net	0.465	0.362	0.171	88.9

illustrated in Fig. 3, the TDS showed strong positive based correlation with sodium, sulfate and chloride while moderate correlation ($r \approx 0.6$) with bicarbonate and alkalinity. Similarly, graph of distance-based correlation (cross-variogram) showed spatial correlation among six selected variables (bicarbonate, alkalinity, sodium, sulfate, magnesium, and chloride with TDS (shown in Fig. 5). All variables have strong distance-based correlation except magnesium with other variables.

In Fig. 6, the box-cox transformed variable has been shown using the estimated value of $\lambda = -0.2203$ which is approximately normally distributed. Transformation is considered using the expression:

$$Z^* = \begin{cases} \dfrac{Z^{\lambda}-1}{\lambda} : \lambda \neq 0 \\ \log(Z) : \lambda = 0 \end{cases} \quad (10)$$

where λ is the transformed parameter. Further, we used krige.control function of geoR Package [26] to get the predicted values and used the leave on out cross validation statistics to

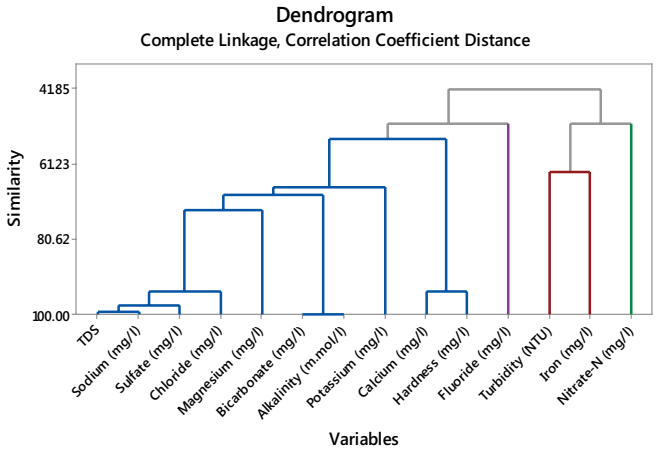

Fig. 4. Dendrogram based on complete linkage methods for the assessment of most correlated factors.

evaluate the prediction errors. The map of bicarbonate and alkalinity showed that the area between longitude 73.5–73.7 and latitude 30.6–30.8 is alarming.

Since, the WHO permissible limit of sodium, sulfate, and chloride is 250, 200, and 250 mg/L respectively; therefore, the areas that lie between longitude 73.5–73.7 and latitude 30.6–30.8 is not acceptable. Ordinary kriging can also predict the same variables but it does not consider the regression settings and use one variable at a time. Thus, we compared the spatial cokriging with ordinary kriging on the basis of performance measures (RMSPE and MAE). Both measures indicates that cokriging is best technique for the underlying data.

Contour maps also showed a good view of the spatial distribution of all WQPs. Our findings were approximately

Table 3
Consistently selected variables in 100 simulations by different regression methods

Regression method	Variables selected in each regression approach
PLSR (Component 1)	Magnesium, bicarbonate, alkalinity, chloride, sodium, sulfate
PCR (Component 1)	Magnesium, bicarbonate, alkalinity, chloride, sodium, sulfate
Lasso	Magnesium, alkalinity, sulfate
Adaptive lasso	Magnesium, alkalinity, sulfate
Elastic net	Magnesium, bicarbonate, alkalinity, chloride, sodium, sulfate

Table 4
Regression coefficients based on lasso, adaptive lasso, elastic net, PCR, and PLS in one simulation

	Lasso	A-lasso	El-net	PCR	PLS
(Intercept)	0.0193	0.0245	0.0238	0.0193	0.0193
Turbidity	*	*	*	0.0049	–0.0055
Calcium	*	*	*	–0.0099	0.0212
Magnesium	0.0789	0.0462	0.0326	0.1326	0.1537
Hardness	*	*	*	0.0603	0.0928
Bicarbonate	*	*	0.0984	0.2143	0.1932
Alkalinity	0.3396	0.3694	0.1066	0.2148	0.1936
Chloride	*	*	0.0021	0.1467	0.1304
Potassium	*	*	*	0.0755	0.1561
Sodium	*	*	0.1403	0.2031	0.1529
Sulfate	0.4639	0.5240	0.1843	0.1999	0.1470
Iron	*	*	*	–0.0488	–0.0042
Nitrate	*	*	*	0.0037	0.0577
Fluoride	*	*	*	0.0411	0.0268

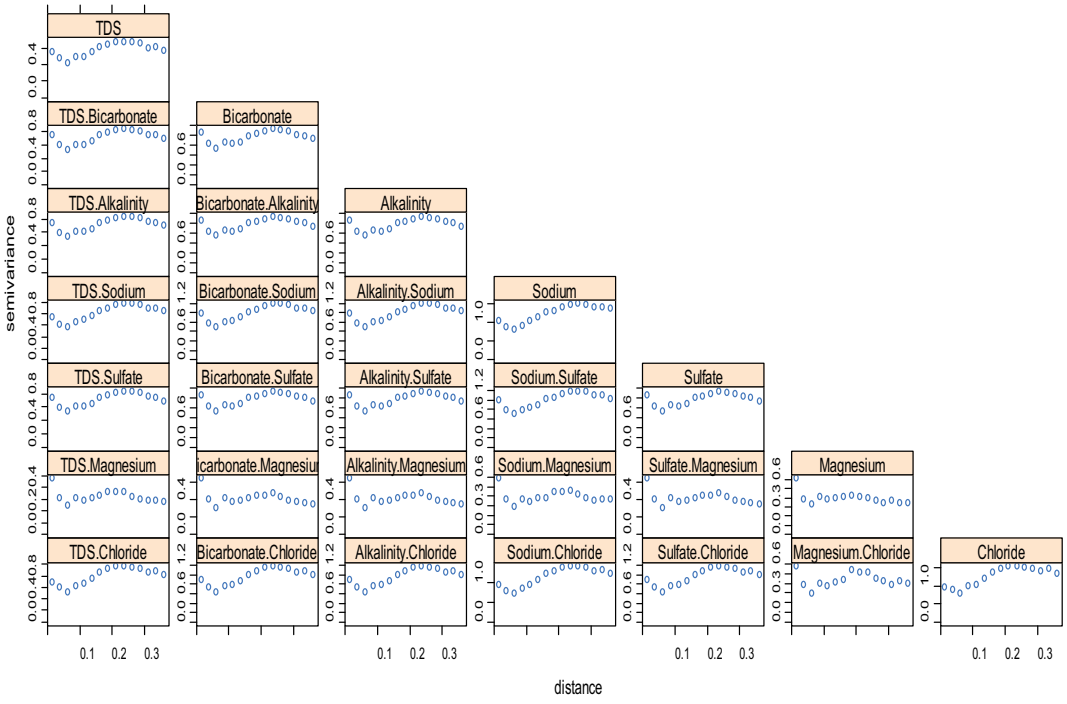

Fig. 5. Graph of cross-variogram to assess the distance-based spatial correlation among six selected variables (bicarbonate, alkalinity, sodium, sulfate, magnesium, and chloride with TDS).

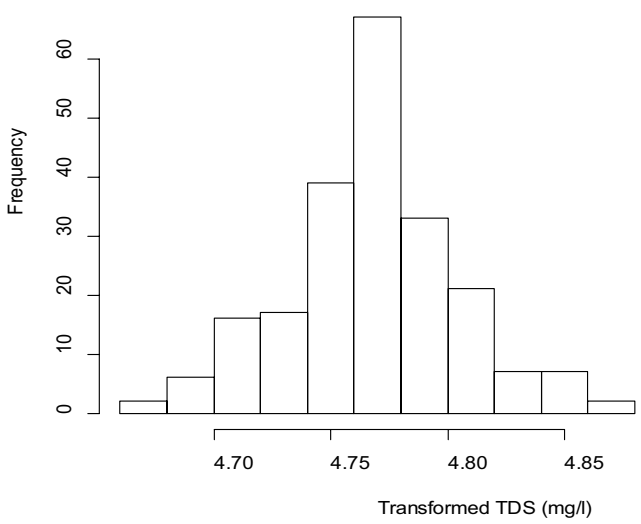

Fig. 6. Histogram of Box-Cox transformed response variable (TDS) with estimated $\lambda = -0.2203$.

matched with the researches of Triki [10] and Ahmad et al. [4]. Although, we used and compared different dimension reduction techniques, while Triki [10] and Ahmad et al. [4] used only principal component analysis for selection of the important variables.

4. Conclusions

In this paper, predictive approaches based on regularized regression and latent variables regression methods identified the most important and influencing WQPs. lasso, EL, AL, PLSR, and PCR suggested 6 variables out of 14 WQPs as most effecting to TDS. Correlation matrix demonstrated high positive correlation and at the same time cross-variogram map showed strong distance-based correlation of TDS with sodium, sulfate, chloride, bicarbonate, and alkalinity except magnesium. Prediction maps showed that the areas that fall between latitude 30.6 –30.8 and longitude 73.5 –73.7 are alarming. The results showed to be a valuable mean for quick observing of water quality with the help of regression analysis. Further, it is recommended to the residents of these

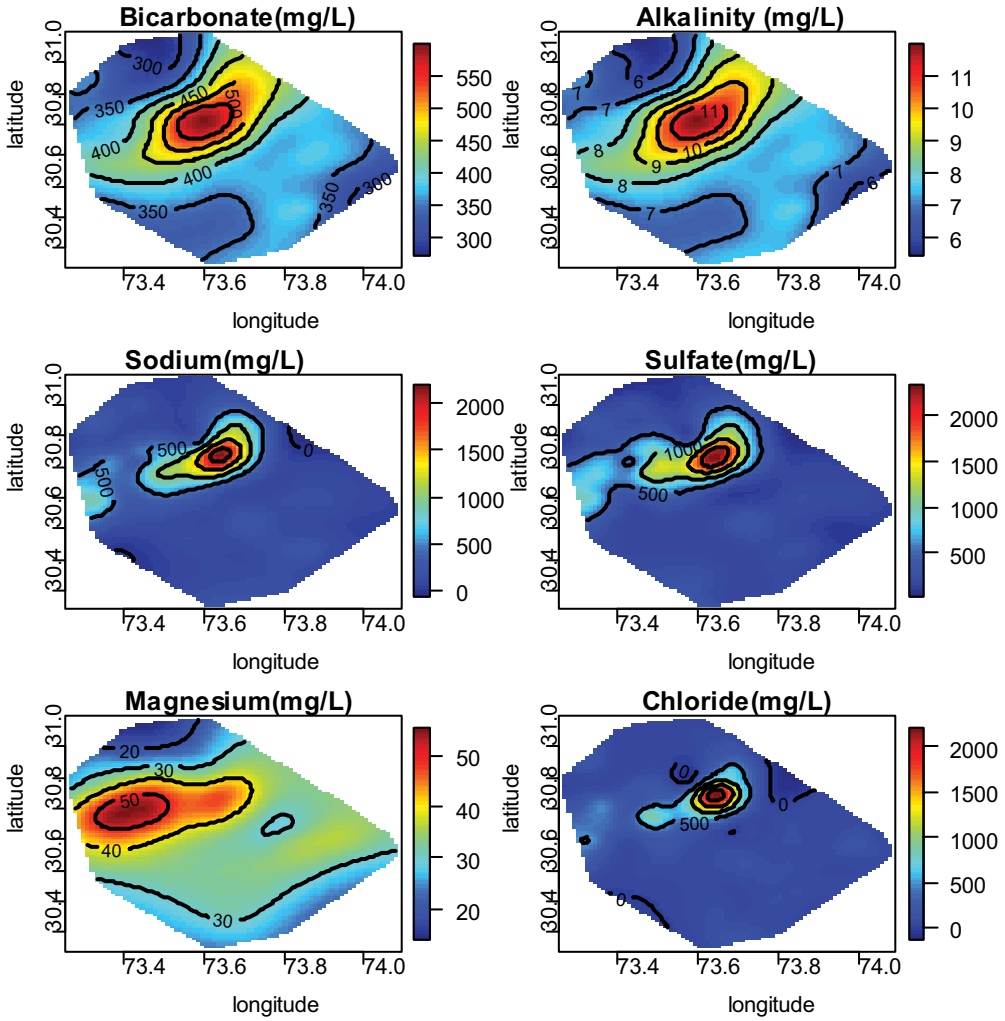

Fig. 7. Prediction maps of most significant water quality parameters (bicarbonate, alkalinity, sodium, sulfate, magnesium, and chloride) using the cokriging.

areas that they must avoid drinking water without purification and government should install water purification plants to save the lives of these residents. On the basis of results, the inhabitants must see the distribution of each parameter and use the water after purification if it is not falling under permissible limits. For methodological point of view, we wish to use the Bayesian kriging methods along with these dimension reduction techniques in future for more valid prediction results.

Acknowledgments

Authors are thankful to HEC, Pakistan for funding, and PCRWR for providing the research data under consideration.

References

[1] H. Li, C.D. Smith, L. Wang, Z. Li, C. Xiong, R. Zhang, Combining spatial analysis and a drinking water quality index to evaluate monitoring data, Int. J. Environ. Res. Public Health, 16 (2019) 357.
[2] S. Muhammad, M. Shah, S. Khan, Arsenic health risk assessment in drinking water and source apportionment using multivariate statistical techniques in Kohistan region, northern Pakistan, Food Chem. Toxicol., 48 (2010) 2855–2864.
[3] F. Sanchez-Martos, R. Jimenez-Espinosa, A. Pulido-Bosch, Mapping groundwater quality variables using PCA and geostatistics: a case study of Bajo Andarax, southeastern Spain, Hydrol. Sci. J., 46 (2001) 227–242.
[4] M. Ahmad, S. Chand, H.M. Rafique, Geostatistical cokriging and multivariate statistical methods to evaluate groundwater salinization in Faisalabad, Pakistan, Desal. Water Treat., 84 (2017) 93–101.
[5] A. Azizullah, M.N.K. Khattak, P. Richter, D.P. Häder, Water pollution in Pakistan and its impact on public health - a review, Environ. Int., 37 (2011) 479–497.
[6] O.S. Von Ehrenstein, D.N. Guha Mazumder, M. Hira-Smith, N. Ghosh, Y. Yuan, G. Windham, A. Ghosh, R. Haque, S. Lahiri, D. Kalman, S. Das, A.H. Smith, Pregnancy outcomes, infant mortality, and arsenic in drinking water in West Bengal, India, Am. J. Epidemiol., 163 (2006) 662–669.
[7] S. Haydar, M. Arshad, J.A. Aziz, Evaluation of drinking water quality in urban areas of Pakistan: a case study of Southern Lahore, Pak. J. Eng. Appl. Sci., 5 (2016) 16–23.
[8] M. Ahmad, S. Chand, H.M. Rafique, Predicting the spatial distribution of sulfate concentration in groundwater of Jampur-Pakistan using geostatistical methods, Desal. Water Treat., 57 (2016) 28195–28204.
[9] M.B. Arain, T.G. Kazi, J.A. Baig, M.K. Jamali, H.I. Afridi, A.Q. Shah, N. Jalbani, R.A. Sarfraz, Determination of arsenic levels in lake water, sediment, and foodstuff from selected area of Sindh, Pakistan: estimation of daily dietary intake, Food Chem. Toxicol., 47 (2009) 242–248.
[10] I. Triki, N. Trabelsi, M. Zairi, H. Ben Dhia, Multivariate statistical and geostatistical techniques for assessing groundwater salinization in Sfax, a coastal region of eastern Tunisia, Desal. Water Treat. 52 (2014) 1980–1989.
[11] H. Eslami, M.H. Ehrampoush, A. Esmaeili, A.A. Ebrahimi, M.H. Salmani, M.T. Ghaneian, H. Falahzadeh, Efficient photocatalytic oxidation of arsenite from contaminated water by Fe_2O_3–Mn_2O_3 nanocomposite under UVA radiation and process optimization with experimental design, Chemosphere, 207 (2018) 303–312.
[12] H. Eslami, M.H. Ehrampoush, A. Esmaeili, M.H. Salmani, A.A. Ebrahimi, M.T. Ghaneian, H. Falahzadeh, R.F. Fard, Enhanced coagulation process by Fe-Mn bimetal nano-oxides in combination with inorganic polymer coagulants for improving As (V) removal from contaminated water, J. Cleaner Prod., 208 (2019) 384–392.
[13] S. Banerjee, B.P. Carlin, A.E. Gelfand, Hierarchical Modeling and Analysis for Spatial Data, Chapman & Hall/CRC Monographs on Statistics and Applied Probability, 2014.
[14] P.P. Adhikary, H. Chandrasekharan, D. Chakraborty, K. Kamble, Assessment of groundwater pollution in West Delhi, India using geostatistical approach, Environ. Monit. Assess., 167 (2010) 599–615.
[15] M. Ahmad, S. Chand, Spatial distribution of TDS in drinking water of Tehsil Jampur using ordinary and Bayesian kriging, Pak. J. Stat. Oper. Res., 11 (2015) 377–386.
[16] R. Tibshirani, Regression shrinkage and selection via the lasso, J. R. Stat. Soc. Ser., B, 58 (1996) 267–288.
[17] H. Zou, T. Hastie, Regularization and variable selection via the elastic net, J. R. Stat. Soc. Ser. B, 67 (2005) 301–320.
[18] H. Zou, The adaptive lasso and its oracle properties, J. Am. Stat. Assoc., 101 (2006) 1418–1429.
[19] I. Jolliffe, Principal Component Analysis, Springer Series in Statistics, New York, 2011.
[20] S. Wold, K. Esbensen, P. Geladi, Principal component analysis, Chemom. Intell. Lab. Syst., 2 (1987) 37–52.
[21] H. Wold, Nonlinear Iterative Partial Least Squares (NIPALS) Modelling: Some Current Developments, Multivariate Analysis–III, Academic Press, Elsevier, New York, 1973, pp. 383–407.
[22] A.M. Subyani, A.M. Al-Dakheel, Multivariate geostatistical methods of mean annual and seasonal rainfall in southwest Saudi Arabia, Arab. J. Geosci., 2 (2009) 19–27.
[23] WHO, Guidelines for Drinking-Water Quality, 1st ed., Volume 1: Recommendations, World Health Organization, Geneva, Switzerland, 2004.
[24] E. Pebesma, B. Graler, Introduction to Spatio-Temporal Variography, ifigi, Institute for Geoinformatics, University of Munster, 2016. Available at: https://cran.r-project.org/web/packages/gstat/vignettes/st.pdf
[25] R.C. Team, R: A Language and Environment for Statistical Computing, Vienna, Austria, R Foundation for Statistical Computing, 2016, 2017.
[26] P.J. Ribeiro, P.J. Diggle, geoR: A Package for Geostatistical Analysis, R-NEWS, (ISSN 1609-3631), 1st ed., A Newsletter of the R project, University of California, Davis, USA, 2006, pp. 1–28.