A Comparison of Genetic Algorithms and Particle Swarm Optimization to Estimate Cluster-Based Kriging Parameters

Carlos Yasojima(✉), Tiago Araújo, Bianchi Meiguins, Nelson Neto, and Jefferson Morais

Faculty of Computer Science, Federal University of Pará, Belém, Pará, Brazil
takeshiyasojima@gmail.com,
{tiagoaraujo,bianchi,nelsonneto,jmorais}@ufpa.br

Abstract. Kriging is one of the most used spatial estimation methods in real-world applications. Some kriging parameters must be estimated in order to reach a good accuracy in the interpolation process, however, this task remains a challenge. Various optimization methods have been tested to find good parameters of the kriging process. In recent years, many authors are using bio-inspired techniques and achieving good results in estimating these parameters in comparison with traditional techniques. This paper presents a comparison between well known bio-inspired techniques such as Genetic Algorithms and Particle Swarm Optimization in the estimation of the essential kriging parameters: nugget, sill, range, angle, and factor. In order to perform the tests, we proposed a methodology based on the cluster-based kriging method. Considering the Friedman test, the results showed no statistical difference between the evaluated algorithms in optimizing kriging parameters. On the other hand, the Particle Swarm Optimization approach presented a faster convergence, which is important in this high-cost computational problem.

Keywords: Bio-inspired algorithms · Artificial Intelligence · Geostatistic · Kriging

1 Introduction

Kriging is a geostatistical interpolation technique that predicts the value of observations in unknown locations, based on previously collected data. The kriging error or interpolation error is minimized by studying and modeling the spatial distribution of points already obtained. This spatial distribution or spatial variation is expressed in the form of an experimental variogram [9].

The variogram is the basis for the application of the kriging method. Thus, the kriging process is defined in three main steps. First, the experimental variogram is calculated. Then, the theoretical variogram is modeled to represent the experimental variogram. Finally, the value of a given point is predicted using the built theoretical model.

P. Moura Oliveira et al. (Eds.): EPIA 2019, LNAI 11804, pp. 750–761, 2019.
https://doi.org/10.1007/978-3-030-30241-2_62

Artificial Intelligence techniques have been used to improve the kriging process as shown in [2,8,10,13,15], however, it is still a challenge to determine what method is better suited for a given database. As stated in [8] and applied in [10], bio-inspired algorithms, in general, are suitable to help define the theoretical variogram parameters. Furthermore, these types of algorithms do not require a single initial seed value as input, but rather an interval.

Several researchers have implemented bio-inspired algorithms to optimize theoretical variogram parameters, such as Genetic Algorithms (GA) [2,10,15], and more recently Particle Swarm Optimization (PSO) [12]. However, which to the best of the authors knowledge, there is no systematic comparison in relation to accuracy and convergence of the kriging parameters estimation as well computational processing effort of these algorithms considering the same scenario of study.

This paper applies the cluster-based kriging method designed by [1]. In this method, the spatial data are divided into different subgroups by the K-means clustering method [14], where each data point is interpolated using only data from the same group. A limitation of the cluster-based method proposed by [1] is the task of defining a single theoretical variogram model for all groups, without considering that distinct regions of the analyzed database may present specific behaviors during the kriging process. Therefore, we propose the estimation and optimization of different parameters to each group. Some problems encountered in this kind of improvement, such as cluster overlapping and unknown point classification, are solved using the K-Nearest Neighbour (KNN) algorithm [14] and data preprocessing.

In this context, the purpose of this paper is to evaluate well known bio-inspired techniques such as Genetic Algorithms and Particle Swarm Optimization when applied in the estimation of the essential kriging parameters. As said, in order to perform tests, the cluster-based kriging method was applied and a GA and PSO model was built for each cluster. The results showed that both algorithms were statistically equal in optimizing the variogram parameters. However, PSO converged faster than GA. This conclusion is important because the adopted methodology (cluster-based kriging, interpolation cost function, and so forth) presents a high computational cost. Therefore, it is essential to reduce the number of iterations of optimization algorithms.

This paper is organized as follows. Section 2 consists of the theoretical background involving important concepts for the understanding of this paper. In Sect. 3, the steps of the proposed methodology are detailed. Section 4 presents the database used in this work besides the results of the experiments performed. Finally, Sect. 5 presents the final considerations.

2 Background

2.1 Kriging

Kriging is an interpolation technique widely used in geostatistics to predict spatial data. This method takes into account the characteristics of regional variables

autocorrelation. These variables have some spatial continuity, which allows the data obtained by the sampling of specific points to be used to parameterize the prediction of points where the value of the variable is unknown [9].

Let Z be a set of observations of a target variable (response variable) denoted as $\{z(s_1),\ z(s_2),\ \ldots,\ z(s_N)\}$, where $s_i = (x_i, y_i)$ is a point in a geographical space; x_i and y_i are its coordinates (primary locations); and N is the number of observations.

Values of the target variable at some new location s_0 can be derived using a spatial prediction model. The standard version of kriging is called ordinary kriging (OK), where the predictions are based on the model:

$$\hat{z}_{OK}(s_0) = \sum_{i=1}^{n} w_i(s_0).z(s_i) = \lambda_0^T.\mathbf{z} \tag{1}$$

where λ_0 is a vector of kriging weights (w_i), and $\mathbf{z}$ is the vector of N observations at primary locations.

So, in order to estimate the weights, we calculate the semivariances $\gamma(h)$ based on the differences between the neighboring values:

$$\gamma(h) = \frac{1}{2}E[(z(s_i) - z(s_i + h))^2] \tag{2}$$

where $z(s_i)$ is the observation of the target variable at some point location, and $z(s_i + h)$ is the observation of the neighbour at a distance $s_i + h$.

Suppose that there are N point observations, this yields $N \times (N - 1)/2$ pairs for which a semivariance can be calculated. If we plot all semivariances versus their separation distances a variogram cloud is produced. For an easier visualization of this variogram cloud, the values are commonly averaged for a standard distance called "lag". If we display such averaged data, then we get the standard experimental variogram, which can be seen in Fig. 1.

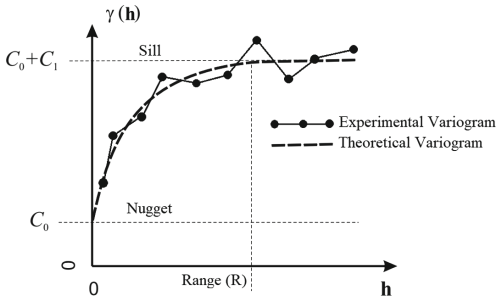

Fig. 1. Example of a final variogram model.

Once we calculate the experimental variogram, we can fit it using a theoretical model, such as linear, spherical, exponential, Gaussian, among others.

The variograms are commonly fitted using a cost function (e.g. weighted least squares [3]). Hence, the main objective is to minimize this cost function. In this work, in order to simplify the experiments, we only use the exponential theoretical model, which is given by

$$\gamma(h) = C_0 + C_1 \left(1 - EXP\left(-3\left(\frac{h}{R}\right)^2\right)\right) \tag{3}$$

where R is the range of influence or simply range, which is the coordinate where the model starts to flatten out; C_0 is the nugget effect, which can be attributed to measurements errors or spatial sources of variation at distances smaller than the sampling interval; and $C_O + C_1$ is the sill, which is the value that the model attains at the range R. These parameters, also called coefficients, determine the theoretical variogram as illustrated in Fig. 1.

Once we have estimated the theoretical model, we can use it to derive semivariances at all locations and solve the kriging weights. The ordinary kriging (OK) weights are solved multiplying the covariances:

$$\lambda_0 = \mathbf{C^{-1}}.\mathbf{c_0}; \qquad C(|h| = 0) = C_0 + C_1 \tag{4}$$

where $\mathbf{C}$ is the covariance matrix derived for $N \times N$ observations and $\mathbf{c_0}$ is the vector of covariances at a new location. Note that the $\mathbf{C}$ is in fact a $(N+1) \times (N+1)$ matrix if it is used to derive kriging weights, since one extra row and column are used to ensure that the sum of weights is equal to one:

$$\begin{bmatrix} C_{(s_1,s_1)} & \dots & C_{(s_1,s_N)} & 1 \\ \vdots & \ddots & \vdots & \vdots \\ C_{(s_N,s_1)} & \dots & C_{(s_N,s_N)} & 1 \\ 1 & \dots & 1 & 0 \end{bmatrix}^{-1} \cdot \begin{bmatrix} C_{(s_0,s_1)} \\ \vdots \\ C_{(s_0,s_N)} \\ 1 \end{bmatrix} = \begin{bmatrix} w_1(s_0) \\ \vdots \\ w_N(s_0) \\ \varphi \end{bmatrix} \tag{5}$$

where φ is the *Lagrange multiplier*. After calculating the weights, the prediction is then given by Eq. 1.

When the experimental variogram is distinct for two or more directions, we have an anisotropic phenomenon [9]. The anisotropy is calculated considering a certain angle from 0 to 180°, and a factor given by

$$Anisotropy factor = \frac{a_2}{a_1} \tag{6}$$

where a_1 and a_2 are the biggest and smallest radius of the ellipse (area of effect in the kriging process), respectively. This factor varies between 0 and 1, with 1 being an isotropic model. Therefore, in case of anisotropy, five parameters are used to estimate the theoretical variogram model: nugget, sill, range, angle, and the anisotropy factor.

2.2 Population Diversity Index

The standard population diversity (SPD) describes the level of variation in a certain population. Greater diversity implies greater variability of the population [16]. Considering a population with P individuals (G_1, G_2,...,G_p), given that each individual (or particle) has T parameters (or genes), we can denote G_i = ($G_{i,1}$, $G_{i,2}$,...,$G_{i,T}$). So, the general mean for each gene T is given by

$$G_T^{ave} = \frac{1}{P}\sum_{i=1}^{P} G_{i,T} \tag{7}$$

In the normalization step, the standard deviation for each gene T in relation to the population P is calculated by

$$\sigma(G_T^{ave}) = \sqrt{\frac{1}{P}\sum_{i=1}^{P}(G_{i,T} - G_T^{ave})^2} \tag{8}$$

Finally, the variability of the population P in each generation of the bio-inspired algorithm is given by

$$SPD = \frac{1}{T}\sum_{j=1}^{T}\left(\frac{\sigma(G_j^{ave})}{G_j^{ave}}\right) \tag{9}$$

3 Methodology

The cluster-based kriging method [1] was adopted to evaluate GA and PSO algorithms. In this scenario, we first applied a preprocessing step using standardization algorithms and a statistical measure to remove outliers. After that, the K-means algorithm was used to find U groups of data. In this clustering step, the KNN algorithm was used to minimize the cluster overlapping problem, improving the clustering groups. For each group, the bio-inspired algorithms were used to find the kriging parameters, in other words, a model was built for each cluster u. Then, the unknown points (test data) were allocated via KNN to one of the previously built clusters and interpolated by the respective model.

The flow chart that describes the proposed methodology can be seen in Fig. 2. For each number of cluster (1 to 5), this process was repeated 10 times in the 10-fold cross-validation. Each iteration was performed using different training and test data sets.

3.1 Cost Function and Evaluation Metric

The fitness function used in GA and PSO algorithms was obtained by applying the kriging process at each data point (leave-one-out cross-validation) of the training data (90% of the database). Regarding the evaluation metric, the 10-fold cross-validation was applied. For both cases, fitness and evaluation, the

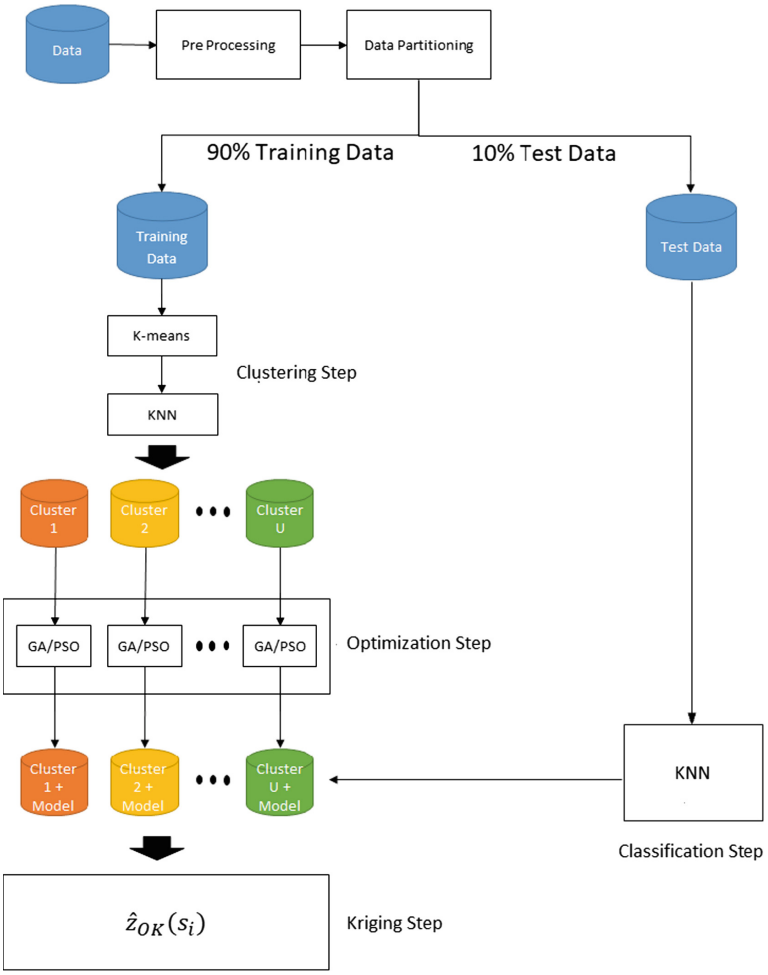

Fig. 2. Flowchart of the proposed methodology.

interpolation cost function (Eq. 10) was employed [10]. More specifically, the normalized mean squared error (NMSE) index was used as figure of merit and calculated by

$$NMSE_u = \frac{1}{\sigma^2.n}\sum_{i=1}^{n}[\hat{z}(s_i) - z(s_i)]^2 \tag{10}$$

where $\hat{z}(s_i)$ is the predicted value of the target variable obtained by the kriging method at the hidden point s_i; $z(s_i)$ is the real value of the target variable at the hidden point s_i; n is the total number of points in the cluster u; and σ^2 is the variance of the target variable considering the cluster u data. A lower NMSE

indicates a better prediction value. The NMSE index of the database is given by

$$NMSE = \sum_{u=1}^{U} NMSE_u \tag{11}$$

It is important to point out that the leave-one-out cross-validation was used only to calculate the fitness function of each solution (GA and PSO). In order to measure the accuracy, we applied the 10-fold cross-validation. So, the studied database was randomly partitioned in 90% for training and 10% for test in each iteration. In all, 10 iterations with different partitions were performed for each number of clusters. The average of these 10 tests was calculated in the end.

3.2 Data Preprocessing

In the data preprocessing step, the spatial information x and y, and the target variable, piezometric wells in this work, were normalized between 0 and 1. This procedure is important to ensure that every variable has the same weight in the clustering process and to avoid cluster overlapping. In the sequel, we used the Z-score measure with 99% confidence level to remove outliers.

3.3 Data Clustering

K-means [14] is one of the simplest unsupervised learning algorithms that solve the clustering problem. This clustering algorithm partitions the database into U clusters, where the user provides the value of U. In this work, the K-means method was chosen based on the solution proposed by [1], however, it is important to observe that any other clustering method can be used in the proposed methodology. Remembering that the main objective is to evaluate the performance and behavior of GA and PSO bio-inspired optimization methods.

After the preprocess and partition tasks, the training data was split into U clusters using the spatial information x and y, and the target variable. As shown in [1], the clustering process often results into overlapped clusters. This problem impairs the correct allocation of unknown (new) data into the clusters. So, in order to minimize the overlapping, all data were previously normalized between 0 and 1, and the KNN algorithm was applied in order to enhance the data grouping by allocating the current point based on the k neighbors. For example, the black circle highlighted in Fig. 3(a) demonstrates data overlapping, which was reduced with the application of the proposed methodology, as can be seen in Fig. 3(b).

3.4 Optimization Phase

For each cluster obtained by the K-means technique, we applied an optimization algorithm (GA and PSO) to find optimal kriging parameters: nugget, sill, range, factor, and angle. The accuracy of the interpolation are directly correlated to how good these parameters are. Each bio-inspired algorithm was evaluated based on the best set of parameters.

Fig. 3. Example of data clustering: (a) Original data clustering with overlapping and (b) Normalization + KNN data clustering.

3.5 Classification

The KNN algorithm was applied in order to classify the test data points into one of the previously built clusters based only on the spatial coordinates. Then, the kriging process was carried out and the error was calculated.

4 Experiments and Results

4.1 Database

The studied database represents the mountainous region of Wolfcamp Aquifer in West Texas/New Mexico [1]. This study area was already employed in other works [1,3] and classified as irregularly spaced with anisotropic data. This database contains 85 data points, including the spatial coordinates (x and y) and piezometric wells information (target variable).

4.2 Experiments

In order to evaluate and compare the results for the two studied bio-inspired algorithms, a manual tuning process was performed. The population size and the number of iterations were the same for both algorithms, and other specific parameters were manually tested. The final values are shown in Table 1. The GA code was designed on the R package GA [11] and the PSO code was also implemented using the R language programming.

For convenience, some parameters of the experimental variogram were fixed, such as the number of lags (= 1), the model type (= exponential), and the nugget effect (= 0). So, the chromosome (GA) and the particle (PSO) had the following variables: sill, range, angle, and factor. Their lower and upper bounds were defined, after the normalization step, as: range (= 0 to d); sill (= 0 to σ^2); angle (= 0 to 180°); and factor (= 0 to 1); where d is the maximum distance between two points and σ^2 is the target variable variance.

Table 1. GA and PSO parameters.

Algorithm	Parameter	Value
GA	Population Size	100
	Generations	50
	Crossover Probability	0.8
	Mutation Probability	0.1
	Elitism	5%
	Selection Method	Roulette
	Crossover Method	Real Coded [4]
	Mutation Method	Real Coded [5]
PSO	Particles	100
	Iterations	50
	Social Constant	2
	Cognitive Constant	2
	Inertia Range	0.4 to 0.8

GA and PSO fitness from one to five clusters in the optimization phase are described in Table 2. For more than one cluster settings, the results were obtained by summing the NMSE errors (Eq. 11). Because of the computational cost, 10 executions for each number of clusters were performed, each one using 90% of the database and the leave-one-out process. Note that the best results (or lowest errors) were obtained with five clusters for both algorithms and the average and standard deviation tend to decrease as the number of clusters increases. PSO always achieved better fitness than GA, but we can infer that GA is more stable, since PSO presented a higher standard deviation.

Table 2. Best fitness/NMSE, fitness average and fitness standard deviation.

		Number of clusters				
		1	2	3	4	5
GA	Best fitness	0.357	0.341	0.353	0.310	**0.287**
	Average	0.458	0.457	0.441	0.407	**0.381**
	Std. Dev.	0.098	0.051	0.036	0.038	0.044
PSO	Best fitness	0.354	0.299	0.314	0.269	**0.222**
	Average	0.739	0.625	0.537	0.513	**0.498**
	Std. Dev.	0.712	0.300	0.114	0.111	0.110

Figure 4 presents GA and PSO convergence curves from one to five clusters in the optimization phase. The results showed that PSO converges faster than

GA, in other words, PSO reached the best result found by GA before the tenth generation in most cases. Since it is a high computational cost process, this kind of information is relevant considering future works. This better convergence is probably explained by the fact that PSO presented populations with higher diversity level, as will be discussed in the sequel.

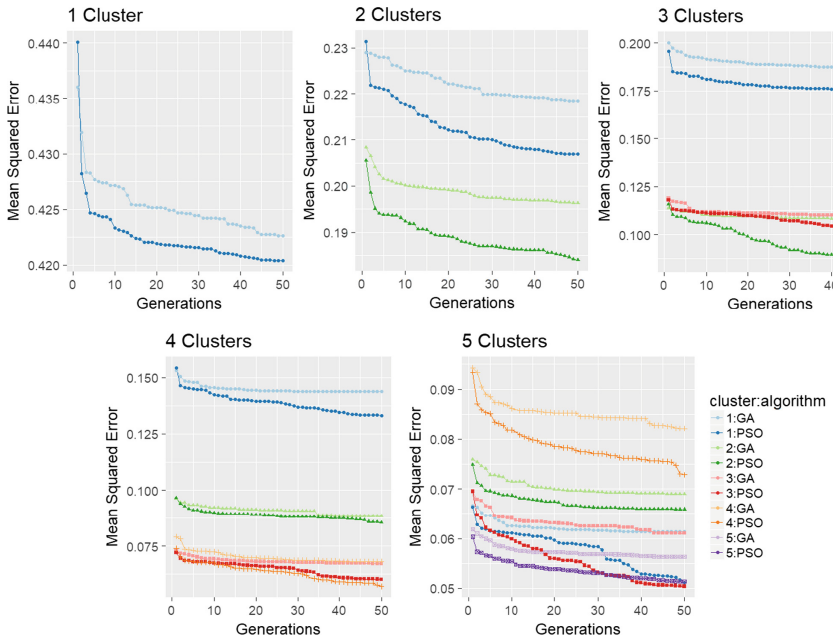

Fig. 4. Convergence charts for one to five clusters. Each line represents the average of 10 executions of GA or PSO.

In Fig. 5, we can see the standard population diversity (SPD) calculated for both algorithms. PSO had a much higher variation in the population than its counterpart GA. Obviously, a high diversity does not guarantee a better result, but it is a good indicator that the population is well spread out in the search space. Figure 5 shows the SPD values for the configuration with one cluster, just as an example, since the behavior was similar in the other settings.

Figure 6 presents the results of the classification step that were calculated using the 10-fold cross-validation process. More specifically, each iteration (= 10 for each number of clusters) used 10% of the database to test the kriging parameters previously estimated with the remaining 90%. The p-value obtained by the Friedman test [7] was 0.17, indicating that the error difference between GA and PSO was not statistically significant.

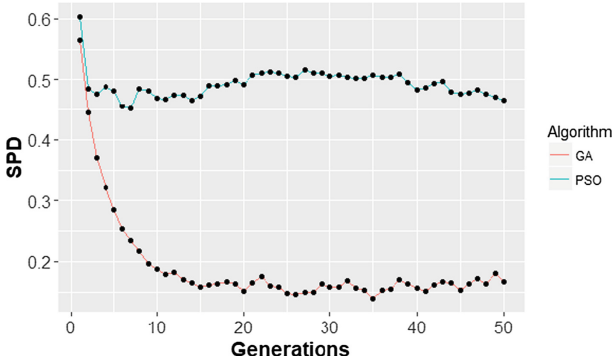

Fig. 5. SPD for GA and PSO. Average of 10 executions for only one cluster.

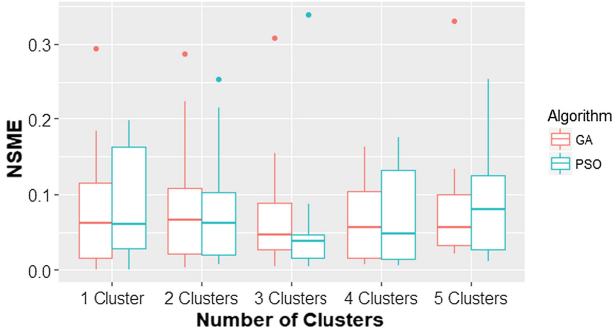

Fig. 6. Boxplot of NMSE considering the classification step for GA and PSO from one to five clusters. Average of 10 executions for each number of clusters.

5 Conclusions

The results obtained with the proposed methodology demonstrated that, based on the Friedman test, the evaluated algorithms (GA and PSO) are statistically equivalents when estimating the kriging parameters on the studied database. However, in the optimization phase, PSO converged faster than GA in all scenarios (1 to 5 clusters), which is an important conclusion.

Furthermore, exploring different parameters and customizing other operators could be interesting tasks to thoroughly assess the strengths of each method. Other topics that would add value to this research are: (i) reduce the computational processing time; (ii) test other techniques and databases in the proposed methodology; and (iii) discuss the impact of the clustering-based method on the stationary hypothesis. Stationary data is one whose statistical properties such as mean, variance, among others, are all constant over the spatial domain, which is suitable for the kriging process [9]. In [6], the author states that including the spatial coordinates in the clustering step, like the proposed methodology, does not guarantee the stationary hypothesis.

References

1. Abedini, M., Nasseri, M., Ansari, A.: Cluster-based ordinary kriging of piezometric head in west texas/new mexico-testing of hypothesis. J. Hydrol. **351**(3–4), 360–367 (2008)
2. Abedini, M., Nasseri, M., Burn, D.: The use of a genetic algorithm-based search strategy in geostatistics: application to a set of anisotropic piezometric head data. Comput. Geosci. **41**, 136–146 (2012)
3. Cressie, N.: Fitting variogram models by weighted least squares. J. Int. Assoc. Math. Geol. **17**(5), 563–586 (1985)
4. Deep, K., Thakur, M.: A new crossover operator for real coded genetic algorithms. Appl. Math. Comput. **188**(1), 895–911 (2007)
5. Deep, K., Thakur, M.: A new mutation operator for real coded genetic algorithms. Appl. Math. Comput. **193**(1), 211–230 (2007)
6. Fouedjio, F.: A spectral clustering approach for multivariate geostatistical data. Int. J. Data Sci. Anal. **4**(4), 301–312 (2017)
7. Gibbons, J.D., Fielden, J.D.G.: Nonparametric Statistics: An Introduction, vol. 90. Sage, Newbury Park (1993)
8. Gonçalves, Í.G., Kumaira, S., Guadagnin, F.: A machine learning approach to the potential-field method for implicit modeling of geological structures. Comput. Geosci. **103**, 173–182 (2017)
9. Hengl, T.: A Practical Guide to Geostatistical Mapping, vol. 52. Hengl, Amsterdam (2009)
10. Li, Z., Zhang, X., Clarke, K.C., Liu, G., Zhu, R.: An automatic variogram modeling method with high reliability fitness and estimates. Comput. Geosci. **120**, 48–59 (2018)
11. Scrucca, L., et al.: GA: a package for genetic algorithms in R. J. Stat. Softw. **53**(4), 1–37 (2013)
12. Wang, Z., Chang, Z., Luo, Q., Hua, S., Zhao, H., Kang, Y.: Optimization of riveting parameters using kriging and particle swarm optimization to improve deformation homogeneity in aircraft assembly. Adv. Mech. Eng. **9**(8) (2017). https://doi.org/10.1177/1687814017719003
13. Wei, Z., Liu, Z., Chen, Q.: GA-based kriging for isoline drawing. In: 2010 International Conference on Environmental Science and Information Application Technology (ESIAT), vol. 2, pp. 170–173. IEEE (2010)
14. Witten, I.H., Frank, E., Hall, M.A., Pal, C.J.: Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann, Cambridge (2016)
15. Xialin, Z., Zhengping, W., Zhanglin, L., Chonglong, W.: An intelligent improvement on the reliability of ordinary kriging estimates by a GA. In: 2010 Second WRI Global Congress on Intelligent Systems (GCIS), vol. 2, pp. 61–64. IEEE (2010)
16. Yasojima, E.K.K., de Oliveira, R.C.L., Teixeira, O.N., Pereira, R.L.: CAM-ADX: a new genetic algorithm with increased intensification and diversification for design optimization problems with real variables. Robotica **37**, 1–46 (2019)