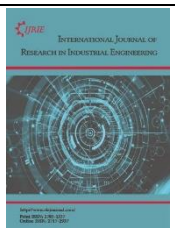
E-ISSN: 2717-2937 | P-ISSN: 2783-1337

International Journal of Research in Industrial Engineering

www.riejournal.com

Int. J. Res. Ind. Eng. Vol. 11, No. 4 (2022) 399–410.



Paper Type: Research Paper

A Genetic Algorithm for Curve Fitting by Spline Regression

Fatemeh Sogandi*

Department of Mechanical Engineering, University of Torbat heydarieh, Torbat Heydarieh, Khorasan Razavi, Iran; f.sogandi@torbath.ac.ir.

Citation:

Sogandi, F. (2022). A genetic algorithm for curve fitting by spline regression. *International journal of research in industrial engineering, 11*(4), 399-410.

Received: 24/01/2022 Reviewed: 24/02/2022 Revised: 09/03/2022 Accepted: 01/05/2022

Abstract

Curve fitting is a computational problem in which we look for a base objective function with a set of data points. Recently, nonparametric regression has received a lot of attention from researchers. Usually, spline functions are used due to the difficulty of the curve fitting. In this regard, the choice of the number and location of knots for regression is a major issue. Therefore, in this study, a Genetic Algorithm (GA) simultaneously determines the number and location of the knots based on two criteria. Those are the least square error and capability process indices. The proposed algorithm performance has been evaluated by some numerical examples. Simulation results and comparisons reveal that the proposed approach in curve fitting has satisfactory performance. Also, an example illustrated a sensitivity analysis of the number of knots. Finally, simulation results from a real case in Statistical Process Control (SPC) show that the proposed GA works well in practice.

Keywords: Capability process index, Genetic algorithm, Least square error, Spline regression.

1 | Introduction

With the development of technology in computation and measurement, scientists usually encounter providing information by curve fitting. If the pattern of a link between the predictor variable and response variable isn't known, or if there is no complete past information on the shape of a data pattern, nonparametric regression models are used. In situations in which there is a complex shape for measured data, piecewise polynomial functions or splines are the most important methods for smoothing in curve fitting. In the spline technique, the spline function must be continuous to higher derivatives at different points in the domain of function f. In this method, it is assumed that function f can be estimated by a continuous sequence of knots. Note that choosing the number and location of the knots is a challenge in data interpolation through spline regression.

As mentioned in Dierckx [1], the distribution of knots is a nonlinear optimization problem. To solve these problems, as one of the first works, Dimatteo et al. [2] used a Bayesian model in Markov Chain Monte Carlo (MCMC) in spline regression. Some studies carried out some computations based on

Corresponding Author: f.sogandi@torbath.ac.ir

https://dx.doi.org/10.22105/riej.2022.326256.1290

nonlinear optimization such as Ahmed et al. [3]. Also, Zhao et al. [4] proposed an adaptive knot placement using a generalized mix model-based continuous optimization algorithm based on a B-Spline curve approximation. Gazioglu et al. [5] developed penalized regression spline methodology which uses all the data and improves the precision of estimation. Also, Lai and Wang [6] applied the asymptotic behavior of penalized spline estimators using bivariate splines over triangulations and an energy functional as the penalty. After that, Seo et al. [7] proposed an outlier detection method in penalized spline regression models.

Furthermore, Schwarz and Krivobokova [8] developed a unified framework to investigate the properties of all periodic spline-based estimators, including regression, penalized, and smoothing splines. Moreover, Montoril [9] suggested a spline estimation of the functional-coefficient in regression models for time series with correlated errors. In addition, for time series nonparametric regression models with discontinuities, Yang and Song [10] used polynomial splines to estimate locations and sizes of jumps in the mean function. Papp and Alizadeh [11] applied a shape-constrained estimation using nonnegative splines. Moreover, Ma et al. [12] proposed a new method in spline regression in the presence of categorical predictors. In recent years, Zhou et al. [13] proposed the polynomial spline method to estimate a partial functional linear model. Recently, Daouia et al. [14] developed a novel constrained approach to the boundary curve achieved from the smoothness of spline approximation.

Usually, in most of these techniques, firstly, spline coefficients are estimated, and then the knots are selected. These methods display a relatively satisfactory performance; however, they are statistically complicated, and sometimes results fall in local solutions. Hence, some researchers have employed omission and addition techniques to estimate the knots. In this regard, Powell [15] produced extra knots in one variable, and Jupp [16] required an initial estimation of the location of the knots that is not feasible in practice. Similarly, Dierckx [1] needed an error tolerance or a smoothing factor to estimate the location of the knots at first. Ma [17] obtained a plug-in formula for the optimal number of interior knots based on the theoretical results of asymptotic optimality and strategies for choosing them in the spline estimator. Wang [18] treated the number and locations of knots as free parameters and used reversible jump MCMC to obtain posterior samples of knot configurations. In this work, second-order programming is used to estimate the remaining parameters based on the number and location of the knots.

On the other hand, metaheuristic algorithms are computational intelligence paradigms especially used for sophisticated solving optimization problems. For example, Engin and İşler [19] proposed a parallel greedy algorithm to solve the fuzzy hybrid flow shop problems with setup time and lot size. Also, Goli et al. [20] proposed a comprehensive model of demand prediction based on hybrid artificial intelligence and metaheuristic algorithms in the dairy industry. Moreover, Shahsavari et al. [21] suggested a novel GA for a flow shop scheduling problem with fuzzy processing time. On this subject, Sanagooy Aghdam et al. [22] proposed a heuristic method of GA and Simulated Annealing (SA) for the purpose of placing readers in an emergency department of a hospital. Recently, Khalili and Mosadegh Khah [23] presented a new mathematical optimization model using queuing theory to determine the hotel capacity in an optimal manner. On this subject, Rezaee and Pilevari [24] presented a mathematical model of sustainable multilevel supply chain using a meta-heuristic algorithm approach. Also, Alizadeh Firozi et al. [25] used an uncapacitated single allocation hub location problem for uncapacitated single allocation hub location problem.

In the field of application of GA in curve fitting, Irshad et al. [26] proposed a technique to capture the outline of planar objects based on two rational cubic functions for approximating the boundary curve using GA. It is worth mentioning that GA is an approach for an optimal selection of the number and location of the knots. This issue was introduced by Holland [27] for the first time. Afterward, Lee [28] changed the search space of the GA to a discontinuous type and used one-hot encoding to show the knots. In addition, Yoshimoto et al. [29] proposed a coded GA in curve fitting. In this regard, Pittman [30] showed the knots by integer coding. In his method, the number of the knots is assumed to be fixed

and only their location must be optimized. In addition, Tongur and Ülker [31] estimated curve knot points, which are found for curves by using Niched Pareto GA. Garcia et al. [32] proposed a hierarchical GA to counter the B-Spline curve interpolation problem. Their proposed approach helps identify the number and location of the knots, and it is capable of simultaneous determination of coefficients of the B-Spline function. Gálvez et al. [33] introduced an adapted elitist clonal selection algorithm for automatic knot adjustment of B-Spline curves that determines the number and location of knots to obtain accurate data. Garcia et al. [32] applied a hierarchical GA to tackle the B-Spline curve-fitting problem. Fengler and Hin [34] proposed a simple and general approach to fitting the discount curve under no-arbitrage constraints based on a penalized shape-constrained B-spline. Liu et al. [35] suggested jump-detection and curve estimation methods for the discontinuous regression function. Wu et al. [36] surveyed the problem of fitting scattered data points with ball B-Spline curves and then proposed a corresponding fitting algorithm based on the particle swarm optimization algorithm.

On this subject, Karadede and Özdemir [37] suggested a hierarchical soft computing model for estimating the parameter of curve fitting problems consisting of three stages. Afterward, Ramirez et al. [38] applied a parallel hierarchical Genetic Algorithm (GA) and B-splines to solve the curve-fitting problem of noisy scattered data using a multi-objective function. Recently, Li and Lily [39] proposed an approach based on an extreme learning machine algorithm to solve nonlinear curve fitting problems. Also, Yeh et al. [40] provided a new algorithm for curve fitting by a B-Spline of arbitrary order to determine the knot vector. They utilized a feature function that describes the amount and spatial distribution of the input curve.

Generally, the distribution of knots in splines is a nonlinear optimization problem. To solve this problem, researchers used some methods such as the Bayesian model, MCMC, generalized mix model-based continuous optimization algorithm, and penalized regression spline, the constrained approach to the boundary curve. Recently, some works used the GA algorithm to counter the spline curve interpolation problem. Choosing the number and location of the knots is an important challenge in data interpolation through spline regression. As aforementioned, a lot of attention has been given to estimating the number and location of the knots. Hence, in this study, a new GA has been employed based on three approaches, including Least Squares Error (LSE), Capability Process Index (CPI), and a combination of these two functions for curve fitting. In the rest of the paper, at first, the proposed GA is discussed in detail.

In the third section, the performance of the proposed approaches is evaluated, and three estimation methods of the proposed algorithm are compared. Afterward, a sensitivity analysis of the number of knots is illustrated by an example. Section 4 provides one of the applicability of the proposed method applied in change point estimation in the monitoring curve of the cooling equipment sales. Finally, the conclusion and further researches are given in the last section.

2 | Proposed Method

Consider a vector $x=\left[x_1, x_2, ..., x_n\right]$ fitted within different intervals with different response variables $y=\left[y_1, y_2, ..., y_p\right], p \le n$. This paper focuses on the estimation of the regression parameters. Assume that $\Delta=\left\{a=x_0<x_1<...<x_n=b\right\}$ is a partition on $I=[a,b]$ interval in which distances between points are not necessarily equal. A spline is a function that is constructed piecewise from polynomial functions. To provide a visual interpretation, a schematic concept of spline regression is shown below:

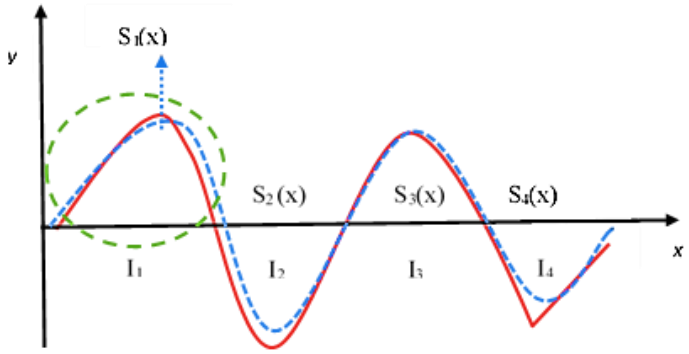


Fig. 1. A demonstration of spline regression for curve fitting.

"In general, the setup of fixed knots is an arbitrary restriction of the set of available spline curves" [29]. Therefore, it is first considered that fixed knots to solve the problem. It is worth mentioning that the proposed method is a flexible model for normal data in which residuals are normally distributed. To construct the chromosomes showing knots vector, Haupt and Haupt [41] is appllied. In the GA scheme, initially, a population of chromosomes is randomly produced in which each chromosome is an answer vector for the knots vector. After that, the selected chromosomes in the initial population are replaced with new chromosomes obtained from mutation and crossover operators. We estimate the regression parameters in each interval of x. Then, for population, the objective function is calculated using three approaches including least squared error and CPI and the combination of them, which are explained briefly in the following. Then, a certain number of the parent chromosomes are selected from the initial population. The selection and replacement processes keep on till the completion of the algorithm.

The coefficients of the spline functions can be estimated by least squares regression. This method is used when the type of distribution is exclusively normal. In this method, a certain value for x is the best-predicted value for y and $f(x)$.

$$y = f(x) + \text{noise}. \tag{1}$$

In which the function f is called regression. Now, the parameters of the distribution are based on the minimizing sum of $(f(x)-y)^2$ for all observations. It is also important to verify, in a residual analysis that if the assumptions of the white noise residuals are satisfied, we will be sure that the model well fitted.

On another side, the detection and the examination of outliers are important parts of data analysis because some outliers in the data may have a detrimental effect on statistical analysis. Many authors have discussed outlier detection methods. In this regard, we utilized CPI method. In this method, the estimated regression parameters are obtained by applying CPI to the residual. In other words, using this index, the process compares the output of the controlled process with the quality specification limits. The comparison is made by the ratio of the standard variation of samples from the residuals to 6 times of the standard variation residual. There are several statistics can be used to measure the capability of a process: Cp, Cpk, and Cpm. For the sake of simplicity, the CPI of Cp is used as one of the objective functions of GA. Assume that there is a two-sided specification, and USL and LSL are the upper and lower specification limits value, respectively. In Statistical Process Control (SPC), Cp is calculated as follows:

$$C_p = \left(\frac{USL - LSL}{6\sigma}\right). \tag{2}$$

As can be seen in simulation studies, using applying separately two mentioned objective functions leads to satisfactory results. It is worth mentioning that considering simultaneously these approaches have more desirable answers than separately. Applying the least-squares criterion may remove a chromosome

that has only a few outliers. On the other hand, GA based on the objective function of CPI may give an answer in which all response variables are near to the control limit, while total error is unneglectable. According to this concept and conducting some simulation results, it can be easily concluded two approaches act against each other. Hence combining them with a weighting method leads to a better solution. The proposed method based on the two estimation functions is as follows:

$$\alpha\left(f\left(x\right)-y\right)^{2}-\beta\left(\frac{USL-LSL}{6\sigma}\right). \quad (3)$$

Where two parameters of α and β are the weights for two objective functions of least squares criterion and the CPI, respectively.

2.1 | Proposed GA Pseudo-Code to Estimate the Spline Regression

Succinctly, the recommended pseudo-code of the GA algorithm is given in the following.

I. Determining the input parameters.
II. Producing the primary population from 0 and 1 values.
III. Transforming the 0 and 1 values to D vector or knots vector. (D vector is defined clearly in [34]).
IV. Estimating regression parameters using LSE for each interval of D vector of the population.
V. Calculating the fitting function according to *Eq. (3)*.
VI. Comparing the minimum function as the best answer with the best answer in the previous replication.
VII. Creating a new population according to the following method:

- *Select two vectors from D vectors from the current population using the selection operator with the occurrence probability of pc.*
- *Transforming the selected vectors D to 0 and 1 vectors and applying the crossover operator to create new vectors.*
- *Select a vector from D vectors using the selection operator with the occurrence probability of pm.*
- *Transforming the selected D vectors to 0 and 1 vectors and applying mutation operator on the new vector.*
- *Adding the new population to the initial population, as well as pr times the initial population.*

3 | Performance Evaluation of the Proposed Method

In this subsection, we present some simulation results to evaluate the performance of the proposed method. In this regard, we use some multiple linear regression examples. In this method, we should obtain parameters such as percentage of crossover, and mutation operators to enhance the efficiency of the proposed GA. To achieve this goal, the trial and error method is employed to adjust parameters. In this regard, the given parameter including crossover operation percentage is tuned 65 percent and mutation operation is 20 percent (pm=0.2, pc=0.65). Similarly, the rate of selecting superior answers from the initial population is 5 percent, (pr=0.05). The chromosomes of the population are assumed 100. Genes in each chromosome and the particles in each gene are equal to 20 and 7, respectively. In addition, the replications in each run are equal to 3000. The assumed functions with determined knots are shown in *Table 1*. In this regard, we minimize the total error based on only the least squared error. Afterward, it is similarly assumed that maximizing the CPI is the objective function. Then, the combination of the two functions is considered. It should be noted that we use a Mean Squared Error (MSE) criterion to appraise and compare proposed methods with different objective functions

Table 1. The assumed functions in each interval created by knots.

Number	Functions
1	$Y_1=x_1+3x_2+5x_3+8x_4$
2	$Y_2=x_1+8x_2+7x_3+3x_4$
3	$Y_3=2x_1+4x_2+6x_3+6x_4$
4	$Y_4=2x_1+3x_2+1x_3+9x_4$
5	$Y_5=2x_1+5x_2+7x_3+3x_4$
6	$Y_6=2x_1+3x_2+6x_3+5x_4$
7	$Y_7=x_1+4x_2+2x_3+8x_4$
8	$Y_8=x_1+2x_2+7x_3+2x_4$
9	$Y_9=4x_1+8x_2+9x_3+8x_4$
10	$Y_{10}=x_1+3x_2+6x_3+6x_4$
Knot vector	[17,36,43,43,54,58,62,69,93,100]

Objective functions based on the LSE and the CPI separately may have satisfactory results. However, they may not achieve the appropriate fitting. Because, as expected, in the least squared error, the emphasis is on the total deviations. On the other hand, in some cases, we may witness excessive error while in the majority of points; the fitting has been well achieved. Therefore, in this situation, this approach is not able to select an appropriate model for curve fitting. On the contrary, when the CPI is used, may there be considerable errors in the majority of points with fewer outliers. With fewer outliers. As a result, combining two functions and employing them simultaneously is acceptable to achieve more appropriate fitting and lower MSE criterion. As mentioned before, the assumed knot vector is considered in *Table 1*. Also, the knot vector obtained from the proposed method is illustrated as an estimated knot vector in 11 examples in *Tables 2, 3*, and *4*.

In *Tables 2* and *3*, an example has been provided with a determined input variable in each run to obtain the optimal solution for one of the objective functions (CPI and LSE, respectively) and the other objective function calculated with the optimal population. In this regard, its final population has been assumed as the population of another objective function to perform once again. Moreover, the three approaches including objective function based on the LSE, CPI, and the combination of them are applied to the mentioned method in *Table 4*.

Comparing the results in *Tables 2, 3,* and *4* shows that the obtained vector of the solution is close to the initial knot vector. We can confirm that considering simultaneous two objective functions improves substantially the performance of the algorithm. In summary, the simulation results indicate that the proposed method is capable of handling behaviors of a wide range of observations on the sub-intervals.

Table 2. Superior performance of CPI approach to the LSE approach.

	CPI Function	Corresponding LSE Function
MSE	58.70	69.90
Knot points	[17, 36, 43, 43, 54, 58, 62, 69, 93, 100]	[17, 25, 43, 55, 60, 70, 78, 87, 91, 100]
MSE	299.70	305.40
Knot points	[17, 26, 58, 65, 67, 89, 97, 100, 100, 100]	[15, 25, 59, 68, 69, 82, 100, 100, 100, 100]
MSE	121.60	769.90
Knot points	[1, 6, 22, 46, 48, 60, 86, 91, 100, 100]	[12, 60, 67, 87, 91, 92, 100, 100, 100, 100]

Table 3. Superior performance of the LSE approach to the CPI approach.

	LSE Function	Corresponding CPI Function
MSE	152.20	243.10
Knot Points	[33, 35, 39, 52, 55, 80, 90, 100, 100, 100]	[33, 44, 58, 63, 69, 78, 86, 90, 100, 100]
MSE	147.50	194.10
Knot Points	[1, 9, 12, 28, 45, 68, 74, 82, 92, 100]	[1, 9, 12 28, 34, 45, 69, 85, 92, 100]
MSE	64.40	86.80
Knot Points	[12, 20, 39, 45, 55, 76, 77, 100, 100, 100]	[12, 34, 45, 54, 69, 73, 82, 83, 86, 100]

Table 4. Comparison of the proposed methods.

	CPI Function	LSE Function	CPI+LSE
MSE	79.90	0.1	0
Knot points	[8, 20, 27, 36, 68, 71, 85, 87, 95, 100]	[17, 25, 35, 43, 55, 60, 70, 87, 91, 100]	[17, 25, 35, 43, 55, 60, 70, 87, 90, 100]
MSE	41.60	0.3	0
Knot points	[15, 40, 41, 44, 59, 69, 71, 93, 94, 100]	[17, 25, 35, 43, 56, 61, 70, 87, 91, 100]	[17, 25, 35, 43, 55, 60, 70, 87, 90, 100]
MSE	26.70	0.1	0
Knot points	[9, 25, 29, 39, 54, 58, 71, 79, 81, 100]	[17, 25, 35, 43, 55, 60, 70, 86, 90, 100]	[17, 25, 35, 43, 55, 60, 70, 87, 90, 100]
MSE	970.90	2.7	0.1
Knot points	[12, 68, 84, 89,91, 92, 97, 100, 100, 100]	[17, 25, 34, 38, 55, 60, 70, 87, 91,100]	[17, 25, 35, 43, 55, 60, 70, 87, 91, 100]
MSE	67.50	0.1	0
Knot points	[31, 33, 36, 40, 58, 59, 61, 70, 85, 100]	[17, 25, 35, 43, 55, 60, 70, 87, 91, 100]	[17, 25, 35, 43, 55, 60, 70, 87, 90, 100]

Table 5. The effect of the number of knots on total square error (MSE values).

Number of Knots	CPI Function	LSE Function	Combining the Two Functions
2	297.00	249.60	220.40
3	256.20	266.70	200.60
4	184.30	189.00	166.20
5	172.40	184.0	92.30
6	162.80	147.70	86.10
7	102.10	81.60	68.50
8	97.30	60.90	59.70
9	45.90	48.50	36.50
10	26.70	0.1	0

3.1 | Sensitivity Analysis of the Number of Knots

First, we use simulations to demonstrate that our method is not sensitive to some knots. We assume the numerical example and the mentioned parameters in the previous section, the results of *Table 5* indicate that the combined method has better performance than the other methods. As we expected, as the number of knots increases, the performance of all the proposed methods increases. In addition, as shown in *Fig. 4*, the more the number of knots is assumed, the fewer differences between the methods LSE and CPI. Note that when the number of knots is more than one, the proposed method can be used.

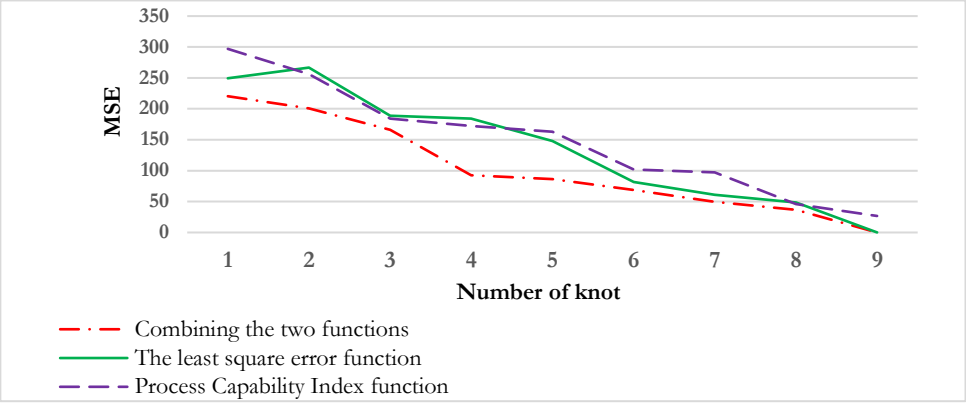


Fig. 4. Comparison of the performance of the proposed methods.

Real world numerical data is usually difficult to analyze. To this end, the idea of the applying GA in spline is developed. As shown in *Tables 2-4*, simulation results confirm that proposed GA help substantially to spline curve fitting. Using the proposed algorithm, a series of unique polynomials are fitted between each of the data points, with the stipulation that the curve obtained be continuous and appear smooth. There are many cases in which spline plays an important role in data analysis. Wherever spline is used, the proposed algorithm can be attractive and help improve splines. Hence, the proposed method can be used in various fields from a managerial point of view.

4 | A Comparative Study

Due to the lack of an analytic expression for optimal knot locations, different methodologies in the specialized literature have been demonstrated for the selection and optimization of knot vectors. Some fast deterministic methods employ. However, in the case of complex point clouds, the results are far away from the optimum. Alternatively, metaheuristic methods especially the GA algorithm yield knot vectors which are very close to the optimum, but only converge slowly and are, therefore, time- and computing power-consuming. Furthermore, the performance of these algorithms is seriously affected by the occurrence of data gaps. Recently, Bureick et al. [42] proposed an elitist GA to solve the knot adjustment problem for B-Spline curves despite the possible occurrence of data gaps. It is worth mentioning that we focused on the determination of knot location and knot vector size. In reality, we try to realize model selection and knot vector determination simultaneously. By contrast, Bureick et al. [42] focused solely on knot vector determination.

To evaluate the efficiency of proposed algorithm, our method and the elitist GA are applied to a test function. To evaluate the capability of the proposed algorithm, the test functions are introduced in Yoshimoto et al. [29] according to *Eq. (4)*. The chosen parameters for both algorithms to obtain the subsequent results are gathered mainly from the literature.

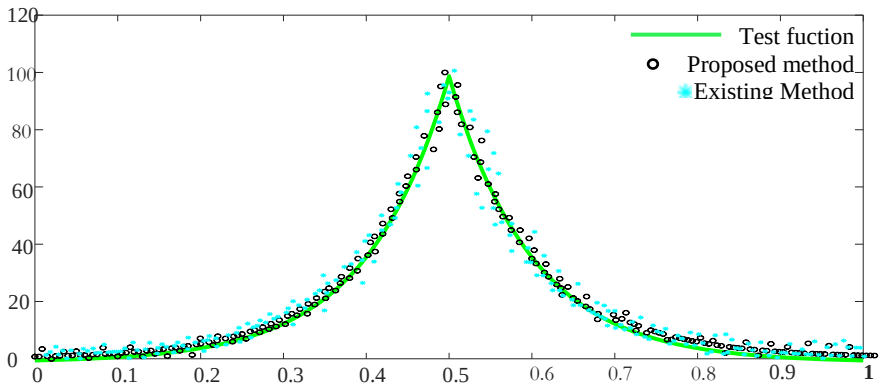


Fig. 5. Comparison results for test function with given parameters.

$$y = \frac{100}{e^{|10x-5|}} + \frac{(10x-5)^5}{500}. \quad (4)$$

Generally, the simulation results show that the proposed GA algorithm is a simple method and performs slightly better than comparative methods. However, elitist GA solves the knot adjustment problem in a faster manner than the proposed approach.

5 | Application

The estimation of locations of the knots in spline functions can be used in different applications. Recently, Toutounji and Durstewitz [43] have utilized this concept and detected multiple change points using adaptive regression splines with application to neural recordings. One another application is in SPC. In this respect, control charts are one of the most important tools in monitoring a process, but most control charts delay warning alarm of a change in the process. The real-time in which the process changes is called a change point. To save time and cost, its estimation is an important issue in SPC. In this regard, if the knots are assumed as the change points, the multiple change points in monitoring the qualitative characteristic can be estimated through estimation of the number and location of the knots. Hence, multiple change points can be considered an interesting application. So far, some studies have been conducted to estimate the multiple change points that are restricting assumptions such as fixity in the number of change points.

In this subsection, we specifically illustrate the implementation of our method in the estimation of multiple change points. Note that the knot numbers obtained from the proposed method are illustrated as the estimated change point vector. To validate the algorithm in this example, we consider the multiple change point vectors within the given ranges as the determined change point vector in *Table 6*. As can be seen, different functions of the simple linear regression model are assumed in this table. Now, simulation results are calculated using the proposed algorithm and the values of pm=0.2 and pc=0.65 are obtained by adjusting the algorithm parameters. The MSE for the CPI, the least squared error, and the combination of the two functions are equal to 72.40, 39.20, and 37.6, respectively. The change points are illustrated in *Table 6* to show the applicability of the proposed method in this real example. As shown in this table, the estimations of change points obtained by the proposed method are close to the determined change points.

Table 6. A real example of the multiple change points.

	Simple Linear Profile Functions within Different Ranges of the Independent Variable
1	$y_1 = 3 + x$
2	$y_2 = 8 + 6x$
3	$y_3 = 4 + 2x$
4	$y_4 = 2 + 3x$
5	$y_5 = 5 + 2x$
6	$y_6 = 3 + 2x$
7	$y_7 = 4 + x$
8	$y_8 = 2 + 7x$
9	$y_9 = 8 + 4x$
10	$y_{10} = 3 + 9x$
	Determined change points
	[17, 25, 35, 43, 55, 60, 70, 87, 90, 100]
	The estimations of change points by the proposed method
	[7, 19, 31, 47, 56, 62, 69, 81, 91, 100]

6 | Conclusion

Among nonparametric, a spline model is one of the regression models with considerable statistical interpretation. However, this method requires the identification of knots. In this regard, we proposed a

regression spline based on a GA that is intuitively appealing and simple. The proposed algorithm is provided to estimate the number and the location of the knots simultaneously. The proposed algorithm has the specified ability to handle data with changeable behaviors on certain sub-intervals. The proposed GA is based on the LSE considering CPI. The performance of the proposed method was evaluated using several numerical examples. Also, it was shown that the more the number of knots is assumed, the fewer differences between the methods LSE and CPI. Note that this algorithm can be used for Normal-type of observations with normal residuals. In the end, the applicability of the proposed algorithm was shown using a real example.

For further researche, functions and methods defined in spline regression may be applied in the objective function to minimize total error. The dependence of the proposed method on residuals distribution may also be eradicated by applying other nonparametric regression. In addition, a new GA can be provided for non-Normal data.

Funding

This work was financially supported by the Research Deputy of Education and Research, University of Torbat Heydarieh. The grant number is 122 22/01/2022.

Conflicts of Interest

All co-authors have seen and agree with the contents of the manuscript and there is no financial interest to report. We certify that the submission is original work and is not under review at any other publication.

References

[1] Dierckx, P. (1995). *Curve and surface with splines*. Oxford University Press.

[2] DiMatteo, I., Genovese, C. R., & Kass, R. E. (2001). Bayesian curve-fitting with free-knot splines. *Biometrika*, *88*(4), 1055-1071. https://doi.org/10.1093/biomet/88.4.1055

[3] Ahmed, S. M., Biswas, T. K., & Nundy, C. K. (2019). An optimization model for aggregate production planning and control: a genetic algorithm approach. *International journal of research in industrial engineering*, *8(3)*, 203-224. http://dx.doi.org/10.22105/riej.2019.192936.1090

[4] Zhao, X., Zhang, C., Yang, B., & Li, P. (2011). Adaptive knot placement using a GMM-based continuous optimization algorithm in B-Spline curve approximation. *Computer-aided design*, *43*(6), 598-604. https://doi.org/10.1016/j.cad.2011.01.015

[5] Gazioglu, S., Wei, J., Jennings, E. M., & Carroll, R. J. (2013). A note on penalized regression spline estimation in the secondary analysis of case-control data. *Statistics in biosciences*, *5(2)*, 250-260. https://doi.org/10.1007/s12561-013-9094-9

[6] Lai, M. J., & Wang, L. (2013). Bivariate penalized splines for regression. *Statistica sinica*, 1399-1417. http://dx.doi.org/10.5705/ss.2010.278

[7] Seo, H. S., Song, J. E., & Yoon, M. (2013). An outlier detection method in penalized spline regression models. *Korean journal of applied statistics*, *26*(4), 687-696. https://doi.org/10.5351/KJAS.2013.26.4.687

[8] Schwarz, K., & Krivobokova, T. (2016). A unified framework for spline estimators. *Biometrika*, *103*(1), 121-131. https://doi.org/10.1093/biomet/asv070

[9] Montoril, M. H., Morettin, P. A., & Chiann, C. (2014). Spline estimation of functional coefficient regression models for time series with correlated errors. *Statistics & probability petters*, *92*(1), 226-231. https://doi.org/10.1016/j.spl.2014.05.021

[10] Yang, Y., & Song, Q. (2014). Jump detection in time series nonparametric regression models: a polynomial spline approach. *Annals of the institute of statistical mathematics*, *66*(2), 325-344. https://doi.org/10.1007/s10463-013-0411-3

[11] Papp, D., & Alizadeh, F. (2014). Shape-constrained estimation using nonnegative splines. *Journal of computational and graphical statistics*, *23*(1), 211-231. https://doi.org/10.1080/10618600.2012.707343

[12] Ma, S., Racine, J. S., & Yang, L. (2015). Spline regression in the presence of categorical predictors. *Journal of applied econometrics*, *30*(5), 705-717. https://doi.org/10.1002/jae.2410

[13] Zhou, J., Chen, Z., & Peng, Q. (2016). Polynomial spline estimation for partial functional linear regression models. *Computational statistics*, *31*(3), 1-23. https://doi.org/10.1007/s00180-015-0636-0

[14] Daouia, A., Noh, H., & Park, B. U. (2016). Data envelope with constrained polynomial splines. *Journal of the royal statistical society: series b (statistical methodology)*, *78*(1), 3-30. https://doi.org/10.1111/rssb.12098

[15] Powell, J. D. (1970). Curve fitting by splines in one variable. *Numerical approximation to functions and data*, *12*(1), 65-83. https://doi.org/10.2307/2316601

[16] Jupp, D. L. (1978). Approximation to data by splines with free knots. *SIAM journal on numerical analysis*, *15(2)*, 328-343. https://doi.org/10.1137/0715022

[17] Ma, S. (2014). A plug-in the number of knots selector for polynomial spline regression. *Journal of nonparametric statistics*, *26*(3), 489-507. https://doi.org/10.1080/10485252.2014.930143

[18] Wang, X. (2008). Bayesian free-knot monotone cubic spline regression. *Journal of computational and graphical statistics*, *17*(2), 373-387. https://doi.org/10.1198/106186008X321077

[19] Engin, O., & İşler, M. (2021). An efficient parallel greedy algorithm for fuzzy hybrid flow shop scheduling with setup time and lot size: a case study in apparel process. *Journal of fuzzy extension and applications*, *3*(3), 249-262. http://dx.doi.org/10.22105/jfea.2021.314312.1169

[20] Goli, A., Zare, H. K., Moghaddam, R., & Sadeghieh, A. (2018). A comprehensive model of demand prediction based on hybrid artificial intelligence and metaheuristic algorithms: a case study in dairy industry. *Journal of industrial and systems engineering*, *11*(1), 190-203. https://doi.org/10.1080/10485252.2014.930143

[21] Shahsavari, N., Abolhasani, M. H., Sheikhi, H., Mohammadi Andargoli, H., & Abolhasani, H. (2014). A novel Genetic algorithm for a flow shop scheduling problem with fuzzy processing time. *International journal of research in industrial engineering*, *3*(4), 1-12. https://doi.org/10.1007/s10463-013-0411-3

[22] Sanagooy Aghdam, A., Kazemi, M. A. A., & Eshlaghy, A. T. (2021). A hybrid GA–SA multiobjective optimization and simulation for RFID network planning problem. *Journal of applied research on industrial engineering*, *8*(Spec. Issue), 1-25. http://dx.doi.org/10.22105/jarie.2021.295762.1357

[23] Khalili, S. & Mosadegh Khah, M. (2020). A new queuing-based mathematical model for hotel capacity planning: a Genetic algorithm solution. *Journal of applied research on industrial engineering*, *7*(3), 203-220. http://dx.doi.org/10.22105/jarie.2020.244708.1187

[24] Rezaee, F., & Pilevari, N. (In Press). Mathematical model of sustainable multilevel supply chain with meta-heuristic algorithm approach (case study: atmosphere group: industrial and manufacturing power plant). *Journal of decisions and operations research*, *7*(Spec. Issue), 1-17. DOI: 10.22105/dmor.2021.270853.1310

[25] Alizadeh Firozi, M., Kiani, V., & Karimi, H. (2022). Improved Genetic algorithm with diversity and local search for uncapacitated single allocation hub location problem. *Journal of decisions and operations research*, *6*(4), 536-552. http://dx.doi.org/10.22105/DMOR.2021.272989.1325

[26] Irshad, M., Khalid, S., Hussain, M. Z., & Sarfraz, M. (2016). Outline capturing using rational functions with the help of GA. *Applied mathematics and computation*, *274*(1), 661-678. https://doi.org/10.1016/j.amc.2015.10.014

[27] Holland, J. H. (1975). *Adaptation in natural and artificial systems*. University of Michigan Press. https://doi.org/10.7551/mitpress/1090.003.0007

[28] Lee, T. (2002). On algorithms for ordinary least squares regression spline: a comparative study. *Journal of statistical computation and simulation*, *72(8)*, 647-663. https://doi.org/10.1080/00949650213743

[29] Yoshimoto, F., Harada, T., & Yoshimoto, Y. (2003). Data with a spline using a real-coded Genetic Algorithm. *Computer-aided design*, *35*(8), 751-760. https://doi.org/10.1016/S0010-4485(03)00006-X

[30] Pittman, J. (2002). Adaptive splines and Genetic Algorithms. *Journal of computational and graphical statistics*, *11(3)*, 615-638. https://doi.org/10.1198/106186002448

[31] Tongur, V., & Ülker, E. (2016). B-Spline curve knot estimation by using niched Pareto genetic algorithm (npga). In *Intelligent and evolutionary systems* (pp. 305-316). Springer, Cham.

[32] Garcia, C. H., Cuevas, F. J., Trejo-Caballero, G., & Rostro-Gonzalez, H. (2015). A hierarchical GA approach for curve fittingwith B-splines. *Genetic programming and evolvable machines*, *16(2)*, 151-166. https://doi.org/10.1007/s10710-014-9231-3

[33] Gálvez, A., Iglesias, A., Avila, A., Otero, C., Arias, R., & Manchado, C. (2015). Elitist clonal selection algorithm for optimal choice of free knots in B-Spline data. *Applied soft computing*, *26(1)*, 90-106. https://doi.org/10.1016/j.asoc.2014.09.030

[34] Fengler, M. R., & Hin, L. Y. (2015). A simple and general approach to fitting the discount curve under no-arbitrage constraints. *Finance research letters, 15*(1), 78-84. https://doi.org/10.1016/j.frl.2015.08.006

[35] Liu, G. X., Wang, M. M., Du, X. L., Lin, J. G., & Gao, Q. B. (2018). Jump-detection and curve estimation methods for discontinuous regression functions based on the piecewise B-Spline function. *Communications in statistics-theory and methods*, *47*(23), 5729-5749. https://doi.org/10.1080/03610926.2017.1400061

[36] Wu, Z., Wang, X., Fu, Y., Shen, J., Jiang, Q., Zhu, Y., & Zhou, M. (2018). Fitting scattered data points with ball B-Spline curves using particle swarm optimization. *Computers & graphics, 72(1),* 1-11. https://doi.org/10.1016/j.cag.2018.01.006

[37] Karadede, Y., & Özdemir, G. (2018). A hierarchical soft computing model for parameter estimation of curve fitting problems. *Soft computing*, *22*(20), 6937-6964. https://doi.org/10.1007/s00500-018-3413-5

[38] Ramirez, L., Edgar, J., Capulin, C. H., Estudillo-Ayala, M., Avina-Cervantes, J. G., Sanchez-Yanez, R. E., & Gonzalez, H. R. (2019). Parallel hierarchical Genetic algorithm for scattered data fitting through B-Splines. *Applied sciences*, *9*(11), 2336. https://doi.org/10.3390/app9112336

[39] Li, M., & Lily, D. L. (2020). A novel method of curve fitting based on optimized extreme learning machine. *Applied artificial intelligence*, *34*(12). 849-865. https://doi.org/10.1080/08839514.2020.1787677

[40] Yeh, R., Youssef, S. N., Peterka, T., & Tricoche, X. (2020). Fast automatic knot placement method for accurate B-Spline curve fitting. *Computer-aided design, 128*(1). 102905. https://doi.org/10.1016/j.cad.2020.102905

[41] Haupt, R. L., & Haupt, S. E. (2004). *Practical genetic algorithms*. John Wiley & Sons, New York.

[42] Bureick, J., Alkhatib, H., & Neumann, I. (2019). Fast converging elitist genetic algorithm for knot adjustment in B-Spline curve approximation. *Journal of applied geodesy*, *13*(4), 317-328. https://doi.org/10.1515/jag-2018-0015

[43] Toutounji, H., & Durstewitz, D. (2018). Detecting multiple step changes using adaptive regression splines with application to neural recordings. *Frontiers in neuroinformatics*, *12*(1), 210-231. https://doi.org/10.3389/fninf.2018.00067