JEAS ,Vol. 31 No. 149 (2025) pp (58-72)
P-ISSN 2518-5764 - E-ISSN 2227-703X
DOI: https://doi.org/10.33095/8mjedd16

Journal of Economics and Administrative Sciences

Using some artificial intelligence algorithms to estimate the parametric regression function for spatially dependent data of water pollution of the Euphrates River

Ons Edin Musa *
College of Physical Education and Sports Science,
Mustansiriyah university, Baghdad, Iraq.

***Corresponding author**

Sabah Manfi Redha
Department of Statistics
College of Administration and Economics
University of Baghdad, Baghdad, Iraq.

Received:18/3/2025 **Accepted:15/7/2025** **Published: 1/10/2025**

Abstract:

These models account for the spatial effects resulting from the proximity of events. A compromise exists in the mathematical accuracy of model parameters when spatial correlations are present in the data of the phenomenon. Data reliant on spatial correlations are crucial in statistical modelling, especially in environmental science, economics, epidemiology, and various other disciplines.

This study employs and compares three artificial intelligence approaches—the genetic algorithm (GA), the TABU search algorithm (TSA), and the binary firefly algorithm (Binary FFA)—to determine which is the most efficient for estimating the parametric regression function for spatially dependent data.

The Mean Absolute Percentage Error numbers derived from the simulation demonstrated that the Binary FFA method yielded the most accurate estimations. This illustrates the superiority of the algorithm compared to conventional methods, as well as Genetic Algorithms (GA) and Tabu Search Algorithms (TSA), in environmental assessments (particularly, water pollution in the Euphrates River) and the estimation of regression models for geographically dependent data.

The regression parameter analysis for spatially dependent environmental data about Euphrates River pollution indicates that the temperature variablity exerted little influence on total dissolved salts. Conversely, the variables calcium (Ca), magnesium (Mg), and potassium (K) exhibited a significant and advantageous influence on total dissolved salts. In contrast, the variable sodium (Na) displayed a distinctly detrimental effect simultaneously.

Keywords: Artificial intelligence, Genetic algorithm, Tabu search algorithm, Binary Firefly algorithm, spatially dependent data, Euphrates Riner water pollution.

1. Introduction:

Artificial intelligence has many applications in practical, natural, scientific, and human life, as artificial intelligence algorithms depend on the principles and concepts of artificial intelligence. Artificial intelligence algorithms rely on the idea of artificial intelligence, as they are distinguished by their ability to create dynamic methods that suit the nature of the problem to be studied and determine the appropriate practical method to find the optimal solution from among the set of solutions to the problem. These algorithms improve the solution's value according to the obstacles and variables.

The topic of spatial cross-sectional series of spatial economic measurement models is one of the types of standard models that have the same characteristics as time series. These models are concerned with the spatial effects between observations of phenomena due to proximity in place. In contrast, traditional economic measurement models are concerned with the reliability of observations at a certain period. The presence of spatial correlations in phenomena data is a significant problem due to the presence of spatial juxtapositions, which are not considered when studying these phenomena, as these correlations affect the accuracy of the estimated model parameters. We must search for standard models that consider spatial correlations and use estimation methods that address this problem and give efficient estimates.

2. Literature Review and Hypothesis Development:

Some authors who have written on spatially dependent economic data will be mentioned. Dubin (1988) estimated the correlation function parameters and regression coefficients using a maximum likelihood method, and spatial autocorrelation was tested. The study found spatial autocorrelation between community members and their locations (Dubin, 1988). Prucha and Kelejian (1998) Employed Generalized Spatial Two-Stage Least Squares (GS2SLS) to estimate the parameters of a linear regression model utilizing a spatially lagged dependent variable. This study on the distribution of large samples in an estimator analysis (GS2SLS) demonstrated that the estimator is consistent and exhibits natural convergence (Kelejian & Prucha, 1998). Duczmal et al. (2007) utilized a genetic methodology to deduce irregular geographical groupings. Rapid atomic creation and Kuldrov's spatial scanning statistic assessment diminish graph-related work. The geometric incompressibility penalty function reduces irregularities in cluster geometric shapes. The approach was significantly faster, less variable, and more versatile than elliptical scanning. It was administered to Brazilian patients with breast cancer (Duczmal et al., 2007). Martins-Fillo & Yao (2009) Subsequent to the discourse on the parametric linear regression model, they computed the LLE convergence distribution. Given the reliability and consistency of the trace distribution, they identified the bias-covariance matrix of its parameters. They proposed employing a two-stage approach to estimate and ascertain the linear estimator's relevance instead (Martins-Filho & Yao, 2009). Suhad (2016) examined parametric estimation techniques for the spatial dynamic model with lengthy data in a stable state, incorporating fixed geographic and temporal influences. She commenced with Spatial Fixed Effects—repeated fixed effects model. A maximum likelihood estimate is employed. The transformation method diminished estimator bias (Al-Tamimi, n.d.). Hoshino (2017) An estimator based on moments assessed a semi-parametric spatial autoregressive model featuring autoregressive independent variables. The model and proposed strategy in Tokyo forecasted the impact of a police substation in residential areas on home burglaries. Demonstrated consistency and inherent convergence of the proposed estimator (Hoshino, 2018). Zakariya (2019) suggests that the firefly strategy is used to count data regression variable Selection. The suggested method's prediction accuracy and variable selection criteria were tested using simulations and real data. The proposed methods outperformed traditional methods (Algamal, 2019). Anwar & Sabah (2020) utilised observational data to estimate the survival function by employing a genetic algorithm to ascertain the optimal values of the Weibull distribution. MWLS utilises GA, maximum likelihood, moments, and least squares.

Comparisons of mean square error in survival functions. Information regarding lung cancer and bronchitis will be used. Research indicated a hybridisation of genetic algorithms with least squares methods (Abdel Hadia & Redha, 2020). Ons & Sabah (2022) Environmental flower pollination yielded a semiparametric regression function with explanatory and dependent variable measurement errors. Wald, Bartlett, and Durbin developed the parametric model; Nadaraya Watson, K-Nearest Neighbour, and median assessed the nonparametric model. The flower pollination algorithm estimated semi-parametric regression utilizing an ecological model and measurement errors, finding the optimal ecological scale measurement error model using MSE. Median-Durbin enhanced Baghdad air and environmental pollution data, and Median-Wald increased flower pollination. FP outperformed FPA in all statistical and semi-parametric models (Ons Edin Musa & Sabah Manfi Ridha, 2022). Ons & Sabah (2023) The researchers used an organized TABU algorithm to estimate the semi-parametric regression function with measurement errors in both the explanatory and dependent variables and then compared the models to choose the optimal one. The models are compared using the mean square error (MSE), with the parametric model generated via the instrumental variables approach (Wald method, Bartlett's method, and Durbin's method). The nonparametric model was calculated utilizing kernel smoothing, K-nearest neighbour smoothing, and mean smoothing (Musa & Ridha, 2023). Alain, Emmanuel (2023) Sampled SLR-REML, RF, RFRK, and Bayesian models. The evaluation employed WoSI-ISRIC SoilGrid 250-meter data and simulations. Model validation used RMSE, R, L, spatial autocorrelation, range parameter, response intensity, and covariate correlation to determine if the estimated period could be covered. With low spatial autocorrelation, SLR-REML accuracy and bias were attained. Linear mixed modelling with INLA-SPDE and SLR models shows soil and environmental component dynamics. Abubaker & Fatima (2024) The correlations were validated through RMSE, R, L, probability coverage density, autocorrelation, range parameter, strength of adaptation, and variable-to-variable correlation. SLR-REML ensures precise autocorrelation and classical birth termination. INLA-SPDE and SLR models tackle temporal, mixed, and linguistic heterogeneity (Younis et al., 2024). Zeheng, Xinyu & et al. (2024) Using the convolution algorithm and genetic algorithm, sequence analysis in bioinformatics engineering, we investigate the experiments of the K-gene set analysis model to demonstrate the benefits of the gene algorithm and its reference importance for current gene information statistics, as it can identify single nucleotide polymorphisms (SNPS) associated with contact dermatitis, and establish close links between genetic polymorphisms and disease (He et al., 2024). The researchers Ahmed & Emad (2024) evaluated a mixed Poisson regression model for a latent class utilizing observations from various sources and categories. Parameters computed using the traditional Expectation-Maximization methodology. MSE simulations evaluated the EM method and GA using sample sizes (n = 50, 90, 120) and three standard parameter combinations (S1, S2, S3); GA surpassed EM (Eleas & Aboudi, 2024). Matthew, Andrew, et al. (2025) developed a neural Bayesian estimator to estimate marginal subscales, hence accelerating uncertainty quantification. Maximum stability, Gaussian processes, and global sea surface temperature data were employed. We estimate Gaussian process model parameters at 2.161 sites within minutes on a GPU utilizing thousands of irregular data points (Matazi et al., 2024). The study will be divided into two components: theoretical and applied. The theoretical aspect includes a comprehensive overview of the spatially dependent data and the artificial intelligence techniques used in the research, specifically the Genetic Algorithm, the Artificial TABU Search Algorithm, and the Binary Firefly Algorithm. The application component was divided into two sections. The initial component used an experimental approach that utilized simulation to identify the best approach for estimating the regression function for spatially dependent data. The objective was to examine real data and estimate the regression function for spatially dependent pollution data of the Euphrates River. Ultimately, it encompassed the principal results we derived and subsequently articulated several recommendations.

This research aims to:

1. Estimating the coefficients of a parametric regression model for spatial data inside an ecosystem.
2. This research employs three artificial intelligence algorithms to estimate spatial data coefficients for studies involving spatially dependent data. The results of these algorithms will be compared to identify the most effective algorithm for representing spatial environmental pollution data, utilizing the Mean Absolute Relative Error criterion.

3. Research Methodology:

3.1 Spatial Analysis:

Spatial Analysis is a method of measuring spatial relationships between phenomena based on measurements of location, proximity, shape, dimensions, areas, etc., to interpret spatial relationships and predict the behaviour of those phenomena in the future. This type of Analysis aims to reveal mutual spatial relationships and connections between the components of the phenomenon. Spatial Analysis has several levels, including 1) Two-Dimensional, 2) Three-Dimensional, 3) Four-Dimensional. (Dawood, 2012).

When analyzing a two-dimensional plane, only the horizontal locations (for example, longitude and latitude) that express the geographical location of the components of the phenomenon under study are analyzed. Suppose the values of the third dimension (height) are available for the items. In that case, the data is analyzed in three dimensions: Surface Analysis (for example, a three-dimensional representation of the study area and its different topography, and the creation of vertical cross-sections between various locations in the region). A four-dimensional spatial analysis is performed if several periods are available for the same data (Anselin, 1988).

3.1.1 Spatial Dependence:

The spatial dependence of the sample data observations at location *(i)* depends on other observations at location *(j)* when $i \neq j$ and according to the following formula:

$$Y_i = f(Y_j) \qquad , i = 1,2,\dots,n \ , \qquad i \neq j \ \dots (1)$$

That is, the values of the sample data at one point depend on the values of the data at other locations, and the strength of the spatial dependence between observations of spatial units decreases the farther the distance is. That is, the economic measurement depends on the spatial effect and the gradient of distance and location, and it can be estimated for small areas using data on a number of Contiguous counties to mitigate the scarcity of time series. Spatial dependence occurs when the administrative boundaries of counties, states, and census tracts do not accurately reflect the underlying nature that generates the sample data (LeSage, 2015).

3.1.2 Spatial Lag Operator:

The primary goal of using a spatial weight matrix in formulating spatial economic, econometric models is to link a variable at one point in space to other spatial units in the system. In time series, this is achieved by using a lag factor that shifts the variable by one or more periods. The backward lag factor is shown in the following formula (Sainsbury-Dale et al., 2025).:

$$Y_{t-K} = L^K Y \ \dots (2) \qquad \text{Delayed by grade K}$$

The spatial lag operator finds the weighted average of adjacent observations. The spatial lag of the dependent variable *(Y)* at *(i)* is written in the following form (Duczmal et al., 2007):

$$L^K Y_i = \sum_j W_{ij} Y_j \quad , \qquad j = 1, 2, \dots, n \ \dots (3)$$

In the language of matrices, it is written as follows:

$$L^K Y = W_K \underline{Y}$$

Where:

$\underline{Y}$: represents a vector (n*1) that includes all observations.

W : represents the spatial weights matrix.

The set of row scores in a W matrix often equals one (LeSage, 2015).

3.1.3 Spatial Contiguity Matrices:

Natural measurements of spatial dependency or autocorrelation depend on the bilateral juxtaposition between spatial units, so the expression for the basic juxtaposition structure is (1,0) of values. That is, if two spatial units have common boundaries of non-zero length, they are considered contiguous and assigned a value (1). Still, they are deemed non-contiguous and assigned a value (0) if they do not have common boundaries. There are several types of spatial contiguity matrices, which are Binary Contiguity Matrices, Weights Based on Distance and Adjusted Weights Matrix (Matazi et al., 2024).

This research will use the Binary Contiguity Matrix Criterion because it suits the research objectives and environmental data.

3.1.4 Binary Contiguity Matrix:

Data availability allows the construction of spatial weight matrices based on adjacency. Let n be the Number of spatial units, W: be the spatial weight matrix with dimension $n * n$, which is a square, symmetrical, positive, non-random matrix with elements. W_{ij} At location ij, which represents each element in the matrix W, the weights W_{ij} Are determined for each pair of adjacent and non-adjacent sites by a set of pre-defined rules that show the spatial relationships between the sites, and the general formula of this matrix is as follows (Sainsbury-Dale et al., 2025):

$$W = \left[\begin{bmatrix} W_{11} & \cdots & W_{1n} \\ \vdots & \ddots & \vdots \\ W_{n1} & \cdots & W_{nn} \end{bmatrix}\right] \quad \dots(4)$$

The values of the W elements are determined as follows:

$$W_{ij} = \begin{bmatrix} 1 & \qquad if\ i\ and\ j\ are\ contiguous \\ 0 & \qquad f\ i\ and\ j\ are\ not\ contiguous \end{bmatrix} \quad \dots(5)$$

The binary contiguity matrix (spatial weight matrix) represents the spatial arrangement of cells. There are essential methods for constructing the contiguity matrix adopted, including (LeSage, 2015):

i. Rook Contiguity Criterion.
ii. Linear Contiguity Criterion.
iii. Bishop Contiguity Criterion.
iv. Queen Contiguity Criterion.
v. Double Linear Contiguity Criterion.
vi. Double Rook Contiguity Criterion.

We will use the Queen Contiguity Criterion because it suits the research objectives and environmental data.

3.1.5 Queen Contiguity Criterion:

Contiguity is calculated when two adjacent cells share a common side and a common vertex point. This means that the juxtaposition is formed by achieving the juxtaposition of Rook and the juxtaposition of Bishop. Figure (1) shows the juxtaposition (LeSage, 2015).

A	B	C
D	E	F
G	H	I

Figure (1) shows the juxtaposition of the Queen criterion

Source: (LeSage, 2015)

From Figure (1), the weight matrix of Queen contiguity (W) will be as follows:

$$W = \begin{bmatrix} 0 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 0 \end{bmatrix} \quad \ldots (6)$$

3.1.6 Estimating Spatial Linear Regression Models for Cross-Section Data:

The Maximum Likelihood method and Jacobian will be used to estimate spatial models (LeSage, 2015). The starting point for this topic is:

$$Y = \rho WY + X\beta + \epsilon \quad \ldots (7)$$

$$\epsilon = \lambda W\epsilon + u \quad \ldots (8)$$

Where:

β : Represents a vector of parameters (k*1).

X : represents explanatory variables (n*k).

ρ: Represents the spatial dependence parameter. The coefficient of the spatially lagged dependent variables.

λ: is the coefficient in a spatial autoregressive structure for the disturbance ϵ . The spatial correlation parameter represents the coefficient for spatially correlated errors.

Y: Represents the vector of the dependent variable $n * 1$.

WY: The response variable represents the spatial regressor.

W: Represents the spatial Contiguity matrix.

ϵ: The error vector with n*1 dimension.

$u \sim N(0, \sigma^2 I_n)$, ϵ *and* X are independent variables (Sainsbury-Dale et al., 2025).

3.2 Genetic Algorithm (GA):

Genetic algorithms are optimization algorithms that find the maximum or minimum of a function. GA is based on the mechanism of natural selection and natural genetics. It is classified as one of the evolutionary algorithms based on simulating the work of nature from Darwin's perspective. GA is used as a random search method to find optimal or close to optimal solutions using natural biological mechanisms such as inheritance, mating, and genetic Mutation by passing on good traits to successive generation processes, producing optimal offspring, and repeating genetic cycles to improve the offspring with modern phases and patterns (Safe et al., 2004).

3.2.1 Methodology of the Genetic Algorithm:

Chromosomes represent individuals in computer simulations utilizing evolutionary methods to identify optimal solutions. Chromosomes are frequently illustrated in binary form (0,1). The evolution process commences by randomly selecting and replicating chromosomes from the initial population. In each generation, the Fitness Function is computed for every chromosome to identify which one is best. She selected ideal chromosomes for crossing (Safe et al., 2004). Mating generates a genetic mutation that continues until the trade-off function produces the ideal response, concluding the genetic algorithm. Certain evolutionary algorithms terminate after reaching the maximum Number of generations, resulting in suboptimal outcomes (Duczmal et al., 2007).

3.2.2 Steps of Genetic Algorithm:

The researcher Reeves presented this algorithm. (Reeves, 2010)

```
Choose an initial population of chromosomes;
while termination condition not satisfied do
    repeat
         if crossover condition satisfied then
        {select parent chromosomes;
        choose crossover parameters;
        perform crossover};
        if mutation condition satisfied then
       {choose mutation points;
        perform mutation};
       evaluate fitness of offspring
   until sufficient offspring created;
select new population;
 endwhile
```

3.2.3 Components, Structure, & Terminology:

GA is designed to simulate a biological process, so the terminology is borrowed from biology. The standard components in all genetic algorithms are a fitness function for optimization, the number of chromosomes, selecting the chromosomes to be reproduced, Crossover to produce the next generation, and Mutation. The first step in GA is to generate a set of random solutions (chromosomes). The length of the chromosome and how it is represented depend on the nature of the problem (Sivanandam SN, 2008). There are several ways to define and encode chromosomes:

1. Binary Encoding: The chromosome is represented as a series of numbers consisting of zeros and ones.
2. Real Value Encoding: Chromosomes are represented as fractional numbers.
3. Integer Value Encoding: Representing chromosomes as a series of integer numbers.
4. Character Representation Encoding: Chromosomes are represented as a string of letters.
5. Tree Representation Encoding: Chromosomes are represented as a tree (Carr, 2014).

3.2.4 Selection:

In each generation, a subset of chromosomes is chosen based on a specific ratio to produce a new generation, determined by the Fitness Function. There exist various ways we use roulette.

1. Roulette Wheel Selection: The trade-off function of each member of the generation dictates their distribution over the 100 sectors of the roulette wheel. A person is selected after the wheel is spun randomly until it halts at a pointer. As the trade-off value escalates, the quantity of sectors on the wheel and the probability of Selection for the subsequent generation also increase. This method adheres to the following equation:

$$P_i = \frac{F_i}{\sum_{i=1}^{n} F_i} \quad \dots (9)$$

Where:

P_i: Represents the individual's probability. F_i: Represents the fitness function value. n: Number of members of the generation.

2. Elitist Selection.
3. Tournament Selection (Sivanandam SN, 2008).

3.2.5 Reproduction:

The process of generating a new generation of selected individuals.

Crossover: Crossover transpires between chosen progenitors to generate two novel people, and this procedure persists until the most recent generation is established. Various crossover processes encompass (Reeves, 2010):

1. One-point Crossover.
2. K-Points Crossover.
3. Cut and Splice.

Mutation: This method is not attributable to the parents but rather to a swift alteration in the progeny resulting from mating, which modifies one or more genes inside the chromosome. Numerous genetic mutations are present (Sivanandam SN, 2008):

1. Exchange: In this process, only the change is made.
2. Shaft and Exchange: Two processes co-occur: change and displacement.

3.3 The Artificial TABU Search Algorithm (TSA):

The name tabu came from a list within the work of this algorithm, which contains a set of solutions called the "tabu list," that is, the list of prohibited things that there is no need to return to and use again. This algorithm is characterized by great flexibility and falls within the category of random algorithms, an experimental research method that uses local search methods to find solutions to optimization and improvement problems (Piniganti L., n.d.).

3.3.1 TSA Description:

This local search algorithm comprises three fundamental concepts. It oversees the dynamics of search solutions, employs adaptable memory architectures, and modifies the intensification and diversification of research. To prevent redundancy in solutions, the algorithm incorporates the least favorable response into the block list if no superior alternatives are present in the neighborhood. If a solution has been recently accessed or contravenes a rule, it will be designated as "blocked" or "Tabu-Forbidden," and the algorithm will not yield (Hertz et al., 1995). The TSA examined the vicinity of each solution. Memory structures identify appropriate solutions in each new neighborhood. $N^*(X)$ throughout the search process. In the present neighborhood $N^*(X)$, memory structures facilitate the transition from the current solution X to the enhanced solution X_best. Tabu lists are temporary collections of solutions that have been previously explored n times (tabu tenure, retained solutions) (Glover F et al., 2007).

3.3.2 Basic elements of the TS algorithm:

1. Local search: The following solution is usually the best.
2. Tabu List or Tabu Forbidden.
3. A mechanism for randomly changing the path of solutions when there is no progress for a long time (Musa & Ridha, 2023).

3.3.3 Types of Memories:

There are three types of memory structures used in TSA:

- Short-term memory: Recently acquired solutions reside in short-term memory. The answer will be rejected unless it is eliminated from the prohibited list. - Medium-term memory: Condensation principles in medium-term memory direct the search towards optimal sites. - Long-term memory: Diversification guidelines direct research towards novel areas within the discipline to broaden answers.

Short-term, medium-term, and long-term memory may intersect based on the phenomenon or context. In certain studies, short-term memory may demonstrate superior efficacy.

The TSA circumvents tabu list limitations using aspiration criteria (Musa & Ridha, 2023).

3.3.4 Tabu Algorithm Strategies:

The TSA has three strategies:

1. Intensive memory strategy: This approach enhances the Selection and transition of benefit-based solutions, facilitating optimal solution retrieval. Midterm memory retrieval induces condensation.
2. Diversification strategy: This approach employs long-term memory retrieval. Diversification entails exploring novel research avenues.
Path reconnection, strategic fluctuation, reinforcement through restriction, and solution appraisal are employed to execute intensification and diversification strategies.
3. Aspiration criteria strategy: This approach assesses research adaptability at TAPU and is essential. Ambition facilitates progress and resolution, even in the face of prohibition (Hertz et al., 1995).

3.3.5 Steps of TSA:

The researchers (Glover F & Laguna M., 1998) presented**:**

A simplified form of the TS algorithm using short-term memory.

```
Tabu list
Current solution ← Initial solution
Best solution S
While fitness (best solution) = best candidate
  For candidate ∈ current solution. Get Neighbourhood do
            If (tabu list contains (candidate)) then
                 If (not tabu list. contains (S candidate)) and
                (fitness (candidate)>fitness (best candidate)) then
                      Best candidate
                 End if
    Else if (fitness (candidate)>fitness (best solution)) then
              Best candidate
                   End if
         End for
Current solution ←  best solution
        If (fitness (best candidate)>fitness (best solution)) then
              best solution ←  best scandidate
       end if
             tabu list.push (best candidate)
       if  tabu list. size > tabu tenure  , then
            tabu list. remove first ( )
       end if
end while
return best solution
```

3.4 Firefly Algorithm (FFA):

The FFA is a nature-inspired stochastic global optimization technique. It emulates the mating and communication mechanism of fireflies through light flashes. The majority of fireflies produce brief, recurrent lights to attract partners, entice prey, or provide warnings.

Two factors in most fireflies make them visible at a limited distance. The first is that light intensity is inversely proportional to the square distance. The second is that light weakens because the air absorbs it (Yang XS., 2010).

3.4.1 Artificial fireflies:

Yang devised this technology employing three ideal principles for artificial fireflies. Fireflies are unisex, hence attracting one another irrespective of gender. The dimmer Firefly will move towards the brighter one, as gravitational attraction correlates with luminosity. The attractiveness of fireflies diminishes with increasing distance (Ezugwu et al., 2020).

The movement of fireflies attracted to another, brighter Firefly is determined by:

$$x_i = x_i + \beta_0 e^{-\gamma r_{ij}^2}\left(x_j - x_i\right) + \alpha(rand - 0.5) \;\ldots (10)$$

Where:

$\beta_0 e^{-\gamma r_{ij}^2}\left(x_j - x_i\right)$: is attraction.

α : being the random parameter, $\alpha \in [0,1]$

$rand$: is a random number $\sim U[0,1]$, so the expression (rand - 0.5) ranges from [-0.5,0.5]. To allow for positive and negative contrast.

β_0 : is always set to 1 and $\alpha \in [0,1]$.

A random distribution extends to a normal distribution with mean =0 and variance=1 to allow the noise rate in the environment to vary.

γ : is the distance between $Firefly_i, Firefly_j$. It characterizes the gravitational anisotropy, and its value is essential in determining the convergence speed and how the FA algorithm behaves. Its value often ranges from 0.01 to 100.

$$\gamma_{ij} = \left|\left|x_i - x_j\right|\right| \;\ldots (11)$$

x_i : represents the position of Firefly i (Yang, 2009).

2.4.2 Steps of FFA:

The researchers (Hassanien & Emary, 2018) presented:

```
input : $n$ (Number of Fireflies)
        $N$ (Iter Number of iterations for optimization)
        $\gamma$ (attractiveness parameter)
        $\alpha$ (contribution of the random term (environment noise))
output: Optimal firefly position and its fitness
   Initialize a population of n fireflies’ positions at random
   Find the best solution based on fitness;
   while Stopping criteria not met do
        foreach $Firefly_i$ do
        foreach $Firefly_j$ do
         if $Firefly_j$ is better than $Firefly_i$ then
              Move $Firefly_i$ towards $Firefly_j$ using equation
 $x_i = x_i + \beta_0 e^{-\gamma r_{ij}}(x_j - x_i) + \alpha(rand - 0.5)$
         else end
   Evaluate The positions of individual fireflies end
```

3.4.3 FFA variant:

There are several variants of this algorithm, including Discrete FFA, Binary FFA, Chaotic Firefly algorithm, Parallel Firefly, FFA for constrained problems, L'evy flight FFA (LFA), Intelligent Firefly Algorithm, Gaussian firefly algorithm (GDFF), Network-structured firefly algorithm (NS-A) and FFA with adaptive parameters (Hassanien & Emary, 2018).

We will use the Binary Firefly algorithm because it suits the research objectives and environmental data.

3.4.4 Binary FFA:

This algorithm forces the firefly positions to be on the binary grid, and the FFA update equation (Equation 11) allows the position of the binary FFA to be updated, which may lead to continuous values. Continuous values are mapped to binary values by the following equation (Hassanien & Emary, 2018):

$$X_{id} = \begin{cases} 1 & if\ rand\ < \dfrac{1}{1 + \exp\left(-x_{id}\right)} \\ 0 & otherwise \end{cases} \quad \ldots (12)$$

Where X_{id} is the position of the ith agent in dimension d, and rand is a random number. It follows a uniform distribution in the range from 0 to 1. The above equation requires that the position/solutions of the Firefly belong to 0, 1. Binary encoding is used as in the following equation.

$$\tanh(x_p) = \frac{\exp(2 * |x_p|) - 1}{\exp(2 * |x_p|) + 1} \quad \ldots (13)$$

Where x_p Is the continuous value of the solution x in dimension d. The solution's final binary value is calculated with the following equation:

$$X_{id} = \begin{cases} 1 & if\ rand\ < \tanh\left(x_p\right) \\ 0 & otherwise \end{cases} \quad \ldots (14)$$

Where tanh is the hyperbolic tan function (Hassanien & Emary, 2018).

3.5 Mean Absolute Percentage Error (MAPE):

In this research, we used the criterion of the average absolute error of proportions to compare the results and its mathematical formula (Chai & Draxler, 2014).

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{Y_i - \hat{Y}_i}{Y_i}\right| \quad , i = 1,2, \ldots , n \quad \ldots (15)$$

Where :

Y : the real value.

$\hat{Y}$: estimated value.

n : Number of observations used in each experiment.

The best estimator is the one that has the lowest value for this criterion.

4. Results and Discussion:

4.1 Experimental Aspect:

We will use an experimental methodology to verify the theoretical structure and show the superiority of one estimation method among the following: classical, GA, TSA, and Binary FFA. We will compare their outcomes using the MAPE criterion to identify the best approach for estimating real data. The simulation method has been used for this purpose. The theoretical component will be implemented in the simulation utilizing various sample sizes (N=20, N=50, N=100, and N=150). The data results were derived utilizing MATLAB 2016a.

4.1.1 Analysis of the results of the simulation experiment:

MAPE evaluates the data using classical methods alongside three artificial intelligence algorithms (GA, TSA, Binary FFA) throughout different sample sizes.

Table 1: The Mean Absolute Percentage Error values (MAPE)

N= 20				
Classic	**GA**	**TSA**	**Binary FFA**	**Best**
4.0236	0.2569	0.1147	0.1365	TSA
N= 50				
2.0025	0.0987	0.0257	0.0147	Binary FFA
N= 100				
0.2571	0.0136	0.0119	0.0018	Binary FFA
N= 150				
0.0879	0.0064	0.0011	0.0002	Binary FFA

Source: Prepared by the Researchers.

Table No. (1) Showes, When N = 20, the TSA is the best, exhibiting the lowest MAPE of 0.1147. For N=50, the Binary FFA exhibited its best performance with a MAPE value of 0.0147; for N=100, it again showed superiority with a MAPE value of 0.0018; and for N=150, it maintained its preeminence with a MAPE value of 0.0002. The simulation results in Table (1) indicate that the best approach for estimating spatially dependent environmental variables is the Binary FAA algorithm.

4.2 Analyze Real Data:

The binary FFA algorithm will be applied to real data regarding the environmental conditions in Iraq, obtained from the Iraqi Ministry of Environment - Water Pollution Department statistics for the year 2022, specifically concerning the Euphrates River, where a series of variables affecting the water quality of this river have been identified. Our sample has 185 members.

Variables included in the research

1. Total dissolved salts (TDS) is the dependent variable.
2. Temperature (temp.), an independent variable.
3. Calcium (Ca), an independent variable.
4. Magnesium (Mg), an independent variable.
5. Potassium (K), an independent variable.
6. Sodium (Na), an independent variable.

4.2.1 Analyze real data results:

Parameter values of the regression function for spatially dependent water pollution data of the Euphrates River using Binary FFA.

Table 2: The Parameter values of the regression function for spatially dependent data.

parameter	result	Test values	P-value
B_0	107.198	0.804	0.423
B_1	-0.028	0.777	0.439
B_2	0.405	6.858	0.001
B_3	0.212	4.512	0.001
B_4	-0.535	6.001	0.001
B_5	0.747	6.863	0.001

Source: Prepared by the Researchers.

$$TDS = 107.198 - 0.028\,Temp, +0.405\,Ca + 0.212\,Mg - 0.535\,K + 0.747\,Na$$

Results Discussion: From the results presented in Table (2), we observe the following:

Upon comparing the p-value against the significance level of 0.05 for each variable in the study, we observed the following:

1. The first explanatory variable is temperature. The parameter value was $\beta_1 = -0.028$. The observed p-value of 0.439, exceeding the significance level, shows that this variable is not significant. Temperature does not affect total dissolved salts (TDS).
2. The second explanatory variable, Ca, has a parameter value of $\beta_2 = 0.405$. Upon comparing its p-value with the significance level, we observed that the p-value is less than the significance threshold, indicating the significance and effect of the variable Ca on TDS. Ca has a direct positive effect on TDS with a value of 0.405.
3. The third explanatory variable, Mg, has a parameter value of$\beta_3 = 0.212$, indicating a significant and positive effect on TDS of 0.212. This inference was determined by comparing its p-value with 0.05.
4. The fourth explanatory variable, K, has a parameter value of $\beta_4 = -0.535$. Upon comparing its p-value with the significance level, we determined that the p-value was less than the significance level, indicating the significance and effect of variable K on TDS. K contains a negative effect value of 0.535 on TDS directly.
5. The fifth explanatory variable, Na, has a parameter value of $\beta_5 = 0.747$. Upon comparing its p-value with the significance level, we determined that this value was less than the significance level, indicating the significance and effect of the variable Na on TDS. Na has a direct positive effect value of 0.747 on TDS.

5. Conclusions:

The MAPE values obtained from the simulation showed that the Binary FFA algorithm provided the best estimates, indicating that it is best over classical methods, as well as GA and TSA, in environmental considerations and in estimating regression models for spatially dependent data.
Analysis of the regression parameters for the spatially dependent data about the environmental aspect (pollution of the Euphrates River water) revealed that the temperature variable did not affect the total dissolved salts. Conversely, the variables (Calcium (Ca), Magnesium (Mg), and Potassium (K)) demonstrated a significant and positive effect on the total dissolved salts. The Sodium (Na) variable exerted significant negative effects on the total dissolved salts.

Authors Declaration:

Conflicts of Interest: None
-We Hereby Confirm That All The Figures and Tables In The Manuscript Are Mine and Ours. Besides, The Figures and Images, which are Not Mine, Have Been Permitted Republication and Attached to The Manuscript.
- Ethical Clearance: The Research Was Approved by The Local Ethical Committee in The University.

References:

Abdel Hadia, A. T., & Redha, S. M. (2020). Estimate The Survival Function by Using the Genetic Algorithm. *Journal of Economics and Administrative Sciences*, *26*(122), 440–454. https://doi.org/10.33095/jeas.v26i122.2018

Algamal, Z. (2019). Variable Selection in Count Data Regression Model based on Firefly Algorithm. *Statistics, Optimization & Information Computing*, *7*(2). https://doi.org/10.19139/soic.v7i2.566

Al-Tamimi, S. A. S. (n.d.). *Compared Study to Some of The Application Practical With (Data Panel) for Model Spatial Dynamic the Estimating of Methods*. [Doctoral dissertation, Mustansiriyah university].

Anselin, L. (1988). *Spatial Econometrics: Methods and Models* (Vol. 4). Springer Netherlands. https://doi.org/10.1007/978-94-015-7799-1

Carr, J. (2014). An introduction to genetic algorithms. *Computers & Mathematics with Applications*, *32*(6), 133. https://doi.org/10.1016/S0898-1221(96)90227-8

Chai, T., & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature. *Geoscientific Model Development*, *7*(3), 1247–1250. https://doi.org/10.5194/gmd-7-1247-2014

Dawood, J. M. (2012). *Foundations of spatial analysis*. 59–80.

Dubin, R. A. (1988). Estimation of Regression Coefficients in the Presence of Spatially Autocorrelated Error Terms. *The Review of Economics and Statistics*, *70*(3), 466. https://doi.org/10.2307/1926785

Duczmal, L., Cançado, A. L. F., Takahashi, R. H. C., & Bessegato, L. F. (2007). A genetic algorithm for irregularly shaped spatial scan statistics. *Computational Statistics & Data Analysis*, *52*(1), 43–52. https://doi.org/10.1016/j.csda.2007.01.016

Eleas, A. K., & Aboudi, E. H. (2024). Expectation Parameters in the Poisson Mixture Regression Model for Latent Class by Applying Genetic Algorithm and Maximization Algorithm. *Journal of Economics and Administrative Sciences*, *30*(140), 434–449. https://doi.org/10.33095/jeas.v26i122.2018

Ezugwu, A. E.-S., Agbaje, M. B., Aljojo, N., Els, R., Chiroma, H., & Elaziz, M. A. (2020). A Comparative Performance Study of Hybrid Firefly Algorithms for Automatic Data Clustering. *IEEE Access*, *8*, 121089–121118. https://doi.org/10.1109/ACCESS.2020.3006173

Glover F, & Laguna M. (1998). *Tabu search.* Springer US.

Glover F, Laguna M, & Marti R. (2007). Principles of Tabu Search. *Approximation Algorithms and Metaheuristics.*, *23*(1), 1–12.

Hassanien, A. E., & Emary, E. (2018). *Swarm Intelligence*. CRC Press. https://doi.org/10.1201/9781315222455

He, Z., Shen, X., Zhou, Y., & Wang, Y. (2024). Application of K-means clustering based on artificial intelligence in gene statistics of biological information engineering. *Proceedings of the 2024 4th International Conference on Bioinformatics and Intelligent Computing*, 468–473. https://doi.org/10.1145/3665689.3665767

Hertz, A., Taillard, E., & De Werra, D. (1995). *A Tutorial on Tabu Search*.

Hoshino, T. (2018). Semiparametric Spatial Autoregressive Models with Endogenous Regressors: With an Application to Crime Data. *Journal of Business & Economic Statistics*, *36*(1), 160–172. https://doi.org/10.1080/07350015.2016.1146145

Kelejian, H. H., & Prucha, I. R. (1998). A Generalized Spatial Two-Stage Least Squares Procedure for Estimating a Spatial Autoregressive Model with Autoregressive Disturbances. *The Journal of Real Estate Finance and Economics*, *17*(1), 99–121. https://doi.org/10.1023/A:1007707430416

LeSage, J. P. (2015). “Theory and Practice of Spatial Econometrics.” *Spatial Economic Analysis*, *10*(3), 400–400. https://doi.org/10.1080/17421772.2015.1062285

Martins-Filho, C., & Yao, F. (2009). Nonparametric regression estimation with general parametric error covariance. *Journal of Multivariate Analysis*, *100*(3), 309–333. https://doi.org/10.1016/j.jmva.2008.04.013

Matazi, A. K., Gognet, E. E., & Kakaï, R. G. (2024). Digital soil mapping: a predictive performance assessment of spatial linear regression, Bayesian and ML-based models. *Modeling Earth Systems and Environment*, *10*(1), 595–618. https://doi.org/10.1007/s40808-023-01788-1

Musa, O. E., & Ridha, S. M. (2023). *Using the artificial TABU algorithm to estimate the semi-parametric regression function with measurement errors*. 040037. https://doi.org/10.1063/5.0126025

Ons Edin Musa, & Sabah Manfi Ridha. (2022). Semi-parametric regression function estimation for environmental pollution with measurement error using artificial flower pollination algorithm. *International Journal of Nonlinear Analysis and Applications.*, *13*(1), 1375–1389.

Piniganti L. (n.d.). *A Survey of Tabu Search in Combinatorial Optimization*. UNLV Theses, Dissertations, Professional Papers, and Capstones. 2132. http://dx.doi.org/10.34917/5836151

Reeves, C. R. (2010). Genetic algorithms. In Handbook of metaheuristics (pp. 109-139). Springer, Boston, MA.

Safe, M., Carballido, J., Ponzoni, I., & Brignole, N. (2004). *On Stopping Criteria for Genetic Algorithms* (pp. 405–413). https://doi.org/10.1007/978-3-540-28645-5_41

Sainsbury-Dale, M., Zammit-Mangion, A., Richards, J., & Huser, R. (2025). Neural Bayes Estimators for Irregular Spatial Data Using Graph Neural Networks. *Journal of Computational and Graphical Statistics*, 1–16. https://doi.org/10.1080/10618600.2024.2433671

Sivanandam SN, D. S. S. S. D. S. (2008). *Introduction to Genetic Algorithms*. Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-73190-0

Yang, X. S. (2009). Firefly algorithms for multimodal optimization. In International symposium on stochastic algorithms (pp. 169-178). Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-642-04944-6_14

Yang XS. (2010). *Nature-inspired metaheuristic algorithms.* Luniver press.

Younis, A., Belabbes, F., Cotfas, P. A., & Cotfas, D. T. (2024). Utilizing the Honeybees Mating-Inspired Firefly Algorithm to Extract Parameters of the Wind Speed Weibull Model. *Forecasting*, *6*(2), 357–377. https://doi.org/10.3390/forecast6020020