Int. J. Nonlinear Anal. Appl. 13 (2022) No. 1, 1375–1389
ISSN: 2008-6822 (electronic)
http://dx.doi.org/10.22075/ijnaa.2022.5744

Semi-parametric regression function estimation for environmental pollution with measurement error using artificial flower pollination algorithm

Ons Edin Musa[a,*], Sabah Manfi Ridha[b]

[a] *Mustansiriyah University , College of physical education and sports science, Iraq*
[b] *Baghdad University, Department of statistics,Iraq*

(Communicated by Madjid Eshaghi Gordji)

Abstract

Artificial Intelligence Algorithms have been used in recent years in many scientific fields. We suggest employing flower pollination algorithm in the environmental field to find the best estimate of the semi-parametric regression function with measurement errors in the explanatory variables and the dependent variable, where measurement errors appear frequently in fields such as chemistry, biological sciences, medicine, and epidemiological studies, rather than an exact measurement.

We estimate the regression function of the semi-parametric model by estimating the parametric model and estimating the non-parametric model, the parametric model is estimated by using an instrumental variables method (Wald method, Bartlett's method, and Durbin's method), The non-parametric model is estimated by using kernel smoothing (Nadaraya Watson), K-Nearest Neighbor smoothing and Median smoothing. The Flower Pollination algorithms were employed and structured in building the ecological model and estimating the semi-parametric regression function with measurement errors in the explanatory and dependent variables, then compare the models to choose the best model used in the environmental scope measurement errors, where the comparison between the models is done using the mean square error (MSE).

These methods were applied to real data on environmental pollution/ air pollution in the city of Baghdad, and the most important conclusions that we reached when using statistical methods in estimating parameters and choosing the best model, we found that the Median-Durbin model is the best as it has less MSE, but when using flower The pollination algorithm showed that the Median-Wald model is the best because it has the lowest MSE, and when we compare the statistical methods

[*] Corresponding author

Email addresses: `ons.edin1001a@coadec.uobaghdad.edu.iq` (Ons Edin Musa), `drsabah@coadec.uobaghdad.edu.iq` (Sabah Manfi Ridha)

Received: May 2021 *Accepted:* October 2021

with the FPA in selecting semi-parametric models, we notice the superiority of the FP algorithm in all methods and for all models.

Keywords: Semi-parametric, Measurement error, flower Pollination algorithm, instrument variables method, kernel smoothing, Nadaraya Watson, K-Nearest neighbor smoothing, median smoothing.

1. Introduction

Artificial Intelligence (AI) is considered one of the most important fields that are developing rapidly, and it has many important applications in practical life, and in general, AI includes thinking, Knowledge, planning, learning, communication, perception, and the ability to deal with a goal. Artificial Intelligence Algorithms are based on the principles and concepts of artificial intelligence. These algorithms are characterized by their ability to devise dynamic methods appropriate to the nature of the problem to be studied and to determine a practical way to find an appropriate solution from among a set of possible solutions to the problem. It is also characterized by optimizing the value of the solution according to the constraints and variables.

To solve statistical problems whose data involve measurement errors, it is important to distinguish between the common error model and the measurement error model, where the amount and type of measurement error affect estimates of the parameters of phenomena in terms of bias and consistency in general and on health impact estimates in the epidemiology of environmental pollution in a particular search. The measurement error of the explanatory variable and the dependent variable is a common problem.

What are the consequences of analyzes that ignore measurement error?

Measurement error is common in epidemiological studies and can have significant effects on the characteristics of environmental risk assessments, especially those related to biota.

Therefore, it is wise to include measurement error considerations in planning epidemiological studies and environmental pollution to ensure human health and the environment, as classic measurement error models will be studied for environmental pollution in Baghdad.

In this research, we will use semi-parametric regression models to build and estimate the parameters of an environmental model for air pollution, through which the non-parametric and parametric regression models are combined at the same time, and then a regression model is characterized by the possibility of dealing with the multi-dimensional problem of non-parametric models that occurs when the number of the explanatory variables included in the analysis and then the decreasing accuracy of the estimate, as well as the advantage of this type of models with more flexibility than the parametric models that adhere to certain conditions. Some authors who wrote with errors of measurement with semi-parametric models- (Lixing Zhu and Hengjian Cui, 2003) examine the effect of measurement errors in both the parametric and the non-parametric part, simultaneously. consider a partial linear regression model with measurement errors [14], (Hamidul, Howard, Raymond & Louise, 2016) suggested a semi-spherical regression approach in the event of measurement errors to obtain bias-corrected estimates of regression parameters and derive the characteristics of the large sample [7], (Mengyan & Yanyuna, 2019) considered a general parameter regression model that allows a covariate to be measured with heterogeneous errors, where the variance was dealt with using the B-spline approximation [9], (Virgelio, Joseph & Erniel, 2019) assumed a mixed semi-parametric analysis of the covariance model [1], estimated the parameters of this model by the constrained maximum potential method and spline smoothing.

It is necessary to find methods that give optimal results in the fields of science whose data (data of explanatory variables and the dependent variable) are accompanied by measurement errors without the need to correct the data.

2. Research goal

1. Estimating the parameters of a semi-parametric regression model in the presence of the problem of measurement errors in the explanatory variables and the dependent variable in the ecosystem using several methods of estimation and choosing the best among them through the mean square error (MSE).

2. Employing and suggesting mathematical formulas and functions for the Flower Pollination algorithm to be used in studies whose data contain measurement errors.

3. Comparing the models with statistical methods used as well as their comparison when using the artificial flower Pollination algorithm in environmental aspects through the mean square error (MSE) standard.

Theoretical side

3. Measurement error

The measurement error is simply defined as the difference between the value of the observed variable and the value of the correctly measured variable. This error occurs as a result of a defect in the device used in the measurement process [3], although laboratories routinely calibrate their measuring instruments using the standards and values that It is returned based on the use of the resulting calibration curve, spatial or temporal fluctuation during measurement or inattention and negligence by the person performing the measurement. The value of this error is defined as an added value on the log scale [12].

The measurement error model for the explanatory variable and the dependent variable is as follows:

$$X^* = X + u \tag{3.1}$$

$$Y^* = Y + v \tag{3.2}$$

Where Y^* : response variable (dependent variable) with measurement errors, (measured value).
Y : represents the real value vector of the dependent variable.
X : represents the real value vector of the explanatory variable.
X^* : explanatory variable vector with measurement error (measured value).
u : measurement error of the explanatory variable with mean 0 and variance σ_u^2.
v : measurement error of the dependent variable of degree n*1 with mean 0 and variance σ_v^2
u , v , X :are mutually independent.

4. The Semi-parametric Regression Model With Measurement Errors

This model will be written in the following form because it contains measurement errors in the three vectors, the explanatory variable vector for the parameter part, the explanatory variable vector for the non-parametric part, and the response variable:

$$Y^* = X^*\ \beta + g\ (\ t^*\ ) + \varepsilon \tag{4.1}$$

$$X^* = X + u \tag{4.2}$$

$$T^* = T + s \tag{4.3}$$

$$Y^* = Y + v + u + s + \varepsilon \tag{4.4}$$

Where

Y^* **:** the response variable (the dependent variable) with measurement errors for the variable itself, for the parameter and non-parametric variable, and for the random error of the model, which is of degree n * 1.

Y **:** represents the real value vector of the dependent variable.

X^* **:** the explanatory variable vector with a measurement error of the parameter model (measured observed value) with dimension (n*p).

X **:** represents the real value vector of the explanatory variable of the parameter part of the degree (n*p).

P **:** the number of parametric explanatory variables.

t^* **:** the explanatory variable vector with a measurement error of the non-parametric model (measured observed value) with dimension (n*q).

t **:** represents the real value vector of the explanatory variable of the non-parameter part of the degree (n*q).

q **:** the number of non-parametric explanatory variables.

β **:** the vector represents the unknown parameter in the parametric part of the degree p*1.

$g\ (\ t^*\ )$ **:** It is an unknown smoothing function of degree n*1.

u **:** the measurement error of the explanatory variable in the parametric model has a mean of 0 and a variance of σ_u^2 .

s **:** the measurement error of the explanatory variable in the non-parametric model has a mean of 0 and a variance of σ_s^2 .

v **:** the Measurement error of the dependent variable with mean 0 and variance σ_v^2.

ε **:** An independent random error vector of degree n*1 with mean 0 and variance σ^2.

$u,\ v, s,\ X,\ T,$ and e**:** are mutually independent.

5. Parametric regression estimator with Measurement Errors

The parametric model is estimated using the instrumental variable method, which the method is used to estimate B in an unbiased and consistent manner in the event of measurement errors. The basis of this method is to find a set of variables associated with the explanatory variables in the model but not associated with errors [12].
Let Z be a matrix of degree $n{*}k$ for k instrument variables $Z_1, Z_2, \ldots, Z_k$ each variable contains n observations, so we have

$$plim\left(\frac{1}{n}\acute{Z}X^*\right) = \varSigma_{ZX^*} \tag{5.1}$$

The parameter estimator is as follows:

$$\widehat{B_{IV}} = \left(\acute{Z}\ X^*\right)^{-1} \acute{Z}\ Y^* \tag{5.2}$$

$\widehat{B_{IV}}$ is an unbiased and consistent estimator of β.

5.1. Methods for choosing instrument

5.1.1. Wald's method

We find the median of the observations of the variable X^* $(x_1^*, x_2^*, \ldots, x_n^*\ )$, then classify the observations by defining the instrument variable Z as follows [12]:

$$Z = \begin{cases} 1 & if \quad X_i^* \ > median\ (x_1^*, x_2^*, \ldots, x_n^*) \\ -1 & if \quad X_i^* \ < median\ (x_1^*, x_2^*, \ldots, x_n^*) \end{cases} \tag{5.3}$$

$$Z = \begin{bmatrix} 1 & z_1 \\ \vdots & \vdots \\ 1 & z_n \end{bmatrix} \quad , \quad X = \begin{bmatrix} 1 & x_1^* \\ \vdots & \vdots \\ 1 & x_n^* \end{bmatrix} \tag{5.4}$$

Now for the two sets of Z observations, we calculate the following:

The first group: It is the set of observations with a value less than the median of the variable X^*, for which we find the arithmetic mean of Y^* and X^*, $\overline{Y}_1^*$, $\overline{X}_1^*$ respectively.

The second group: It is the group of observations with a higher value than the median of the variable X^*, for which we find the arithmetic mean of Y^* and X^*, $\overline{Y}_2^*$, $\overline{X}_2^*$ respectively.

To calculate $\widehat{B_{IV}}$ we use the equation (5.2)

We apply equation number (5.2).

If n is odd, the middle observations can be omitted.

5.1.2. Bartlett's method

We arrange the observations of the variable X^* in ascending or descending order, then we form three groups for the instrument variable Z so that each group contains $n/3$ of the observations as follows [12] :

$$Z = \begin{cases} 1 & if\ observation\ is\ in\ the\ top\ group \\ 0 & if\ observation\ is\ in\ the\ middle\ group \\ -1 & if\ observation\ is\ in\ the\ bottom\ group \end{cases} \tag{5.5}$$

Now we ignore the observations in the middle group, then calculate the following for the lower and upper groups.

The lower group: we find it $\overline{Y}_1^*$, $\overline{X}_1^*$.

The upper group: we find it $\overline{Y}_3^*$, $\overline{X}_3^*$.

To calculate $\widehat{B_{IV}}$ we use the equation (5.2)

$$\widehat{\beta}_{1IV} = \left(\frac{\overline{Y}_3^* - \overline{Y}_1^*}{\overline{X}_3^* - \overline{X}_1^*} \right) \tag{5.6}$$

$$\widehat{\beta}_{0IV} = \overline{Y}^* - \widehat{\beta}_{1IV}\ \overline{X}^* \tag{5.7}$$

5.1.3. Durbin's method

We arrange the observations of the variable X^* in ascending order, then define the values of the observations of the instrument variable Z as ranks of the variable X^* , and then apply the parameter estimation equation [12].
We apply equation number (5.2).

$$\widehat{\beta}_{1IV} = \left(\frac{\sum_{i=1}^{n} Z_i \left(Y_i^* - \overline{Y}^* \right)}{\sum_{i=1}^{n} Z_i \left(X_i^* - \overline{X}^* \right)} \right) \tag{5.8}$$

To calculate $\widehat{B_{0IV}}$ we use the equation (5.7)

6. Non parametric regression estimator

The nonparametric model is estimated by using three smoothing methods, the first is kernel smoothing using (Nadaraya Watson), the second is K-Nearest neighbor smoothing, and the last is median smoothing.
The semi-parametric regression functions are estimated according to the following formula:

$$g_n \left(t^*, \widehat{\beta}_{1IV} \right) = \sum_{i=1}^{n} W_{ni}(t^*)(Y_i^* - \acute{X}_i^* \widehat{\beta}_{1IV} \ ) \tag{6.1}$$

Kernel smoothing : Kernel smoothing is weighted average estimates that use the kernel function as weights. The kernel's weight sequence (in the case of X is one-dimensional) is defined as [6]:

$$W_{ni}(t^*) = \frac{k_h(t^* - t_i^*)}{\hat{f}_h(t)} \tag{6.2}$$

Where

$$\hat{f}_h \left(t^* \right) = n^{-1} \sum_{i=1}^{n} K_h(t^* - t_i^*) \tag{6.3}$$

$\hat{f}_h \left(t^* \right)$: It is a density estimator for Rosenblatt-Parzen kernel.

Nadaraya – Watson Estimator : This estimator was proposed by Nadaraya 1965 and Watson 1964 based on the weight series method. This estimator is one of the most widely used and most common estimators in estimating the nonparametric regression function $g(t)$. This estimator has many characteristics, including the possibility of using it, whether the design is fixed or random [6**?**].
It is calculated according to the following formula :

$$W_{ni} \left(t^* \right) = \frac{K \left(\frac{t^* - t_i^*}{h} \right)}{\sum_{i=1}^{n} K \left(\frac{t^* - t_i^*}{h} \right)} \tag{6.4}$$

Where

$W_{ni}(X)$: Weight series whose sum = 1, which is a real, non-negative, continuous, finite, symmetric function, and it's integral = 1.
K : kernel function.
Kernel function properties :

$$\int_{-\infty}^{\infty} K(u)\, du = 1 \tag{6.5}$$

$$\int_{-\infty}^{\infty} uK(u)\, du = 0 \tag{6.6}$$

The Gaussian function will be used in this estimator.
h: bandwidth $h > 0$.

$$h = 1.06\ s\ n^{-1/5} \tag{6.7}$$

and that shape is destined Nadaraya – Watson be like this :

$$\hat{g}_h(t^*) = \frac{\sum_{i=1}^{n} K_h(t^* - t_i^*)\left(Y_i^* - \acute{X}_i^* \widehat{\beta}_{1IV}\right)}{\sum_{i=1}^{n} K_h(t^* - t_i^*)} \tag{6.8}$$

K-Nearest neighbor K-NN : It is a weighted average of different neighbors, i.e. it is the weighted average of the response variables for adjacency around x, this neighborhood is defined by the nearest k of the x variables in the Euclidean distance. Its shape is determined by kernel functions and bandwidth h [6, 2].
The weight sequence for K-NN was presented by Loftsgaarden & Quesenberry (1965) in the field of density estimation. In regression, K-NN is defined as :

$$\hat{g}_n(t^*) = n^{-1} \sum_{i=1}^{n} W_{ni}(t^*)\ \left(Y_i^* - \acute{X}_i^* \widehat{\beta}_{1IV}\right) \tag{6.9}$$

Where

$\{W_{ni}(t^*)\}_{i=1}^{n}$: weight sequence.

$$J_{t^*} = \{i\ :\ t_i^*\ is\ one\ of\ the\ k\ nearest\ observations\ to\ t^*\} \tag{6.10}$$

$$W_{ni}(t^*) = \begin{cases} \frac{n}{k} & if\ i\ \in J_t \\ 0 & otherwise \end{cases} \tag{6.11}$$

The Gaussian function will be used in this estimator.

Median smoothing : Smoothing the median is the closest technique to solving the problem of estimating the conditional median function, that is mean using the median function instead of the conditional expectation function. The conditional $med(Y\,|X = x)$ is more immune to the conditional expectation $E(Y\,|X = x)$. It is defined mathematically as follows [2] :

$$\hat{m}(x) = med\{Y_i^* : i \in J_{t^*}\} \tag{6.12}$$

$$J_{t^*} = as\ in\ the\ equation \tag{6.13}$$

The median of Y corresponding to X is calculated and K is the nearest neighbor of X.

7. The artificial flower pollination algorithm FPA

The flower pollination algorithm is one of the nature-inspired algorithms inspired by the process of flower pollination. It was created by Yang in 2012.
Its applications: It works with non-linear models, in image processing, in computer science, engineering, operations research, education (Selecting university academic credits), ...
Pollination takes two main forms: Biotic and Abiotic.
Biotic pollination: It means that pollen is spread by pollinators such as insects and animals, 90% of flowers belong to this group, meaning that pollen grains are transmitted by pollinators, such as swarms of insects and animals. Abiotic pollination: it means that wind and water carry out the process of pollination, 10% of the flowers take the abiotic form by pollination [13, 5].
The pollination process is carried out in two ways:
Pollination-Self occurs when a flower pollinates itself or another flower on the same plant. Pollination-Cross occurs when pollen grains are transferred from one flower of one plant to another flower of another plant.

8. Rules of FPA

The constancy of a flower and the behavior of pollinators in the pollination process can be described by the following four rules [13, 5]:

1. Bio Pollination and cross pollination: It is a global pollination process with pollinators that carry pollination and that make levy trips.

2. Abiotic self pollination: It is considered local pollination.

3. The constancy of a flower can be considered as the probability of reproduction which is proportional to the similarity of the two flowers concerned.

4. A switch or interaction between global pollination and local pollination can be controlled by the possibility of switch $p \in [0, 1]$.

Can be used (0.5) as an initial value for p ($p = 0.5$), and studies have indicated that ($p = 0.8$) works best as an optimum value for most applications.

9. Mathematical Representation Of Global And Local Pollination

Global pollination and local pollination are the main steps of the flower pollination algorithm. In global pollination, pollen grains are carried by pollinators such as insects, and pollen grains can travel long distances because insects can often fly and move for a long range. Therefore, rule No. (1) and rule No.(3) are mathematical as follows [13, 8]:

$$X_i^{t+1} = X_i^t + \gamma \, L(\lambda) \left(X_{best} - X_i^t \right) \tag{9.1}$$

Where
X_i^t : the solution for X_i in cycle t .
X_{best} : the best solution obtained is the solution in cycle t . Which is the best solution that has been found among all the solutions in the current generation.
γ : scaling factor to control step size.
$L(\lambda)$: pollination strength parameter.

X_i^{t+1} : the new solution.
L is the standard gamma function, and this distribution is used because it is suitable for large steps of swarms.

$$L \sim \frac{\lambda\, \Gamma(\lambda) \sin(\frac{\pi\lambda}{2})}{\pi} \frac{1}{s^{1+\lambda}} \qquad , \quad (s > s_0 > 0) \tag{9.2}$$

The value of s is calculated as follows:

$$s = \frac{u}{|v|^{\lambda-1}} \tag{9.3}$$

$$u \sim N\left(0,\ \sigma^2\right) \quad , \quad v \ \sim\ N(0,\ 1) \tag{9.4}$$

$$\sigma^2 = \left[\frac{\Gamma(1+\lambda)}{\lambda\, \Gamma\left(\frac{1+\lambda}{2}\right)} \cdot \frac{\sin(\pi\ \lambda/2)}{2^{(\lambda-1)/2}} \right]^{1/\lambda} \tag{9.5}$$

$$S\left(X_i^j(t)\right) = \frac{1}{1+\ e^{-X_i^j(t)}} \tag{9.6}$$

Usually the value of $s_0 = 0.1$.
In local pollination, the pollination is self-pollinating. It represents rule No. (2) and rule No. (3) mathematically as follows:

$$X_i^{t+1} =\ X_i^t +\ \epsilon\left(X_j^t - X_k^t\ \right) \tag{9.7}$$

Where
X_i^t : the solution for X_i in cycle t .
X_i^{t+1} : the new solution.
X_j^t , X_k^t : pollen from different flowers on the same plant, where k , j are randomly selected indices.
ϵ: a random variable that follows a uniform distribution $U(0,1)$.
After the global and local round of pollination, intensive exploitation was made and the best flower was taken, as shown in the following equation:

$$X_i^{t+1} =\ X_{best} + H\ (\epsilon_1 -\ [(\epsilon_2 -\ \epsilon_3)\, X_{best}] \tag{9.8}$$

Where
H : it is a control parameter, which is calculated as follows :

$$H = \begin{cases} 1, & if \quad \epsilon_4 < p \\ 0, & otherwise \end{cases} \tag{9.9}$$

where
ϵ_1 , ϵ_2 , ϵ_3 , ϵ_4 : random variables that follow a uniform distribution $U(0,1)$.

10. Structure of the Flower Pollination algorithm FPA [13, 11]

Objective minimize or maximize $f(x)$, $x = (x_1, x_2, \ldots, x_n)$
Initialize a population of n flowers/pollen gametes with random solutions

```
Find the best solution $X_{best}$ in the initial population
Define a switch probability $p \in [0, 1]$
While ( $t < MaxGeneration$)
  For $i = 1 \ : n \quad (all\ n\ flowers\ in\ the\ population)$
  If $rand < p$
    Draw a (d – dimensional) step vector L from a Levy distribution
    Global pollination via $X_i^{t+1} = X_i^t + \gamma\, L(\lambda) \left(X_{best} - X_i^t\right)$
  Else
    Draw $\in$ from a uniform distribution in $[0, 1]$
    Randomly choose j and k among all solution
    Do local pollination via $X_i^{t+1} = X_i^t + \epsilon \left(X_j^t - X_k^t\right)$
  End if
    Evaluate new solutions
    Calculate the fitness of the new solution $(f(X^{t+1}))$
    If new solutions are better, update them in the population
    If $f(X^{t+1}) \leq f(X^t)$ then
    $X^t = X^{t+1}$
  End if
   Find the current best solution $X_{best}$
End while
Output the best solution found
```

11. Comparison standard

Many criteria measure the amount of efficiency in estimating the quasi-parametric regression function and choosing the best model, but in this research, we will use the mean square error criterion for comparison [4].

In statistics, the mean squared error of an estimator measures the average of the squares of the errors - that is, the average squared difference between the estimated values and the actual value.

Application side

In this aspect, what was mentioned in the theoretical side will be studied and applied to real data related to the environmental aspect (environmental pollution/air pollution) in the city of Baghdad for the period from (3 June to 17 July) 2018.

Air pollution was studied with NO_2 gas and what are the factors affecting it, where the influencing factors were: NO_x is the explanatory variable with measurement error in the parameter part, and O_3 is the explanatory variable with measurement error in the non-parametric part, and NO_2 is the dependent variable with measurement error (response variable).
This data was taken from the Iraqi Ministry of Health and Environment/ air quality monitoring station/ HORIBA APNA-370 Ambient NOx Monitor.

12. Results and discussion

The real data results were obtained using MATLAB 2016a.

Table 1: shows the mean square error values of the models (MSE)

Semi-parametric model		MSE	MSE (FP)	Best
Nadaraya – Watson	**Wald**	**0.225229**	**0.003306**	**FP**
	Bartlett	**0.357334**	**0.003385**	**FP**
	Durbin	**0.175333**	**0.003316**	**FP**
K-Nearest neighbor	**Wald**	**0.013574**	**0.00321352**	**FP**
	Bartlett	**0.648600**	**0.004246**	**FP**
	Durbin	**0.010202**	**0.003231**	**FP**
Median	**Wald**	**0.013574**	**0.00321326**	**FP**
	Bartlett	**0.586634**	**0.004197**	**FP**
	Durbin	**0.010201**	**0.003231**	**FP**
Best		**Median-Durbin**	**Median-Wald**	

Through Table No. (1), we notice that the Median-Durbin model in the semi-parametric models appeared the best as it had the lowest MSE of 0.010201.

When employing semi-parametric models in the flower pollination algorithm, the Median-Wald model appeared to be the best, as its MSE value was 0.00321326.

When comparing the ordinary methods in semi-parametric models with the FP algorithm, we notice the superiority of the FP algorithm in all methods and for all models.
as shown in the following figures :

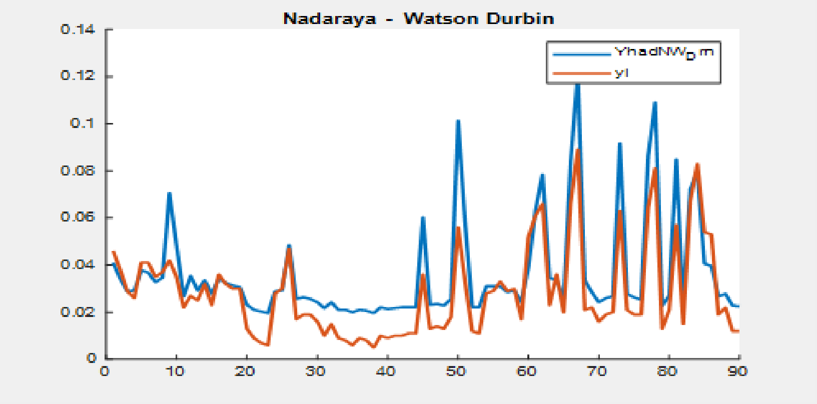

Figure 1: Nadaraya – Watson Durbin

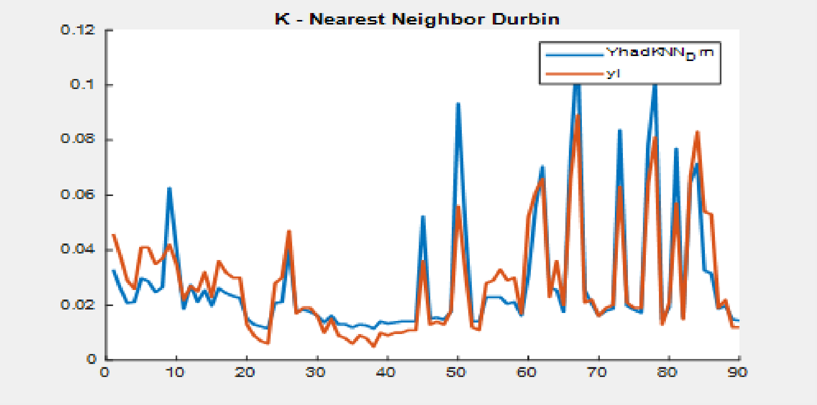

Figure 2: K-Nearest Neighbor Durbin

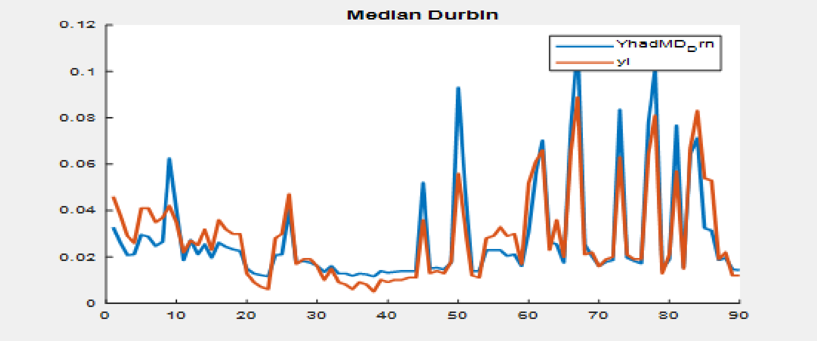

Figure 3: Median Durbin

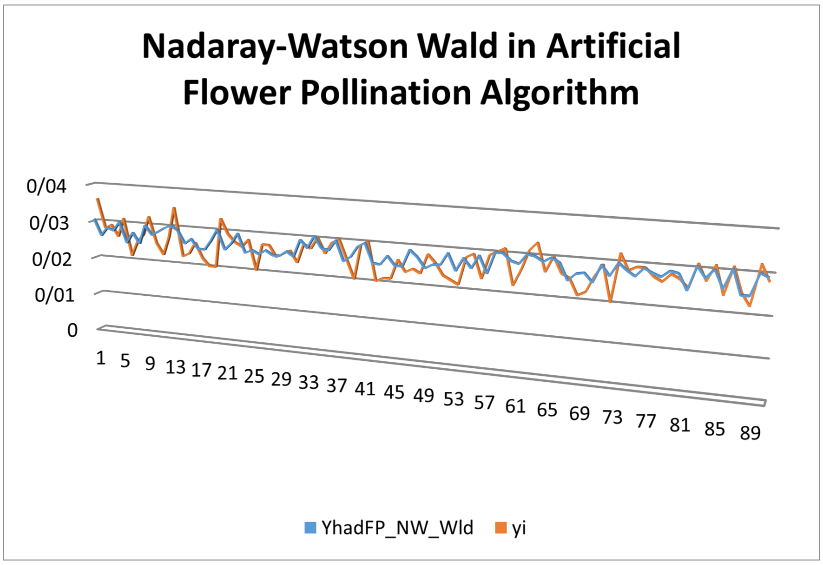

Figure 4: Nadaray-Watson Wald in Artificial Flower Pollination Algorithm

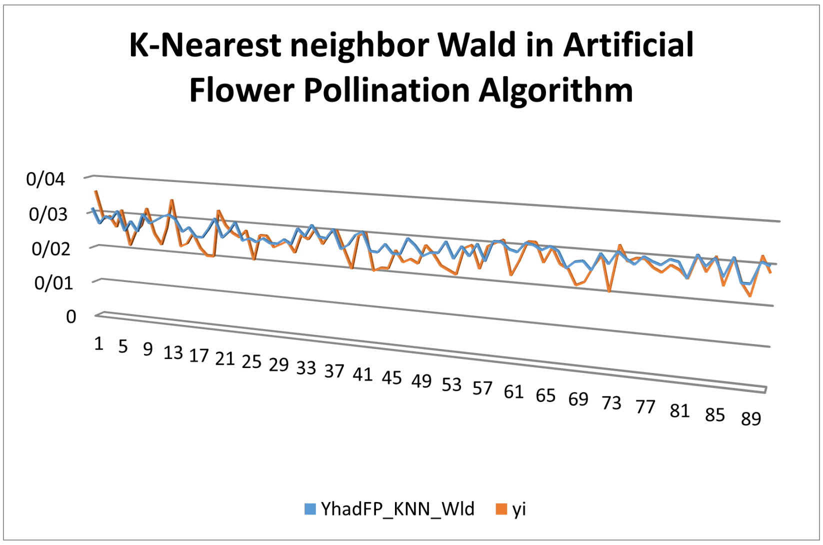

Figure 5: K-Nearest neighbor Wald in Artificial Flower Pollination Algorithm

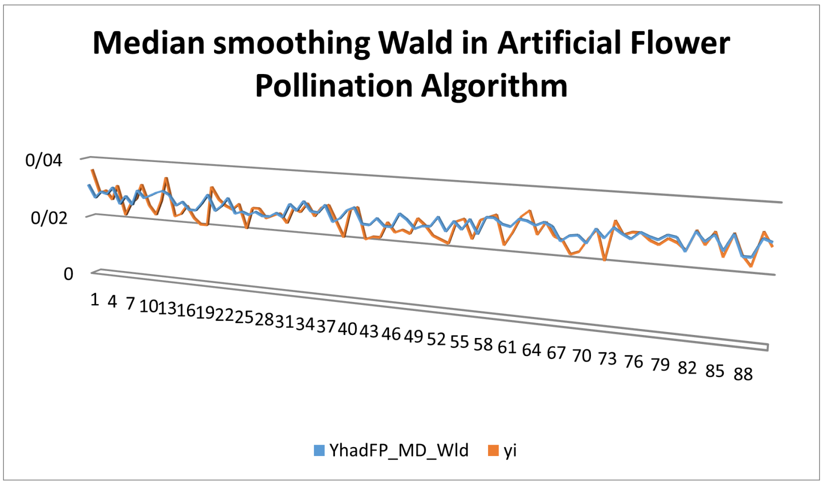

Figure 6: Median Wald –in Artificial Flower Pollination Algorithm

Table 2: Parameter values of the median-Durbin best model in statistical methods

parameter	result	Test values
B_0	0.0089902	0.789
B_1	0.3564154	17.803 *
G(t)	0.00000529	0.000089

Through Table 2, which represents the parameter values of the best model, the value of $\beta_1 = 0.3564154$ appeared, which is an influential and positive value, and this means that there is an effect of NO_x by 36% on NO_2 and directly.

The value of $G(t) = 0.00000529$ appeared, which is an ineffective and positive value, and this means that there is no effect for O_3 on NO_2 .

Table 3: Parameter values of the median-wald best model in employing - the flower pollination algorithm

	result	**Test values**
B_0	**0.0254972**	**7.28 ***
B_1	**0.0061816**	**10.04 ***
G(t)	**-0.0003071**	**0.052**

Through Table 3, which represents the parameter values of the best model in the FP algorithm, the value of $\beta_1 = 0.0061816$ appeared, which is an influential and positive value, and this means that there is an effect of NO_x by 6% on NO_2 and directly.

The value of $G(t) = -0.0003071$ appeared, which is an ineffective and positive value, and this means that there is an effect of O_3 by 0.03%.

13. Conclusions and Recommendations

In light of the theoretical side and based on the results of its application to the real data, a set of conclusions and recommendations were reached:

13.1. Conclusions

After executing the experiment on environmental pollution data/air pollution in the presence of measurement errors and the presented MSE results and the results of the parameters of the best model, the researcher concluded the following:

1. The results of MSE when using statistical methods in semi-parametric models with measurement errors indicated that the best model is Median-Durbin. When using the kernel smoothing (Nadaraya – Watson) method with the three parametric methods, we found that the best model is Nadaraya - Watson - Durbin. But when using the K-Nearest neighbor with the three parametric methods, we conclude that the best model is K-Nearest neighbor - Durbin. Finally, when using Median smoothing with the three parametric methods, we concluded that the best model is Median smoothing - Durbin.

2. By observing the results of MSE when employing the flower pollination algorithm, we conclude that the best semi-parametric model with measurement errors is Median-Wald. When using the kernel smoothing (Nadaraya – Watson) method with the three parametric methods using FPA, we concluded that the best model is Nadaraya - Watson - Wald. But when using the K-Nearest neighbor with the three parametric methods with FPA, we conclude that the best model is K-Nearest neighbor - Wald. Finally, when using Median smoothing with the three parametric methods, we conclude that the best model is Median smoothing – Wald.

3. When comparing the statistic methods in semi-parametric models with the FP algorithm, we notice the superiority of the FP algorithm in all methods and for all models.

4. By observing the parameter values in the best semi-parametric model when using the statistical methods, we concluded that NO_x has effect on NO_2 while O_3 does not affect on NO_2 .

Also, by observing the parameter values in the best semi-parametric model when employing FPA, we concluded that NO_x affects on NO_2 while O_3 has no effect NO_2 .

13.2. Recommendations

In light of the conclusions we reached through the research, the following recommendations can be summarized:

1. The importance of taking into consideration enough errors accompanying the measurement of different variables in any scientific, economic or social phenomenon.. and not neglecting or overlooking them due to their clear impact on modeling those phenomena and the accuracy of their results.

2. Conducting more studies on estimating the semi-parametric regression function with the presence of measurement errors using different semi-parametric models and developing them.

3. The use of artificial intelligence algorithms in estimating semi-parametric regression functions in scientific, social, and economic studies when there are measurement errors in their variables because of their great role in determining the best model or optimum estimate.

References

[1] V. M. Alao, J. R. G. Lansangan, and E. B. Barrios, *Estimation of semiparametric mixed analysis of covariance model* , Commun. Stat. Simul. Comput., (2019) 1-17.

[2] P. Bickel, P. Diggle, S. Fienberg, U. Gather, I. Olkin and S. Zeger, *Springer series in statistics* , New York: Springer, 2009.

[3] B. S. Everitt and A. Skrondal, *The Cambridge dictionary of statistics* , fourth edition, United States of America by Cambridge University Press, New York, 2010.

[4] Great Learning Team, *Mean Squared Error – Explained | What is Mean Square Error?* , Aug 8, (2020), https://www.mygreatlearning.com/blog/mean-square-error-explained/.

[5] A. E. Hassanien and E. Emary , *Swarm intelligence: principles, advances, and applications*, CRC Press, 2018.

[6] W. Härdle , *Applied nonparametric regression* , Universityat zu Berlin, Spandauer Str., D–10178 Berlin, 1994.

[7] Huque, M. H. Huque, H. D. Bondell, R. J. Carroll and L. M. Ryan, *Spatial regression with covariate measurement error: A semiparametric approach* , Biometrics,72(3)(2016) 678-686.

[8] A. E. Kayabekir, G. Bekdaş, S. M. Nigdeli and X. S. Yang, *A comprehensive review of the flower pollination algorithm for solving engineering problems* , Nat. Inspired Algorithms Appl, Optim., (2018) 171-188.

[9] M. Li, Y. Ma and R. Li , *Semiparametric regression for measurement error model with heteroscedastic error* , J. Multivar. Anal. , 171 (2019) 320-338.

[10] H. F. F. Mahmoud , *Parametric versus Semi and Nonparametric Regression Models* , arXiv preprint arXiv:1906.10221, (2019).

[11] E. Nabil , *A modified flower pollination algorithm for global optimization* , Expert Syst. Appl., 57 (2016) 192-203.

[12] Shalabh, IIT Kanpur, *Measurement Error Models* , Econometrics, Chapter 16, 2012.

[13] X. S. Yang , *Flower Pollination Algorithm for Global Optimization* , In: Unconv. Comput. Nat. Comput., Lecture Notes in Computer Science, 7445 (2013) 240-249 .

[14] L. Zhu and H. Cui, *A semi-parametric regression model with errors in variables* , Scandinavian, 30 (2)(2003) 429-442.