Periodicals of Engineering and Natural Sciences ISSN 2303-4521
Vol. 8, No. 1, March 2020, pp.263-271

Robust non-parametric regression models for some petroleum products

Raad Fadhel Hasan[1], Nabaa Naeem Mahdi[2], Aseel Abdul Razak rasheed[3]
[1,2,3] Statistics Department/Collage of administration and Economics, Mustansiriyah University

ABSTRACT

The importance of statistics appears in trying of demonstrating different phenomena by models that are nearer to the reality. These models may be causative and are built on the basis of reason and result this is called Regression Model. It has function shape and based on specific assumptions. But, sometimes we come up with a more flexible approach in case of absence of the knowledge of the studied phenomena or the first time made experiment or when we can't mention the causative function between the variables. This type of models is called Non parametric Regression. It is a type of regression in which the value of the independent variable doesn't take a specific shape, but built from information taken from the data and this requires a sample of volume bigger than the usual volume. For the parametric regression, the data suggest the structure of model and parameter estimation. Due to the importance of the petroleum products in lives of people that is considered as a source of civil and civilized development. Three petroleum products are taken (white oil, diesel oil, and fuel oil) through the non-parametric regression application. Besides, five non-parametric methods are taken (Lowes, Robust Lowes, Mean, Median, and Polynomial). They have been compared between each other, and we concluded that the robust non-parametric regression is the best way to compare the value of mean square error.

Keywords: Kernel Regression, Epanechikov Kernel function, Quadratic Kernel function, Bandwidth, Gaussian Kernel

Corresponding Author:

Aseel Abdul Razak Rasheed
Statistics Department, Collage of administration and Economics, Mustansiriyah University
E-mail: aseelstat@uomustansiriyah.edu.iq

1. Introduction

When modeling the parametric regression model that corresponds to the dependent variable (y), the explanatory variable (x) or several explanatory variables, these models are used when there is information in the form of this relationship and a set of assumptions is made. The model parameters are analyzed and evaluated using known estimation methods such as the maximum likelihood estimation, the least squares method or Bayesian Estimation and other methods. The estimation function of the regression function is the estimation results of a curve chosen from a set of curves to match the data. This test is subjected to several conditions to match the expected forms. Another method to match the curves of the studied data is the non-parametric regression methods. These methods are more flexible, so that they allow high flexibility in the forms of the slope curve. The assumption using these methods is that the function that links the relationship between the two variables. It is derived and dependent on the main data since the data type explains the actual shape of the regression curve and forms the properties of these methods. This is useful to discover the general relationship between the two variables and to be able to give predictions to the new observations without referring to the fixed non-linear model.

As well as the ability to find false views through the study of isolated points or remote, it represents a flexible way that is able to replace the missing values or the values shown between the values next to the point.
The aim of the study is to focus on the methods of non-parametric and robust non-parametric regression by taking a series of time series and aligning them to construct a non-parametric regression model. Then, comparing between them to know which of these models is the most precise using criterion goodness of fit.

2. Non - parametric regression

The regression model is described as follows: $y_i = m(\beta x_i') + \xi_i \qquad i = 1,2,.....,n$

y_i is the value of the dependent variable , β is the value of the parameters of the model, x_i stands for matrix of independent variables and ξ_i is the limit of the random error that is supposed to be distributed naturally. The parameters of the model are often estimated in the maximum likelihood (ML), the OLS or other methods, by using the following relationship [3-5]:

$$\hat{\beta} = (x'x)^{-1}x'y$$

The model's estimator and efficiency are subject to the availability of normal distribution properties to model errors. Since this and other properties are not met in most data, more precise models are constructed and the distribution is not taken into account. These are the non-scientific models which are described as follows:

$$y_i = g(x_i') + \xi_i$$

The model is characterized by its non-compliance with the parameter, so the estimation is directly related to the function g(.), The function should be continuous. Below is a review of some of the estimation methods for the non- parametric regression model.

3. Lowes's normal method

The following algorithm is used to estimate the non- parametric regression model [6-8]:

1- Find the Euclidean distance (d (i, j)) between the observations (i and j).
2- We choose the ratio of (h) from (N) chosen by the close-to-view observations (i) according to the relationship [0.40N], where [ ] means the largest positive integer determining the number of nearby observations.
3- Calculate the relative dimension according to the following relationship:

$$D(i,j) = \frac{d(i,j)}{\max_j (d(i,j))}$$

It is possible to calculate the weight associated with the Kernel function through W (j) and through the use of the Tricube function.

4- Weight function:

$$w(u) = \begin{cases} (1-|u|^3)^3 & |u| < 1 \\ 0 & |u| \geq 1 \end{cases}$$

As a result, the weight accompanying the observation (j) is described by the relationship:

$$w_j = g\left(\frac{x_j - x_0}{\max_j (d(i,j))}\right)$$

5- The weighted regression equation is estimated to obtain the parameters of the model, and then to predict or estimate after reducing the following relationship:

$$s = \sum_{j=1}^{n} g\left(\frac{y_j - y_0}{\max(d_{i,j})}\right)\left(y_j - \alpha_0 - \alpha_1 x_j - \alpha_2 x_j^2\right)$$

Or to reduce the sum of the following weighted squares:

$$s = \sum_{j=1}^{n} w_j (y_j - \alpha_0 - \alpha_1 x_j - \alpha_2 x_j^2)^2$$

When we obtain a regression model if ($\alpha_2 = 0$) in the case of liner regression and when it does not equal to zero, we get quadratic regression. The equation is estimated through the ordinary less square weighted method at which the value is found at the point (x_0), according to the following relationship:

$$\hat{Y}_{(x_0)} = \hat{\alpha}_0 + \hat{\alpha}_1 \hat{x}_0 + \hat{\alpha}_2 x_0$$

Thus, a separate equation is estimated for all values (x).

4. Robust Lowes's model

New weights are calculated according to the following relationship [8]:

$$D'(i,j) = \frac{|r_{(j)}|}{6.ncdian_j(|r_j|)}$$

Where $r_{(j)} = y_j - \hat{y}_{loess}(x_j)$, the error is estimated from Lowes method habit and the new weight is calculated based on previous weights through quadratic function.

$$w_j^* = \begin{cases} \frac{15}{16}(1-u^2)^2 & |u| < 1 \\ 0 & |u| \geq 1 \end{cases}$$

5. The Kernel function

The main characteristic of the Kernel regression models is the use of the kernel function, which assigns weight to all data observations based on the distance dimension of the predictive observations. The larger weight for the close and distant observations has the lowest weight. The multiple kernel function has the following functions [4, 5, 9]:

1- The uniform kernel function and its formula:

$$K(u) = 0.5 \qquad I_{|u|} \leq 1$$

2- Triangle kernel function and its formula is:

$$K(u)=(1-|u|) \qquad I_{|u|} \le 1$$

3- Epanechikov kernel function and its formula is:

$$K(u)=0.75(1-u^2) \qquad I_{|u|} \le 1$$

4- Quadratic Kernel function and the following formula is:

$$K(u)=\frac{15}{16}(1-u^2)^2 \qquad I_{|u|} \le 1$$

5- Tri weight kernel function and its formula are:

$$K(u)=\frac{35}{32}(1-u^2)^3 \qquad I_{|u|} \le 1$$

6- Tri cube kernel function is:

$$K(u)=(1-|u|^3)^3 \qquad I_{|u|} \le 1$$

The second side that we rely on in the kernel regression model is the width of the bandwidth that accompanies each variable. The Y is a variable dependent on ($x_1,x_2,\ldots\ldots x_{k)}$ of the independent explanatory variables and for the prediction of observation (y_i) where ($1 \le i \le n_{vaild}$) based on the independent observations (j) where ($1 \le j \le n_{learn}$). The weight through Gaussian kernel function and the width of a fixed package is equal to (h_I). An independent variable (x_I) in which (l = 1,2, ..., k), can be described as follows:

$$w_{ij}=\frac{1}{(\sqrt{2\pi})^k \prod_{\ell=1}^{k} h_\ell}.\exp\left[-\sum_{\ell=1}^{k}\left(\frac{x_{j\ell}-x_{i\ell}}{h_\ell}\right)^2\right]$$

For estimation, we rely on a polynomial rank. If the rank is equal to zero, it is called Constant polynomial from which we obtain a systematic estimator Nadaraya-Watson. The calculation of the prediction of the observation (i^{th}), is according to the following relation:

$$y_i=\sum_{j=1}^{u_{\ell ewm}} w_{ij}.y_i \Big/ \sum_{j=1}^{u_{\ell earn}} w_{ij}$$

In this case, only the explanatory variables are used to calculate the weights of the learning sample. The model of class (1, 2) predictive values can be calculated by the following relationships:

$$y_i=a_0+\sum_{\ell=1}^{k} a_\ell x_{i\ell}^{\ell} \qquad Level\ (1)$$

$$y_i = a_0 + \sum_{\ell=1}^{k} a_\ell x_{i\ell}^{\ell} + \sum_{\ell=1}^{k}\sum_{m=1}^{k} b_{\ell m} x_{i\ell} x_{im} \qquad Level \ (2)$$

Before estimating the polynomial models, the observations are weighted through the use of Nadaraya-Watson that the sample size for a stage is known by two phases:

1- For the estimation of (yi) a fixed number is taken.

2- The closest neighbor (k nearest neighbors) which is a complementary function by limiting the size of the sample in the learning phase to the value of (k) of observation

6. Application side

Due to the importance of local oil products and their impact on the progress of the country and the well-being of the people, the data, which represent three time series of local sales of petroleum products, were taken as monthly data from (2005 -2017) was obtained from the refinery Dora which represents (154) observations. The first series represents white oil, one of the oil derivatives resulting from the process of refining crude oil. It is called kerosene, and it is known since long time ago. It was first extracted by the scientist Al-Razi and it was called white oil in his book known (Kitab Al Asrar). In the nineteenth century, kerosene became one of the most important derivatives. In 1854, the name of kerosene was registered as a brand name in America, and this company has been monopolizing this name for many years until it is widely used on every tongue. Kerosene is one of the most important refinery products is within a range of thermal ranges ranging from (150-250) C. It contains a number of paraffin's and naphthalene, which is used in domestic as fuel for cooking, heating and lighting. It is also a major component of jet fuel. In addition, it is used in the manufacture of some industrial solvents and thinner solvents in paints, which are mixed in different proportions to reduce the degree of viscosity. The second series represents diesel oil. It is named after the German inventor Rudolf Diesel. In 1892, he invented a fuel-powered engine that is less capable of ignition and explosion than gasoline, so it is used in military vehicles, trucks and some cars. The third series represents fuel oil, which is widely used in industry. Many industrialized systems have replaced coal since the beginning of the 20th century and is currently used in heavy industries (mining, cement and glass).

6.1. Data Analysis

The first series is for the white oil after the application of the program (XLSTAT) on the data and using the non-parametric regression. Five methods were applied to the model of the non-parametric regression based on Lowes, Robust Lowes, Mean, Median, and Polynomial criteria. The following results are obtained as shown in Table 1.

Table 1. The five methods used with goodness of fit for white oil

	Method	Kernel	R^2	MSE	RMSE
A	Lowes	Tricube	0.616	2858659596.015	53466.434
B	Robust Lowes	Tricube	0.617	2853569715.704	53418.814
C	Mean	Tricube	0.459	4030810032.348	63488.661
D	Median	Tricube	0.454	4096694928.362	64005.429
E	Polynomial	Tricube	0.376	4652636949.052	68210.241

From the observation of the table above, we find that the second method is the best where the coefficient of identification (R^2) was the greatest compared to the other four methods. In addition, it has achieved the lowest MSE and RMSE with values of 2853569715.704 and 53418.814. Then, the third, fourth and fifth ones have come respectively. The chart of the above-mentioned methods can be seen as depicted by Figure 1.

The second series which represents diesel oil for has used for the previous model of non-parametric regression, and the results have shown in Table 2.

Table 2. The five methods used with the goodness of fit for diesel oil

	Method	Kernel	R^2	MSE	RMSE
A	Lowes	Tricube	0.737	4835783.021	2199.041
B	Robust Lowes	Tricube	0.730	4956362.057	2226.289
C	Mean	Tricube	0.600	7343813.476	2709.947
D	Median	Tricube	0.592	7540740.770	2746.041
E	Polynomial	Tricube	0.179	15083714.879	3883.776

The results of the diesel oil series showed that the first method was the best to give it the largest coefficient of determination (R^2) with the value of 0.737 with less MSE and RMSE values of 4835783.021 and 2199.041 respectively. The second method ranked second by noting Good match criteria and then methods 3, 4 and 5, respectively. Third series represents fuel oil after applying the non- parametric regression model on the series, the following results were obtained in Table 3.

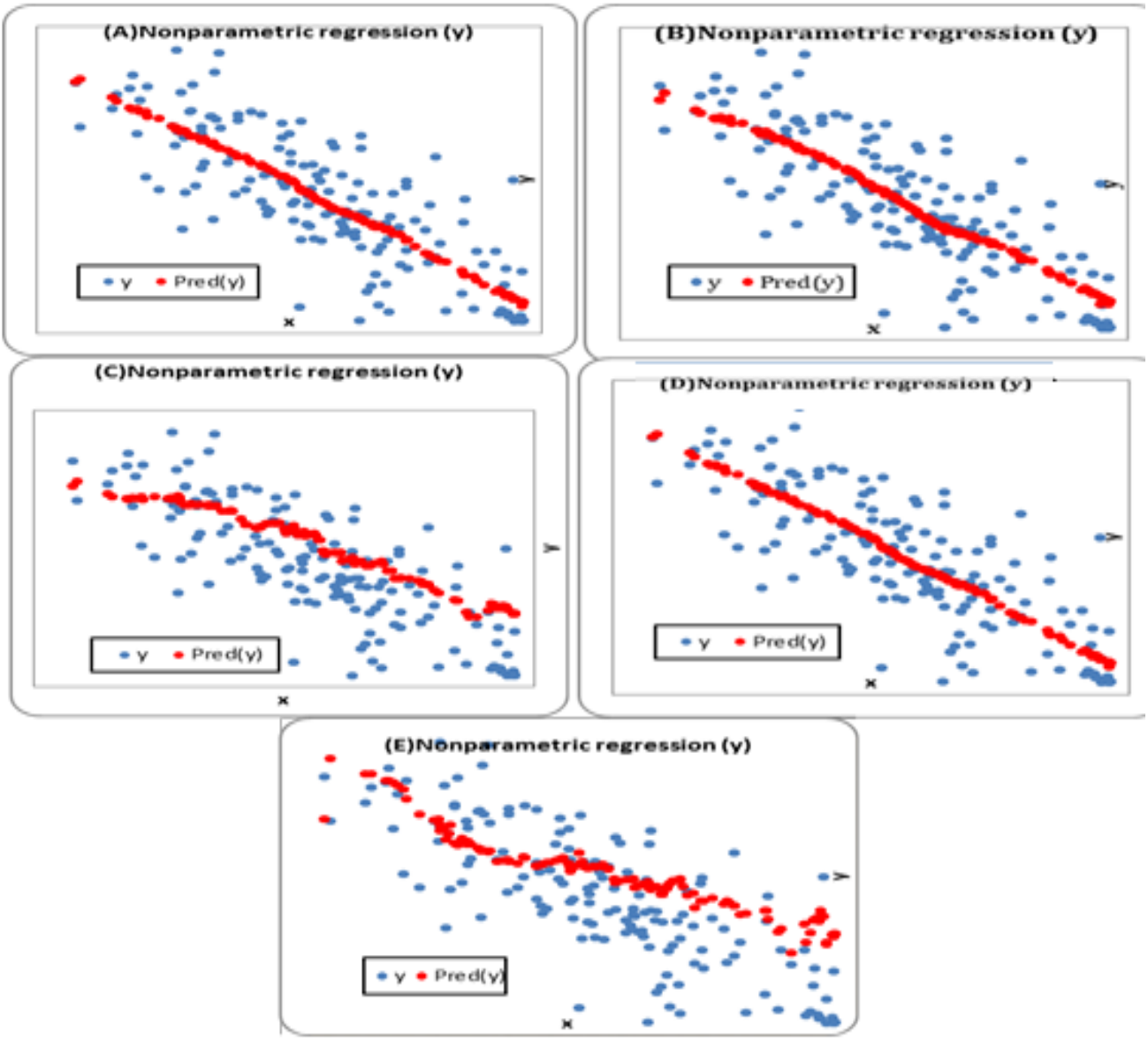

Figure 1. The first time series for each method

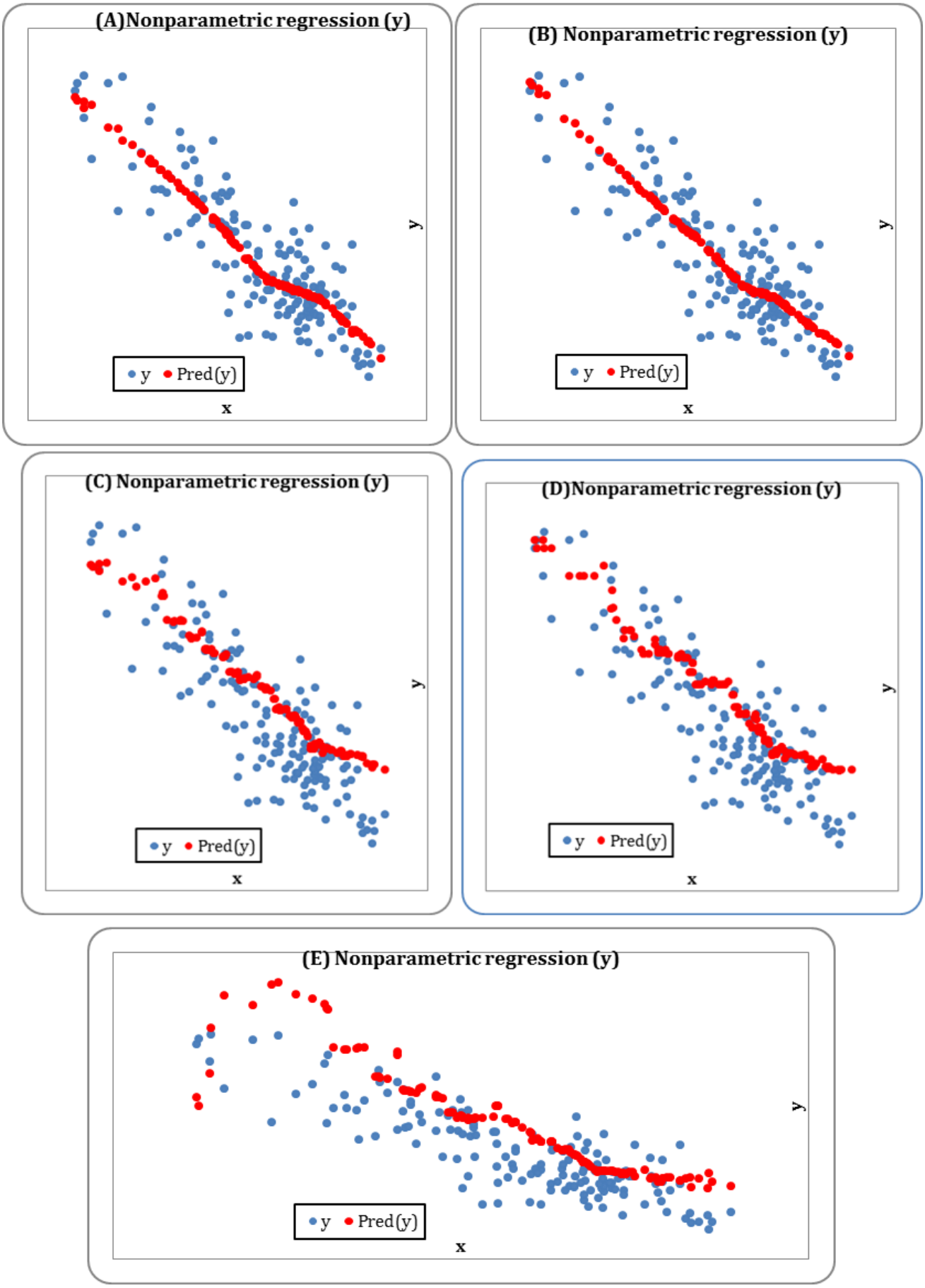

Figure 2. The second time series for each method

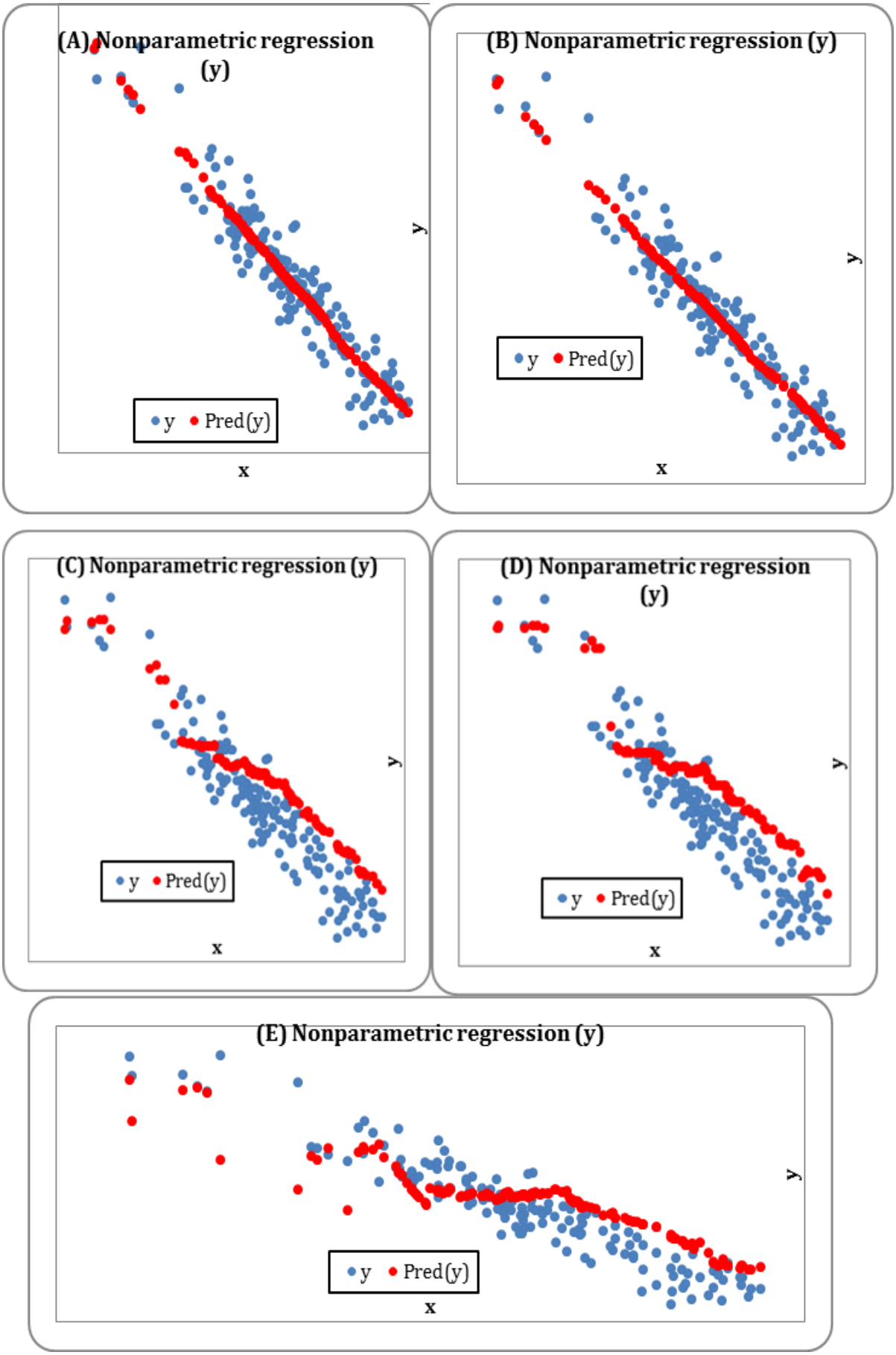

Figure 3. The third time series for each method

Table 3. The five methods used with the goodness of fit for fuel oil

	Method	Kernel	R^2	MSE	RMSE
A	Lowes	Tricube	0.896	4666497898.845	68311.770
B	Robust Lowes	Tricube	0.896	4645214576.896	68155.811
C	Mean	Tricube	0.698	13533996984.007	116335.708
D	Median	Tricube	0.634	16479309506.040	128371.763
E	Polynomial	Tricube	0.695	13660336307.828	116877.441

The results showed that the first and second methods were equal in R^2 with the value in 0.896. The other criteria for goodness of fit MSE and RMSE for the second method were better than those in the first method. Then the third, fifth and fourth are come respectively.

7. Conclusions and recommendations

- All of the estimated non-parametric models indicate the superiority of the Robust Lowes's model, which takes into account the nature and behavior of outlier data.
- The data used contain extreme values that cannot be treated according to the usual non-parametric normal estimation methods because of their low efficiency in estimating the model.
- Most data and even those collected from government sources are still inefficient in terms of accuracy in assembly, which reduces the efficiency of the estimated models, whether parametric or non-parametric.
- Working on the followers of estimation methods is recommended taking into account the nature of inaccuracy in the Iraqi data and to address these data were used more concrete methods.
- Building a data bank (database) for data in a high accuracy and making it in the researcher's field.

References

[1] John Fox, "Nonparametric Regression", Mc Master, Hamilton, Canada, 2004.
[2] H. Denies and L. Yan, "Cross-Validation in Nonparametric Regression with Outliers", *The Annals of Statistics*, Vol. 33, No. 5, pp.2291-2310,2005.
[3] A. Azzalini, A.W. Bowman and W. Hardle "On the use of nonparametric regression for model checking", *Biometrika*, vol.76, pp.1-11,1989.
[4] John Fox, "Nonparametric Regression", Appendix to an R and S-Plus Companion to Applied Regression, 2002
[5] R.W. Schucany, "Kernel smoother: an overview of curve Estimators in Nonparametric Statistics", Department of Statistical Science, SMU, Dallas TX, 2004.
[6] W. Hardle , "Applied Nonparametric Regression", Cambridge MA : Cambridge University Press,1994.
[7] T. Hastie and R.J. Tibshirani, "Generalized Additive Model", Chapman & Hall, London 1990.
[8] D. Aydi, "A Comparison of the Nonparametric Regression Models Using Smoothing Spline and Kernel Regression", World Academy of Science, Engineering and Technology, 2007.
[9] Q. Li, J. Racine, "Nonparametric Econometrics Theory and Practice", Princeton University, 2007.