AL- Rafidain
University College

PISSN: (1681-6870); EISSN: (2790-2293)

Journal of AL-Rafidain
University College for Sciences

Available online at: https://www.jrucs.iq

JRUCS

Journal of AL-Rafidain
University College for
Sciences

# Using Nonparametric Procedure to Develop an OCMT Estimator for Big Data Linear Regression Model with Application Chemical Pollution in the Tigris River

| Assist. Prof. Dr. Ahmed M. Salih | Prof. Dr. Munaf Y. Hmood |
|---|---|
| amahdi@uowasit.edu.iq | munaf.yousif@coadec.uobaghdad.edu.iq |
| Department of Statistics - College of Administration and Economic – University of Wasit, Wasit, Iraq | Department of Statistics - College of Administration and Economic - Baghdad University, Baghdad, Iraq |

**Article Information**

**Article History:**
Received: December, 24, 2022
Accepted: March, 3, 2023
Available Online: December, 31, 2023

**Keywords:**
Big Data, SICA, OCMT, Chemical Pollution, Nonparametric

**Correspondence:**
Assist. Prof. Dr. Ahmed M. Salih
amahdi@uowasit.edu.iq

**Abstract**

***Chemical pollution*** *is a very important issue that people suffer from and it often affects the nature of health of society and the future of the health of future generations. Consequently, it must be considered in order to discover suitable models and find descriptions to predict the performance of it in the forthcoming years. Chemical pollution data in Iraq take a great scope and manifold sources and kinds, which brands it as* ***Big Data*** *that need to be studied using novel statistical methods. The research object on using* ***Proposed Nonparametric Procedure NP*** *Method to develop* ***an (OCMT)*** *test procedure to estimate parameters of linear regression model with large size of data (Big Data) which comprises many indicators associated with chemical pollution and profoundly have an effect on the life of the Iraqi people. The* ***SICA*** *estimator were chosen to analyze data and the* ***MSE*** *were used to make a comparison between the two methods and we determine that* ***NP*** *estimator is more effective than the other estimators under* ***Big Data*** *circumstances.*

https://doi.org/10.55562/jrucs.v54i1.620

## 1. Introduction

The analysis of Big Data sets comes to be a substantial matter for many researches, it represents an abundant challenge for academics and data analyzers to improve more of the effective analyzing methods in order to explain the difficulties that appear in data when dimensions grow bigger. Scientists adopted algorithms schemes to present their techniques and methods in data analysis that should be easy to comprehend. Researchers offered definitions for Big Data such like [3].

Boyd and Crawford [10] elucidated it as “A cultural, technological, and scholarly phenomenon that rests on the interplay of Technology, Analysis and Mythology”. Chen et al [10] define it as “The data sets and analytical techniques in applications that are so large and complex that they require advanced and unique data storage, management, analysis, and visualization technologies”. Big Data is used in numerous scientific applications such as marketing, healthcare, demography, and various other fields [2]. Collecting information and summarizing them is the primary stage in any of the statistical methods with information or without and choosing the model relying basically on the knowledge behind high dimensions and their nature [6]. A number of methods have been presented to summarize be use in information summary from Big Data as a primary step before taking action in the analyzing process, including Factorial Analysis, Factor Models, and Panelized Least Squares. [7]. We selected a linear regression model with various instructive variables in the form:

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon} \qquad (1)$$

Where $\boldsymbol{Y}$ symbolizes ***(n × 1)*** variable vector (Chemical Pollution Ratio CPR in Tigress River Water) and $\boldsymbol{X}$ is ***(n × p)*** dependent variables matrix having a great size of dimensions, that have an effect on CPR , $\boldsymbol{\varepsilon}$ is ***(n × 1)*** random vector of errors and $\boldsymbol{\beta}$ is (p × 1) represent the vector of coefficients that we wish to estimate.

## 2. SICA

**SICA** which is Smooth Integration of Counting and Absolute Deviation technique presented by Lv &Fan 2009 [2] who presented penalized kind that defies model selection complications. They began with model in (1) and supposed that $\boldsymbol{\partial}$ is original vector of parameters, which hypothetically energies zero, but could be nonzero vector that develops penalty function. Penalized regressions characteristically employ the first order of the Lp-norm. Lv & Fan acclaimed a mix of Lp-norms which are L1 and L2-norm as a parameter of penalty to brand shrinkage for $\beta$. They had chosen a penalty parameter that was reflected by [3] Nikolova as listed below:

$$\boldsymbol{f(\lambda,\beta)} = \boldsymbol{\lambda}\frac{\boldsymbol{(m+1)\|\beta\|_1}}{\boldsymbol{m+\|\beta\|_1}} \qquad (2)$$

$f(\lambda,\beta)$ is the penalty parameter employed to reduce totality errors and $0 < \lambda < 1$ the value of $\lambda$ is set by researcher and $\boldsymbol{m}$ is constant $m > 0$ a tiny number that is allegedly to be positive small number [2]. The scholars proposed to promote denominator of (2) to quadratic and employ the original regression parameters $\boldsymbol{\partial}$ in place of $\beta$. The outcome is a penalty parameter that equals a ratio that lies amid of L1, L2 norms and attains the scattered retrieval over $\beta$ as:

$$\boldsymbol{f(\lambda,\beta)} = \boldsymbol{\lambda}\frac{\boldsymbol{(a+1)\|\beta\|_1}}{\boldsymbol{(a+\|\partial\|_1)^2}} \qquad (3)$$

Various methods were presented to estimate $\boldsymbol{\partial}$ . Lv and Fan commended $\theta = \hat{\beta}^{OLS} = (X'X)^{+}X'y$ .Where + Pseudoinverse that allows solving any least squares system [4] that uses to avoid singularity of the matrix $X'X$ . **SICA** estimator is the justification of following optimization [2].

$$\boldsymbol{\hat{\beta}^{SICA}} = \mathbf{arg}\min_{\boldsymbol{\beta}}\left(\boldsymbol{2^{-1}\varepsilon'\varepsilon + \lambda k\|\beta\|_1}\right) \qquad (4)$$

Where $k = \frac{\boldsymbol{(m+1)}}{\boldsymbol{(m+\|\partial\|_1)^2}}$ and via the terms of matrices.

$$\boldsymbol{L(\beta)} = \left(\boldsymbol{2^{-1}}\left(\underline{\boldsymbol{y}} - \boldsymbol{X\beta}\right)'\left(\underline{\boldsymbol{y}} - \boldsymbol{X\beta}\right) + \boldsymbol{\lambda kv\beta'}\right)$$

Where $\boldsymbol{v}$ is a $(p \times 1)$ vector of ones.

$$\boldsymbol{L(\beta)} = \left(\boldsymbol{2^{-1}}(\underline{\boldsymbol{y}}'\underline{\boldsymbol{y}} - \boldsymbol{2X'\beta} + \boldsymbol{\beta'X'X\beta}) + \boldsymbol{\lambda kv\beta'}\right)$$

By discriminating regarding $\beta$ and equalizing to zero we estimate [10]:

$$\boldsymbol{\hat{\beta}^{SICA}} = \boldsymbol{(X'X)^{+}(X'Y - \lambda kv)} \qquad (5)$$

## 3. Proposed Estimator

Our suggested estimator symbolizes an adaption of OCMT estimator which is an iterative process of selecting variables, first we discuss the OCMT then we propose our suggested process to improve it.

***OCMT*** denotes another scheme of the penalized regression method that elasticities courtesy with the predictive power of some covariates in its place of taking all the $\boldsymbol{p}$ covariates that rely upon selecting some regressors upon their competence to explain the explanatory variable. method was acquiesced by Chudik, Kapetanios and Pesaran in 2016 [8]. They designed a process of new model selection for High dimension-sets. The notion is to exam every regression coefficient $\boldsymbol{\beta_j}$ separately one after another and giving a focus on the impact and the marginal t of regressors. Next the operation is repeated till all major covariates are intricate statistically in regression. They secondhand notions of multiple testing to have a control over likelihood of electing the factual model. They called the novel method of estimation as **"One covariate at a time multiple testing"** ***OCMT***.

The individuality of $\boldsymbol{X_j = X_1, X_2, \dots, X_p}$ is set to be signal variables and for abridging the process and signify the entire of covariates in the regression model as $\boldsymbol{S_p = \{x_{ij}, j = 1, 2, \dots k, \dots . k^*, \dots, p\}}$ . There are collections of covariates; $\boldsymbol{k}$ symbolizes entirely covariates with $\boldsymbol{(\beta_j \neq 0)}$,

2nd collection $\{\boldsymbol{k+1, \dots, k^*}\}$ indicates the covariates with $\boldsymbol{(\beta_j = 0)}$ but they retain an influence of net impact of regression model, although the residual $\boldsymbol{p - k^*}$covariates stand for the covariates having peripheral impact on the linear model [5]. The notion is to improve several covariates from the 2nd collection to the 1st collection by some test statistic repetitions that choose the greatest noteworthy regressors.

As k is limited but unidentified, we set $\boldsymbol{h = 1, 2, .. p.}$ Pesaran & Smith [9] expressed the mean of the net impact as:

$$\boldsymbol{\theta_j = \sum_{h=1}^{p} \beta_h \sigma_{jh} \quad , \quad j = 1, 2, \dots, p} \qquad (6)$$

The denotation of the net influence is effect that can be created by one covariate. to test hypothesis $\boldsymbol{H_0 : \beta_j = 0 \quad Vs \quad H_0 : \beta_j \neq 0}$ as we focus on net impact.

The parameter $\theta_j$ has a major role in the procedure of selecting regressors, by assuming we have $Y$ and $p$ covariates of $X$. Initially, we may write the model in (1) as follows [5].

$$\boldsymbol{Y = X\emptyset + \varepsilon} \qquad (7)$$

Where $\emptyset_j = \frac{\theta_j}{\sigma_{jj}}$ and $\theta_j$ as in (6), now the parameter $\emptyset$ reflect the peripheral and the total net impact, the notion of $\beta$ by $\emptyset$ is that the final one displays the net and marginal effects on the dependent-variable. $\boldsymbol{Y}$, and force be an enhanced measure that alters from repetition to other because the lasting variable after selections. The t test of $\emptyset$ of the model in (7) can be symbolized to test the hypothesis $\boldsymbol{H_0 : \emptyset_j = 0 \quad Vs \quad H_1 : \emptyset_j \neq 0}$ as $\boldsymbol{t_{\emptyset j}}$ in the following form [7].

$$\boldsymbol{t_{\emptyset j} = \frac{\widehat{\emptyset}_j}{s.e.(\widehat{\emptyset}_j)}} \qquad (8)$$

And parameter $\emptyset$ can be expressed as:

$$\boldsymbol{\widehat{\emptyset}_j = \gamma_j \left(X_j{}' M X_j\right)^{-1} X_j{}' M Y} \qquad (9)$$

Where $var(\widehat{\emptyset}_j) = \boldsymbol{\gamma_j \left(X_j{}' M X_j\right)^{-1}}$ and $\gamma_j$ indicate the variance stricture by means of net impact formula, then by estimating the coefficients, (8) that expressed as [11]:

$$\boldsymbol{t_{\emptyset j} = \sqrt{\gamma_j \left(X_j' M X_j\right)^{-1}}\, X_j' M Y} \qquad (10)$$

Where $M = I_n - \tau_n \tau_n{}'/n$ , $\tau_n$is $n \times 1$ ones vector and $\boldsymbol{\gamma_j = \sum_{h=1}^{p} \widehat{\sigma}_{jh}^2}$ and the estimation of the marginal covariance as:

$$\boldsymbol{\widehat{\sigma}_{jh}^2 = \frac{R_j' R_h}{n}} \qquad (11)$$

Where $\boldsymbol{R_j = \left[I_n - X_j \left(X_j' X_j\right)^{-1} X_j'\right] Y}$ and for each covariate in (10) we reject $\boldsymbol{H_0 : \emptyset_j = 0}$ if $\boldsymbol{|t_{\emptyset j}| > C_{(p,\alpha)}}$ where $C_{(p,\alpha)}$ is the critical value as follow:

$$\boldsymbol{C_{(p,\alpha)} = \Phi^{-1}\left(1 - \frac{p_r}{2cp^{\alpha}}\right)} \qquad (12)$$

Where $\boldsymbol{\Phi^{-1}}$ measure of the standard normal inversions, where $c, \alpha$ are constants. The selection of critical value is vital meanwhile it owns a control over the influence of the procedure of selection.

In the conclusion of the first step of multiple selection process, all covariates that embrace $\emptyset \neq \mathbf{0}$ are selected and presume k is the amount of the selected covariates in the $1_{st}$ phase. Let $\boldsymbol{X_{k1}}$be the matrix that consumes all the preferred covariates in $1_{st}$ stage, while $\boldsymbol{p-k1}$ covariate hold in matrix $\boldsymbol{X_R}$.

In the following step, the model for $\boldsymbol{p-k1}$ covariates is written in a similar manner of the first step, and the mean net impact was rewritten [10].

$$\boldsymbol{\theta_j = \sum_{h=k1+1}^{p} \beta_h \sigma_{jh}} \qquad , \quad \boldsymbol{j = 1, 2, , \ldots, p-k} \tag{13}$$

For the rest covariates, similarly the regression model is written as similarly to the $1_{st}$ stage [20].

$$\boldsymbol{Y = X_R \emptyset + \varepsilon} \tag{14}$$

Assume that $\boldsymbol{X_{k2}}$ will be matrix includes all designated covariates of additional stage, the t test for the extra stages will be repeated till there is no covariates selected by the t test that disregards covariates that do not private peripheral impact on the explanatory variable. Then, we established $X_k$ to be the matrix of entire covariates elected in all phases of test procedure

$\boldsymbol{X_k = [X_{k1}, X_{k2}, X_{k3}, \ldots.]}$.

In the last stage, the OCMT estimator for model which covers completely the nominated covariates, will be OLS estimator as.

$$\boldsymbol{\hat{\beta}^{OLS} = (X_k' X_k)^{-1} X_k' Y} \tag{15}$$

The estimator which we suggested is a nonparametric that stands for a substitute for student-t test in the OCMT estimator. Thus, by substituting the test in (10) by the subsequent formula firstly suggested by Xu [12]

$$\boldsymbol{T_n = \frac{\sum_{i,j,l=1}^{n} (Z_i - Z_j)' D_s^{-1} (Z_j - Z_l)(e_i - e_j)'(e_j - e_l)}{4n(n-1)(n-2)(n-3)}} \tag{16}$$

Where $\boldsymbol{i \neq j, i \neq l, j \neq l}$ and $\boldsymbol{D_s = diag(S_{11}, S_{22}, , , S_{pp})}$ a diagonal matrix of the covariance matrix of $\boldsymbol{X}$ and $\boldsymbol{Z_i}$ is the standardized value of $\boldsymbol{X_i}$**.** Xu proposed to express the random error $\boldsymbol{e_i}$ in the following form.

$$\boldsymbol{e_i = \sqrt{12} \left[ \frac{r(Y_i)}{n+1} - \frac{1}{2} \right]} \tag{17}$$

Where $r(Y_i)$ represent the standardized value of the rank of $Y_i$ then we reject the hypothesis $\boldsymbol{H_0 : \beta_j = 0}$ in any level of significant $\alpha$ if

$$\boldsymbol{T_n \geq \sqrt{2tr(R'R)}\, z_\alpha / n}$$

Where $\boldsymbol{R = D^{\frac{1}{2}} S D^{\frac{1}{2}}}$ , $\boldsymbol{S = \sum_{i=1}^{n} (X_i - \bar{\bar{X}})(X_i - \bar{\bar{X}})' / (n-1)}$ this nonparametric test will be used as a nonparametric procedure substituting the student-t test in (10) to develop a new scheme of OCMT estimator and we will brand it with NP.

## 4. Mean Absolute Percentage Error MAPE

There are numerous statistical comparing strategies that are created on a convinced belief or theoretical basis [1]. We choose **MAPE** as a comparison tool among our three different estimators **SIS, OCMT, NP** and **MAPE** is suitable measure due the variate of variables in Big Data sets, and the formula of MAPE will be as follow[6].

$$\boldsymbol{MAPE = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{\hat{Y} - Y}{Y} \right|} \tag{18}$$

## 5. Simulation

In this section we make a simulation to generate different sizes of samples by using real data that detailed in Appendix. Here we have Y which symbolizes the Chemical Pollution Ratio **CPR** $(\boldsymbol{n \times 1})$ and $\boldsymbol{X}$ which signifies the explanatory variables $(n \times 80)$ . We select different sample sizes as $\boldsymbol{n = 200, 300, 400, 500, 600, 700, 800, 900, 1000}$ and outcomes of MAPE for the three estimators be as follows.

**Table (1): MAPE for SICA, OCMT, NP**

| n | SICA | OCMT | NP |
|---|---|---|---|
| 200 | 0.2451 | 0.2010 | 0.2206 |
| 300 | 0.2786 | 0.2176 | 0.2198 |
| 400 | 0.2976 | 0.2243 | 0.2277 |
| 500 | 0.3002 | 0.2280 | 0.2291 |
| 600 | 0.3279 | 0.2997 | 0.2994 |
| 700 | 0.3298 | 0.3011 | 0.3001 |
| 800 | 0.4023 | 0,3146 | 0.3056 |
| 900 | 0.4156 | 0.3200 | 0.3134 |
| 1000 | 0.4237 | 0.3312 | 0.3179 |

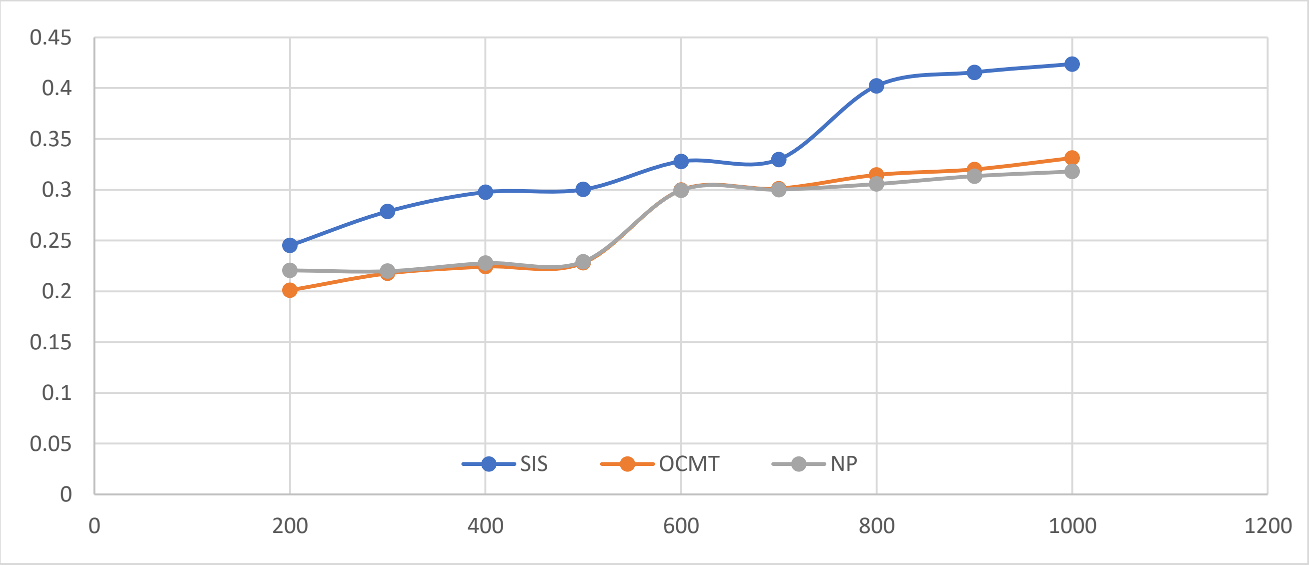

**Figure (1): MAPE for SICA, OCMT, NP**

## 6. Real Data

We collected data concerning Chemical Pollution on the water of Tigress River in middle of Iraq. We inspected 1000 samples of water and obtained results from 3 labs which symbolizes Y the CPR and 80 variable symbolizes the chemical and biological characteristics that affect CPR. Y $(1000 \times 1)$ and matrix of X $(1000 \times 80)$ ,and results of the three estimators will be :

**Table (2): MAPE for SIS, OCMT, NP Real Data**

| Method | MAPE |
|---|---|
| SICA | 0.5632 |
| OCMT | 0.3982 |
| NP | 0.3320 |

## 7. Conclusions and Recommendations

From the simulated and real data tables (1,2) we can conclude the OCMT method is efficient and better than other two methods when sample size is small and NP method is better when sample size is very large, then we can conclude than NP method is very efficient method to evaluate the general linear regression coefficients under BIG Data conditions. We recommend to study other kinds of environmental contaminations in Iraq like soil and air pollution and use many statistical techniques to explain this phenomenon and predict it in forthcoming and also we recommend to use nonparametric and robust methods to make analysis for Big Data sets.

## References

**[1]** Acharjya. D, Kauser. A, “A Survey on Big Data Analytics: Challenges, Open Research Issues and Tools”, International Journal of Advanced Computer Science and Applications. Vol. 7, No 2, pp. 511-518, 2016.

**[2]** Chudik. A, Kapetanios. G, Pesaran. M, “One-Covariate at Time, Multiple Testing Approach to variable selection in High-Dimensional Regression Models”, Econometrica, Vol. 86, Issue. 4, pp. 1479-1512, 2018.

**[3]** Fan. Y, Lv. J, “Sure Independence Screening for Ultra-High Dimensional Feature Space”, Royal Statistical Society, Vol. 70, Issue. 5, pp. 849-911, 2008.

**[4]** Firinguetti. L, “ A Generalized Ridge Regression Estimator and its Finite Sample Properties”, Communication in Statistics – Theory and Methods, Vol. 28, No. 5, pp. 1217-1229, 1999

**[5]** Troskie. C, Chalton. D, “A Bayesian Estimate for the Constants in Ridge Regression” South African Statistical Journal, Vol. 30, pp. 119-137, 1996.

**[6]** Kapetanios. G, Marcellino. M, Petrova. K, “Analysis of the Most Recent Modeling Techniques for Big Data with Particular Attention to Bayesian Ones” Eurostat. Statistical working papers. ISBN 978-92-79-77350-1, 2018.

**[7]** Knight. K, Fu. W, “Asymptotics for Lasso-Type Estimator” The Analysis of Statistics, Vol. 28, No. 5, pp. 1356-1378, 2000.

**[8]** Kumar. R, Moseley. B, Vassilvitskii. S, Vattani. A, “Fast Greedy Algorithms in MapReduce and Streaming”, ACM Transactions on Parallel Computing, Vol. 2, No. 3, pp. 154-170, 2011.

**[9]** Pesaran. M, Smith. R, “Signs of Impact Effects in Time Series Regression Models” Economics Letters, Vol. 122, pp. 150-153, 2014.

**[10]** Salih. A, Hmood. M, “Analyzing big data sets by using different panelized regression methods with application: surveys of multidimensional poverty in Iraq”, Periodicals of Engineering and Natural Sciences. Vol. 8, No. 2, pp. 991-999, 2020.

**[11]** Salih. A, Hmood. M, “Big Data Analysis by Using One Covariate at a Time Multiple Testing (OCMT) Method: Early School Dropout in Iraq” Int. J. Nonlinear Appl. Issue. 12, No. 2, pp. 931-938, 2021.

**[12]** Xu. K, “A new nonparametric test for high-dimensional regression Coefficients’, Journal of Statistical Computation and Simulation, Vol. 5, pp. 855-867, 2017

## Appendix: Variables

| no. | Variable | no. | Variable | no. | Variable |
|---|---|---|---|---|---|
| **Y** | CPR | **29** | DO | **57** | $^{137}Cs$ |
| **1** | $SO_3$ | **30** | CODMn | **58** | $^{90}Sr$ |
| **2** | $SO_2$ | **31** | COD | **59** | Coliforms |
| **3** | pH | **31** | BOD | **60** | Streptococci bacteria |
| **4** | Turbidity (NTU) | **32** | NH4-N | **61** | Phytoplankton |
| **5** | Electrical conductivity | **33** | TP | **62** | Zooplankton, |
| **6** | Alkalinity ( $CaCO_3$) | **34** | TN | **63** | Zoobenthos, |
| **7** | Hardness ( $CaCO_3$) | **35** | TCu | **64** | Macrophytes |
| **8** | Ammonia ($N\text{-}NH_3$) | **36** | TZn | **65** | PbH |
| **9** | Nitrate ($N\text{-}NO_3^-$) | **37** | F− | **66** | $H_2S_2$ |
| **10** | Apparent colors (HU) | **38** | TSe | **67** | $AgS_2$ |
| **11** | Nitrite ($N\text{-}NO_2^-$) | **39** | TAs | **68** | $SO_2$ |
| **12** | Phosphorus (P) | **40** | THg | **69** | $SO_4$ |
| **13** | Fluoride ($F^-$) | **41** | TCd | **70** | ESP |
| **14** | Iron (Fe) | **42** | Cr6+ | **71** | SAR |
| **15** | Escherichia coli | **43** | TPb | **72** | Hco3 |
| **16** | PbO | **44** | TCN | **73** | Cl |
| **17** | $CdO_3$ | **45** | V-ArOH | **74** | EC |
| **18** | $HgO_2$ | **46** | Petroleum | **75** | P |
| **19** | AgO | **47** | Anionic surfactant | **76** | CIS |
| **20** | MgO | **48** | S2− | **77** | CO |
| **21** | KO | **49** | Sulphate | **78** | $CO_3$ |
| **22** | Dissolved Metals | **50** | Mercury | **79** | KCl |
| **23** | Particulate Metals | **51** | Copper | **80** | MgCl |
| **24** | Exchangeable Metals | **52** | Zinc | | |
| **25** | Residual Metals | **53** | PCB | **-** | |
| **26** | Essential Metals | **54** | HCH | **-** | |
| **27** | Dissolved oxygen | **55** | PAH | **-** | |
| **28** | Biochemical Oxygen | **56** | MgS | **-** | |

مجلة كلية الرافدين الجامعة للعلوم (2023)؛ العدد 54؛ 531 - 538

AL- Rafidain University College

PISSN: (1681-6870); EISSN: (2790-2293)
مجلة كلية الرافدين الجامعة للعلوم

Available online at: https://www.jrucs.iq

JRUCS
Journal of AL-Rafidain University College for Sciences

# استخدام الإجراء اللامعلمي لتطوير مقدر OCMT لنموذج الانحدار الخطي للبيانات الضخمة مع تطبيق التلوث الكيميائي في نهر دجلة

| أ.م.د. أحمد مهدي صالح | أ.د. مناف يوسف حمود |
|---|---|
| amahdi@uowasit.edu.iq | munaf.yousif@coadec.uobaghdad.edu.iq |
| قسم الاحصاء - كلية الادارة والاقتصاد - جامعة واسط، واسط، العراق | قسم الاحصاء - كلية الادارة والاقتصاد - جامعة بغداد، بغداد، العراق |

معلومات البحث

**تواريخ البحث:**
تاريخ تقديم البحث: 2022/12/24
تاريخ قبول البحث: 2023/3/3
تاريخ رفع البحث على الموقع: 2023/12/31

**الكلمات المفتاحية:**
البيانات الضخمة، SICA، OCMT، التلوث الكيميائي، اللامعلمي

**للمراسلة:**
أ.م.د. أحمد مهدي صالح
amahdi@uowasit.edu.iq

## المستخلص

يعد التلوث الكيميائي من القضايا المهمة جداً التي يعاني منها الإنسان وغالباً ما يؤثر على طبيعة صحة المجتمع ومستقبل صحة الأجيال القادمة. وبالتالي لا بد من دراستها من أجل اكتشاف النماذج المناسبة وإيجاد أوصاف للتنبؤ بأدائها في السنوات القادمة. تأخذ بيانات التلوث الكيميائي في العراق نطاقا كبيرا ومتعددة المصادر والأنواع، مما يجعلها بيانات ضخمة تحتاج إلى دراسة باستخدام أساليب إحصائية جديدة. يهدف البحث إلى استخدام طريقة NP المقترحة لتطوير إجراء اختبار (OCMT) لتقدير معلمات نموذج الانحدار الخطي ذو الحجم الكبير من البيانات (Big Data) والذي يضم العديد من المؤشرات المرتبطة بالتلوث الكيميائي ولها تأثير عميق على الحياة. من الشعب العراقي. تم اختيار مقدر SICA لتحليل البيانات وتم استخدام MSE لإجراء مقارنة بين الطريقتين ونقرر أن مقدر NP أكثر فعالية من المقدرات الأخرى في ظل ظروف البيانات الضخمة.

https://doi.org/10.55562/jrucs.v54i1.620