توظيف خوارزمية الفراشات المضيئة لأختيار عرض الحزمة في مقدر نداريا- واتسون المتعدد

هادي سلمان محمد
الوقف السني/ دائرة التعليم لديني والدراسات الاسلامية

زكريا يحيى الجمال
قسم الاحصاء والمعلوماتية/ كلية علوم الحاسوب والرياضيات

(قدم للنشر في ٢٠٢٣/٩/٢٤، قبل للنشر في ٢٠٢٣/١١/١)

المستخلص:

إن موضوع تحليل الانحدار يلقى اهتماماً متزايداً وواضحاً في معظم الدراسات وخصوصاً الاقتصادية والطبية منها. ويعد نموذج الانحدار اللامعلمي بصورة عامة والانحدار اللامعلمي المتعدد بوجه خاص أحد أهم وأبرز نماذج الانحدار المستخدمة في السنوات الأخيرة التي شهدت توسعاً كبيراً وخصوصاً في الجانب الاقتصادي والبيئي. إذ يعدّ مقدر نداريا- واتسون المتعدد (Multivariate Nadaraya-Watson estimator**) من أهم المقدرات المستعملة في أنموذج الانحدار اللامعلمي المتعدد. حيث أن هذا المقدر يعتمد بدوره في تقدير نموذج الانحدار اللامعلمي المتعدد على مصفوفة معلمات تسمى بمعلمات التمهيد(**smoothing parameter**) والتي لتقديرها أهمية كبيرة في تحقيق جودة توفيق المنحى المقدر في نموذج الانحدار اللامعلمي المتعدد. تم في هذا البحث اقتراح توظيف خوارزمية مستوحاة من الطبيعة والمتمثلة بخوارزمية الفراشات المضيئة في عملية تقدير مصفوفة معلمات التمهيد (**Bandwidth matrix**) في مقدر نداريا- واتسون المتعدد. كما تمّ استخدام أسلوب محاكاة المونت – كارلو لتوليد بيانات تتبع عدد من نماذج الأنحدار اللامعلمي المتعدد. لقد أظهرت نتائج المحاكاة تفوق الطريقة المقترحة مقارنةً بطرائق التقدير الأخرى معتمدين متوسط مربعات الخطأ بوضعها معياراً للمقارنة.**

الكلمات المفتاحية: مقدرات النواة، مصفوفة عرض الحزمة، مقدر نداريا - واتسون المتعدد، خوارزمية الفراشات المضيئة.

Employ the Firefly algorithm for bandwidth selection in the Multivariate Nadaraya-Watson estimator

Hadi Salman Muhammad
The Sunni Endowment/Department of Religious Education and Islamic Studies

Zakaria Yahya El-Gammal
Department of Statistics and Informatics/ College of Computer Science and Mathematics

Abstract

The topic of regression analysis is receiving increasing and clear attention in most studies, especially economic and medical ones. The nonparametric regression model in general and the multiple nonparametric regression model in particular is one of the most important and prominent regression models used in recent years, which have witnessed great expansion, especially in the economic and environmental aspects. The Multivariate Nadaraya-Watson estimator is one of the most important estimators used in the multiple nonparametric regression model. In estimating the multiple nonparametric regression model, this estimator, in turn, relies on a matrix of parameters called smoothing parameters, the estimation of which is of great importance in achieving good fit of the estimated curve in the multiple nonparametric regression model. In this research, it was proposed to employ an algorithm inspired by nature, represented by the Fireflies algorithm, in the process of estimating the smoothing parameter matrix (Bandwidth matrix) in the Ndaria-Watson multiple estimator. The Monte Carlo simulation method was also used to generate data following a number of multiple nonparametric regression models. The simulation results showed the superiority of the proposed method compared to other estimation methods, using the mean square error as a standard for comparison.

Keywords: kernel estimator; smoothing matrix, Multivariate Nadaraya-Watson estimator, Firefly algorithm.

١ـ المقدمة: Introduction

إنّ تحليل الانحدار(Regression Analysis) يعد أحد الأساليب الإحصائية المهمة لدراسة العديد من الظواهر الطبيعية والاجتماعية والاقتصادية والطبية وغيرها، إذ يستخدم في تمثيل العلاقة بين المتغيرات العشوائية المختلفة بالنسبة إلى عينة معينة أو بالنسبة إلى المجتمع على هيئة معادلة إحصائية لتحقيق الكثير من الأهداف المهمة التي يتوصل إليها من خلال تلك العلاقة. تعد أساليب الانحدار مفيدة في عملية بناء النماذج الإحصائية، إذ تصنف نماذج الانحدار إلى صنفين أساسيين بحسب طبيعة البيانات: نماذج الانحدار المعلمي (Parametric Regression Models) ونماذج الانحدار اللامعلمي ((Nonparametric (Rencher;2002) Regression Models. ان نماذج الانحدار اللامعلمي (Nonparametric Regression Models) تقوم على إيجاد العلاقة بين متغير الاستجابة والمتغيرات التوضيحية من خلال منحنى يصف تلك العلاقة، لذا فإن الباحث يكون مهتماً بإعطاء وصف عام للعلاقة وليس دراسة التفاصيل الدقيقة للعلاقة في الانحدار اللامعلمي. وإن دالة العلاقة تكون غير معروفة وهذه النماذج تكون أكثر مرونة ولاتعتمد على فروض سابقة كما في الانحدار المعلمي، بل تعتمد بشكل أساسي ومباشر على البيانات (Data) حيث إن نوع البيانات يفسر الشكل الفعلي لمنحنى الإنحدار (محمد، ٢٠١١).

إن تحديد مصفوفة عرض الحزمة (Bandwidth) أو ما تسمى بمصفوفة معلمات التمهيد (H) في مقدر مقدر نداريا - واتسون المتعدد لذات أهمية كبيرة في تحديد وتقريب شكل دالة الانحدار اللامعلمي الى الدالة الأصلية، من خلال إيجاد الطريقة المثلى للموازنة بين التباين والتحيز. فعندما تكون القيمة لمعلمة التمهيد صغيرة فإن التحيز يكون صغيراً، والتباين يكون كبيراً، وهذا بدوره يؤدي إلى خشونة شكل الدالة وبعكسه نحصل على تعظيم التحيز، وتصغير التباين ويكون شكل الدالة أكثر تمهيداً (سلاسة) اذ انّ عملية الاختيار الجيد لقيم مصفوفة معلمة التمهيدية (H) يتم من خلال المفاضلة بين التحيز والتباين للحصول على أقل متوسط مربعات خطأ (MSE).

٢ـ هدف البحث وأهميتهُ

ان هذا البحث يهدف الى توظيف إحدى خوارزميات التقنيات الذكائية وهي خوارزمية الفراشات المضيئة (Firefly algorithm) لتقدير قيم مصفوفة معلمة التمهيد (H) بحيث تكون أكثر كفاءة مقارنة مع الطرائق الأخرى، تعمل على تحسين النتائج في عملية تقدير قيم المصفوفة (H) من خلال تجارب المحاكاة وبأستعمال نماذج مختلفة.

وأما اهمية هذا البحث فتكمن في كونه يسلط الضوء على أهمية تطبيق بعض الخوارزميات الذكائية في تقدير مصفوفة عرض الحزمة وتصنيفها كأحدى الطرائق البديلة للطرائق الأحصائية التقليدية التي تخص الموضوع.

٣ـ المقدرات اللامعلمية Nonparametric estimator

إن المرونة العالية التي تتمتع بها المقدرات اللامعلمية نظراً لكونها لا تتطلب توفر فروض بشأن توزيع المجتمع قياساً بالمقدرات المعلمية والتي تتطلب مجموعة من الفروض، ونظراً للتطور الهائل في أجهزة الحواسيب أدى إلى ميل الباحثين في العقود الأخيرة للأهتمام بموضوع الأنحدار اللامعلمي، وطرائق التمهيد االخاصة بهِ، ومن أنواعه: نموذج

الأنحدار اللامعلمي البسيط (simple Nonparametric Regression Model) والذي يقوم على إيجاد العلاقة بين متغير الاستجابة ومتغير توضيحي واحد فقط، وصيغته كما يأتي: (Koláček & Horová;2017)

$$\mathbf{y}_i = m(\mathbf{x}_i) + \boldsymbol{\varepsilon}_i \qquad i = 1,2,\ldots,n \tag{1}$$

أما إذا كان النموذج يقوم على إيجاد العلاقة بين متغير الإستجابة، وعدد من المتغيرات التوضيحية، فعندئذٍ يسمى بنموذج الانحدار اللامعلمي المتعدد (Multivariate Nonparametric Regression Model) إذ أن هذه العلاقة تكون غير معروفة ويتم تقديرها باستعمال طرائق عدة منها: طريقة التقدير اللُبي (Kernel estimation)، وطريقة الشرائح التمهيدية(Smoothing Splines)، وطريقة المويجة (Wavelet estimator)، وصيغة معادلته تكون بالشكل الآتي (محمد و عبدالحسن، ٢٠١٨;عيسى ومناف،٢٠١٢):

$$\mathbf{y}_i = \mathbf{m}(\mathbf{x_{1i}}, \mathbf{x_{2i}}, \ldots, \mathbf{x_{di}}) + \boldsymbol{\varepsilon_i} \qquad i = 1,2,\ldots,n \tag{2}$$

حيث أن: $(\mathbf{x_{1i}}, \mathbf{x_{2i}}, \ldots, \mathbf{x_{di}})$ تمثل مصفوفة متجهات المشاهدات للمتغيرات التوضيحية.

وأختصاراً تكتب:

$$\mathbf{y_i} = \mathbf{m(X)} + \boldsymbol{\varepsilon_i} \qquad i = 1,2,\ldots,n \tag{3}$$

$\mathbf{m(X)}$: تمثل دالة الأنحدار غير المعروفة والمطلوب تقديرها بالطرائق اللامعلمية.

هناك العديد من الطرائق اللامعلمية لتقدير هذه الدالة غير المعروفة والموضحة في المعادلة (١) و(٢)، إذ إن الهدف من التمهيد هو لتقريب دالة الانحدار اللامعلمي التقريبية الى دالة الانحدار اللامعلمية الحقيقية، والعمل على تعديل المشاهدات، إذ أنّ هذه الطرائق بنيت على أساس نموذج مقدر؛ ليعطي نموذجاً مقارباً للواقع وللتنبؤ بالمستقبل، وإن معظم الطرائق اللامعلمية تفترض أن الخطأ يتوزع بمتوسط مساوٍ للصفر وتباين محدد وإن دالته هي دالة مستمرة (Continuous Function) وممهدة (Smoothed)، ومن المقدرات اللامعلمية الإحصائية التي لاتتطلب توفر فروض بشأن توزيع المجتمع، والذي يعد أداة فعالة تعتمد بشكل أساسي على البيانات (Data) هو مقدر مقدر نداريا-واتسون (Nadaraya-Watson estimator) (Hardle;1994).

٤ـ مقدر نداريا-واتسون المتعدد Estimator (MNWE) Multivariate Nadaraya-Watson

أن مقدرNadaraya-Watson يُعد من أكثر المقدرات الشائعة الاستخدام في تقدير دالة الأنحدار اللامعلمي، تم اقتراحه هذا المقدر عام (١٩٦٤) من قبل الباحثين Nadaraya وWatson، كما أنه يعد من أبسط أنواع الممهدات، إذا غالباً ما يستعمل على في العديد من المجالات البحثية الإحصائية بالأعتماد على طريقة متسلسلة الأوزان كما في الصيغة (٤)، إن عملية تقدير دالة الانحدار اللامعلمي $m(\mathbf{X})$ غير المعروفة تتم باستخدام المتوسط الموزون،(Aydin;2007)

وتعرف طريقة المتوسط الموزون بأنها مشابهة لطريقة المربعات الصغرى الموزونة وكما هو مبين في المعادلة الآتية :

$$\hat{m}(\mathbf{x_i}) = n^{-1}\sum_{i=1}^{n} \mathbf{w}_i(\mathbf{x_i})\mathbf{y_i}\,; \qquad i = 1,2,\ldots,n \tag{4}$$

إذ أن $\mathbf{w}_i(\mathbf{x}_i)$: تمثل سلسلة من الأوزان الطبيعية الموجبة التي تعتمد على كل قيم المتجه وإن مجموع هذه الأوزان يساوي الواحد (Boente, et al;1997)

$$\sum_{i=1}^{n} \mathbf{w}_i(\mathbf{x}_i) = 1 \quad ; \quad w_i \geq 0 \tag{5}$$

إذ إن هذه الأوزان تمثل دالة المسافة في فضاء X، والصيغة العامة للأوزان تكتب بالشكل الآتي (Aydin,2007 & Hardle;1990):

$$\mathbf{w}_i(\mathbf{x_i}) = \frac{k_h(\mathbf{x_i} - x_0)}{\hat{f}_h(x)} \qquad i = 1,2,\ldots,n \tag{6}$$

وان:

$$\hat{f}_h(\mathbf{x}) = n^{-1}\sum_{i=1}^{n} k_h(\mathbf{x_i} - x_0) \tag{7}$$

حيث:

$$k_h(u) = h^{-1}k(u/h) \tag{8}$$

إذ أن:

h : المعلمة التمهيدية (Smoothing Parameter)، أو عرض الحزمة Bandwidth، وتكون قيمتها أكبر من الصفر.

حيث إن $k(u/h_n)$ تمثل إحدى الدوال اللُبية، كما إن متسلسلة الأوزان للدالة اللُبية يشار إليها بالمختصر $\{\mathbf{w}_i(\mathbf{x_i})\}_{i=1}^{n}$ ، وتتمثل بالأوزان المؤشرة للمشاهدات i بالنسبة لـ Y_i التي تعتمد على المسافة بين النقطة x_i والنقطة x_0، إذ عادة ما تكون هذه الأوزان كبيرة إذا كانت المسافات قليلة، وتقل في حالة كون المسافات كبيرة، وأبسط طريقة لتمثيل دالة الأوزان هذه هي بوصف شكل دالة الأوزان $\mathbf{w}_i(\mathbf{x})$ بدالة الكثافة وبمعلمة ثابتة، والتي بدورها تقوم بتعديل حجم الأوزان بالقرب من النقطة x_0، وبالتالي سيكون المقدر بالصيغة الآتية :(Hardle;1990)

$$\hat{m}(\mathbf{x}) = \frac{\sum_{i=1}^{n} k_h(\mathbf{x}_i - x_0)\mathbf{y}_i}{\sum_{i=1}^{n} k_h(\mathbf{x}_i - x_0)} = \frac{k(u)}{\sum k(u)} \tag{9}$$

أي يصبح شكل المقدر بالصيغة الآتية:

$$\hat{m}(\mathbf{x}) = \frac{\sum_{i=1}^{n} k\left(\frac{\mathbf{x}_i - x_0}{h}\right)\mathbf{y}_i}{\sum_{i=1}^{n} k\left(\frac{\mathbf{x}_i - x_0}{h}\right)}, \qquad h > 0 \tag{10}$$

وبتعميم ذلك في حالة وجود أكثر من متغير توضيحي أي d من المتغيرات التوضيحية تكون الصيغة لمقدر Nadaraya-Watson كالآتي (Soméa & Kokonendjia, ٢٠١٥):

$$\hat{m}(\mathbf{x})_{MNW} = \frac{\sum_{i=1}^{n} \mathbf{K}\left(\mathbf{H}^{-1}\left(\mathbf{X_i} - \mathbf{x_0}\right)\right)\mathbf{y_i}}{\sum_{i=1}^{n} \mathbf{K}\left(\mathbf{H}^{-1}\left(\mathbf{X_i} - \mathbf{x_0}\right)\right)} \qquad \textbf{(11)}$$

حيث إن : K (.) تمثل الدالة اللُبية المتعددة.

H: مصفوفة عرض الحزمة ذات بعد ($d \times d$) وتكون قطرية ومتماثلة و موجبة التعريف.

أي إن:

$$\mathbf{H} = \mathbf{h}_{1\times d}\mathbf{I}_{d\times d} \;;\; \mathbf{h} = \left[h_1, h_2, \cdots, h_d\right]', \qquad \textbf{(12)}$$

$$\mathbf{H} = \begin{bmatrix} h_1 & 0 & \dots & 0 \\ & h_2 & \dots & 0 \\ & & \ddots & \vdots \\ & & & h_d \end{bmatrix}_{d\times d} \qquad \textbf{(13)}$$

٥- عرض الحزمة Bandwidth

إن عملية اختيار عرض الحزمة (Bandwidth) تعد الخطوة الأكثر أهمية في تقريب دالة الانحدار اللامعلمي إلى الدالة الأصلية، ومن اجل الحصول على التقريب الجيد والملائم لا بد من إيجاد الطريقة المُثلى لأجل الموازنة بين كل من التباين و التحيز بحيث يكون مقدار الخطأ أقل مايمكن والذي عادة يقاس بمعيار متوسط مربعات الخطأ Mean Squared Error(MSE) أو متوسط مربعات الخطأ التكاملي Mean Integrated Squared Error(MISE) تتأتى هذه الموازنة من خلال استخدام أفضل قيمة لعرض الحزمة، إذ إن اختيار هذه القيمة يجب أن يكون بعناية وحذر، وذلك لكون القيمة الصغيرة جداً تؤثر على تمهيد المنحنى، وتكوّن منحنى تمهيد منخفض (Under Smoothing) ، وتكوّن منحنى تمهيد مرتفع(Over Smoothing) في حالة كانت القيمة كبيرة جداً (Schimek, 2013& Chn,1995).

ويرمز لمعلمة عرض الحزمة بالرمز(h) إذا كانت تستخدم لأحادي المتغير، حيث تتضمن اختيار معلمة مفردة، أما في حالة متعدد المتغيرات فتتضمن اختيار مصفوفة عرض الحزمة ويرمز لها بالرمز(H)، وتوجد عدة أساليب لاختيار القيمة المُثلى لعرض الحزمة (Bandwidth) والتي تحاول تقليل مجموع مربعات الخطأ للنموذج وهي: طريقة العبور الشرعي Cross Validation(CV)، وطريقة العبور الشرعي العام Generalized Cross Validation(GCV) وغيرها من الطرائق الأخرى(Mustafa & Algamal; 2022).

٦- طريقة العبور الشرعي Cross Validation(CV)

هذه الطريقة تُعد من الطرائق شائعة الأستخدام؛ لإيجاد أنسب قيمة للمعلمة التمهيدية h، إذ تلعب هذه القيمة دور مهماً في تباين وتحيزالمقدر. إن فكرة هذه الطريقة تقوم على أساس تقسيم البيانات الى L من المجاميع الجزئية $(g_1, g_2, \dots, g_l)$ بحيث أن كل مجموعة تحتوي على عدد متساوي من المشاهدات $n_j = (n_1, n_2, \dots, n_l)$، إذ يتم استبعاد مجموعة واحدة في كل مرة وتكون المجموعة المستبعدة هي g_j بحيث أن $j = 1,2,\dots, l$ ، و كحالة خاصة عندما L=١ تسمى Leave One Out Method، أما في حالة وجود أكثر من متغير توضيحي أي d من المتغيرات التوضيحية نستعمل الصيغة كالآتي: (Hardle, 1990 ; Koláček & Horová,2017)

$$CV(\mathbf{H}) = \frac{1}{n}\sum_{i=1}^{n}\left[\mathbf{y_i} - \hat{m}_{-1}(\mathbf{X_i};\mathbf{k})\right]^2 \quad \textbf{(14)}$$

$\hat{m}_{-j}(\mathbf{x}_i)$: تمثل مقدرات Nadaraya-Watson في حالة استبعاد مشاهدة.

ولإيجاد مصفوفة قيم المعلمات التمهيدية المُثلى H نستعمل الصيغة الآتية:

$$\hat{\mathbf{H}} = \arg\min_{H \in H} \mathbf{H} \quad \textbf{(15)}$$

٧- طريقة العبورالشرعي العام Generalized Cross Validation (GCV)

هذه الطريقة GCV مستمدة من طريقة العبور الشرعي CV، حيث يتم الحصول عليها من صيغة CV الموضحة في الصيغة (14) وذلك عن طريق استبدال عناصر القطر الرئيسي $\hat{m}(x_i)$ لمصفوفة التمهيد S_h بمعدلها أي إن(Kauermann, & Opsomer, 2004,; Mustafa& Algamal;2022)

$$GCV_{\mathbf{H}} = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{\hat{m}(\mathbf{x_i}) - \mathbf{y_i}}{1 - tr(S_{\mathbf{H}})/n}\right\}^2 \quad \textbf{(16)}$$

ولايجاد المصفوفة **H** الخاصة بقيم المعلمات التمهيدية المُثلى نستعمل الآتي:

$$\hat{\mathbf{H}}_{Gcv} = \underset{H \in H}{argmin}\, \mathrm{GCV}_H \quad \textbf{(17)}$$

٨- خوارزمية الفراشات المضيئة: Firefly algorithm

في الآونةِ الأخيرةِ أزدادَ أهتمام الباحثون بتصميم وتطويرالعديد من خوارزميات التحسين وخاصةً تلك المُستوحاة من الطبيعة، إذ أنّ فكرة هذه الخوارزميات أُستُنتِجت عن طريق ملاحظة سلوك بعض مصادر الطبيعة المختلفة: كالنحل والنمل والفراشات والاسماك والطيور والنباتات وأنظمة الامواج والأنهار وغيرها، ومنها ما هو مستوحى من ذكاء السرب والتي تُعد واحدة من الخوارزميات المهمة لحل العديد من المشكلات المعقدة في البحث العلمي وتحسينها ، إذ تمّت دراسة خوارزميات ذكاء السرب على نطاق واسع حيث تم تطبيقها بنجاح على مجموعة

متنوعة من مشكلات التحسين المعقدة نظراً لتمتعها بالبساطة والمرونة والكفاءة العالية; Kahya Altamir, & (Algamal, 2020).

تعتمد معظم خوارزميات التحسين المستوحاة من الطبيعة على ذكاء السرب، إذ يشكل هذا النوع من الخوارزميات جزءاً كبيرا من الخوارزميات المعاصرة، وأصبحت هذه الخوارزميات مستخدمة على نطاق واسع في التحسين وتحليل البيانات وكذلك في التعلم الآلي والذكاء الإصطناعي. وتعُّد خوارزمية الفراشات المضيئة (Firefly Algorithm FA) أو ما تسمى بخوارزمية اليراعات واحدة من أحدث أساليب ذكاء السرب الجديدة واقوى خوارزميات التحسين التي تم تطويرها لأول مرة من قبل الباحث Yangفي بداية عام ٢٠٠٨(Yang;2013)

أثبتت الخوارزمية أنها فعالة وذات أداء جيد في حل مشكلات التحسين المختلفة. تم إيجاد خوارزمية الفراشات من محاكاة السلوك الاجتماعي للفراشات المضيئة على أساس جاذبية الفلاش (الأضواء الساطعة) من خلال تمثيل ميزة بعض الخصائص الوامضة للفراشات وكيفية التفاعل معها، إذ أن وميض الفراشة هو نظام إشارة يستخدم لجذب فراشة أخرى.

حيث يمكن حساب المسافة بين اثنين من الفراشات في المواقع بالمسافة الديكارتية والتي يمكن حسابها بإستخدام المعادلة الآتية :

$$r_{ij} = \|xi - xj\| = \sqrt{\sum_{d=1}^{D} \left(x_{id} - x_{jd}\right)^2} \quad \textbf{(18)}$$

إذ أن x_i: هو موقع اليراعة iإذ $x_i = \{x_{1i}, x_{2i}, ..., x_{id}\}$، وأن x_j: موقع اليراعة j إذ $x_i = \{x_{1i}, x_{2i}, ..., x_{id}\}$ ، D عدد الابعاد وأن $d \in D$

يمكننا تلخيص آلية عمل خوارزمية الفراشات(FA) بالخطوات الآتية:(Yang, 2010)

١- جميع الفراشات ومن كلا الجنسين يمكن أن تنجذب كل فراشةٍإلى كل الفراشات الأخرى. إذ أنّ الفراشات الأكثر جاذبية (أشراقا) تَنجذِبُ إليها الفراشات الأقل جاذبية (إشراقاً).
٢- تتناسب جاذبية الفراشة مع شدة الضوء الذي يتناقص كلما زادت المسافة عن الفراشات الأخرى.
٣- يتم تحديد جاذبية الفراشة من خلال موقعها داخل مساحة البحث.
٤- تؤدي القيمة الأفضل لوظيفة اللياقة في موقع معينٍإلى زيادة جاذبية الفراشة.
لكل فراشة شدة ضوء أو سطوع يتم إستخدام قيمتهُ لتقييم جودتها. إن سطوع اليراعة i في موقع معين x نستطيع أن نشير إليه بالآتي:

$$I(x_i) = f(x_i) \quad \textbf{(19)}$$

حيث أن شدة ضوء الفراشة تتناسب طردياً مع سطوعها وترتبط بالقيم الموضوعية. عند المقارنة بين الفراشات، تنجذب الفراشة التي لها شدة ضوء منخفض بأتجاه الفراشة الأخرى ذات الضوء الأعلى شدة، إذ أنّ شدة ضوء الفراشة يعتمد على I_o من الضوء المنبعث من الفراشة والمسافة r_{ij}بين زوج من الفراشات. يمكن وصف شدة الضوء $I(r)$من خلال دالة متناقصة بشكل رتيب لـ r_{ij} والتي يمكن صياغتها كالآتي:

$$I(r) = I_o e^{-(\gamma rij)2} \quad \textbf{(20)}$$

γ:هو عامل امتصاص تأثير الضوء.

ونظراً لأن الجاذبية لكل فراشة تتناسب مع شدة الضوء التي تراها الفراشات المجاورة، لذلك يجب السماح للجاذبية بالتنوع بإختلاف درجة الإمتصاص، حيث يمكن تحديد الشكل الرئيسي لتباين الجاذبيةZ بالمعادلة التالية: (Al Radhwan, & Algamal, 2021)

$$Z(r) = Z_o e^{-(\gamma rij)2} \quad \textbf{(21)}$$

إذ أن $Z(r)$تمثل دالة جاذبية الفراشة عند المسافة r و Z_0هي الجاذبية الأولية لليراعة عند مسافة (r = 0)، وعند الشروع بالتنفيذ نفترض أن (Z_0=1) تساوي الواحد . حيث يتم تحديث الحركة للفراشات حسب المعادلة الآتية:

$$x_i^{(t+1)} = x_i^{(t)} + Z_o e^{-\gamma r^2_{ij}} (x_j^{(t)} - x_i^{t)}) + \alpha_t \varepsilon_i^t \quad \textbf{(22)}$$

إذ أن α_t: هو معامل التوزيع العشوائي. ε_i^t: متجه لأرقام عشوائية مأخوذة من توزيع Uniform.

يعتمد تأثير هذه الحركة العشوائية في المعلمة α_t فيما إذا تم إختياره ليكون كبيراً فإن الحل x_i سيتحرك بشكل عشوائي مبتعداً عن الموقع، بخلاف إذا كان α_t صغيرة جداً، فستتحرك في الموقع وقد تصبح ضئيلة مقارنة بالحركة نحو الفراشات الأكثر إشراقاً (yong;2013).

ولأن كفاءة مقدر (MNW) تعتمد إلى حد كبير على الاختيار المناسب لمصفوفة معلمة التجانس(H). والتي تعتد في تكوينها على اختيار القيم المناسبة لعرض الحزمة. تم إقتراح توظيف خوارزمية FA لتقدير مصفوفة عرض الحزمة (H)، وبالإعتماد على هذه التقنية فان كل عنصر (فراشة) في المجموعة سوف يكون لديه موقع واحد فقط التي سوف يقوم بالبحث عنه. إذ يمثل هذا الموقع القيمة لعرض الحزمة. بناًء على ذلك فان توظيف هذه الخوارزمية سيكون على النحو التالي :

الخطوة الاولى: تحديد حجم المجموعة بـ٢٠ وهو (عدد الفراشات). إذ أن كل فرشة سوف تكون لها قيمة واحدة لمعلمة التمهيد. فضلاً عن ذلك تحديد عدد التكرارات داخل الخوارزمية وهو ٢٥٠.

الخطوة الثانية: توليد القيم الأولية التي تحتاجها الخوارزمية التي سوف تمثل القيم الأولية الإفتراضية لـ h ولكون قيم موجبة فإنه سيتم توليدها من التوزيع المنتظم المستمر وفق الفترة [1, 0]. بناًء على إختيار الباحث بعد إجراء التجارب العديدة.

الخطوة الثالثة: تحديد قيم معلمات خوارزمية الفراشات المضيئة وهي $Z_0 = 1$ ،: $\alpha = 0.1$ ، $\gamma = 0.2$.

الخطوة الرابعة: لغرض إختيار القيم المثلى، تم الإعتماد على دالة اللياقة (Fitness function) والمعرفة كالآتي:

$$\text{Fitness} = \min \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2, \quad \textbf{(23)}$$

الخطوة الخامسة: بالإعتماد على أقل قيمة تحصل عليها أي فراشة سيتم تحديث باقي مواقع الفراشات وفق الصيغة (٢١).

الخطوة السادسة: نستمر بالحل لحين الوصول إلى أعلى تكرار للخوارزمية والذي تم تحديده بالخطوة الأولى والذي سوف يمثل الحل الامثل.

٩ـ نتائج المحاكاة

لاختبار مدى جودة أداء الطريقة المقترحة تم تصميم العديد من التجارب ومحاكاتها باستعمال لغة البرمجة (R)، إذ تم الأخذ بنظر الاعتبار ثلاثة أحجام للعينات (250 , 100 , 50)=n، وتم إجراء المقارنات للطرق المختلفة المستخدمةCV، GCV، مع الطريقة المقترحة لخوارزمية (FA) وباستخدام دالة Epanechnikov كدالة لُبية، مع دراسة أربعة نماذج مختلفة هي:

النموذج الأول: تم توليد هذا الأنموذج وفق المعادلة الآتية: (Lijian & Rolf, 1999) :

$$\mathbf{y_i} = \left\{(\mathbf{x_1} - 0.5)^2 + \mathbf{x_2^2}\right\}\sin(2\pi \mathbf{x_3}) + \boldsymbol{\varepsilon_i} \qquad \textbf{(24)}$$

حيث تم توليد المتغيرات التوضيحية X_1 وX_2 وX_3 من التوزيع المنتظم ضمن الفترة [1 , 0]، أما $\boldsymbol{\varepsilon_i}$ فيتوزع بالشكل الآتي: $\boldsymbol{\varepsilon_i} \sim N(0, 0.025)$

النموذج الثاني: تم توليد هذا النموذج وفق المعادلة الآتية: (Goutte,et.al,2000) :

$$\mathbf{y_i} = 10\sin(\pi \mathbf{x_1 x_2}) + 20\left(\mathbf{x_3} - \frac{1}{2}\right)^2 + 10\mathbf{x_4} + 5\mathbf{x_5} + \boldsymbol{\varepsilon_i} \qquad \textbf{(25)}$$

حيث تم توليد المتغيرات التوضيحية x_1 وX_2 و X_3و X_4وX_5 من التوزيع المنتظم ضمن الفترة [1 , 0]، أما $\boldsymbol{\varepsilon_i}$ فيتوزع بالشكل الآتي $\boldsymbol{\varepsilon_i} \sim N(0,1)$:

النموذج الثالث: تم توليد هذا النموذج وفق المعادلة الآتية: (Shang,et.al,2014) :

$$\mathbf{y_i} = \sin(2\pi\, \mathbf{x_1}) + 4(1 - \mathbf{x_2})(1 + \mathbf{x_2}) + \frac{2\mathbf{x_3}}{1 + 0.8\,\mathbf{x_3^2}} + \boldsymbol{\varepsilon_i} \qquad \textbf{(26)}$$

حيث تم توليد المتغيرات التوضيحية X_1 وX_2 وX_3 من التوزيع المنتظم ضمن الفترة [1 , 0]،أما $\boldsymbol{\varepsilon_i}$ فيتوزع بالشكل الآتي: $\boldsymbol{\varepsilon_i} \sim N(0, 0.02)$

النموذج الرابع: تم توليد هذا النموذج وفق المعادلة الآتية(Koláček & Horová, 2017):

$$\mathbf{y_i} = (\mathbf{x_1} - 0.5)^3 + (\mathbf{x_2} - 0.5) + \boldsymbol{\varepsilon_i} \qquad \textbf{(27)}$$

كما وتم توليد المتغيرات التوضيحية X_1 وX_2 من التوزيع المنتظم ضمن الفترة [1 , 0]، أما $\boldsymbol{\varepsilon_i}$ فيتوزع بالشكل الآتي: $\boldsymbol{\varepsilon_i} \sim N(0, 0.02)$

حيث تم تكرار اجراء التجربة(250) مرة، وتم الأعتماد على متوسط مربعات الخطأ (MSE) كمعيار للمقارنة بين طرائق التقدير المستخدمة وبيان الطريقة الأفضل وتلخيص نتائج الطرق المستخدمة في الجداول من (١-٤)

اظهرت النتائج تفوق طريقة (FA) على باقي الطرائق الاخرى (CV,GCV) من حيث معاير MSE فعلى سبيل المثال جدول (٢) اعطت طريقة (FA) اقل قيمة عندما n=50، حيث بلغت 0.0131 مقارنة بـ 2.989 لطريقة CV

و3.134 لطريقة GCV. بالاضافة الى ذلك الطريقة المقترحة حافظت على افضليتها عند تغير حجم العينة ولجميع النماذج المستخدمة.

الجدول(١): معدل قيمة **MSE** وقيمة الانحراف المعياري للأنموذج الأول في مقدر **Nadaraya-Watson**

	CV	GCV	FA
n=50	0.047 ± 0.008	0.046 ± 0.008	0.0130 ± 0.0023
n=100	0.047 ± 0.006	0.044 ± 0.007	0.0013 ± 0.0016
n=250	0.044 ± 0.003	0.012 ± 0.040	0.0012 ± 0.0014

الجدول(٢): معدل قيمة **MSE** وقيمة الانحراف المعياري للأنموذج الثاني في مقدر **Nadaraya-Watson**

	CV	GCV	FA
n=50	3.134 ± 0.649	2.989 ± 0.921	0.0131 ± 0.0018
n=100	3.125 ± 0.534	2.834 ± 0.783	0.0079 ± 0.00021
n=250	2.687 ± 0.301	2.081 ± 0.242	0.0048 ± 0.0001

الجدول(٣): معدل قيمة **MSE** وقيمة الانحراف المعياري للأنموذج الثالث في مقدر مقدر **Nadaraya-Watson**

	CV	GCV	FA
n=50	0.326 ± 0.057	0.319 ± 0.062	0.061 ± 0.040
n=100	0.313 ± 0.043	0.291 ± 0.050	0.049 ± 0.022
n=250	0.309 ± 0.018	0.275 ± 0.021	0.035 ± 0.013

الجدول(٤): معدل قيمة **MSE** وقيمة الانحراف المعياري للأنموذج الرابع في مقدر مقدر **Nadaraya-Watson**

	CV	GCV	FA
n=50	0.714 ± 0.143	0.353 ± 0.079	0.0049 ± 0.0019
n=100	0.531 ± 0.134	0.335 ± 0.056	0.0056 ± 0.0014
n=250	0.468 ± 0.054	0.219 ± 0.036	0.0045 ± 0.0013

10- الاستنتاجات:

١- من خلال مقارنة الطريقة المقترحة FA وطرائق التقدير اللامعلمية المعتمدة لتقدير مصفوفة معلمات التمهيد (H)في نموذج الإنحدار اللامعلمي، أن أفضل طريقة كانت الطريقة المقترحة FA ، لتحقيقها أقل قيمة لمعيار متوسط مربعات الخطأ (MSE)ولجميع حجوم العينات الثلاثة(250 , 100 , 50) ، ولجميع النماذج الأربعة المعتمدة في تجارب المحاكاة.

٢- جاءت طريقة GCV أفضل طريقة لتقدير مصفوفة (H) بعد الطريقة المقترحة FA لإعطائها نتائج جيدة في تقدير مصفوفة(H)، وكانت طريقة CV بالمرتبة الاخيرة لإعطائها أعلى قيم لـ MSE مقارنة بالطرائق .

11- المصادر:

١. عيسى، أسيل مسلم و حمود، مناف يوسف،(٢٠١٢)، "مقارنة بعض المقدرات شبه المعلمية لتقدير دالة الانحدار"، مجلة العلوم الاقتصادية والإدارية، المجلد (١٨)، العدد (٦٧) الصفحات [٢٧٣- ٢٨٨].

٢.محمد، لقاء علي وعبد الحسن، ميسم عبد النبي، (٢٠١٨)، "مقارنة المقدرات اللامعلمية في تحليل الانحدار المتعدد لدالتي كاما وبيتا" ، مجلة العلوم الاقتصادية والإدارية، المجلد (٢٤)، العدد (١٠٨)، الصفحات[٤٩٧- ٤٨٨]

٣. محمد، محمد عبد الحسين، (٢٠١١)، "استخدام مقدر كيرنل Nadaraya-Watson في تقدير دالة الانحدار اللامعلمي"، مجلة القادسية للعلوم الإدارية والاقتصادية، المجلد (١٣)، العدد (١) ١٥.

المصادر الأجنبية:

4. Al Radhwan, A.M.N.,& Algamal, Z.Y. ,(2021), "Improving K-means clustering based on firefly algorithm" Journal of Physics: Conference Series, 1897 (2021) 012004 IOP Publishing doi:10.1088/1742-6596
5. Aydin, D., (2007), "A comparison of the nonparametric regression models using smoothing spline and kernel regression", World Academy of Science, Engineering and Technology, (36), PP[253-257].
6. Boente, G., Fraiman, R., & Meloche, J., (1997), "Robust plug-in bandwidth estimators in nonparametric regression", Journal of Statistical Planning and Inference, 57(1), pp[109-142].
7. C.K. Chn (1995), " Bandwidth selection in Nonparametric Regression With general errors", Journal of Statistic Planning and Inference, 44 ,PP. 265-275.
8. Goutte, C., Larsen, J., & technology, v., (2000), "Adaptive metric kernel regression", Journal of VLSI signal processing systems for signal, image, 26(1-2), pp[155-167].
9. Hardle, W., (1990), "Applied Nonparametric Regression", Cambridge MA : Cambridge University press.
10. Harldle , W. , (1994)," Applied non parametric regression ".
11. Kahya, M. A., Altamir, S. A., & Algamal, Z. Y. (2020). Improving firefly algorithm-based logistic regression for feature selection. Journal of Interdisciplinary Mathematics, 22(8), 1577-1581. doi:10.1080/09720502.2019.1706861.
12. Kauermann, G., & Opsomer, J., (2004), "Generalized cross-validation for bandwidth selection of backfitting estimates in generalized additive models", Journal of Computational and Graphical Statistics, 13(1),PP[66-89].
13. Koláček, J., & Horová, I., (2017), "Bandwidth matrix selectors for kernel regression", Computational Statistics, 32(3), [1027-1046].
14. Lijian, Y., & Rolf, T., (1999), "Multiverait bendwidth selection for local linear regression", Statist, 61, pp[793-815].
15. Mustafa, M. Y. & Algamal, Z. Y.,(2022), "Bandwidth Selection in Multivariate Nadaraya-Watson Estimator based on Meta-Heuristic Optimization Algorithms: A Simulation Study" Mathematical Statistician and Engineering Applications, 71(4), [4877-4887].
16. Schimek, M. G., (2013)," Smoothing and regression: approaches, computation, and application", John Wiley & Sons.

17. Soméa, S. M., & Kokonendjia, C. C., (2015), "Effects of associated kernels in nonparametric multiple regressions", arXiv preprint arXiv:1502.01488.
18. Shang, H. L., Zhang, X., & Shang, M. H. L., (2014), "Package bbemkr".
19. Yang, X. S. (2010)," Firefly algorithms for multimodal optimization", In 5th symposium on stochastic algorithms, foundations and applications, SAGA 2009 (pp 169–178).
20. Yang, X. S. (Ed.). (2013). "Cuckoo search and firefly algorithm: Theory and applications" (Vol. 516). Springer.