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

توظيف خوارزمية الاعشاب الضارة لاختيار عرض الحزمة في مقدر نداريا- واتسون المتعدد

أ. د. زكريا يحيى الجمال	م.م. مروة يحيى مصطفى العبيدي
zakariya.algamal@uomosul.edu.iq	marwa.yahya@uomosul.edu.iq
قسم الاحصاء والمعلوماتية، كلية علوم الحاسوب والرياضيات، جامعة الموصل، الموصل، العراق	

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 الاعشاب الضارة 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 الانحدار اللامعلمي المتعدد. لقد أظهرت نتائج المحاكاة تفوق الطريقة المقترحة مقارنةً بطرائق التقدير الأخرى معتمدين متوسط مربعات الخطأ بوضعها معياراً للمقارنة.

معلومات البحث

تواريخ البحث:

تاريخ تقديم البحث: 19/2/2024
تاريخ قبول البحث: 12/4/2024
تاريخ رفع البحث على الموقع: 31/12/2024

الكلمات المفتاحية:

مقدرات النواة، مصفوفة عرض الحزمة، مقدر نداريا - واتسون المتعدد، خوارزمية الاعشاب الضارة.

للمراسلة:

أ. د. زكريا يحيى الجمال
zakariya.algamal@uomosul.edu.iq

DOI: https://doi.org/10.55562/jrucs.v56i1.10

1. المقدمة

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 Regression (Models) [1]. 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 [2].

إن تحديد مصفوقة عرض الحزمة (Bandwidth) أو ما تسمى بمصفوفة معلمات التمهيد (H) في 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).

2. هدف البحث وأهميتهُ

ان هذا البحث يهدف الى توظيف إحدى خوارزميات التقنيات الذكائية وهي خوارزمية الاعشاب الضارة (invasive weed optimization 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 وباستعمال نماذج مختلفة.

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

3. 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 للاهتمام بموضوع الانحدار اللامعلمي، وطرائق التمهيد الخاصة بهِ، ومن أنواعه: نموذج الانحدار اللامعلمي البسيط (simple Nonparametric Regression Model) والذي يقوم على إيجاد العلاقة بين متغير الاستجابة ومتغير توضيحي واحد فقط، وصيغته كما في المعادلة (1): [3].

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

أما إذا كان النموذج يقوم على إيجاد العلاقة بين متغير الاستجابة، 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)، وصيغته كما في المعادلة (2): [4, 5]

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

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

وأختصاراً تكتب كما في المعادلة (3):

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

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

هناك العديد من الطرائق اللامعلمية لتقدير هذه الدالة غير المعروفة والموضحة في المعادلتين (1) و(2)، 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 نداريا-واتسون (Nadaraya-Watson estimator) [6].

4. مقدر نداريا-واتسون المتعدد Multivariate Nadaraya-Watson Estimator (MNWE)

أن مقدرNadaraya-Watson يُعد من أكثر المقدرات الشائعة الاستخدام في تقدير دالة الانحدار اللامعلمي، تم اقتراحه هذا المقدر عام (1964) من قبل الباحثين Nadaraya وWatson، كما أنه يعد من أبسط أنواع الممهدات، إذا غالباً ما يستعمل في العديد من المجالات البحثية الإحصائية بالاعتماد على طريقة متسلسلة الأوزان كما في المعادلة (4)، إن عملية تقدير دالة الانحدار اللامعلمي $\mathbf{m(X)}$ غير المعروفة تتم باستخدام المتوسط الموزون، [7].

وتعرف طريقة المتوسط الموزون بأنها مشابهة لطريقة المربعات الصغرى الموزونة وكما هو مبين في المعادلة (4):

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

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 كما توضحه المعادلة (5): [8].

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

وإن دالة الوزن لها الخصائص الآتية: [6]

1) $\mathbf{w(x)} > 0 \quad for \ |\mathbf{x}| < 1$

2) $\mathbf{x} > 0$ إذا كانت دالة الوزن $\mathbf{w(x)}$ متزايدة

$$\mathbf{w}(\mathbf{x}) = 0 \quad for\ |\mathbf{x}| \geq 1 \quad (3$$

إذ إن هذه الأوزان تمثل دالة المسافة في فضاء X، والصيغة العامة للأوزان تكتب كما في المعادلة (6) : [7,9]

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

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 $k(u/h)$ 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 بالمعادلة (9): [9]

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

أي يصبح شكل المقدر كما في المعادلة (10):

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

وبتعميم ذلك في حالة وجود أكثر من متغير توضيحي أي d من المتغيرات التوضيحية تكون الصيغة لمقدر Nadaraya-Watson كما في المعادلة (11): [10].

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

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

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

أي إن:

$$\mathbf{h} = [h_1, h_2, \cdots, h_d]' \qquad \mathbf{H} = \mathbf{h}_{1\times d}\mathbf{I}_{d\times d} \tag{12}$$

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

5. 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 والتحيز 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) تأتى 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 [11, 12].

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 [13].

6. طريقة العبور الشرعي Cross Validation (CV)

هذه الطريقة تُعد من الطرائق شائعة الأستخدام؛ لإيجاد أنسب قيمة للمعلمة التمهيدية h، إذ تلعب هذه القيمة دوراً 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=1 تسمى (Leave One Out Method)، أما في حالة وجود أكثر من متغير توضيحي أي d من المتغيرات التوضيحية نستعمل صيغة المعادلة (14): [3, 9]

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

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

ولإيجاد مصفوفة قيم المعلمات التمهيدية المُثلى H نستعمل المعادلة (15):

$$CV(\mathbf{H}) = \frac{1}{n}\sum_{i=1}^{n}\left[\mathbf{y_i} - \hat{m}_{-1}(\mathbf{X_i};\mathbf{k})\right]^2 \tag{15}$$

7. طريقة العبور الشرعي العام Generalized Cross Validation (GCV)

هذه الطريقة GCV مستمدة من طريقة العبور الشرعي CV، حيث يتم الحصول عليها من صيغة CV الموضحة في المعادلة (14) وذلك عن طريق استبدال عناصر القطر الرئيسي $\hat{m}(x_i)$ لمصفوفة التمهيد S_h بمعدلها أي إن: [13, 14]

$$CV(\mathbf{H}) = \frac{1}{n}\sum_{i=1}^{n}\left[\mathbf{y_i} - \hat{m}_{-1}(\mathbf{X_i};\mathbf{k})\right]^2 \tag{16}$$

ولإيجاد المصفوفة H الخاصة بقيم المعلمات التمهيدية المُثلى نستعمل المعادلة (17):

$$\hat{\mathbf{H}}_{Gcv} = \underset{H\in H}{argmin}\,\mathrm{GCV}_H \tag{17}$$

8. خوارزمية الاعشاب الضارة: Invasive Weed Optimization Algorithm

خوارزمية أمثلة الأعشاب الضارة ((Invasive Weed Optimization Algorithm (IWO) هي خوارزمية التحسين العشوائي العددي المستوحاة بيولوجيا من الأعشاب الضارة والتي اقترحت لأول مرة من قبل Mehrabian و Lucas في عام (2006 م) [15]. وهذه الخوارزمية ببساطة تحاكي السلوك الطبيعي للأعشاب الضارة في الاستعمار وايجاد مكان مناسب للنمو والتكاثر. لمحاكاة السلوك الاستعماري للأعشاب الضارة يجب ان تؤخذ بعض الخصائص الاساسية لهذه العملية بنظر الاعتبار:

1. يتم نشر عدد محدود من البذور على منطقة البحث (تهيئة عدد السكان).
2. كل البذور تنمو الى نباتات مزهرة وتنتج البذور اعتمادا على دالة اللياقة (التكاثر).
3. البذور المنتجة يتم نشرها عشوائيا على منطقة البحث لتنمو وتصبح نباتات جديدة (التشتت المكاني).
4. تستمر هذه العملية الى ان يتم الوصول الى الحد الأقصى من عدد النباتات.

وفقط النباتات ذات دالة اللياقة العالية يمكنها البقاء على قيد الحياة وإنتاج البذور، ويجري القضاء على الاخرين (الإقصاء التنافسي). تستمر العملية الى ان يتم الوصول الى الحد الأقصى من التكرارات على امل ان النبات الذي يحمل أفضل دالة لياقه سيكون هو الاقرب الى الحل الامثل تتضمن خوارزمية أمثلة الأعشاب الضارة (IWO) عدد من الخطوات الاساسية، هذه الخطوات مترابطة مع بعضها البعض ولا يمكن تطبيق هذه الخوارزمية على اي مسالة مالم تطبق هذه الخطوات جميعها والا ستفقد خوارزمية أمثلة الأعشاب الضارة (IWO) قيمتها وفائدتها في ايجاد وتحسين الحل، ويمكن توضيح خطوات الخوارزمية على النحو الاتي:

- **الخطوة الاولى: تهيئة المجتمع الابتدائي Initialize A Population**

يتم توليد مجتمع ابتدائي من الحلول ونشرها على d من الابعاد من مساحة المشكلة مع مواقع عشوائية وحساب قيمة دالة اللياقة لهذه المجتمع.

- **الخطوة الثانية: التكاثر Reproduction**

يسمح للنبات في مجتمع النباتات بإنتاج البذور seed (التكاثر) وذلك اعتمادا على قيمة دالة اللياقة الخاصة به وكذلك الحد الأعلى والأدنى لدالة اللياقة في المستعمرة، اذ يزداد عدد البذور التي ينتجها النبات خطياً من الحد الأدنى الممكن لإنتاج البذور الى أقصى حد ممكن، وبعبارة اخرى فان النبات ينتج البذور اعتمادًا على قيمة دالة اللياقة الخاصة به واقل دالة لياقة للمستعمرة واعلى دالة لياقة للمستعمرة وذلك للتاكد من ان الزيادة تكون خطية

المعادلة (18) ادناه توضح عملية التكاثر للأعشاب الضارة :

$$seed_i = floor\left(\frac{f_i - f_{\min}}{f_{\max} - f_{\min}}\left(S_{\max} - S_{\min}\right)\right) + S_{\min}\mathrm{GCV}_H \qquad (18)$$

حيث ان Floor تدل على ان البذور تقرب لأقرب عدد صحيح، f_i تمثل دالة اللياقة ل i من الأعشاب الضارة، $f_{\min} and f_{\max}$: تمثل الحد الأقصى والأدنى لقيمة دالة اللياقة في لمستعمرة وان $S_{\max} and S_{\min}$ تمثل الحد الأقصى والأدنى لعدد البذور التي سوف تنتج في المستعمرة.

تمثل المعادلة (18) اعلاه العلاقة الرياضية بين عدد البذور وقيمة دالة اللياقة للأعشاب الضارة اذ ينخفض عدد البذور مع زيادة قيمة دالة اللياقة وعدد البذور يتراوح بين ال $S_{\max}$ و $S_{\min}$. تعتبر الأفراد القابلة للتكاثر هي تلك الأفراد ذوات أفضل قيمة لدالة اللياقة من الأفراد غير الملائمة للاستخدام وتعني كلمة "أفضل" هنا هي ان لهذه الأفراد فرصة اكبر للبقاء على قيد الحياة والتكاثر. لذا لا يسمح للأفراد غير الملائمة للاستخدام بالتكاثر. ومع ذلك فان وجهة النظر هذه تتجاهل شيء مهما الا وهو ان الخوارزمية التطورية هي طريقة احتمالية وتكرارية، فمن الممكن ان بعض الأفراد غير الملائمة للاستعمال تحمل في داخلها معلومات اكثر فائدة من الأفراد الملائمة خلال عملية التطور. علاوة على ذلك غالبا ما يستطيع النظام الوصول الى النقطة المثلى اذا كان بالإمكان عبور المنطقة غير قابلة للتطبيق (وخاصة في فضاء البحث غير المحدب). لذا اقترحت تقنية التكاثر اعلاه لإعطاء فرصة اكبر للأفراد غير الملائمة للاستخدام للبقاء على قيد الحياة، وهذه العملية مماثلة للآلية التي تحدث في الطبيعة.

- **الخطوة الثالثة: التشتت المكاني Spatial Dispersal**

توفر هذه الخطوة لخوارزمية الأعشاب الضارة خاصيتي العشوائية والتكيف, اذ يتم توزيع البذور المتولدة عشوائيا على d من الابعاد في فضاء البحث بواسطة ارقام عشوائية تتوزع توزيعا طبيعيا بمعدل (μ=0) وتباين متغير. وهذا يعني ان البذور سيتم توزيعها عشوائيا بحيث انها تقع بالقرب من النبات الام. الا ان الانحراف المعياري (SD) Standard deviation (σ) للدالة العشوائية سيخفض من قيمة اولية محددة مسبقا $(\sigma_{initial})$ الى قيمة نهائية (σ_{final}) في كل خطوة (كل جيل)، من خلال المعادلة (19):

$$\sigma_{iter} = \frac{\left(iter_{\max} - iter\right)^n}{\left(iter_{\max}\right)^n}\left(\sigma_{initial} - \sigma_{final}\right) + \sigma_{final} \qquad (19)$$

اذ ان σ_{iter} يمثل الانحراف المعياري في الخطوة الحالية، $iter_{\max}$ يمثل الحد الأقصى من التكرارات، وان n يمثل معدل التأشير اللاخطي. يضمن هذا التحويل ان احتمالية اسقاط البذور في منطقة بعيدة ينخفض بشكل غير خطي في كل خطوة زمنية مما يؤدي الى تجميع النباتات المجربة وازالة النباتات غير الملائمة.

يتم حساب موقع البذور الجديدة باستخدام المعادلة (20):

$$x_{son} = x_{parent} + sd = x_{parent} + random * \sigma_{iter} \quad (20)$$

حيث ان x_{son} يمثل موقع الذرية وان x_{parent} يمثل موقع الاباء في حين Random يمثل توليد اعداد عشوائية من التوزيع الطبيعي القياسي محصورة ضمن الفترة [0,1].

- **الخطوة الرابعة: الإقصاء التنافسي Competitive Exclusion**

إذا كان النبات لا يترك أي نسل فسوف ينقرض من الوجود، لذا دعت الحاجة الى نوع من التنافس بين النباتات للحد من العدد الأقصى من النباتات في المستعمرة. بعد مرور بعض التكرارات فان عدد النباتات في المستعمرة تصل الى الحد الأقصى عن طريق التكاثر السريع ومع ذلك فمن المتوقع ان يتم استنساخ النباتات المجربة اكثر من النباتات غير الملائم . عند الوصول الى الحد الأقصى لعدد النباتات في المستعمرة Pmax فسوف تنشط آلية إقصاء النباتات ذات دالة اللياقة الضعيفة لذلك الجيل. اذ تعمل آلية الإقصاء على النحو التالي: عندما يتم الوصول الى الحد الأقصى لعدد الأعشاب في المستعمرة يسمح لكل عشب بإنتاج البذور وذلك وفقا للآلية المذكورة في الخطوة (2) (التكاثر)، ثم يتم السماح للبذور المنتجة بالانتشار في منطقة البحث وذلك وفقا للخطوة (3) (التشتت المكاني). عندما تجد جميع البذور مواقعها في منطقة البحث يتم ترتيبها مع آبائها (كمستعمرة من الأعشاب الضارة). بعد ذلك يتم القضاء على الأعشاب الضارة ذات دالة اللياقة المنخفضة للوصول الى الحد الأقصى المسموح به للمجتمع في المستعمرة. وبهذه الطريقة ترتب النباتات وذريتها معا والعنصر ذو أفضل دالة لياقة سينجو ويبقى على قيد الحياة مع السماح لعملية التكرار داخل الخوارزمية. وكما ذكر سابقا في الخطوة (2) فإن هذه الآلية تعطي فرصة للنباتات ذات دالة اللياقة المنخفضة لإعادة الإنتاج فإن كانت ذريتها ذات دالة لياقة جيدة في المستعمرة فستنجو وتبقى على قيد الحياة بعبارة اخرى لا يتم إقصائها. وتطبق آلية التحكم بالمجتمع على الذرية ايضاً لحين انتهاء مرحلة معينة مما يحقق الإقصاء التنافسي. (في الشكل 1) يوضح الية عمل الخوارزمية في اختيار المتغيرات.

x_1	x_2	.....	x_{p-1}	x_p
1	0	.....	1	0

شكل (1): الية اختيار المتغيرات حسب خوارزمية الاعشاب الضارة

9. 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 (50,100,200)=n، وتم إجراء المقارنات للطرق المختلفة المستخدمة CV، GCV، مع الطريقة المقترحة لخوارزمية (IWO) وباستخدام دالة Epanechnikov كدالة لُبية، مع دراسة أربعة نماذج مختلفة هي:

- **النموذج الأول**

تم توليد هذا الأنموذج وفق المعادلة (21): [16]

$$\mathbf{y_i} = \left\{(\mathbf{x_1} - 0.5)^2 + \mathbf{x_2^2}\right\}\sin(2\pi\mathbf{x_3}) + \boldsymbol{\varepsilon_i} \quad (21)$$

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

$$\boldsymbol{\varepsilon_i} \sim N(0, 0.025)$$

- **النموذج الثاني**

تم توليد هذا النموذج وفق المعادلة (22): [17]

$$\mathbf{y_i} = 10\sin(\pi\mathbf{x_1}\mathbf{x_2}) + 20\left(\mathbf{x_3} - \frac{1}{2}\right)^2 + 10\mathbf{x_4} + 5\mathbf{x_5} + \boldsymbol{\varepsilon_i} \quad (22)$$

حيث تم توليد المتغيرات التوضيحية X_1 وX_2 وX_3 وX_4 وX_5 من التوزيع المنتظم ضمن الفترة [0,1]، أما $\boldsymbol{\varepsilon_i}$ فيتوزع بالشكل الآتي :

$$\boldsymbol{\varepsilon_i} \sim N(0,1)$$

- **النموذج الثالث**

تم توليد هذا النموذج وفق المعادلة (23): [18]

$$\mathbf{y_i} = \sin(2\pi\,\mathbf{x_1}) + 4(1-\mathbf{x_2})(1+\mathbf{x_2}) + \frac{2\mathbf{x_3}}{1+0.8\,\mathbf{x_3^2}} + \boldsymbol{\varepsilon_i} \qquad (23)$$

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

$\boldsymbol{\varepsilon_i} \sim N(0, 0.02)$

- **النموذج الرابع**

تم توليد هذا النموذج وفق المعادلة (24) [3]:

$$\mathbf{y_i} = (\mathbf{x_1} - 0.5)^3 + (\mathbf{x_2} - 0.5) + \boldsymbol{\varepsilon_i} \qquad (24)$$

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

$\boldsymbol{\varepsilon_i} \sim N(0, 0.02)$

حيث تم تكرار اجراء التجربة 250 مرة، وتم الاعتماد على متوسط مربعات الخطأ (MSE) كمعيار للمقارنة بين طرائق التقدير المستخدمة وبيان الطريقة الأفضل وتلخيص نتائج الطرق المستخدمة في (الجداول 1-4).

اظهرت النتائج تفوق طريقة (IWO) على باقي الطرائق الاخرى (CV,GCV) من حيث معاير MSE فعلى سبيل المثال (جدول 2) اعطت طريقة (IWO) اقل قيمة عندما n=50 ، حيث بلغت0.0131 مقارنة بـــ 3.136 لطريقة CV و2.979 لطريقة GCV بالإضافة الى ذلك الطريقة المقترحة حافظت على افضليتها عند تغير حجم العينة ولجميع النماذج المستخدمة.

جدول (1): معدل قيمة MSE وقيمة الانحراف المعياري للأنموذج الأول في مقدر Nadaraya-Watson

	CV	GCV	IWO
n=50	0.008 ± 0.046	0.008 ± 0.045	0.0023 ± 0.0130
n=100	0.005 ± 0.046	0.006 ± 0.043	0.0017 ± 0.0012
n=250	0.003 ± 0.043	0.040 ± 0.013	0.0014 ± 0.0013

جدول (2): معدل قيمة MSE وقيمة الانحراف المعياري للأنموذج الثاني في مقدر Nadaraya-Watson

	CV	GCV	IWO
n=50	0.649 ± 3.136	0.921 ± 2.979	0.0016 ± 0.0131
n=100	0.532 ± 3.122	0.783 ± 2.833	0.00021 ± 0.0069
n=250	0.303 ± 2.687	0.242 ± 2.080	0.0001 ± 0.0047

جدول (3): معدل قيمة MSE وقيمة الانحراف المعياري للأنموذج الثالث في مقدر Nadaraya-Watson

	CV	GCV	IWO
n=50	0.057 ± 0.326	0.062 ± 0.319	0.040 ± 0.061
n=100	0.043 ± 0.310	0.040 ± 0.281	0.027 ± 0.047
n=250	0.016 ± 0.304	0.021 ± 0.265	0.013 ± 0.032

جدول (4): معدل قيمة MSE وقيمة الانحراف المعياري للأنموذج الرابع في مقدر Nadaraya-Watson

	CV	GCV	IWO
n=50	0.143 ± 0.713	0.089 ± 0.343	0.0019 ± 0.0069
n=100	0.134 ± 0.521	0.056 ± 0.335	0.0016 ± 0.0046
n=250	0.044 ± 0.467	0.036 ± 0.216	0.0013 ± 0.0045

10. الاستنتاجات

1. من خلال مقارنة الطريقة المقترحة IWO وطرائق التقدير اللامعلمية المعتمدة لتقدير مصفوفة معلمات التمهيد (H) في نموذج الانحدار اللامعلمي، أن أفضل طريقة كانت الطريقة المقترحة IWO، لتحقيقها أقل قيمة لمعيار متوسط مربعات الخطأ (MSE) ولجميع حجوم العينات الثلاثة (250, 100, 50)، ولجميع النماذج الأربعة المعتمدة في تجارب المحاكاة.
2. جاءت طريقة GCV أفضل طريقة لتقدير مصفوفة (H) بعد الطريقة المقترحة IWO لإعطائها نتائج جيدة في تقدير مصفوفة (H)، وكانت طريقة CV بالمرتبة الاخيرة لإعطائها أعلى قيم لـ MSE مقارنة بالطرائق الاخرى.

المصادر

[1] Rencher, A. C., (2002), "Methods of Multivariate Analysis", Second Edition, John Wiley & Sons, Inc., Canada.

[2] محمد، لقاء علي وعبد الحسن، ميسم عبد النبي، (2018)، "مقارنة المقدرات اللامعلمية في تحليل الانحدار المتعدد لدالتي كاما وبيتا" ، مجلة العلوم الاقتصادية والإدارية، المجلد (24)، العدد (108)، الصفحات [497- 488]

[3] Koláček, J., & Horová, I., (2017), "Bandwidth matrix selectors for kernel regression", Computational Statistics, 32(3), [1027-1046].

[4] عيسى، أسيل مسلم و حمود، مناف يوسف،(2012)، "مقارنة بعض المقدرات شبه المعلمية لتقدير دالة الانحدار"، مجلة العلوم الاقتصادية والإدارية، المجلد (18)، العدد (67) الصفحات [273- 288].

[5] محمد، لقاء علي وعبد الحسن، ميسم عبد النبي، (2018)، "مقارنة المقدرات اللامعلمية في تحليل الانحدار المتعدد لدالتي كاما وبيتا" ، مجلة العلوم الاقتصادية والإدارية، المجلد (24)، العدد (108)، الصفحات [497- 488]

[6] Harldle , W. , (1994)," Applied non parametric regression ".

[7] 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].

[8] 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].

[9] Hardle, W., (1990), "Applied Nonparametric Regression", Cambridge MA : Cambridge University press.

[10] Soméa, S. M., & Kokonendjia, C. C., (2015), "Effects of associated kernels in nonparametric multiple regressions", arXiv preprint arXiv:1502.01488.

[11] Schimek, M. G., (2013)," Smoothing and regression: approaches, computation, and application", John Wiley & Sons.

[12] C.K. Chn (1995), " Bandwidth selection in Nonparametric Regression With general errors", Journal of Statistic Planning and Inference, 44 ,PP. 265-275.

[13] 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].

[14] 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].

[15] Mehrabian, A. R., & Lucas, C. (2006). A novel numerical optimization algorithm inspired from weed colonization. Ecological informatics, 1(4), 355-366.

[16] Lijian, Y., & Rolf, T., (1999), "Multiverait bendwidth selection for local linear regression", Statist, 61, pp[793-815].

[17] 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].

[18] Shang, H. L., Zhang, X., & Shang, M. H. L., (2014), "Package bbemkr".

Journal of AL-Rafidain University College for Sciences (2024); Issue 56; pp. 111-119

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

Employ the Algorithm Invasive Weed Optimization for Bandwidth Selection in the Multivariate Nadaraya-Watson Estimator

Zakariya Y. Algamal	Marwah Y. Mustafa
zakariya.algamal@uomosul.edu.iq	marwa.yahya@uomosul.edu.iq
Department of Statistics and Informatics, College of Computer Sciences and Mathematics, University of Mosul, Mosul, Iraq	

Article Information

Article History:
Received: February, 19, 2024
Accepted: April, 12, 2024
Available Online: December, 31, 2024

Keywords:
kernel estimator, smoothing matrix, Multivariate Nadaraya-Watson estimator, invasive weed optimization algorithm.

Correspondence:
Zakariya Y. Algamal
zakariya.algamal@uomosul.edu.iq

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 invasive weed optimization 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.

DOI: https://doi.org/10.55562/jrucs.v56i1.10