التنبؤ باستخدام انحدار العملية الغاوصية وانحدار مُتَّجه الدعم

رائد قراحسن * مصطفى مظهر رنة**

(الإيداع:28 آب 2018 ، القبول: 13شباط 2019)

الملخص

تم في هذا البحث دراسة انحدار العملية الغاوصية (Gaussian Process Regression) (GPR) وانحدار مُتَّجه الدَّعم (Support Vector Regression) (SVR) اللذين يعتبران من أحد أهم تقنيات التعلم الآلي، ويستخدمان في تحليل بيانات مختلف الأنظمة والتنبؤ بسلوكها بدقة عالية.

قُمنا في هذا البحث باقتراح طريقة جديدة لتحسين التنبُؤ عن طريق دمج تنبُؤات طريقة انحدار مُتَّجه الدَّعم وطريقة انحدار العملية الغاوصية، وتم التحقق من جودتها عبر تطبيقها على بيانات واقعية ومولدة. كما تم مقارنة هذه الطريقة مع طريقة انحدار العملية الغاوصية وانحدار مُتَّجه الدَّعم باستخدام مقاييس دقة التنبؤ (MSE، RMSE، MAPE)، بهدف الوصول لأفضل طريقة لتحسين دقة التنبؤ.

ودلَّت نتائج المقارنة أن الطريقة المقترحة تعطي أفضل دقة تنبؤ وأفضل نتائج وذلك لتكرار عدد الأفضلية بالاعتماد على أصغر قيمة من قيم مقاييس الأخطاء المستخدمة وبسبب قدرة منحني الانحدار المثل لها على ملائمة وتمثيل البيانات بشكل أفضل.

الكلمات المفتاحية: انحدار مُتَّجه الدعم، انحدار العملية الغاوصية، مقاييس دقة التنبؤ.

* طالب دراسات عليا (دكتوراه)–قسم الإحصاء الرياضي–كلية العلوم–جامعة حلب

**أستاذ مساعد–قسم الإحصاء الرياضي–كلية العلوم–جامعة حلب

Prediction using Gaussian Process Regression and Support Vector Regression

Raed Kara Hasan * **Moustafa Mazhar Rene** **

(Received: 28 August 2019,Accepted:13 February 2019)

Abstract

In this research study Gaussian Process Regression (GPR) and Support Vector Regression (SVR) are considered one of the most important techniques of automated learning, they are used to analyze various data sets and generate predictions with high prediction accuracy.

In this study, we proposed new method to improve prediction by integrating predictions Support vector regression method and Gaussian Process Regression method and their quality was verified by applying them on both artificial and realistic data. This method was also compared with the Support vector regression method and Gaussian Process Regression using the measurements of prediction error explanation (MSE, RMSE, MAPE), in order to obtain the ideal method to improve prediction accuracy.

The proposed method gives the best predictive accuracy and better results in order to replicate the number of preference based on the smallest value of the values of the error measures used , because of the ability of the regression curve ideals have an appropriate and better data representation.

Keywords: Support Vector Regression, Gaussian Process regression, the measurements of prediction error explanation.

***Postgraduate Student (PhD)–Dept. of Mathematical Statistics –Faculty of Science–University of Aleppo**

****Assistant Professor–Dept. of Mathematical Statistics–Faculty of Science**University of Aleppo**

1–مقدمة: Introduction

يشكل تحليل وتقييم العلاقات بين مجموعة من المتغيرات الهدف الأساسي لمعظم الأبحاث بغرض الوصول إلى نموذج رياضي يصف هذه العلاقات، وتضم هذه النماذج متغيرات تابعة (Dependent Variables) وتسمى أيضاً متغيرات الهدف (Target Variables) يمكن التنبؤ بها بوساطة متغيرات أخرى تعرف بالمتغيرات المستقلة (Independent Variables) أو المتغيرات التفسيرية (Explanatory Variables). تكون غاية هذه النماذج جعل الفرق بين القيم المقدرة (المتنبئ بها) والقيم الفعلية للمتغير الهدف أصغرياً.

يُستخدم النموذج الناتج لوصف وتحليل المشكلة والتنبؤ بمسارها بهدف الخروج بحلول ومقترحات وتوصيات بشأنها الأمر الذي يُساعد في عمليات التنمية والتَّخطيط وإعداد السياسات والاستراتيجيات للتحكُّم في أحداث مُستقبليّة ممكنة الوقوع (Izenman ،2008؛ Nielsen، 2009).

تم تطوير العديد من النّماذج لمُعالجة مسائل التنبؤ، كطريقة المربعات الصغرى، ونماذج الشَّبكات العُصبونيّة، وكان آخرها نموذجي انحدار العملية الغاوصية (GPR) وانحدار مُتَّجه الدَّعم (SVR) (Rasmussen وWilliams،2006).

قَدَّمَ العديد من الباحثين دراسات تتضمَّن دمج طرائق التنبؤ مع بعضها أو مع طرائق الذكاء الاصطناعي، وقد أثبتت هذه النماذج فاعليتها في تحسين دقة التنبُّؤ، ويعدُّ العالم Zhang وزملاؤه عام 2003 أول من درسوا دمج طرائق التنبُّؤ، بحيث قاموا بدمج طرائق الانحدار الذاتي مع طريقة الشبكات العصبية للتنبؤ بالسلاسل الزمنية (Zhang وزملاؤه ،2003).

وقام Shi وزملائه بالعام 2012 باقتراح دمج طرائق الانحدار الذاتي مع انحدار متجه الدعم للتنبؤ بالسلاسل الزمنية (Shi وزملاؤه ،2003).

تكمن أهمية البحث في عرض آخر طرائق التنبؤ وأكثرها استعمالاً وتطوير طرائق التنبؤ وكيفية الحصول على أفضل أداء لهذه الطرائق، كما أن تطبيقات هذه الطرائق كثيرة فهي تدخل في مجالات العلوم المختلفة وتطويرها يساعد على تقدم عملية البحث العلمي.

يكمن الهدف الرئيسي لهذا البحث في دراسة طرائق التنبؤ من خلال دراسة طريقتي انحدار العملية الغاوصية (GPR) وانحدار مُتَّجه الدَّعم (SVR) وتحسين عملية التنبؤ عبر اقتراح طريقة جديدة ومقارنتها بهدف الوصول لأفضل طريقة لتحسين دقة التنبؤ.

2–الموادُّ وطرائقُ البحث: Materials and Methods

2–1–التنبؤ باستخدام انحدار العملية الغاوصية:

يستخدم انحدار العملية الغاوصية (Gaussian Process Regression) أو اختصاراً (GPR) في تقنيات التعلم الآلي. قُدِّمَت طَّريقة العملية الغاوصية كأداة للانحدار(Regression) في مجال التعلم الآلي، لأول مرّة من قبل العالمين Rasmussen وWilliams عام 1996 حيث قاموا بوصف تحسين المعلمات في دالة التغاير والتي كانت مستوحاة من استخدام العملية الغاوصية مع الشبكات العصبونية، وقد تم استخدامها في تطبيقات مختلفة مثل التنبؤ بالنفوذية الجلدية من المواد الكيميائية والتنبؤ بتركيز الأوزون في الهواء (Bishop،2007؛Rasmussen وWilliams، 2006).

ليكن لدينا $g=(g(x_1), g(x_2), g(x_3), \dots, g(x_d))^T$ مُتَّجه ذو d بُعد من الدوال عندئذ تسمى العملية العشوائية $\{g(x): x \in \chi\}$ بعملية غاوص (بحيث أن χ هو فضاء المدخلات) إذا كان مُتَّجه المتغيرات العشوائية $X_1, X_2, \dots, X_d$ يتوزع وفق التوزيع الطبيعي المتعدد بمتوسط $\boldsymbol{\mu}$ و مصفوفة تغاير $\mathbf{K}$، تُعرف عملية غاوص كتوزيع على الدوال $P(g(x))$ بحيث أن $g(x)$ هي دالة معرفة على فضاء المدخلات χ كما يلي: $g: \chi \rightarrow \mathbb{R}$

أي أن العملية الغاوصية هي مجموعة من المتغيرات العشوائية المستمرة محدودة الأبعاد والتي كل منها يخضع للتوزيع الطبيعي وتكون جميع توزيعاتها هي توزيعات طبيعية، وتعتبرعملية غاوص (GP) من أهم تقنيات التعلم الآلي (Rasmussen وWilliams،2006؛ Liu وزملاؤه، 2017).

لتكن لدينا $\mu(x)$ دالة متوسط و $\mathrm{k}(x,\mathrm{x}')$ دالة تغاير معرفتان كما يلي:

$$\mu(x) = E[g(x)]$$

$$\mathrm{k}(x,\mathrm{x}') = Cov\big(g(x), g(\mathrm{x}')\big) = E\big[\big(g(x)-\mu(x)\big)\big(g(\mathrm{x}')-\mu(\mathrm{x}')\big)\big]$$

بحيث $x, x' \in \chi$ عندئذٍ العملية الغاوصية (GP) تأخذ الشكل التالي:

$$\begin{bmatrix} g(x_1) \\ \vdots \\ g(x_d) \end{bmatrix} \sim N_d\left(\begin{bmatrix} \mu(x_1) \\ \vdots \\ \mu(x_d) \end{bmatrix}, \begin{bmatrix} k(x_1,x_1) & k(x_1,x_2) & \cdots & k(x_1,x_d) \\ k(x_2,x_1) & k(x_2,x_2) & \cdots & k(x_2,x_d) \\ \vdots & \vdots & \ddots & \vdots \\ k(x_d,x_1) & k(x_d,x_2) & \cdots & k(x_d,x_d) \end{bmatrix}\right) \qquad (1)$$

ونرمز لذلك بالرمز:

$$P\big(g(x)\big) = \mathcal{GP}(\,\mu(x), k(x,\mathrm{x}')) \qquad (2)$$

نسمي الدالة $k(x,x')$ بدالة التغاير أو دالة النواة (نواة التغاير) وهي دالة موجبة محدودة ولها عدة أنواع (Bishop،2007).

ليكن Y متغير تابع وX متغيرات عشوائية ذو d بُعد، يعطى نموذج الانحدار اللامعلمي وفق العلاقة:

$$\boldsymbol{y} = g(\mathrm{x}) + \boldsymbol{\varepsilon} \qquad , \boldsymbol{\varepsilon} \sim \boldsymbol{N(0,\sigma^2)} \qquad (3)$$

بحيث أن: $g(x)$ هي دالة مجهولة أما في الانحدار المعلمي تكون معلومة، تعاني الطرائق اللامعلمية من مشكلة تعدد الأبعاد (curse of dimensionality) عندما يتم تطبيقها مع المتغيرات المتعددة (أي عندما تكون d كبيرة)، لقد تم تطوير مجموعة متنوعة من النماذج البديلة للتغلب على هذه المشكلة منها نموذج انحدار العملية الغاوصية (GPR) .

إن نموذج انحدار العملية الغاوصية هو نموذج لامعلمي، وهذا يعني بأنه لا يفترض شكل معين للدالة المدروسة ولكن يتم تحديد شكل العلاقة بين المدخلات والأهداف بالكامل من خلال البيانات التي قد تتضمن عدد غير محدود من الدوال، وتكون الدالة الأساسية التي تنتج البيانات مجهولة ولكن يتم توليد التنبؤات من خلال مجموعة من الدوال التي تخضع لتوزيع غاوص في فضاء الدوال، ويعتبر نموذج انحدار العملية الغاوصية من أحدث طرائق التنبؤ، وهو من نماذج بايز الاحتمالية، ففي معظم طرائق انحدار بايز يتم إيجاد معلومات مسبقة عن معلمات النموذج، وبعد ذلك يتم وضع شروط على البيانات لإعطاء معلمات النموذج اللاحق(البعدي)، حيث يمكن صياغة هذه المعلومات المسبقة بشكل توزيع احتمالي يسمى التوزيع القبلي و يحدد نموذج بايز المعلمات المجهولة للنموذج القبلي بينما يحدد نموذج عملية غاوص علاقات الدوال القبلية مباشرة بين مدخلات الاختبار ومدخلات ومخرجات التدريب (Rasmussen وWilliams،2006؛ Liu وزملاؤه ،2017).

لنفترض لدينا مجموعة من البيانات $\{(x_i, y_i)\}_{i=1}^{n}$ بحيث تشير $x_i \in \mathbb{R}^d$ إلى المدخلات والتي لها d بُعد وتشير $y_i \in \mathbb{R}$ إلى القيم الحقيقية للنواتج و n إلى عدد البيانات،عندئذٍ يأخذ نموذج انحدار العملية الغاوصية (GPR) الشكل التالي:

$$\boldsymbol{y_i} = g(\boldsymbol{x_i}) + \boldsymbol{\varepsilon_i} \quad ; \boldsymbol{i = 1, \dots, n} \quad , \boldsymbol{\varepsilon_i} \sim \boldsymbol{N(0,\sigma^2)} \qquad (4)$$

بحيث أن: $g(x) \sim \mathcal{GP}(\,\mu(x), k(x,\mathrm{x}'))$

و $\mathcal{GP}(\,\mu(x), k(x,\mathrm{x}'))$ هي عملية غاوص القبلية (Gaussian process prior) مع دالة متوسط $\mu(x)$ ودالة تغاير $k(x,\mathrm{x}')$ وبالتالي يعطى نموذج انحدار العملية الغاوصية وفق العلاقة:

$$\boldsymbol{y} = \mathcal{GP}(\,\mu(x), k(x,\mathrm{x}')) + \sigma_n^2\,\delta(x,\mathrm{x}') \qquad (5)$$

بحيث أن: $\delta(x,\mathrm{x}')$ دالة دلتا كرونكير (Kronecker delta) و $\delta(x,\mathrm{x}') = 0$ عندما $x \neq \mathrm{x}'$

و $\delta(x,x') = 1$ عندما $x = x'$ و σ_n^2 تباين الضجيج العشوائي ومن الشائع أيضاً أن نفترض $\mu(x)$ =0 أي (دالة المتوسط للعملية الغاوصية القبلية معدومة)عندئذٍ يأخذ نموذج انحدار العملية الغاوصية الشكل التالي:

$$\boldsymbol{y} \sim \mathcal{GP}(0, k(x,x')) + \sigma_n^2\, \delta(x,x') \qquad (6)$$

تم تصميم مجموعة متنوعة من دوال النواة، وسيتم في هذا البحث استخدام دالة النواة الغاوصية والموضحة وفق العلاقة الآتية:

$$k(x,x') = e^{-\frac{\|x-x'\|^2}{2\sigma^2}} \qquad (7)$$

بحيث أن: $\|x-x'\| = \sqrt{(x-x')^T(x-x')}$ تشير إلى طويلة الشعاع $(x-x')$ أو نظيم الفرق بين قيمتين x, x' و σ معامل دالة نواة غاوص (Rasmussen و Williams، 2006).

2-2-التنبؤ باستخدام انحدار متجه الدعم:

يقع انحدار متجه الدعم ضمن نطاق نظرية التعلم الاحصائي أو نظرية VC نسبة إلى العالمين Vapnik و Chervonenkis وتصف هذه النظرية خواص التعلم الآلي التي تمكنها من تعميم البيانات بشكل جيد.

قُدِّمَت هذه الطَّريقة كأداة للتصنيف (classification) لأول مرّة من قبل العالم Vapnik عام 1995 باسم آلية مُتَّجه الدَّعم (Support vector machine) أو اختصاراً (SVM) وقد نالت أهمية كبيرة في المجال التطبيقي. وفي عام (1998) قام Vapnik بتعديل طريقة (SVM) لتُعالج مسائل توفيق التوابع، وأُطلِقَ على هذه الطريقة تسمية طريقة انحدار مُتَّجه الدَّعم (Support vector Regression) والتي يرمز لها اختصاراً بـ (SVR) (Hastie وزملاؤه ،2009).

تعتمد طريقة انحدار مُتَّجه الدَّعم على مبدأ تعلُّم إحصائي غير موجود في نماذج الانحدار التّقليدية يُعرف هذا المبدأ بتخفيض المُخاطرة البنيوية (structural risk minimization) حيثُ يُقدِّم هذا المبدأ حدوداً لانحراف المُخاطرة التَّجريبية (empirical risk) عن المُخاطرة المُتوقَّعة (expected risk) بحيث يتم قياس جودة التقدير في انحدار متجه الدعم باستخدام نوعاً جديداً من دوال الخسارة يدعى بالدالة غير الحساسة لـ ε المقترحة من قبل Vapnik وفق العلاقة التالية:

$$L_\varepsilon\big(y, f(x,w)\big) = \begin{cases} 0 & ; \ |y - f(x,w)| \le \varepsilon \\ |y - f(x,w)| - \varepsilon & otherwise \end{cases} \qquad (8)$$

بحيث w تمثل معاملات الدالة f ويعبر المعامل ε عن الحد الأعلى من الانحرافات المسموح بها وبكلمات أخرى فإننا لا نهتم بالأخطاء مادامت أقل من ε لكننا لن نقبل أي انحراف أكبر من ε.

يتم في انحدار متجه الدعم بداية تنظيم المدخلات x إلى فضاء سمات ذو m بُعد باستخدام مجموعة تحويلات غير خطية محددة $\phi_i(x)$ (إي الانتقال بمجموعة بيانات التدريب من فضاء الادخال إلى فضاء السمات والذي يكون له عدد أبعاد أكبر) ومن ثم يتم بناء نموذج خطي ضمن فضاء السمات (Alex وزملاؤه ،2004).

ويعطى النموذج الخطي (في فضاء السمات) وفق العلاقة:

$$f(x,w) = \sum_{i=1}^{m} w_i \phi_i(x) + b \qquad (9)$$

بحيث يمثل الحد b الانحياز (bias)، وتشير $\phi_i(x)\,; i = 1, \dots, m$ إلى مجموعة تحويلات غير خطية. مثلاً يمكن افتراض التحويل التالي $\Phi: \mathbb{R}^2 \longrightarrow \mathbb{R}^3$ حيث $\Phi(x_1, x_2) = \left(x_1^2, \sqrt{2}x_1x_2, x_2^2\right)$ تشير الرموز السفلية إلى مكونات $x \in \mathbb{R}^2$.

ويتم بالوقت نفسه تخفيض تعقيد النموذج باستخدام دالة الخسارة غير الحساسة لـ ε عبر تصغير $\|w\|^2$. ولذلك يتم ادخال متغيرات مهملة (غير سالبة) $\xi_i, \xi_i^* \; ; \; i = 1, \dots, n$ لقياس انحراف عينات التدريب خارج المنطقة المحددة بـ ε. وتعطى صيغة انحدار متجه الدعم بشكل دالة التصغير الآتية:

$$\min_{w,\xi_i,\xi_i^*} \quad \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*)$$
$$s.t. \quad \begin{cases} y_i - f(x_i, w) \le \varepsilon + \xi_i \\ f(x_i, w) - y_i \le \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \ge 0 \qquad i = 1, \dots, n \end{cases} \tag{10}$$

حيث أن C ثابت موجب (معامل التعميم) (Cherkassky و Ma، 2004؛ Alex وزملاؤه، 2004).

إن ايجاد القيمة المثلى لتابع غير خطي يخضع لقيود بصيغة متراجحات يتم بإضافة شروط جديدة تدعى بشروط KKT (Karush–Kuhn–Tucker) على طريقة مضاريب لاغرانج ، من خلال تقديم مجموعة من المتغيرات الثنوية α_i, α_i^*. بحيث تم إثبات أن لهذه الدالة نقطة سرجية (saddle point) فيما يتعلق بالمتغيرات الأولية primal والثنوية dual عند الحل (Alex وزملاؤه، 2004).

يعطينا حل المسألة السابقة القيم الأمثلية لـ b و w بدلالة المتغيرات الثنوية والتي من خلالها نستطيع تقدير قيم التنبؤ بشكل عددي ويكون لمسألة الأمثلية الحل التالي:

$$w = \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)x_i \quad \Longrightarrow \quad f(x) = \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)K(x, x_i) + b \tag{11}$$

وهذا ما ندعوه بمنشور متجه الدعم (Support Vector expansion).

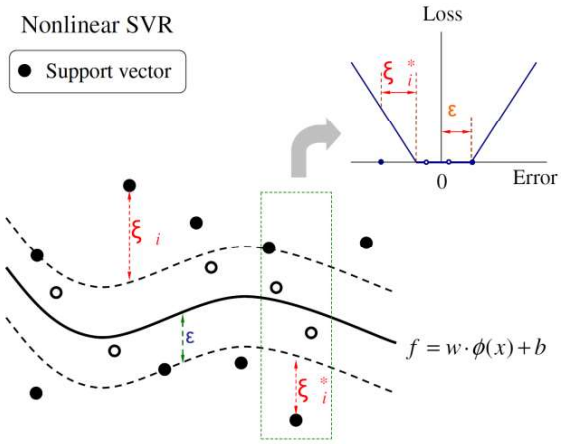
بحيث يتم تقييد المتغيرات الثنوية $\alpha_i^* \; , \alpha_i$ بالشروط $0 \le \alpha_i^* \; , \alpha_i \le C$ ويشير $K(x, x_i)$ إلى دالة النواة بحيث يمكن الإشارة إلى ناتج الجداء الداخلي $\langle \phi(x), \phi(x_i) \rangle$ على أنه مقياس للتشابه بين x, x' في الفضاء الذي تم التحويل إليه (فضاء السمات). أي أننا سنقوم بحساب التشابه في فضاء السمات باستخدام الأنماط x_i (فضاء المدخلات). وبالتالي يمكننا الاكتفاء بمعرفة دالة التشابه k بحيث $K(x, x_i) = \sum_{j=1}^{m}\langle \phi_j(x)\phi_j(x_i) \rangle$ بدلاً من ϕ. وتُعرَّف دالة التشابه k بأنها دالة النواة التي تحقق شروط Mercer (Alex وزملاؤه، 2004).

تم تصميم مجموعة متنوعة من دوال النواة، وسيتم في هذا البحث استخدام دالة نواة لابلاس والموضحة وفق العلاقة الآتية:

$$k(x, x') = e^{\frac{\|x - x'\|}{\sigma}} \qquad ; \; \sigma \ge 0 \tag{12}$$

بحيث أن: $\|x - x'\| = \sqrt{(x - x')^T(x - x')}$ تشير إلى طويلة الشعاع $(x - x')$ و σ معامل دالة نواة لابلاس. وتدعى نقاط البيانات التي تظهر مع معاملات غير معدومة بالعلاقة (11) بمتجهات الدعم SV_s وسنرمز بـ n_{SV} لعدد متجهات الدعم.

يوضح الشكل رقم (1) آلية عمل انحدار متجه الدعم بالحالة غير الخطية، بحيث تمثل النقاط باللون الأسود متجهات الدعم، كما يظهر كيفية تحكم قيمة ε بعرض الشريط غير الحساس، ويظهر قيم المتغيرات المهملة ξ_i, ξ_i^* والتي يتم التحكم بها باستخدام المعامل C .



الشكل رقم (1): انحدار متجه الدعم بالحالة غير الخطية باستخدام دالة الخسارة غير الحساسة لـ ε

تعتمد دقة التقدير لنموذج انحدار متجه الدعم على الضبط الجيد لضبط قيم المعاملات C, ε ومعامل دالة النواة σ (Cherkassky و Ma ،2004؛ Alex وزملاؤه ،2004 ؛ Yuوزملاؤه ،2006).

يحدد المعامل C الموازنة بين تعقيد النموذج (التسطح) (flatness) ودرجة الانحرافات الأكبر من ε المسموح بها في صيغة الأمثلية بالعلاقة (13). بينما يتحكم المعامل ε بعرض المنطقة غير الحساسة لـ ε المستخدمة لملائمة البيانات (Cherkassky و Ma ،2004؛ Alex وزملاؤه ،2004).

سنعتمد على طريقة الاختيار التحليلي AS (analytic selection) للمعاملات C, ε ومعامل دالة النواة σ مباشرة من بيانات التدريب وهي طريقة مقترحة من قبل Cherkassky و Ma عام (2004).

وتعطي هذه الطريقة نتائجاً جيدة وتتصف بسهولة حساب قيمها وتعطى قيم المعاملات بهذه الطريقة وفق العلاقة الآتية:

$$C = max(|\bar{y} + 3S_y|, |\bar{y} - 3S_y|)$$
$$\varepsilon = 3\hat{\sigma}\sqrt{\frac{\ln n}{n}} \qquad (13)$$
$$\sigma \approx \tau \times range(x)$$

بحيث أن $S_y, \bar{y}$ تشير على التوالي إلى المتوسط الحسابي والانحراف المعياري لقيم y

و $range(x) = |max(x) - min(x)|$ و τ عدد ضمن المجال $[0.1,0.5]$ وتم اختيار القيمة $\tau = 0.30$ في هذا البحث و $\hat{\sigma}$ هو تقدير مستوى الضجيج لبيانات التدريب ويتم حسابه عبر الصيغة المقدمة بوساطة نموذج الجار الأقرب $k\_nearest\_neighbour$ كما يلي :

$$\hat{\sigma}^2 = \frac{n^{\frac{1}{5}}k}{n^{\frac{1}{5}}k - 1} \cdot \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \qquad (14)$$

بحيث أن $\hat{y}_i$ هي قيم الانحدار بوساطة طريقة الجار الأقرب و k معامل يمثل عدد نقاط الجوار في خوارزمية الجار الأقرب

(Cherkassky و Ma،2004).

2-3-الطريقة المُقترَحة لتحسين التنبؤ:

تعتبر أساليب تحليل الانحدار من أهم وأقوى أساليب التحليل الإحصائي الذي يُقيم العلاقات بين مجموعة من المتغيرات بغرض الوصول إلى صيغة تصف هذه العلاقات التي تمكننا من التنبؤ عن حصول تغير واحد أو أكثر في ضوء التغيرات الأخرى التي تتعلق بها، أي أن تحليل الانحدار طريق لتوقع نتيجة معينة اعتماداً على متحول أو عدة متحولات مستقلة.

حيث أننا في تحليل الانحدار نجري توافقاً بين نموذج تنبؤي والبيانات المتوفرة لدينا أي أننا سنستخدم البيانات لتقدير نموذج يمكنه أن يصف الظاهرة بشكل جيد، ونستخدم هذا النموذج لنتوقع قيماً للمتحول التابع اعتماداً على متحول أو أكثر من المتحولات المستقلة (التنبؤية)، هذا ويمكننا التنبؤ بأية بيانات اعتماداً على المعادلة العامة التالية:

$$\mathrm{Outcome_i} = \mathrm{model_i} + \mathrm{error_i} \qquad (15)$$

وهذا يعني أن النتيجة يمكننا الحصول عليها باستخدام نموذج ملائم لبيانات مع إضافة نوع من الخطأ، تتخذ شكل المعادلة وفقاً لنوع العلاقة التي نشاهدها ومن واقع البيانات الإحصائية الخاصة بهذه المتغيرات والتي يجب أن تتصف بالدقة وذلك حتى يلائم النموذج طبيعة الظاهرة (Izenman، 2008؛ Nielsen، 2009).

يملك كل من نموذجي الانحدار انحدار العملية الغاوصية GPR وانحدار متجه الدعم SVR إمكانات وخواص مُختلفة عند وصف سلوك وسمات منحني الانحدار ضمن الأنماط الخطية وغير الخطية، لذا فإن النموذج المُقترح في هذا البحث يتُكوَّن من مُركبات كلا النموذجين بحيث نستطيع باستخدام النموذج المقترح نمذجة الأنماط المختلفة لنموذج الانحدار وتحسين مُجمل سلوك التنبُّؤ.

تتميز طريقة انحدار متجه الدعم بمقدرتها على الملائمة الشاملة للبيانات لذلك فهي تستفيد من معلومات كامل مجموعة البيانات لتوليد التنبؤات، بينما تتصف طريقة انحدار العملية الغاوصية GPR بقدرتها على الملائمة الموضعية فهي قادرة على نمذجة أخطاء التنبُّؤ (Hastie وزملاؤه ،2009).

ليكن لدينا مجموعة من البيانات $\{(x_i, y_i)\}_{i=1}^{n}$ عندئذ يتكون نموذج الانحدار من جزء يمثل دالة الانحدار r_i وجزء يمثل قيم الأخطاء e_i، بحيث تَتَضمَّن الأخطاء e_i علاقة غير خطية تربط بين المُشاهدات (Shi وزملاؤه ،2003).

وبالتالي يُمكن التعبير عن y_i (مجموعة البيانات الأصلية) كما يلي:

$$y_i = r_i + e_i \qquad (16)$$

ويتم تقدير $\hat{y}_i$ من خلال مجموعة البيانات المدروسة بثلاث مراحل:

أولاً: يتم تقدير قيم r_i باستخدام انحدار متجه الدعم (أي أننا سنقوم بالتنبُّؤ بالبيانات الأصلية التي لدينا باستخدام نماذج انحدار متجه الدعم)، عندها تكون قيم البواقي عند المشاهدة i هي: $e_i = y_i - r_i$.

ثانياً: يتم نمذجة البواقي باستخدام طريقة انحدار العملية الغاوصية GPR:

$$\hat{e}_i = g(x_i) \qquad (17)$$

علماً أن g دالة غير خطية تمَّ نمذجته باستخدام طريقة انحدار العملية الغاوصية GPR .

ثالثاً: يتم جمع التقديرين اللذان حصلنا عليهما باستخدام نموذجي التنبُّؤ المُستخدمين، وبالتالي فإن النموذج المقترح للتنبُّؤ بالبيانات هو:

$$\hat{y}_i = r_i + \hat{e}_i \qquad (18)$$

ويمكن من العلاقتين (11) و(6) مع استبدال قيمة y_i في العلاقة (6) بالبواقي e_i والتعويض في العلاقة (18) لينتج لدينا العلاقة التالية التي تمثل النموذج المقترح التالي:

$$\hat{y}_i = \sum_{i=1}^{n} (\alpha_i - \alpha_i^*) K(x, x_i) + b + \mathcal{GP}(0, k(x, \mathrm{x}')) + \sigma_n^2 \, \delta(x, \mathrm{x}') \qquad (19)$$

بحيث $k(x, \mathrm{x}')$, $\delta(x, \mathrm{x}'), \sigma_n^2$ قد تم شرحها ضمن الفقرة(**2–1**)

وتمثل e_i قيم البواقي و K دالة النواة و b, α_i, α_i^* التي قد تم شرحها ضمن الفقرة(**2–2**).

وبهذا نكون قد استفدنا من خاصية التنبؤ الشامل لانحدار متجه الدعم في تقدير خط الانحدار وخاصية التنبؤ الموضعي لانحدار العملية الغاوصية في نمذجة وتقدير قيم الأخطاء.

2–4–الجانب التطبيقي:

بغرض اختبار أداء الطريقة المُقترحة في تحسين التنبؤ قُمنا بتطبيق الطريقة المُقترحة وطريقتي انحدار مُتَّجه الدَّعم وانحدار العملية الغاوصية على مجموعتي بيانات واقعية ومولدة.

2–4–1–البيانات الواقعية:

تمثل مجموعة البيانات الواقعية أعمار وأوزان وأطوال مجموعة من الأطفال أعمارهم من لحظة الولادة وحتى ست سنوات ونصف بحيث اعتمدنا بالقياس على الواحدات التالية السنة للعمر والكيلو للوزن والمتر للطول.

يهدف نموذج التَّنبُؤ المراد بناءه إلى تقدير وزن الطفل عن طريق معرفة عمره وطوله بحيث يمثل وزن الطفل المتغير التابع Y، بينما يمثل عمر الطفل المتغير الأول المستقل X_1 ويمثل طول الطفل المتغير الثاني المستقل $\boldsymbol{X_2}$ ،

(مصدر البيانات: نورالدين، (2013)).

2–4–2–البيانات المولدة:

تم توليد مجموعة البيانات المولدة بحجم 100 باستخدام دالة من النمط $sinc$ معطاة وفق العلاقة (Hu وزملاؤه ،2010):

$$y = sinc(x) = \begin{cases} 1 & ; x = 0 \\ \dfrac{sinx}{x} & ; otherwise \end{cases} \qquad (20)$$

وتم إضافة ضجيج عشوائي(تشويش) وفق التوزيع الطبيعي بتوقع رياضي معدوم وبانحراف معياري قدره 0.1σ= و $\rho = 0.5$ و $\rho = 1$.

2–4–3–منهجية اختبار الطريقة المُقترَحة:

تعتبر الدّقة معياراً لاختيار النّموذج الأمثل للتَّنبؤ، ويُقصد بالدّقة قُدرة نموذج التَّنبؤ على إعادة إنتاج البيانات الأصل للعينة المدروسة، ومنه فإن الاختيار المُناسب لمقياس دقة التَّنبؤ يؤثر إيجاباً في تحديد فاعلية نموذج التنبؤ المُستخدم وتعمل مقاييس دقّة التَّنبُؤ القياسيّة بشكل عام على مفهوم الفرق بين القيم الأصلية والقيم المُتوقَّعة أو المُتنبئ بها، وهو ما ندعوه بخطأ التَّنبؤ، وكُلما كان مقدار الفرق قليلاً كانت دالة التنبؤ أفضل وأدقّ، يوجد العديد من مقاييس دقّة التنبؤ، وعادةً لا يتم الاعتماد على مقياس واحد في عمليّة ضبط نموذج التنبؤ. تعتبر المقاييس التّالية: MAPE،MSE, RMSE من أفضل مقاييس المقارنة بين نماذج تنبؤ مختلفة تم بناؤها باستخدام نفس مجموعة بيانات التدريب (Hyndman و Koehler، 2006).

وقد اعتمدنا في بحثنا هذا على هذه المقاييس كونها تلائم طبيعة البحث ويتم حسابهم كما يلي:

$MSE = mean(e_i^2)$ متوسط مربعات الأخطاء Mean squared error

$RMSE = \sqrt{MSE}$	جذر متوسط مربعات الأخطاء	Root mean squared error
$MAPE = mean(\lvert e_i/y_i \rvert)$	متوسط القيم المطلقة للأخطاء النسبية	*Mean absolute percentage error*

بحيث يحسب خطأ التنبؤ من العلاقة: $e_i = y_i - \hat{y}_i$

(يُدعى خطأ التنبؤ في بعض المراجع بالبواقي residuals) (Hardle وزملاؤه ، 2004).

وبهدف تطبيق الطريقتين تم كتابة برنامج نصي باللغة البرمجية **R (لغة برمجية إحصائية)**، حيث تستطيع هذه اللغة القيام بالعديد من تحليلات البيانات بحيث يتم تنظيم هذه التحليلات ضمن ما يسمى بالحزم Packages مما يعني قدرة الباحثين على تطوير البرامج المختلفة الأمر الذي ساهم بانتشار استخدامها في المجالات الأكاديمية (Cotton، 2013 ؛Matloff، 2011).

لتنفيذ هذا البحث تم الاستفادة من الحزم التالية:

الحزمة caret بهدف تطبيق خوارزمية الجار الأقرب لتقدير قيمة الضجيج وحساب مصفوفة المسافات الإقليدية لحساب معاملات انحدار متجه الدعم بالطريقة التحليلية، والحزمة kernlab بهدف تطبيق انحدار العملية الغاوصية وانحدار متجه الدعم.

في المرحلة الأولى قُمنا بتطبيق نموذج انحدار العملية الغاوصية على مجموعتي البيانات الواقعية والمولدة وتم الاعتماد على دالة النواة الغاوصية بالعلاقة (7) كدالة نواة وتم حساب قيم مقاييس الأخطاء الثلاثة (MAPE، RMSE،MSE) للحكم على جودة أداء كل طريقة من طرائق التنبؤ.

وفي المرحلة الثانية قُمنا بتطبيق نموذج انحدار متجه الدعم على مجموعتي البيانات وتم حساب المعاملين C, ε ومعامل دالة النواة σ وفق العلاقة (13) وكما تم الاعتماد على دالة نواة لابلاس بالعلاقة (12) كدالة تشابه في فضاء السمات الذي تم التحوير إليه وتم حساب قيم مقاييس الأخطاء الثلاثة (MAPE، RMSE،MSE) .

وفي المرحلة الثالثة قُمنا بتطبيق النموذج المقترح على مجموعتي البيانات الواقعية والمولدة وتم اعتماد على نفس قيم المعاملات المحسوبة في المرحلتين السابقتين. وتم أيضاً حساب قيم مقاييس الأخطاء الثلاثة (MAPE، RMSE،MSE) .

3–النتائج ومناقشتها:

سنقوم الآن بعرض النتائج التطبيقية على كل من مجموعتي البيانات:

بعد التطبيق العملي ظهرت لدينا النتائج الموضحة بالجداول والأشكال كما يلي:

الجدول رقم (1) قيم المعاملات المستخدمة ضمن مجموعة البيانات الواقعية

معامل انحدار العملية الغاوصية σ_n	معامل دالة النواة الغاوصية σ	معامل الجار الأقرب k	معامل انحدار متجه الدعم ε	معامل انحدار متجه الدعم C	معامل دالة نواة لابلاس σ
1	40.89324	5	0.5108546	26.74253	35.3994

المصدر: من إعداد الباحث اعتماداً على مخرجات اللغة البرمجية R

بعد تطبيق الطريقة المُقترحة وطريقة انحدار مُتَّجه الدَّعم وطريقة انحدار العملية الغاوصية وفق المعاملات الموضحة بالجدول رقم (1) على مجموعة بيانات الأطفال الواقعية وحساب قيم مقاييس الأخطاء لكل طريقة ظهرت لدينا النتائج التالية:

الجدول رقم (2): نتائج تطبيق الطرائق الثلاثة على مجموعة البيانات الواقعية

الطريقة المقترحة			طريقة انحدار مُتَّجه الدَّعم			طريقة انحدار العملية الغاوصية		
MAPE	RMSE	MSE	MAPE	RMSE	MSE	MAPE	RMSE	MSE
0.0581	1.0590	1.1215	0.3627	2.2859	5.2254	0.2917	1.7798	3.1680

المصدر: من إعداد الباحث اعتماداً على مخرجات اللغة البرمجية R

يُظهر الجدول رقم (2) من اليسار إلى اليمين اسم الطريقة المستخدمة وقيم مقاييس الأخطاء الثلاثة (MSE،RMSE ، MAPE) على مجموعة البيانات الواقعية.

نلاحظ من الجدول رقم (2) بأنه كان للطريقة المقترحة قيماً أصغر لمقاييس الأخطاء الثلاثة (MAPE، RMSE،MSE) من قيم مقاييس الأخطاء الناتجة عن التطبيق المفرد لطريقتي انحدار العملية الغاوصية وانحدار متجه الدعم على مجموعة البيانات الواقعية.

وبإعادة الحسابات بالنسبة لمجموعة البيانات المولدة وعند الحالات المختلفة للضجيج العشوائي 0.1σ= و$\rho = 0.5$ و$\rho = 1$ ظهرت لدينا النتائج التالية:

الجدول رقم (3) قيم المعاملات المستخدمة ضمن مجموعة البيانات المولدة

	معامل دالة نواة لابلاس σ	معامل انحدار متجه الدعم C	معامل انحدار متجه الدعم ε	معامل الجار الأقرب k	معامل دالة النواة الغاوصية σ	معامل انحدار العملية الغاوصية σ_n
Sinc 0.1	6	1.294957	0.037591	5	5.860907	0.063164737
Sinc 0.5	6	2.016488	0.185948	5	5.860907	0.525084891
Sinc 1	6	3.268399	0.371603	5	5.860907	0.750846655

المصدر: من إعداد الباحث اعتماداً على مخرجات اللغة البرمجية R

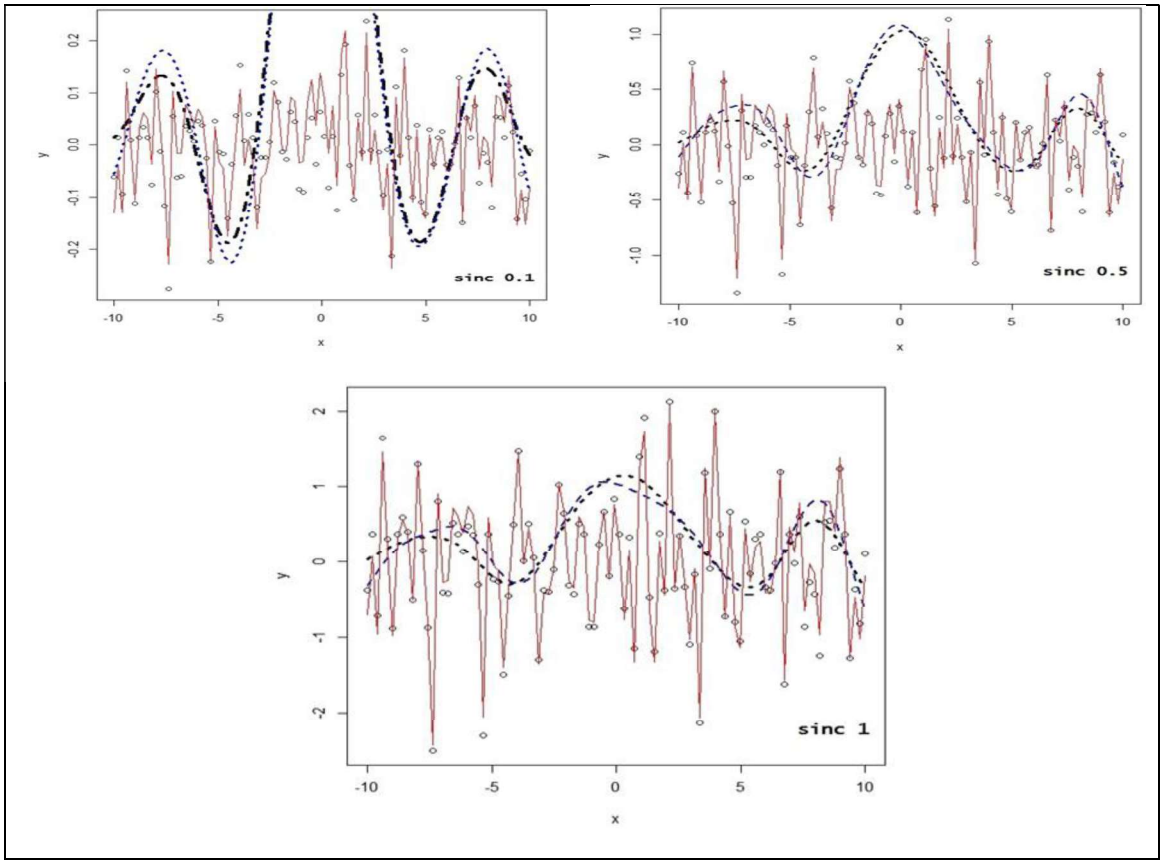
بعد تطبيق الطرائق الثلاثة وفق المعاملات الموضحة بالجدول رقم (3) على مجموعة البيانات المولدة وعند الحالات المختلفة للضجيج العشوائي 0.1σ= و$\rho = 0.5$ و$\rho = 1$ وحساب قيم مقاييس الأخطاء لكل طريقة ظهرت لدينا النتائج التالية:

الجدول رقم (4): نتائج تطبيق الطرائق الثلاثة على مجموعة البيانات المولدة

	طريقة انحدار العملية الغاوصية			طريقة انحدار مُتَّجه الدَّعم			الطريقة المقترحة		
	MSE	RMSE	MAPE	MSE	RMSE	MAPE	MSE	RMSE	MAPE
Sinc 0.1	0.00881	0.09386	1.19612	0.00766	0.08753	0.99279	0.00125	0.03544	0.04065
Sinc 0.5	0.18870	0.43439	1.52227	0.18960	0.43543	1.73138	0.00888	0.09426	1.41863
Sinc 1	0.74879	0.86533	1.44718	0.7468	0.86420	1.89874	0.02159	0.14696	0.48680

المصدر: من إعداد الباحث اعتماداً على مخرجات اللغة البرمجية R

يُظهر الجدول رقم (4) من اليسار إلى اليمين اسم الطريقة المستخدمة وقيم مقاييس الأخطاء الثلاثة على مجموعة البيانات المولدة، حيث أن قيم مقاييس الأخطاء الثلاثة المحسوبة باستخدام الدالة $sinc$ مع إضافة ضجيج عشوائي قدره $0.1\sigma=$ في السطر الأول، وقيم مقاييس الأخطاء الثلاثة المحسوبة باستخدام الدالة $sinc$ مع إضافة ضجيج عشوائي قدره $0.5\sigma=$ في السطر الثاني، وباستخدام الدالة $sinc$ مع إضافة ضجيج عشوائي قدره $1\sigma=$ في السطر الثالث.



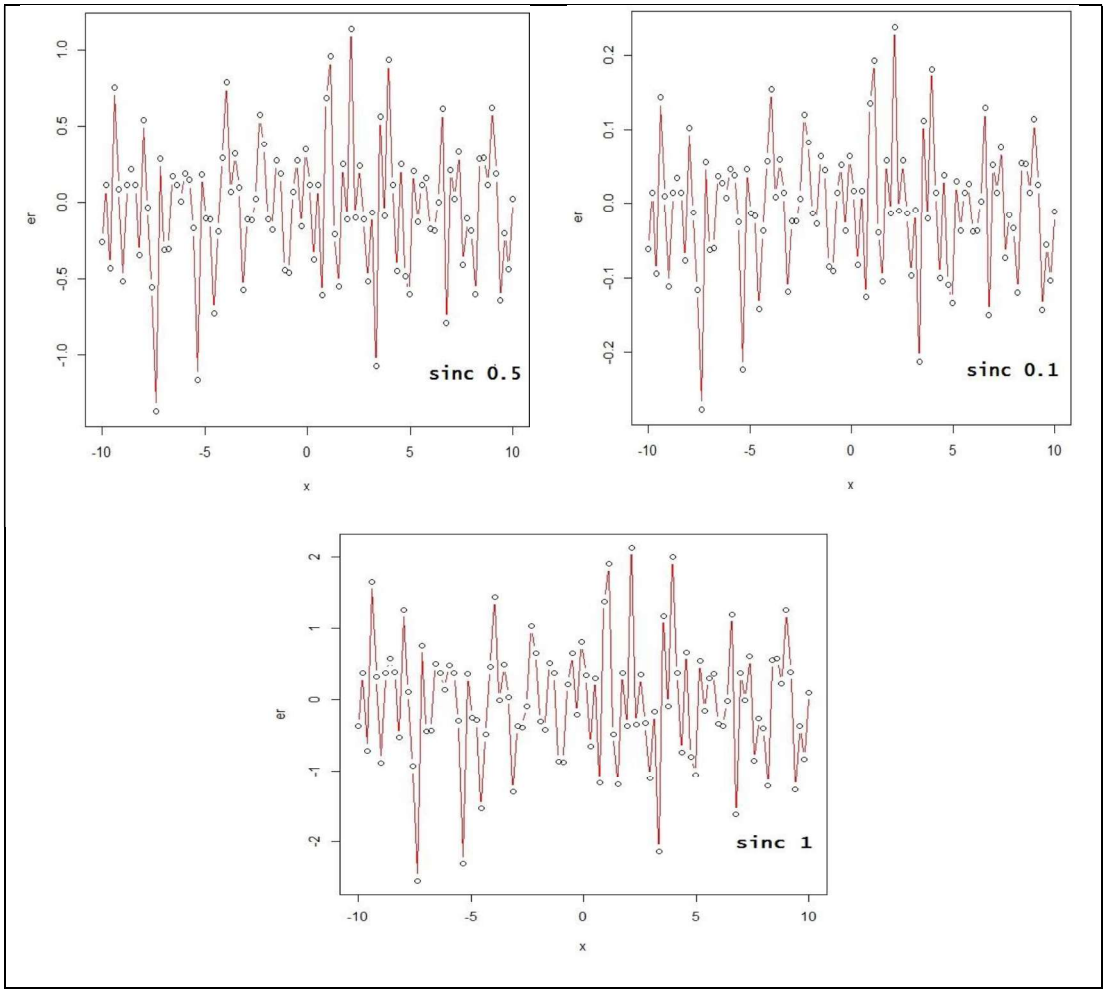
الشكل رقم (2): مقارنة خطوط الانحدار بالطرائق الثلاثة في حالة البيانات المولدة باستخدام الدالة Sinc مع إضافة ثلاث حالات مختلفة للضجيج العشوائي

كما يظهر الشكل رقم (2) مقارنة لخطوط الانحدار المحسوبة بالطرائق الثلاثة (انحدار العملية الغاوصية بالخط المتقطع باللون الأسود وانحدار متجه الدعم بالخط المنقط باللون الأزرق والطريقة المقترحة بالخط المتواصل باللون الأحمر) في حالة البيانات المولدة باستخدام الدالة Sinc مع إضافة ثلاث حالات مختلفة للضجيج العشوائي $\rho=0.1$ و$\rho=0.5$ و$\rho=1$

وتم الرسم باستخدام لغة R باستخدام التعليمتين plot و lines.

نلاحظ من الجدول رقم (4) تفوق الطريقة المقترحة على طريقتي انحدار العملية الغاوصية و انحدار متجه الدعم لأنها حققت قيماً أصغر لمقاييس الأخطاء الثلاثة (MAPE، RMSE،MSE) عند تطبيقها على مجموعة البيانات المولدة باستخدام الدالة ***Sinc*** في الحالات الثلاث المختلفة للضجيج العشوائي $\rho=0.1$ و$\rho=0.5$ و$\rho=1$ ونلاحظ أيضاً أنه كلما ازدادت قيمة الضجيج العشوائي كلما ازدادت قيمة مقياس الخطأ المستخدم.

ويُظهر الشكل رقم (3) نمذجة الأخطاء باستخدام انحدار العملية الغاوصية وهي العملية الموضحة بالخطوة الثانية من النموذج المقترح على مجموعة البيانات المولدة باستخدام الدالة ***Sinc*** مع إضافة ثلاث حالات مختلفة للضجيج العشوائي $\rho=0.1$ و $\rho=0.5$و$\rho=1$.



الشكل رقم (3): نمذجة الأخطاء على مجموعة البيانات المولدة باستخدام الدالة $Sinc$ مع إضافة ثلاث حالات مختلفة للضجيج العشوائي

نلاحظ من الشكل (2) بأن النموذج المقترح يمتلك أفضل تمثيل للبيانات ومن خلال الشكل رقم (3) نجد قدرة نموذج انحدار العملية الغاوصية على ملائمة قيم أخطاء التنبؤ.

4–الاستنتاجات والتوصيات:

4–1–الاستنتاجات: Conclusions

من خلال ما ذكر سابقاً وما سُجل من نتائج نورد ما يلي:

1–أظهرت الدراسة التي أجريناها على مجموعتي البيانات الواقعية والمولدة تفوق الطريقة المقترحة على طريقتي انحدار مُتَّجه الدَّعم وانحدار العملية الغاوصية، وذلك من خلال تحقيقها لأصغر قيمة من قيم مقاييس الأخطاء الثلاثة (MSE، RMSE، MAPE).

2–من خلال نتائج هذه الدراسة تبيّن بأن النموذج المُقترح قد أعطى دقة تنبُؤ أفضل من دقة التنبُؤ التي حصلنا عليها باستعمال نماذج التنبُؤ المُفردة لطريقتي انحدار العملية الغاوصية وانحدار متجه الدعم وذلك لتكرار عدد الأفضلية بالاعتماد على أصغر قيمة من قيم مقاييس الأخطاء المستخدمة وبسبب قدرة منحني الانحدار المثل لها على ملائمة وتمثيل البيانات بشكل أفضل.

4-2-التوصيات: Recommendations

1-مما سبق نوصي باستخدام الطريقة المقترحة كونها أعطت أفضل النتائج على مجموعات بيانات أُخرى.

2- نوصي الباحثين بإجراء عمليات دمج بين طرائق الانحدار الأخرى وتطبيقها في مجالات العلوم المختلفة.

3-إجراء دراسات مستقبلية حول تحسين دقة التنبؤ باستخدام طرائق أخرى غير الطريقة المقترحة في هذا البحث.

5-المراجع العربية:

1) **نورالدين، محمد مالك، (2013)**. تطوير تقنيات تنقيب المعطيات للقيام بمهام التنبؤ. رسالة ماجستير في الرياضيات- كلية العلوم-جامعة حلب.

6-References:

12- **Alex, J., Smol, A., Scholkope, B., (2004).** A Tutorial on Support Vector Regression, Statistics and Computing. 14, 199–222.

13- **Bishop, C. M., (2007).**Pattern Recognition and Machine Learning. Springer.

14- **Cherkassky, V., Ma, Y., (2004).**Practical Selection of SVM Parameters and Noise Estimation for SVM Regression, Neural Networks, 17, 113–126.

15- **Cotton, R., (2013).** Learning R, O'Reilly Media, Inc., United States of America, 377.

16- **Hardle, W., Muller, M., Sperlich, S.,and Werwatz A., (2004)** .Nonparametric and Semiparametric Models, Springer, Berlin, 301.

17- **Hastie, T., Tibshirani, R., Friedman, J., (2009).** The Elements of Statistical Learning Data Mining, Inference, and Prediction. Springer, 2th ED, Berlin, 764.

18- **Hu, Z.,Min, W, Huang, X., (2010).**Parameter Selection of Support Vector Regression Based on Particle Swarm Optimization, International Conference on Granular Computing, 5, 251 – 256.

19- **Hyndman, R., J., Koehler, A. B., (2006).** Another Look at Measures of Forecast Accuracy, International Journal of Forecasting, 22, 679–688.

20- Izenman, A.J., (2008). Multivariate Statistical Techniques: Regression,Classification and manifold learning. New York: Springer.

21- **Liu, Y., Keller, Y., Song, PH., Bond, J., Jiang, G., (2017).**Prediction of concrete corrosion in sewers with hybrid Gaussian processes regression model. RSC Advances, 7, 30894–30903.

22- **Matloff, N., (2011).**The Art of R Programming, Malloy Incorporated, United States of America, 373.

23- **Nielsen, A.,**(2009) .Least Squares Adjustment Linear and Nonlinear Weighted Regression Analysis Informatics and Mathematical Modelling.Technical University, Denmark.

24- **Rasmussen, C. E., Williams C. K. I., (2006).** Gaussian Processes for Machine Learning. MIT, Press.

25- **Shi, J., GUO, J., Zheng, S., (2012).** Evaluation of Hybrid Forecasting Approaches for Wind Speed and Power Generation Time Series. Sustainable Energy Reviews, 16, 3471-3480.

26- **Yu, P., Chen, S., Chang, I., (2006).** Support Vector Regression for Real-time Flood Stage Forecasting, Hydrology, 328, 704- 716.

27- **Zhang, G.; Patuwo, B., Hu, M., (1998).** Time Series Forecasting Using a Hybrid ARIMA and Neural Network Model. Neuro computing, 50, 159-175.