Journal of Education and Science (ISSN 1812-125X)

A Comparison Between Estimating The Parameters of The Gaussian Process Regression Model Using The Maximum Likelihood and The Restricted Maximum Likelihood Methods

A. K. Ilyas *(1), Y. H. Al-Taweel (2)

(1,2) Department of Mathematics, College of Education for Pure Science, University of Mosul, Mosul, Iraq

Article information

Article history:
Received: June 21, 2024
Revised: August 25, 2024
Accepted: September 06, 2024
Available online: December 01, 2024

Keywords:
Computer Model
Gaussian Process
Maximum likelihood estimation (MLE)
Restricted Maximum Likelihood estimation (RMLE)
Borehole Function

Correspondence:
Amena Khalid ilyas
amena.22esp4@2student.uomosul.edu.iq

Abstract

Gaussian process regression models are used as statistical representations of computational models, due to their flexibility in capturing the shape of smooth functions. The Gaussian process regression model has a number of parameters, the estimation of which is an essential step towards building the model. The parameters considered are the regression coefficients β, the scaling parameter σ^2 and the correlation lengths ψ. Estimating these parameters is the problem we address in this paper. The main contribution of this work is a comparison between estimating the parameters of the Gauss process regression model using the maximum likelihood method and the restricted maximum likelihood method. This comparison was made based on some validation measures. The Gauss process regression model, whose parameters were estimated using the two methods above, was applied to a real eight-dimensional example represented by the borehole function model, and all mathematical and graphical operations were carried out using the R program.

DOI: 10.33899/edusj.2024.151002.1472, ©Authors, 2024, College of Education for Pure Science, University of Mosul.
This is an open access article under the CC BY 4.0 license (http://creativecommons.org/licenses/by/4.0/).

1. المقدمة

تتم دراسة عدد متزايد من الأنظمة الطبيعية أو التي من صنع الإنسان بمساعدة النماذج الحاسوبية، في محاولة لتعزيز فهمنا للعمليات الأساسية وقدرتنا على التنبؤ بسلوكها. النماذج الحاسوبية الحديثة يمكن ان تكون معقدة بشكل كبير جدًا، لدرجة أن تنفيذاً واحداً قد يتطلب قدرًا كبيرًا من الوقت لإكماله. علاوة على ذلك، قد يتطلب تنفيذ مهام مثل معايرة الأنموذج أو تحليل حساسية المدخلات الآلاف من عمليات التنفيذ، والتي قد تكون غير عملية بالنسبة لبعض فئات النماذج أو ببساطة غير معقولة[1].

أنموذج انحدار العملية الكاوسية Gaussian Process regression model هو تقنية تساعد في معالجة بعض المشكلات المذكورة أعلاه حيث يعتبر كبديل فعال حسابيًا للأنموذج الحاسوبي الأصلي. تم تقديم نماذج انحدار العملية الكاوسية لأول مرة بواسطة [2]، وهي تعتمد على العمليات الكاوسية Gaussian Processes(GP). تستخدم هذه الفئة من نماذج انحدار العملية الكاوسية GP كتوزيع أولي لمخرجات الأنموذج الحاسوبي، أو بمعنى آخر، تفترض أن مخرجات الأنموذج الحاسوبي عبارة عن عينة عشوائية مأخوذة من توزيع كاوسي مشترك. يسمح أنموذج انحدار العملية الكاوسية بالتنبؤ بمخرجات الأنموذج الحاسوبي عند المدخلات غير المشاهدة، ويوفر في الوقت ذاته فترات ثقة لهذا التنبؤ. بالإضافة إلى ذلك، أدت الأبحاث ذات الصلة إلى ظهور طرق تؤدي إلى معايرة الأنموذج الاحتمالي [3] أو تحليل الحساسية [4] بطريقة فعالة للغاية.

تحتوي العمليات الكاوسية على عدد من المعلمات غير المعلومة بشكل عام. ولذلك، فهي إما بحاجة إلى التقدير، أو إيجاد توزيعاتها الهامشية حسب المنظور البيزي. يوفر التوزيع الهامشي ميزة تحديد مقدار عدم اليقين حول القيمة الحقيقية للمعلمات، وأيضًا يجعل تقديرات المعلمات أكثر وضوحًا واستقرارًا [5]. ومع ذلك، قد لا يكون إيجاد التوزيع الهامشي دائمًا ممكنًا أو عمليًا لواحد أو أكثر من معلمات الأنموذج.

في هذه الحالة، من الشائع إيجاد أفضل تقدير للمعلمة واعتباره هو القيمة الحقيقية، وهي طريقة تُعرف باسم "طريقة الملء" plug-in Method [3].

تحتوي العملية الكاوسية التي نأخذها في الاعتبار في هذا البحث على ثلاثة انواع من المعلمات وهي: معاملات الانحدار β، ومعلمة القياس σ^2 ومتجه المعلمات ψ التي تتحكم في نعومة المخرجات، والمعروفة باسم معلمات الارتباط. في ظل افتراض دالة الوسط الخطية، من الممكن إيجاد التوزيع الهامشي بطريقة تحليلية لمعاملات الانحدار ومعلمة القياس. ومع ذلك، فإن إيجاد التوزيع الهامشي لمعلمات الارتباط غير ممكن تحليليا، حتى بالنسبة لأبسط أشكال دالة الارتباط.

يمكن بناء نماذج انحدار عملية كاوس باستخدام طريقتين مختلفتين. ونعني ببناء نموذج انحدار عملية كاوس تقدير معلماته. الطريقة الأولى لتقدير هذه المعلمات هي طريقة الإمكان الأعظم (MLE). الطريقة الثانية هي طريقة الإمكان الأعظم المقيد(RMLE). قد تؤثر مجموعة معلمات أنموذج انحدار عملية كاوس على أداءها خاصة مع وجود عدد صغير من نقاط التصميم. وهذا يعني أنه إذا لم يتم تقدير المعلمات بشكل جيد، فقد لا نحصل على تقديرات تقريبية جيدة لمخرجات النموذج الحاسوبي أو قد لا يتم تحديد الفروق بشكل جيد.

ولذلك فإن الهدف من بحثنا هو مقارنة وتقييم الطريقتين لبناء نماذج انحدار عملية كاوس. سيتم ذلك باستخدام بعض المقاييس التي تعتمد على الاختلافات بين التنبؤات والقيم الحقيقية للنموذج الحاسوبي. ويتم تنظيم هيكل البحث على النحو التالي: يعرض القسم 2 تفاصيل أنموذج انحدار العملية الكاوسية كبديل للأنموذج الحاسوبي ويناقش القسم 3 الأساليب المختلفة لمعاملة معلماته. وتم عرض في القسم 4 التصميم المستخدم في توليد النقاط. كما يوضح القسم 5 و6 نتائج تطبيق طرق تقدير معلمات أنموذج انحدار العملية الكاوسية على أنموذج دالة البئر.

2. أنموذج انحدار العملية الكاوسية للأنموذج الحاسوبي

لتكن $f(\mathbf{x})$ تشير إلى مخرجات الأنموذج الحاسوبي لمتجه الإدخال $\mathbf{x}$. نفترض أن الأنموذج الحاسوبي يحتوي على مدخلات ذات الأبعاد p اي ان $\mathbf{x} = [x_1, x_2, \cdots, x_p] \in \Re^p$ ومخرجات ذات بعد واحد ((scalar المتمثلة بـ $f(\mathbf{x}) \in \Re$. الفكرة الرئيسة لنماذج انحدار العملية الكاوسية هي استخدام التوزيع الطبيعي المشترك كتوزيع أولي لمخرجات الأنموذج الحاسوبي[6]. نفرض ان مخرجات الأنموذج الحاسوبي هي:

$$p(f(\mathbf{x})|\boldsymbol{\beta}, \sigma^2, \boldsymbol{\psi}) \sim \mathcal{N}\big(h(\mathbf{x})^T\boldsymbol{\beta}, \sigma^2 k(\mathbf{x}, \mathbf{x})\big). \qquad (1)$$

ان دالة الوسط لهذا الأنموذج تعطى بواسطة $h(\mathbf{x})^T\boldsymbol{\beta}$، إذ إن $h(\mathbf{x})$ هو متجه $(q \times 1)$ للدوال المحددة مسبقا لـ $\mathbf{x}$، والمعروفة باسم دوال الانحدار، و $\boldsymbol{\beta}$ هو متجه $(\mathrm{q} \times 1)$ لمعاملات الانحدار. يسمى هذا الشكل من دالة الوسط بالخطي، ويمكن أن يكون مرنا إلى حد ما بسبب الخيارات التي يمكن إجراؤها لدوال الانحدار؛ في الوقت نفسه، فإن خطيته فيما يتعلق بمعاملات الانحدار ييسط التحليل اللاحق.

يتم إعطاء التغاير للأنموذج بواسطة $\sigma^2 k(\mathbf{x}, \mathbf{x}')$ إذ إن σ^2 هو التباين العام. إنّ الارتباط بين مخرجات الأنموذج الحاسوبي يحسب باستخدام دالة الارتباط $k(\mathbf{x}, \mathbf{x})$ [7]. عائلة دوال الارتباط التي سنأخذها في الاعتبار مستقرة وتعتمد قيمتها على النسب $\frac{|x_i - x_i'|}{\psi_i}, i \in [1 \ldots p]$. ان المعلمات الموجبة $\boldsymbol{\psi} = \{\psi_i : i \in [1 \ldots p]\}$ تسمى بمعلمات الارتباط وتتحكم في نعومة المخرجات. دوال الارتباط التي تنتمي إلى الفئة المذكورة أعلاه هي الدوال الأسية، والتربيعية النسبية و دالة ماتيرن Matern [8].

بعد تحديد العملية الكاوسية الأولية للأنموذج الحاسوبي، فإن الخطوة التالية هي الحصول على مخرجات الأنموذج الحاسوبي لعدد من المدخلات المختلفة. تسمى هذه المدخلات نقاط التصميم Design Points ونشير إليها بشكل جماعي بواسطة $D = [\mathbf{x}_1 ، \mathbf{x}_2 ، \cdots ، \mathbf{x}_n]$. يتم الإشارة إلى مخرجات الأنموذج الحاسوبي في هذه النقاط بواسطة متجه $y = [y_1, y_2, \cdots, y_n]$ ذات بعد $(n \times 1)$.

باستخدام نظرية بيز Bayes'Theorem [9] ، والخواص القياسية للتوزيعات الكاوسية متعددة المتغيرات فان توزيع مخرجات الأنموذج الحاسوبي عند نقطة إدخال جديدة $\mathbf{x}$ مشروطة على قيم الأنموذج الحاسوبي المنفذة سابقا وعلى المعلمات $(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\psi})$ هو:

$$p(f(\mathbf{x})|y, \boldsymbol{\beta}, \sigma^2, \boldsymbol{\psi}) = \mathcal{N}\big(m_0(\mathbf{x}), \sigma^2 u_0(\mathbf{x}, \mathbf{x})\big), \qquad (2)$$

حيث

$$m_0(\mathbf{x}) = h(\mathbf{x})^{\mathrm{T}}\boldsymbol{\beta} + k(\mathbf{x}, D)\mathrm{A}^{-1}(y - H\boldsymbol{\beta}),$$

$$u_0(\mathbf{x}, \mathbf{x}) = k(\mathbf{x}, \mathbf{x}) - k(\mathbf{x}, D)\mathrm{A}^{-1}k(D, \mathbf{x}),$$

في المعادلات المذكورة أعلاه، تشير D إلى أن الدوال المعنية تنطبق على جميع عناصر D، أي

$$k(D ، \mathbf{x}) \equiv k(\mathbf{x}, D)^T \text{ و } k(\mathbf{x}, D) \equiv [k(\mathbf{x} ، \mathbf{x}_1)، k(\mathbf{x} ، \mathbf{x}_2)، \cdots ، k(\mathbf{x} ، \mathbf{x}_n)],$$

ان دالة الارتباط بين نقاط التصميم والتوقع تعطى كالاتي:

$$\boldsymbol{k}(\mathbf{x}) = [k(\mathbf{x}, \mathbf{x}_1), k(\mathbf{x}, \mathbf{x}_2), \cdots, k(\mathbf{x}, \mathbf{x}_n)]^T,$$

و

$$H \equiv h(D)^T = [h(\mathbf{x}_1)، h(\mathbf{x}_2)، \cdots ، h(\mathbf{x}_n)]^T,$$

نعرف أيضًا مصفوفة الارتباط لنقاط التصميم D كما يلي:

$$, \ [10]A \equiv \ k\ (D, D) = \begin{pmatrix} 1 & k(\mathbf{x}_1, \mathbf{x}_2) \cdots & k(\mathbf{x}_1, \mathbf{x}_n) \\ k(\mathbf{x}_2, \mathbf{x}_1) & 1 \cdots & k(\mathbf{x}_2, \mathbf{x}_n) \\ \vdots & \vdots & \vdots \\ k(\mathbf{x}_n, \mathbf{x}_1) & \ldots \quad \ldots & 1 \end{pmatrix}$$

3. معاملة المعلمات

نناقش الآن طرقًا مختلفة لتقدير المعلمات $\boldsymbol{\beta}$ وσ^2 و$\boldsymbol{\psi}$ وكذلك إيجاد معادلات التنبؤ بمخرجات أنموذج انحدار عملية كاوس عند أية نقطة إدخال عشوائية $\mathbf{x}$.

3. 1 طريقة الإمكان الأعظم (Maximum Likelihood) [11], [12], [13]

التوزيع التنبئي الوارد في معادلة 2 يكون مشروط على المعلمات الثلاثة غير المعلومة، وأبسط طريقة لمعاملة هذه المعلمات هي تعظيم دالة الإمكان $p\ (y \mid \boldsymbol{\beta}\ ،\ \sigma^2\ ،\ \boldsymbol{\psi})$، بالنسبة لـ $\boldsymbol{\beta}$ ، σ^2 و $\boldsymbol{\psi}$، والتي تعطى من خلال المعادلة التالية:

$$p(y|\boldsymbol{\beta}, \sigma^2, \boldsymbol{\psi}) = \frac{|A|^{-1/2}}{(2\pi\sigma^2)^{n/2}} exp\left\{-\frac{1}{2\sigma^2}(y - H\boldsymbol{\beta})^T A^{-1}(y - H\boldsymbol{\beta})\right\}, \qquad (3)$$

ويتم استخدام تلك التقديرات في التوزيع التنبئي لـمعادلة 2 على سبيل المثال، يمكن إيجاد قيم $\boldsymbol{\beta}$ و $\boldsymbol{\sigma}^2$ التي تعظم التعبير أعلاه لقيمة معينة من $\boldsymbol{\psi}$ من خلال المعادلات التالية:

$$\widehat{\boldsymbol{\beta}}_{ML} = (H^T A^{-1} H)^{-1} H^T A^{-1} y, \qquad (4)$$

$$\hat{\sigma}^2_{ML} = \frac{\left(y - H\widehat{\boldsymbol{\beta}}_{ML}\right)^T A^{-1}(y - H\widehat{\boldsymbol{\beta}}_{ML})}{n}, \qquad (5)$$

و يمكن إيجاد تقدير $\boldsymbol{\psi}$ من خلال تعظيم دالة الإمكان بعد ادخال مقدرات ML لـ $\boldsymbol{\beta}$ و σ^2 كالاتي:

$$\widehat{\boldsymbol{\psi}}_{ML} = arg \max_{\psi}[p(y|\boldsymbol{\beta}_{ML}, \sigma^2_{ML}, \boldsymbol{\psi})]. \qquad (6)$$

سيكون توزيع مخرجات $f(\mathbf{x})$ في ضوء المشاهدات والمعلمات كالاتي:

$$p\big(f(\mathbf{x})|\widehat{\boldsymbol{\beta}}_{ML}, \hat{\sigma}^2_{ML}, \widehat{\boldsymbol{\psi}}_{ML}, y\big) = \mathcal{N}\big(m_{ML}(\mathbf{x}), \hat{\sigma}^2_{ML} u_{ML}(\mathbf{x}, \mathbf{x})\big), \qquad (7)$$

حيث

$$m_{ML}(\mathbf{x}) = h(\mathbf{x})^T \widehat{\boldsymbol{\beta}}_{ML} + \boldsymbol{k}(\mathbf{x})^T A^{-1}\big(y - H\widehat{\boldsymbol{\beta}}_{ML}\big),$$

$$u_{ML}(\mathbf{x}, \mathbf{x}) = k(\mathbf{x}, \mathbf{x}) - \boldsymbol{k}(\mathbf{x})^T A^{-1} \boldsymbol{k}(\mathbf{x}).$$

لاحظ أن A و $k(\mathbf{x}, \mathbf{x})$ و $\boldsymbol{k}(\mathbf{x})$ هي دوال لـ $\widehat{\boldsymbol{\psi}}_{ML}$.

3.2 طريقة الإمكان الأعظم المقيد (Restricted Maximum Likelihood)

وفقا لـ [14] ، يمكن الحصول على دالة الإمكان الأعظم المقيدة (REML) من دالة الإمكان الأعظم، عن طريق تكامل $\boldsymbol{\beta}$ باستخدام التوزيع المنتظم الاولي (أي $p(\boldsymbol{\beta}) \propto constant$). التعبير الناتج هو:

$$p(y|\sigma^2, \boldsymbol{\psi}) \propto \frac{|A|^{-1/2}|H'A^{-1}H|^{-1/2}}{(2\pi\sigma^2)^{\frac{n-q}{2}}} exp\left\{-\frac{1}{2\sigma^2}\big(y - H\widehat{\boldsymbol{\beta}}\big)' A^{-1}\big(y - H\widehat{\boldsymbol{\beta}}\big)\right\}, \qquad (8)$$

قيمة σ^2 التي تعظم العلاقة أعلاه هي

$$\hat{\sigma}^2_{RL} = \frac{(y - H\widehat{\boldsymbol{\beta}})^T A^{-1}(y - H\widehat{\boldsymbol{\beta}})}{n - q}, \qquad (9)$$

ثم يعطى تقدير $\boldsymbol{\psi}$ بالعلاقة:

$$\widehat{\boldsymbol{\psi}}_{RL} = arg \max_{\psi}[p(y|\hat{\sigma}^2_{RL}, \boldsymbol{\psi})], \qquad (10)$$

في ما سبق نُعرف أيضاً

$$\widehat{\boldsymbol{\beta}} = (H^T A^{-1} H)^{-1} H^T A^{-1} y. \qquad (11)$$

لا يمكن اعتبار $\widehat{\boldsymbol{\beta}}$ تقديرا لـ $\boldsymbol{\beta}$، لأنه تم تكامل $\boldsymbol{\beta}$. بإمكاننا استخدام أي رمز للعلاقة أعلاه، ولكن لأنها مطابقة لـ$\widehat{\boldsymbol{\beta}}_{ML}$ ندعوها $\widehat{\boldsymbol{\beta}}$ (ولكن ليس $\widehat{\boldsymbol{\beta}}_{\mathrm{RL}}$).

وبالتالي سيكون توزيع مخرجات $f(\mathbf{x})$ مشروطة على المشاهدات والمعلمات كالاتي:

$$p\big(f(\mathbf{x})\big|y, \hat{\sigma}^2_{RL}, \widehat{\boldsymbol{\delta}}_{RL}\big) = \mathcal{N}\big(m_{RL}(\mathbf{x}), \hat{\sigma}^2_{RL} u_{RL}(\mathbf{x}, \mathbf{x})\big), \qquad (12)$$

حيث

$$m_{RL}(\mathbf{x}) = h(\mathbf{x})^{\mathrm{T}}\widehat{\boldsymbol{\beta}} + \boldsymbol{k}(\mathbf{x})^{\mathrm{T}}A^{-1}\big(y - H\widehat{\boldsymbol{\beta}}\big),$$

$$u_{RL}(\mathbf{x},\mathbf{x}) = k(\mathbf{x},\mathbf{x}) - \boldsymbol{k}(\mathbf{x})^{\mathrm{T}}A^{-1}\boldsymbol{k}(\mathbf{x}) + (h(\mathbf{x})^{\mathrm{T}} - \boldsymbol{k}(\mathbf{x})^{\mathrm{T}}A^{-1}H)(H^{\mathrm{T}}A^{-1}H)^{-1}(h(\mathbf{x})^{\mathrm{T}} - \boldsymbol{k}(\mathbf{x})^{\mathrm{T}}A^{-1}H)^{\mathrm{T}}.$$

لاحظ أن $m_{\mathrm{RL}}(\mathbf{x}) = m_{\mathrm{ML}}(\mathbf{x})$ والحدود الموجودة في السطر الأول من $u_{RL}(\mathbf{x},\mathbf{x})$ هي نفس المصطلحات الموجودة في$u_{\mathrm{ML}}(\mathbf{x},\mathbf{x})$ ، باستثناء أن A و$k(\mathbf{x},\mathbf{x})$ و$\boldsymbol{k}(\mathbf{x})$ يتم حسابها الآن باستخدام $\widehat{\boldsymbol{\psi}}_{\mathrm{RL}}$ بدلاً من $\widehat{\boldsymbol{\psi}}_{\mathrm{ML}}$. بخلاف ذلك، يمكن اعتبار $u_{RL}(\mathrm{x},\mathrm{x})$ هي $u_{\mathrm{ML}}(\mathbf{x},\mathbf{x})$ معززًا بتعبير السطر الأخير في المعادلة أعلاه، والذي يمثل الزيادة في التباين، تبعا لعدم اليقين حول قيمة $\boldsymbol{\beta}$. ومع ذلك، كلا التنبؤين لمخرجات $f(\mathbf{x})$ لهما توزيع كاوس.

4. توليد نقاط التصميم

في هذا القسم، نقدم تفاصيل التصميم المستخدم في توليد النقاط. في هذه الدراسة، استخدمنا تصميم المكعب الزائدي اللاتيني الاقصى (Maximin Latin Hypercube Design) والذي يشير اليه اختصارا (Maximin LHD). إن Maximin LHD هو أحد تصاميم ملأ الفراغ ويعتمد على قياس المسافة بين نقاط التصميم، نقوم أولاً بحساب الحد الأدنى للمسافة بين نقاط التصميم. بعد ذلك، يعملMaximin LHD على تعظيم الحد الأدنى للمسافة بين أية نقطتين من نقاط التصميم بحيث لا تكون أية نقطتين من نقاط التصميم قريبة جداً من بعضها. يتم إعطاء Maximin LHD بواسطة

$$\mathbf{X} = \max_{\mathbf{X}\subset\Omega}\ \min_{\{x,x'\}\in X} dis\big(x_i, x_j\big). \qquad (13)$$

لذلك، يضمن Maximin LHD أن نقاط التصميم منتشرة بشكل متساوي وتغطي مجال الإدخال بالكامل[15].

5. مثال دالة البئر

يتم التطبيق عملياً لما تمت دراسته نظرياً في الفقرات السابقة من خلال أخذ مثال حقيقي متمثل بدالة البئر ذي ثمان أبعاد، تم استخدام دالة البئر Borehole Functionعلى نطاق واسع مثالاً على الأنموذج الحاسوبي. أنموذج البئر استخدمه [17] ,[16] لتمثيل معدل تدفق المياه أسفل البئر. يتم تعريف دالة البئر على النحو الآتي:

$$f(\mathbf{x}) = \frac{2\pi T_u(H_u - H_l)}{\ln\left(\frac{r}{r_w}\right)\left(1 + \frac{2LT_u}{\ln\left(\frac{r}{r_w}\right)r_w^2K_w} + \frac{T_u}{T_l}\right)}. \qquad (14)$$

تحتوي دالة البئر على ثمانية متغيرات إدخال وهي $\mathbf{x} = (r_w, r, T_u, H_u, T_l, H_l, L, K_w)$، متغير الاستجابة $f(\mathbf{x})$ يمثل معدل تدفق المياه في حفرة البئر بالغالون في متر مكعب لكل سنة. متغيرات الإدخال ومجالها هي:

1. نصف قطر البئر ويرمز له $\mathrm{r_w}$ ضمن الفترة $[0\cdot05, 0\cdot15]$ ووحدة قياسه المتر (m).
2. نصف قطر التأثير ويرمز له r ضمن الفترة $[100, 50000]$ ووحدة قياسه المتر (m).
3. أعلى نفاذية الخزان الجوفي ويرمز له T_u ضمن الفترة $[63070, 115600]$ ووحدة قياسه متر مربع لكل سنة $(m^2/year)$.
4. رأس الخزان الجوفي العلوي ويرمز له H_u ضمن الفترة $[990, 1110]$ ووحدة قياسه المتر (m).
5. نفاذية الخزان الجوفي السفلي ويرمز له T_l ضمن الفترة $[63\cdot1, 116]$ ووحدة قياسه متر مربع لكل سنة $(m^2/year)$.
6. أسفل رأس خزان المياه الجوفية ويرمز له H_l ضمن الفترة $[700, 820]$ ووحدة قياسه المتر (m).
7. طول البئر ويرمز له L ضمن الفترة $[1120, 1680]$ ووحدة قياسه المتر (m).
8. التوصيل الهيدروليكي للبئر ويرمز له K_w ضمن الفترة $[9855, 12045]$ ووحدة قياسه متر لكل سنة $(m/year)$.

الآن، نقوم بمقارنة أداء أنموذج انحدار عملية كاوس لدالة البئر والتحقيق فيها بدقة باستخدام الطرق المختلفة لتقدير المعلمات، طريقة MLE وطريقة RMLE. تتم هذه المقارنة من خلال بعض مقاييس التحقق من الصحة. استخدمنا مجاميع مختلفة من أحجام نقاط التصميم $(n = 5p, 7.5p, 10p, 15p, 20p)$، حيث p يمثل عدد متغيرات الإدخال. تم ضبط حجم نقاط التحقق على $3p$. تم توليد نقاط التصميم والاختبار في البداية بواسطة (Maximin LHD).

6. مقاييس التحقق من صحة الأنموذج

في هذا القسم يتم دراسة تأثير عدد نقاط التصميم على دقة تقدير معلمات الأنموذج باستخدام عدد من المقاييس للمقارنة بين أنموذج انحدار عملية كاوس بطريقة MLE وأنموذج انحدار عملية كاوس بطريقةREML . من خلال أخذ المثال الحقيقي السابق المتمثل بدالة البئر.

6.1 مسافة ماهالانوبيس (Mahalanobis Distance)

في هذا القسم نقوم بدراسة تأثير عدد نقاط التصميم على تقدير المعلمة باستخدام مسافة المهالانوبيس. يوفر هذا المقياس طريقة فعالة للمقارنة بين مخرجات أنموذج انحدار عملية كاوس ومخرجات الأنموذج الحاسوبي، معبراً عنها كقيمة واحدة. ويتم حسابها بواسطة

$$M_{MD} = (y^* - E[f(\mathbf{x}^*)|y])^T\ (V[f(\mathbf{x}^*)|y])^{-1}(y^* - E[f(\mathbf{x}^*)|y]). \qquad (15)$$

إذ إن$E[f(\mathbf{x}^*)|y]$ تمثل عناصر متجه الوسط التنبؤي و $V[f(\mathbf{x}^*)|y]$ تمثل مصفوفة التغاير التنبؤية لأنموذج انحدار عملية كاوس.

إنّ M_{MD} لها توزيع F مع درجات حرية m و $n-q$ إذ $q=p+1$ [10].

$$\frac{(n-q)}{m(n-q-2)}M_{MD}|y,\boldsymbol{\psi}\sim F_{m,n-q}. \quad (16)$$

إذا كانت قيمة M_{MD} متطرفة (كبيرة أو صغيرة)، فهذا يدل على وجود تعارض بين الأنموذج الحاسوبي وأنموذج انحدار عملية كاوس. اذا ظهرت مثل هذه المشكلة فمن المهم استكشاف الأخطاء المعيارية الفردية من اجل تحديد سبب المشكلة.

تم حساب قيم مسافة مهالانوبيس بحسب المعادلة (15)، لأنموذج انحدار عملية كاوس الذي تم بناؤه على أساس طريقة MLE وRMLE. وتم حساب قيمة ماهالانوبيس لقيم المختلفة لنقاط التصميم ونقاط الاختبار.

حيث يوضح الجدول (1) قيم مسافة مهالانوبيس لأنموذج انحدار عملية كاوس الذي يعتمد على طريقة MLE ولأنموذج انحدار عملية كاوس الذي يعتمد على طريقة RMLE لقيم مختلفة من نقاط التصميم مع 24 نقطة اختبار.

جدول (1): مسافة ماهالانوبيس المحسوبة لأنموذج انحدار عملية كاوس لدالة البئر حسب طريقة MLE وRMLE لقيم مختلفة من نقاط التصميم مع 24 نقطة اختبار.

Expected Mahalanobis distance	RMLE	MLE	n
24	71.22431	63.59114	40
24	32.42809	31.99704	60
24	24.56113	22.03833	80
24	25.23676	20.05221	120
24	25.06764	23.21811	160

يتبين من الجدول (1) انه عندما تكون نقاط التصميم خمسة اضعاف متغيرات الإدخال اي $n=40$، قيم ماهالانوبيس للطريقتين تكون كبيرة وبعيدة جدا عن القيمة المتوقعة لها وهي 24. كخطوة لاحقة من عملية التحقق تمت الزيادة في عدد نقاط التصميم تدريجيا حيث أخذت في البداية $n=60$، حيث نلاحظ أنه بعد حساب مسافة ماهالانوبيس للطريقتين اصبحت تقترب من القيمة المتوقعة لها وبزيادة نقاط التصميم الى $n=80$ اصبحت قريبة جدا من 24 للطريقتين وعند أخذ $n=120$ و$n=160$ نلاحظ أنّ هناك تغييراً طفيفاً في قيم مسافة ماهالانوبيس بالنسبة للطريقتين.

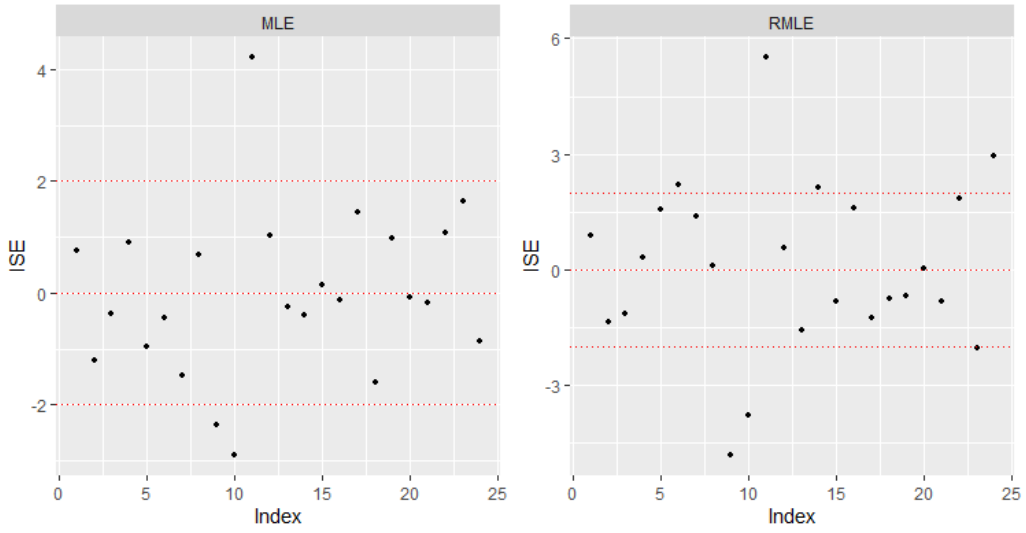
6.2 الأخطاء المعيارية الفردية (Individual Standardized Errors)

المقياس الثاني هو الأخطاء المعيارية الفردية (ISE) التي تعطى بواسطة:

$$ISE=\frac{f(D^*)_i-E[f(\mathbf{x}_i^*)|f(D)]}{\sqrt{V(f(\mathbf{x}_i^*)|f(D)}} \quad (17)$$

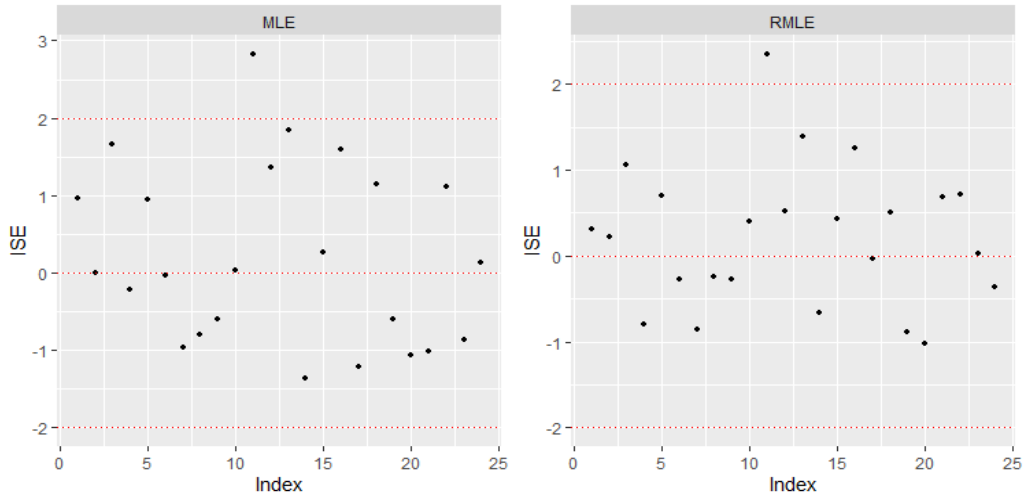
يجب أن تتراوح الأخطاء المعيارية الفردية بين -2 و2 إذا كان أنموذج انحدار عملية كاوس صالحًا [10]. وبالتالي، إذا تم العثور على قيم كبيرة لـ ISE، فإن تنبؤات أنموذج انحدار عملية كاوس لن تكون دقيقة وبالتالي فإن أنموذج انحدار عملية كاوس لن يكون صالحًا. لا تعتمد ISE فقط على الاختلافات بين تنبؤات أنموذج انحدار عملية كاوس والقيم الحقيقية للأنموذج الحاسوبي، ولكنها تعتمد أيضًا على عدم اليقين في تنبؤات أنموذج انحدار عملية كاوس. وذلك لأن الاختلافات بين تنبؤات أنموذج انحدار عملية كاوس والقيم الحقيقية للأنموذج الحاسوبي يتم جعلها معيارية بواسطة الجذر التربيعي لتباين أنموذج انحدار عملية كاوس.

تم حساب الأخطاء المعيارية الفردية بحسب المعادلة (17). وبالتالي، فإننا نعتبر مخطط ISE مقابل ترتيب بيانات الاختبار مقياس لمقارنة أداء أنموذج انحدار عملية كاوس الذي يعتمد على طريقتين مختلفتين لتقدير المعلمات. حيث يوضح الشكل (1) قيم ISE مقابل ترتيب بيانات الاختبار لأنموذج انحدار عملية كاوس الذي يعتمد على طريقة MLE وأنموذج انحدار عملية كاوس الذي يعتمد على طريقة RMLE مع 40 نقطة التصميم و24 نقطة اختبار.



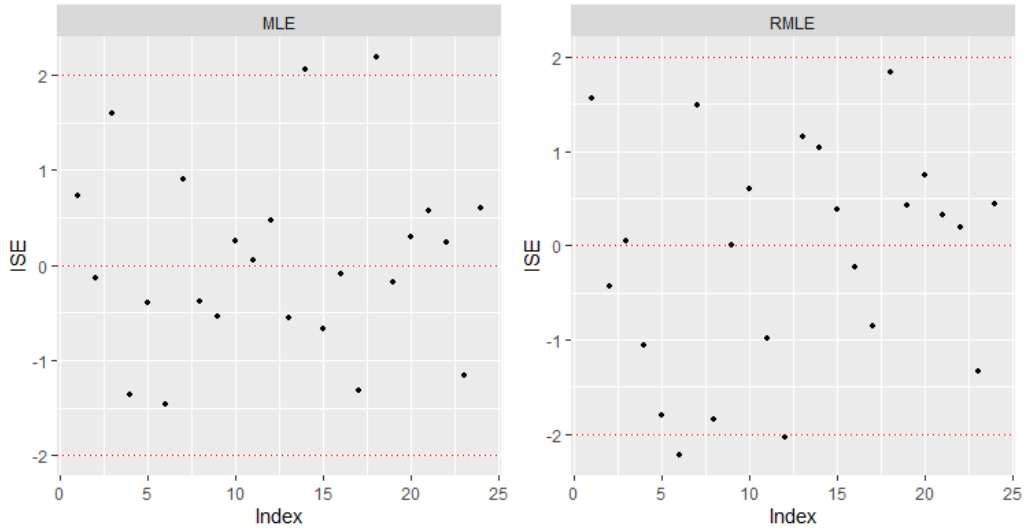
الشكل (1): الأخطاء المعيارية الفردية مقابل ترتيب بيانات الاختبار لـ 40 نقطة تصميم.

يُظهر الشكل (1) ان هناك قسماً من الأخطاء تقع خارج الفترة $[-2,2]$ بالنسبة للطريقتين و هناك قيم منها متطرفة مما يشير الى وجود تعارض بين الأنموذج الحاسوبي وأنموذج عملية كاوس.
وقمنا أيضاً بزيادة الحجم المحدد لنقاط التصميم ليصبح 60 وتم تكرار العمليات مرة أخرى. يوضح الشكل (2) مخططات ISE لأنموذج انحدار عملية كاوس الذي يعتمد على طريقة MLE ولأنموذج انحدار عملية كاوس الذي يعتمد على طريقة RMLE مع 80 نقطة تصميم ومجموعة من 24 نقطة اختبار.



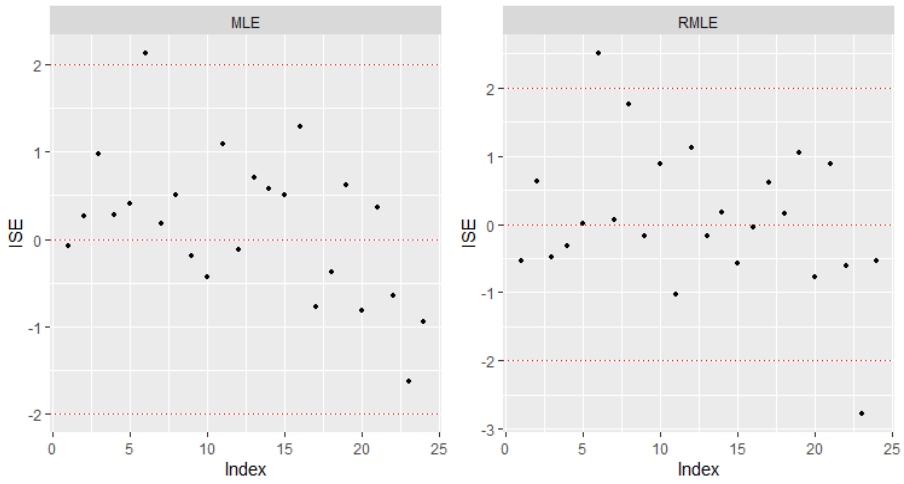
الشكل (2): الأخطاء المعيارية الفردية مقابل ترتيب بيانات الاختبار لـ 60 نقطة تصميم.

نلاحظ من الشكل (2) لأنموذج انحدار عملية كاوس الذي يعتمد على طريقة MLE، أن غالبية ISE صغيرة وأن هناك خطأ واحداً فقط يقع خارج الحدود، ولكنه بعيد من الحدود. بينما أنموذج انحدار عملية كاوس الذي يعتمد على طريقة RMLE يُظهر ان هناك خطأ واحداً فقط يقع خارج الحدود ولكنه قريب منها.
تم زيادة الحجم المحدد لنقاط التصميم ليصبح 80 وتم تكرار العمليات مرة أخرى. يوضح الشكل (3) مخططات ISE لأنموذج انحدار عملية كاوس الذي يعتمد على طريقة MLE ولأنموذج انحدار عملية كاوس الذي يعتمد على طريقة RMLE مع 80 نقطة تصميم ومجموعة من 24 نقطة اختبار.

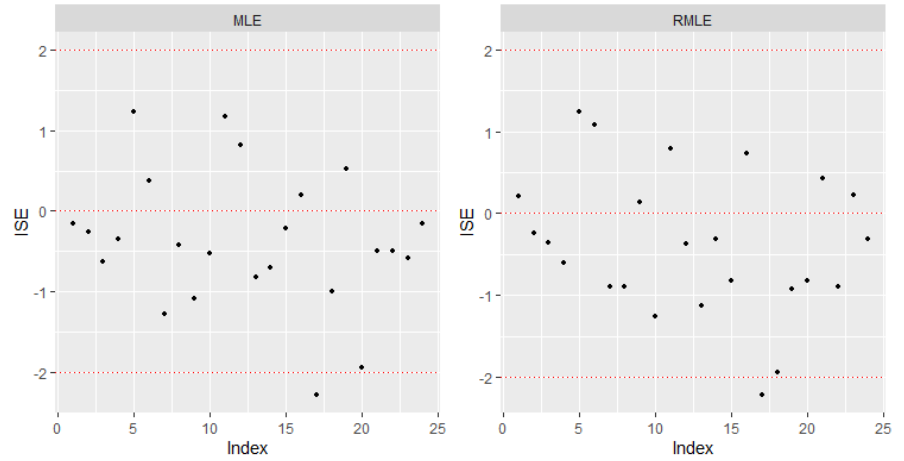


الشكل (3): الأخطاء المعيارية الفردية مقابل ترتيب بيانات الاختبار لـ 80 نقطة تصميم.

من الشكل (3) يمكن ملاحظة ان ISE لأنموذج انحدار عملية كاوس الذي يعتمد على طريقة RMLE و طريقة MLE أصبح الآن صغيرًا مع وجود اثنين فقط من ISE خارج الحدود. ومع ذلك، فإن هذين الخطأين قريبان جدًا من الحدود.
كخطوة لاحقة من عملية التحقق تمت الزيادة في عدد نقاط التصميم تدريجياً حيث تم أخذ $n = 120$ و$n = 160$ كما هو موضح في الشكل (4) والشكل (5):



الشكل (4): الأخطاء المعيارية الفردية مقابل ترتيب بيانات الاختبار لـ120 نقطة تصميم.



الشكل (5): الأخطاء المعيارية الفردية مقابل ترتيب بيانات الاختبار لـ160 نقطة تصميم.

مما تم عرضه في الشكل (4 و5) تبين أن طريقة MLE تتحسن عند زيادة نقاط التصميم حيث يوجد خطأ واحداً فقط من ISE خارج الحدود. ومع ذلك، فإن هذا الخطأ قريب جدًا من الحدود.

يمكننا أن نستنتج من مخططات ISE أن أنموذج انحدار عملية كاوس الذي يعتمد على طريقة MLE قد لا يعمل بشكل جيد مع عدد صغير من نقاط التصميم. قد يصبح أداء أنموذج انحدار عملية كاوس المعتمد على طريقة MLE جيدًا مثل أداء أنموذج انحدار عملية كاوس المعتمد على طريقة RMLE عند استخدام حجماً أكبر من نقاط التصميم.

7.الاستنتاجات

في هذا البحث، تم تقديم طريقة إحصائية شائعة، تسمى أنموذج انحدار عملية كاوس، لتقريب وتحليل النماذج الحاسوبية المعقدة. حيث قمنا بمقارنة أداء أنموذج انحدار عملية كاوس لدالة البئر والتحقيق فيها بدقة باستخدام الطرق المختلفة لتقدير المعلمات، طريقة MLE وطريقة RMLE. حيث يعد تقدير قيم هذه المعلمات هي الخطوة الرئيسة لتحسين أنموذج انحدار عملية كاوس. قمنا ايضا بدراسة تأثير عدد نقاط التصميم في اعطاء تقدير مناسب للمعلمات، إذ إن دقة التقدير تعتمد بدرجة كبيرة على أبعاد المدخلات. وقد استنتجنا ان تقدير معلمات أنموذج انحدار عملية كاوس ليس موثوقا للغاية مع مجموعة صغيرة من نقاط التصميم، وعليه يجب ان يكون عدد نقاط التصميم لبناء أنموذج انحدار عملية كاوس على الأقل عشرة اضعاف بُعد الإدخال لكي تكون التنبؤات باستخدام هذا الأنموذج دقيقة. تم ايضا استخدام مقياس مسافة مهالانوبيس ومقياس الأخطاء المعيارية الفردية للمقارنة بين MLE و RMLE، حيث اظهرت قيم مسافة مهالانوبيس في الجدول (1)، ان طريقة RMLE تعطي تقديرات افضل وقريبة من القيمة المتوقعة لها، كما انها تحتاج الى نقاط تصميم أقل الى حد ما لتعطي تقديرات جيدة للمعلمات، في حين كانت طريقة MLE تتحسن كلما استخدمنا حجماً اكبر من النقاط. كما تبين من الأخطاء المعيارية الفردية، ان اغلب الأخطاء الفردية المعيارية حسب طريقة RMLE كانت ضمن المدى

المطلوب أو قريبة من الحدود، بينما اظهرت الأخطاء أنّ اداء MLE يصبح مثل اداء RMLE عند زيادة نقاط التصميم. وفقا للنتائج المذكورة فأن طريقة RMLE استغرقت أقل قدر من الوقت لبناء الأنموذج ومن ثم فأن لدى RMLE أفضل اداء. يمكن ان تساعد طرق اخرى للتحقق مثل (تجزئة كوليسكي، تجزئة مصفوفة التباين، رسومات Q-Q ،...الخ) في تحسين دقة وكفاءة الطرق الحالية بشكل اكبر خاصة عندما تكون عدد البيانات قليلة. كذلك يمكن استخدام طرق اخرى للتقدير ومقارنتها مع الطرق المذكورة مثل طريقة (خوارزميات MCMC ,طريقة المقاطع CV, طريقة المربعات الصغرى,...الخ).

8. شكر وتقدير (Acknowledgements)
الباحثون ممتنون جدًا لجامعة الموصل/كلية التربية للعلوم الصرفة- قسم الرياضيات على التسهيلات التي قدمتها، والتي ساعدت في تحسين جودة هذا البحث.

9. References

[1] H. Mohammadi, P. Challenor, M. Goodfellow, and D. Williamson, “Emulating computer models with step-discontinuous outputs using gaussian processes,” ArXiv Prepr. ArXiv190302071, 2019.

[2] J. Sacks, W. J. Welch, T. J. Mitchell, and H. P. Wynn, “Design and Analysis of Computer Experiments,” Stat. Sci., vol. 4, no. 4, pp. 409–435, 1989, doi: 10.1214/ss/1177012413.

[3] M. C. Kennedy and A. O’Hagan, “Bayesian Calibration of Computer Models,” J. R. Stat. Soc. Ser. B Stat. Methodol., vol. 63, no. 3, pp. 425–464, Sep. 2001, doi: 10.1111/1467-9868.00294.

[4] J. E. Oakley and A. O’Hagan, “Probabilistic Sensitivity Analysis of Complex Models: A Bayesian Approach,” J. R. Stat. Soc. Ser. B Stat. Methodol., vol. 66, no. 3, pp. 751–769, Aug. 2004, doi: 10.1111/j.1467-9868.2004.05304.x.

[5] R. Paulo, “Default priors for Gaussian processes,” Ann. Stat., vol. 33, no. 2, pp. 556–582, Apr. 2005, doi: 10.1214/009053604000001264.

[6] R. B. Christianson, R. M. Pollyea, and R. B. Gramacy, “Traditional kriging versus modern Gaussian processes for large-scale mining data,” Stat. Anal. Data Min. ASA Data Sci. J., vol. 16, no. 5, pp. 488–506, 2023.

[7] V. Volodina and D. B. Williamson, “Nonstationary Gaussian Process Emulators with Kernel Mixtures,” ArXiv E-Prints, 2018.

[8] A. Marrel and B. Iooss, “Probabilistic surrogate modeling by Gaussian process: A review on recent insights in estimation and validation,” Reliab. Eng. Syst. Saf., p. 110094, 2024.

[9] P. D. Hoff, A first course in Bayesian statistical methods, vol. 580. Springer, 2009.

[10] L. S. Bastos and A. O’hagan, “Diagnostics for gaussian process emulators,” Technometrics, vol. 51, no. 4, pp. 425–438, 2009, doi: 10.1198/TECH.2009.08019.

[11] J. R. Stroud, M. L. Stein, and S. Lysen, “Bayesian and maximum likelihood estimation for Gaussian processes on an incomplete lattice,” J. Comput. Graph. Stat., vol. 26, no. 1, pp. 108–120, 2017.

[12] S. Ameli and S. C. Shadden, “Noise estimation in Gaussian process regression,” ArXiv Prepr. ArXiv220609976, 2022.

[13] C. Lataniotis, S. Marelli, and B. Sudret, “The Gaussian process modelling module in UQLab,” ArXiv Prepr. ArXiv170909382, 2017.

[14] D. A. HARVILLE, “Bayesian inference for variance components using only error contrasts,” Biometrika, vol. 61, no. 2, pp. 383–385, 1974, doi: 10.1093/biomet/61.2.383.

[15] M. D. Morris and T. J. Mitchell, “Exploratory designs for computational experiments,” J. Stat. Plan. Inference, vol. 43, no. 3, pp. 381–402, Feb. 1995, doi: 10.1016/0378-3758(94)00035-T.

[16] B. A. Worley, “Deterministic uncertainty analysis,” Oak Ridge National Lab.(ORNL), Oak Ridge, TN (United States), 1987.

[17] S. Xiong, P. Z. Qian, and C. J. Wu, “Sequential design and analysis of high-accuracy and low-accuracy computer codes,” Technometrics, vol. 55, no. 1, pp. 37–46, 2013.

المقارنة بين تقدير معلمات أنموذج انحدار عملية كاوس باستخدام طريقتي الإمكان الأعظم والإمكان الأعظم المقيد

امنة خالد الياس خضر*[(1)] يونس حازم الطويل [(2)]

[2,1] الرياضيات، كلية التربية للعلوم الصرفة، جامعة الموصل، الموصل، العراق

المستخلص:

تُستخدم نماذج انحدار العملية الكاوسية كتمثيلات إحصائية بديلة للنماذج الحاسوبية المعقدة والمكلفة من الناحية الحسابية، نظرًا لمرونتها في التقاط شكل الدوال الملساء. إنّ أنموذج انحدار العملية الكاوسية له عدد من المعلمات غير المعلومة بشكل عام، والتي يعد تقديرها خطوة أساسية نحو بناء الأنموذج. المعلمات التي يتم النظر فيها هي معاملات الانحدار β، ومعلمة القياس σ^2 ومعلمات الارتباط ψ. إن تقدير هذه المعلمات هي المشكلة التي تناولناها في هذا البحث. المساهمة الرئيسة لهذا العمل هو التقييم والمقارنة بين تقدير معلمات أنموذج انحدار عملية كاوس باستخدام طريقة الإمكان الأعظم وطريقة الإمكان الأعظم المقيد، هذه المقارنة تمت بالاعتماد على بعض مقاييس التحقق من الصحة. نبحث بعد ذلك عن تأثير عدد نقاط التصميم على تقديرها. تم تطبيق أنموذج انحدار عملية كاوس والذي تم تقدير معلماته بالطريقتين اعلاه على مثال حقيقي ذي ثمان أبعاد متمثل بـ أنموذج دالة البئر، وتم تنفيذ العمليات الحسابية والبيانية كافة باستخدام برنامج (R).