مجلة جامعة النهرين العلوم المجلد ١٦ العدد(٣) ايلول، ٢٠١٣ ص 62-77

مقارنة بعض طرائق الانحدار اللامعلمي الجمعي

خلود يوسف خمو* و ساندي قيس**
*** كلية الادرة والاقتصاد، جامعة بغداد.**
**** قسم الاحصاء ، كلية الادارة والاقتصاد، جامعة بغداد.**

الخلاصة

في حالة غياب المعرفة عن الظاهرة كان تقوم التجربة لإول مرة او لا يمكن تحديد العلاقة السببية او السلوكية التي تربط المتغيرات فبدلا من اشتراط ان تأخذ البيانات قالباً او شكلاً دالياً يوصف الظاهرة بشكل مسبق يتم استبدال ذلك بإسلوب اكثر مرونة يدعى التحليل اللامعلمي.

ان التوسع في الشرائح التمهيدية من الحالة أحادية الى الحالة متعددة المتغيرات أظهر مشكلة البعدية Curse of Dimensionality اذ يجب التوسع في حالة الشرائح التمهيدية من التكعيبية Cubic Smoothing Spline الى الحالة Thin Plate Spline. كما ان أسلوب تحليل مثل هذا النوع من الشرائح صعب خاصة لو علم حد التفاعل والذي لا يمكن تمثيله بسهولة كما يحتاج الى مستوى عالي من التحليل والبرمجة ولهذا كانت فكرة النماذج الجمعية Additive Model وخصوصاً وان هناك بعض الخوارزميات التي تتجاوز مشكلة البعدية.

يهدف البحث الى التركيز على طرائق الانحدار اللامعلمي الجمعي (اي حالة ثنائية المتغيرات) بحيث نتلافى مشكلة البعدية وكانت الخوارزميات المستخدمة Backbiting و SIMEX التي تجمع بين المحاكاة والاستكمال ثم المفاضلة بين الخوارزميات باستعمال معايير موجودة واخرى مقترحة كمعيار MGCV وذلك باستخدام العديد من تجارب المحاكاة وتباينات وحجوم عينات مختلفة، إضافة لتطبيق الخوارزميات على بيانات واقعية عن تلوث المياه واعتماد الانموذج الذي يناسب بيانات المياه والذي يعطي اقل معيار مقارنة.

Keywords**:** Nonparametric, Additive Regression, Back fitting, SIMEX, MGCV Extrapolation.

١. المقدمة وهدف البحث

عندما تكون هناك معرفة معرفة مسبقة بالظاهرة المراد دراستها والبيانات غالباً ما تكون كمية كأن تكون دالة سببية او سلوكية تصف مثلاً علاقة الدخل بالانفاق أو التغير في الانتاجية بعد أستعمال نوع معين من الاسمدة أي تكوين صيغة رياضية والتعويل على الشكل من خلال بعض المؤشرات التي من الممكن ان تلخص الشكل الدالي للظاهرة وتدعى المؤشرات بالمعلمات وينصب العمل الاحصائي على هذه المعلمات بافتراض ثبات الظروف الاخرى المحيطة بالظاهرة ويدعى التحليل المعلمي. أما في حالة غياب المعرفة عن الظاهرة كآن تقام التجربة لإول مرة أو لايمكن تحديد العلاقة السببية او السلوكية التي تربط المتغيرات فبدلاً من اشتراط ان تاحذ البيانات قالباً أو شكلا دالياً يوصف الظاهرة بشكل مسبق يتم استبدال ذلك باسلوب اكثر مرونة يدعى التحليل اللامعلمي.

وقد برز استعمال طرائق التمهيد كاداة فعالة تعول على البيانات بدلا من الصيفة الدالية المسبقة للبيانات وان هدف التمهيد هو تقريب دالة الانحدار اللامعلمية التقريبية الى دالة الانحدار اللامعلمية الحقيقية.

واحد انواع الممهدات هي الشرائح وابرزها الشرائح التمهيدية Smoothing Spine اذ تقوم بتقسيم البيانات الى مجاميع ومطابقة كل مجموعة بشكل على حدة ثم تمهيد الاجزاء وربطها مع بعضها عن طريق ما يدعى بالمعلمة التمهيدية.

ان التوسع في الشرائح التمهيدية من حالة أحادي الى متعدد المتغيرات اظهر مشكلة البعدية ففي حالة $p = 2$ يجب التوسع من حالة الشرائح التمهيدية التكعيبية الى حالة Thin

Plate Spline الا ان أسلوب تحليل هذا النوع من الشرائح يبدو امر بالغ الصعوبة خاصة لوعلم ان حد التفاعل لا يمكن تمثيله بسهولة في الظاهرة الا في حالة استعمال نماذج Smoothing Spine ANOVA والتي تحتاج الى مستوى عالي من التحليل العددي والبرمجة العالية ومن الحلول لتجاهل حد التفاعل بين المتغيرات والتركيز على التأثيرات الرئيسية كانت فكرة النماذج الجمعية Additive Models وخصوصا وان هناك بعض الخوارزميات ذو اثر لتجاوز مشكلة البعدية.

يهدف البحث الى التركيز على طرائق الانحدار اللامعلمي الجمعي (اي حالة ثنائية المتغيرات) بحيث نتلافى مشكلة البعدية وكانت الخوارزميات المستخدمة Backbiting و SIMEX التي تجمع بين المحاكاة والاستكمال مع محاولة تطوير الخوارزمية الاخيرة باعتماد ممهدات مقترحة من دوال Kernel ثم المفاضلة بين الخوارزميات باستعمال معايير موجودة واخرى مقترحة كمعيار MGCV وذلك باستخدام العديد من تجارب المحاكاة وتباينات وحجوم عينات مختلفة إضافة لتطبيق الخوارزميات على بيانات واقعية عن تلوث المياه واعتماد الانموذج الذي يناسب بيانات المياه والذي يعطي اقل معيار مقارنة.

وقد كانت هناك العديد من الانجازات للباحثين في مجال الانحدار اللامعلمي الجمعي على سبيل المثال لا الحصر ما قدمه عام 1999 كل من .Carroll, R.J الانحدار اللامعلمي بوجود اخطاء القياس حيث في معظم تطبيقات الانحدار يقاس المتغير المعتمد مع اخطاء وافترض الباحثين اولاً طريقة SIMEX واشتقوا نظرية تقاربية لانحدار Kernel وكذلك افترضوا شرائح الانحدار الجزئية ولعدد ثابت من العقد، الطريقة الثانية تفترض تنبؤ الخطأ له توزيع طبيعي خليط مع عدد غير معروف من المكونات.

وفي عام ٢٠٠٨ قدم كل من .Ruppert, D واخرون النماذج الخطية الجمعية الجزئية عندما التغاير مقاس مع الخطأ وافترضوا طرائق تقدير المعلمات هي SIMEX وناقشوا التقارب لطريقة SIMEX المقترحة.

٢. الجانب النظري

٢,١ الانموذج الجمعي Additive Model [2,3,4,6]:

ان دراسة الانحدار المتعدد بشكل عام والانحدار ثنائي المتغيرات ستفرز عنه مشكلة البعدية حيث عندما يتم إيجاد مقدار تقاربي لبعدين او اكثر يصعب عمل مصفوفة المتغيرات المقاسة بوحدات مختلفة وغيرها من المشاكل لذا كانت الحاجة لظهور الانموذج الجمعي AM كحل عملي حيث يقوم بصفة تجميعية حيث تساعد في سهولة تفسير الظواهر المختلفة.

الانموذج الجمعي كاداة لتحليل البيانات يزود بامتداد منطقي لانموذج الانحدار الخطي القياسي بواسطة السماح لدوال من التغايرات للممهد الاعتباطي. حيث تسمح خوارزمية المطابقة العكسية Back fitting باستخدام مجموعة من الادوات أو النماذج.

وعليه فان الانموذج الجمعي يكتب بالصيغة:

$$Y_i = c + \sum_{i=1}^{n} m_i(x_i) + \xi_i \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots (1)$$

2.1.1 خوارزمية Back fitting [8,9]:

ان نماذج الانحدار اللامعلمي هي صف واسع من النماذج المرنة، كما انها تسمح للباحثين بتحليل البيانات دوم معرفة شكل مفترض للعلاقة بين متغير الاسنجابة والتغايرات. لكن لسوء الحظ طرائق الانحدار اللامعلمي تصبح االكثر تعقيدا في حالة زيادة عدد التغايرات، وإمكانية تقدير ممكن للعلاقات لهذه الحالة من المشااكل مزود في الانموذج الجمعي Additive Model، والذي اقترح بشكل أولي من قبل Friedman & Stuetzle (1981) ثم اصبح اكثر انتشاراً من قبل Hastie & Tibshirani (1990).

الانموذج الجمعي يفترض دالة التوقع الشرطي للمتغير المعتمد Y والتي يمكن ان تكتب كمجموع الحدود الممهدة للتغايرات $x_1, x_2, \ldots, x_D$، كالآتي:

$$E(Y|X = (x_1,\ldots,x_D)) = m(x_1,\ldots,x_D)$$
$$= m_1(x_1) + \ldots + m_D(x_D) \quad \ldots\ldots\ldots\ldots\ldots (2)$$

ان فهم الخصائص النظرية للانموذج الجمعي يتلكىء عند مقارنة التطبيقات العملية ففي حالة الشرائح الجمعية Additive Splines النسبة المثلى للتقارب منجزة لمقدرات الانموذج الجمعي وهي مستقلة عن عدد التفايرات وايضاً افتراض طريقة العبور الشرعي Cross-Validation

لإختيار عدد العقد، ان بقية الباحثين مثل Wahba وآخرون برهنوا نتائج موجودة في سياق الشرائح الممهدة، والاكثر حداثة والشائع ما قدمه Nielsen & Linton (1995) حيث وصفوا إجراء المطابقة للنماذج الجمعية الثنائية بالاعتماد على الانحدار الخطي الموضعي والتكامل الحدي وبينوا الاجراء ينجز للانحدار الجمعي بنسبة تقارب $O_p(n^{-2/5})$ كما للمقدر الخطي الموضعي أحادي المتغيرات.

في حالة النماذج الجمعية ومن خلال المطابقة العكسية البحث النظري معقد تماماً بسبب المقدرات معرفة كحل للخوارزمية التكرارية وفقط في حالة التغايرين تعابير التقديرات تكون واضحة ومتوفرة.

Hastie & Tibshirani (1989) زودوا بشروط كافية والتي توكد مطابقة الخوارزمية العكسية هذه الشروط تكون متحققة فقط لشرائح الانحدار والحدود المعلمية ولكن ليس لانحدار Kernel او الاننحدار المتعدد الموضعي. لكن لسوء الحظ الانحدار المتعدد الموضعي موخراً تبين انه يمتلك العديد من الخصائص النظرية المرغوبة والخصائص العملية وتوافقه مع المطابقة العكسية والتي تكون شائعة لمطابقة النماذج الجمعية.

وبتوظيف نقطتين نظرية مهمة فيما يتعلق بالانموذج الجمعي الثنائي وفي سياق مقدرات المطابقة العكسية وباستخدام الانحدار متعدد الحدود الموضعي وهي:

١. ماهي الشروط الكافية التي تضمن التقارب في المطابقة العكسية ؟.

٢. ما هي الخصائص التقاربية للمقدرات؟.

لتوضيح عمل الخوارزمية نفرض ان (X_1,Z_1,Y_1), (X_2,Z_2,Y_2), …, (X_n,Z_n,Y_n) مجموعة متغيرات عشوائية مستقلة ولها نفس التوزيع ذات بعد $\square^3$ وبفرض الانموذج التالي:

$$Y = \alpha + m_1(X_i) + m_2(Z_i) + \varepsilon_i \quad \ldots\ldots\ldots\ldots (2)$$

حيث ان ε_i مستقل وله نفس التوزيع بمتوسط صفر وتباين σ^2. ولتاكيد مطابقة الدوال m_1 و m_2 تم تضمين الحد الثابت α وبفرض $E(m_1(X_i)) = E(m_2(Z_i)) = 0$

وبفرض $Y = (Y_1,Y_2,\ldots,Y_n)'$ وبشكل مشابه الى X و Z وهي متجهات للدوال الجمعية عند النقاط المشاهدة وكالآتي:

$m_1 = (m_1(X_1),m_1(X_2),\ldots,m_1(X_n))'$,
$m_2 = (m_2(Z_1),m_2(Z_2),\ldots,m_2(Z_n))'$

ولاي ثابت d والتي هي متجه القيم (d, d, ..., d) n وبفرض $S'_{1,x}, S'_{2,z}$ هي Kernel للانحدار المتعدد الموضعي عند X و Z ويمكن ان تكتب بالشكل:

$$\begin{aligned} S'_{1,x} &= e'_1(X'_x W_x X_x)^{-1} X'_x W_x \\ S'_{2,z} &= e'_2(X'_z W_z X_z)^{-1} X'_z W_z \end{aligned} \quad \ldots\ldots\ldots\ldots (3)$$

حيث $e' = (1,0)$ متجه وان:

$$W_x = \text{diag}\left\{\frac{1}{h_1}K\left(\frac{X_1 - x}{h_1}\right)\ldots\frac{1}{h_1}K\left(\frac{X_n - x}{h_1}\right)\right\}$$

$$W_z = \text{diag}\left\{\frac{1}{h_2}K\left(\frac{Z_2 - z}{h_2}\right)\ldots\frac{1}{h_1}K\left(\frac{Z_n - z}{h_2}\right)\right\}$$

ولبعض دالة K Kernel ولعرض حزمة h حيث:

$$X_x = \begin{bmatrix} 1 & (X_1 - x) & \cdots & (X_1 - x)^{p_1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & (X_n - x) & \cdots & (X_n - x)^{p_1} \end{bmatrix}$$

حيث p_1 رتبة المتعدد الموضعية لمطابقة m_1، وبفرض S_1, S_2 تمثل المصفوفات الممهدة حيث الصفوف تكافىء Kernel عند المشاهدات X, Z، بالتتابع. وبفرض:

$$S_1 = \begin{bmatrix} S'_{1,X_1} \\ \vdots \\ S'_{1,X_n} \end{bmatrix},\ S_2 = \begin{bmatrix} S'_{2,Z_1} \\ \vdots \\ S'_{2,Z_n} \end{bmatrix}$$

وعليه يمكن تعريف متجه المطابقة عند نقاط المشاهدات كالآتي:

$$\hat{m} = \hat{\alpha} + \hat{m}_1 + \hat{m}_2 \quad \ldots\ldots\ldots\ldots (4)$$

حيث $\hat{\alpha} = \overline{Y}$ و m_1 و m_2 هي حلول لمجموعة المعادلات التقديرية

$$\begin{bmatrix} I & S_1^* \\ S_2^* & I \end{bmatrix} \begin{bmatrix} \hat{m}_1 \\ \hat{m}_2 \end{bmatrix} = \begin{bmatrix} S_1^* \\ S_2^* \end{bmatrix} Y \quad \ldots\ldots\ldots (5)$$

حيث $S_1^* = (I - 11'/n)S_1$ و $S_2^* = (I - 11'/n)S_2$.

تعديل الممهدات ضرورية لناكيد وحدانية الحلول لتقدير المعادلات وهذا يتطلب جعل:

$$\sum_{i=1}^{n} m_1(X_i) = \sum_{i=1}^{n} m_2(Z_i) = 0$$

في الجانب العملي المعادلات التقديرية تحل بإستخدام خوارزمية المطابقة العكسية لكن في الحالة الثنائية انها تملك حل واضح، حيث:

$$\begin{aligned} \hat{m}_1 &= \left\{ I - (I - S_1^* S_2^*)^{-1} (I - S_1^*) \right\} Y \equiv W_1 Y \\ \hat{m}_2 &= \left\{ I - (I - S_2^* S_1^*)^{-1} (I - S_2^*) \right\} Y \equiv W_2 Y \end{aligned} \quad \ldots (6)$$

بشرط وجود المعكوس المقدر يصبح:

$$\hat{m} = \left\{ \frac{11'}{n} + 2I - (I - S_1^* S_2^*)^{-1} (I - S_1^*)(I - S_2^* S_1^*) (I - S_2^*) \right\} Y \equiv WY \quad \ldots\ldots\ldots (7)$$

2.1.2 خوارزمية SIMEX [5,7]:

ان غالبية التطبيقات العملية تتضمن نماذج الانحدار وتم اعتماد الاستكمال لمسالة التقدير بفرض ان X, U ثوابت، وبفرض البيانات المشاهدة $\{X_i\}_1^n$ هي بالشكل:

$$X_i = U_i + \sigma Z_i \quad \ldots\ldots\ldots (8)$$

حيث Z_i متغير عشوائي طبيعي قياسي مستقل عن U_i، σ تباين الخطأ، U الخطأ العشوائي والى $\lambda \geq 0$ تعرف:

$$X_i(\lambda) = X_i + \lambda^{1/2} \sigma Z_i \quad \ldots\ldots\ldots (9)$$

حيث $\{Z_i\}_1^n$ مستقلة عن $\{X_i, U_i\}_1^n$ والمتغيرات العشوائية مستقلة ولها نفس التوزيع الطبيعي القياسي وبتعريف:

$$\hat{\theta}(\lambda) = E(\hat{\theta}(\lambda) | \{X_i\}_i^n) \quad \ldots\ldots\ldots (10)$$

والى $\lambda \geq 0$ وباستعمال الاستكمال للخلف والى $\lambda \geq -1$، هذا يسبب مقدّر الاستكمال المحاكاة والذي يرمز له $\hat{\theta}_{SIMEX}$.

لتوضيح خوارزمية SIMEX نوضح تطبيقها لإنموذج مكونات التباين المالوف، حيث:

$$\hat{\theta}(\lambda) = \frac{1}{n-1} \sum_{i=1}^{n} \left[X_i(\lambda) - \overline{X}(\lambda) \right]^2 \quad \ldots\ldots\ldots (11)$$

وبالتعويض في (٩) تصبح المعادلة (11) بالشكل:

$$\hat{\theta}(\lambda) = \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i + \lambda^{1/2} \sigma Z_i - \overline{X} - \lambda^{1/2} \sigma \overline{Z} \right)^2 \quad \ldots\ldots\ldots (12)$$

أي ان:

$$\hat{\theta}(\lambda) = E(\hat{\theta}(\lambda) | \{X_i\}_i^n = S_X^2 + \lambda \sigma^2 \quad \ldots\ldots\ldots (13)$$

أي لحالة $\hat{\theta}(\lambda)$ خطية في $\lambda \geq 0$ والاستكمال الى $\lambda = -1$ ينتج:

$$\hat{\theta}_{SIMEX} = S_X^2 - \sigma^2 \quad \ldots\ldots\ldots (14)$$

وإن $\hat{\theta}_{SIMEX}$ مقدّر غير متحيز ومتسق لـ θ.

٢,٢ معايير المقارنة

هناك العديد من المعايير تقيس مقدار الجودة في تقدير دالة الانحدار اللامعلمية التي تم تناولها نظرياً، مع تنوع النماذج التي توظف في التقدير من تلك المعايير.

١. معيار AMSE

ويمثل معدل متوسط مربع الخطأ Average of Mean Square Error ويعطى بالصيغة:

$$AMSE = n^{-1}E\sum_{i=1}^{n}\left[m(x_i)-\hat{m}(x_i)\right]^2 \quad .....(15)$$

٢. معيار MSE

ويمثل متوسط مربع الخطأ Mean Square Error ويعطى بالصيغة:

$$MSE = n^{-1}\sum_{i=1}^{n}\left[m(x_i)-\hat{m}(x_i)\right]^2 \quad .....(16)$$

٣.معيار MGCV المقترح

ان من فوائد معيار GCV هو انه يستخدم في اتجاهين الاول اختيار المعلمة التمهيدية ولحالة الطرائق الشرائحية احادية المتغيرات مثل طريقة الجزاء غير الممهدة Roughness Penalty وطريقة تقليص الجزاء Penalized Shrinkage وطريقة الخطوات المتسلسلة Stepwise POLYMARS، حيث يتم الاعتماد على فكرة الطريقة CV وهو Leave-One-Out حيث تترك كل مرة مشاهدة واحدة خارجاً وخطواته:

i. حساب تقدير $\hat{\mathbf{m}}(\mathbf{x_i})$.

ii. بناء دالة بحذف مشاهدة خارجا في كل مرة.

iii.تعاد الخطوات i و iiلجميع المشاهدات.

اي في كل مرة تحذف مشاهدة ثم نختار المعلمة التمهيدية المقابلة لاصغر CV.

الاتجاه الثاني يستخدم المعيار للمقارنة بين طرائق تقديرات الانحدار اللامعلمي والمعيار يحقق نتائج اقل تحيزاً من معيار MSE وصيغته:

$$MGCV = \frac{\frac{1}{n}\sum_{i=1}^{n}\left\{y_i-\alpha-\sum_{j=1}^{p}\hat{g}_j(t_{ij})\right\}^2}{\left\{1-\left[1+\sum_{j=1}^{p}(trS_j-1)\right]\Big/n\right\}^2} \quad .....(17)$$

حيث ان S_j مصفوفة تمهيد.

المعيار المقترح يعتمد على طريقة تقليص الجزاء (Penalized Shrinkage) PS والتي اثبتت كفاءتها في البحوث اللامعلمية مقارنة مع الطرائق الشرائحية الاخرى عدا طرائق بيز اللامعلمية وطرائق بيز اللامعلمية الحصية والتي كانت الاكفىء دوماً.

اذ تم الاعتماد على معيار MGCV الذي يعتمد على طريقة PS وبذلك يكون معيار مفاضلة بين معايير المعتمدة في البحث وياخذ الصيغة التالية:

$$MGCV = \frac{\frac{1}{n}\sum_{i=1}^{n}\left\{y_i-\alpha-\sum_{j=1}^{p}\hat{g}_j(t_{ij})\right\}^2}{\left\{1-\left[1+\sum_{j=1}^{p}(n-tr(H))/n\right]\right\}^2} \quad ...(18)$$

حيث $H = X[X'X+\alpha D'D]^{-1}X'$ ايضاً تختار α المقابلة لإصغر $MGCV(\alpha)$ وان X مصفوفة التصميم ذات بعد $(n\times(k+p+1))$ و D مصفوفة قطرية ذات بعد $(k+p+1)\times(k+p+1)$ علما بان اول p + 1 من عناصرها اصفار و k من العناصر الباقية قيمتها واحد.

٣. الجانب التجريبي

٣,١ توليد المتغيرات والعينة المستخدمة:

تعرف المحاكاة عموماً بانها تقليد يحاكي الواقع العملي، حيث توظف نماذج ولعدد من الحالات الافتراضية لتكون نتائج التحليل اكثر واقعية وشمولاً وتعميماً، ومن اسباب العمل بالمحاكاة هو للتاكد من تحقق جانب تطبيقي موجود أصلاً أو عند تعذر الحصول على بيانات توفر معلومات دقيقة عن ظاهرة معينة، اوعندما يصعب الاثبات الرياضي النظري لبيان أفضلية طرائق تقدير معينة على حساب أخرى. وتبدي المحاكاة مرونة تجاه اختيار حجوم العينات العشوائية والقدرة العالية بالتنوع بإختيار الاخطاء العشوائية.

تم تنفيذ تجارب المحاكاة بإستخدام ثلاثة حجوم للعينات هي صغيرة (n = 25) ومتوسطة (n = 50) وكبيرة (n = 200) و (n = 400) مع تكرار ٥٠٠ لكل تجربة محاكاة ولحالة التوزيع الطبيعي للخطأ العشوائي وكما يلي:

١. توليد المتغيرات التوضيحية x_j's لتتوزع توزيع منتظم قياسي مستقل اي $x_j \sim U(0,1),\ j=1,2$ ويتم توليدها بإستعمال طريقة Box Muller كالاتي:

يتم توليد متغيرين عشوائيين U_1 و U_2 ليتوزعان توزيع منتظم قياسي U(0,1) ثم يتم تحويل هذين المتغيرين الى

متغيرين عشوائيين مستقلين Z_1, Z_2 ويتبع كل منهما التوزيع الطبيعي القياسي وفق الصيغة:

$$Z_1 = -2(\ln U_1)^{1/2} \cos(2\pi U_2)$$

$$Z_2 = -2(\ln U_1)^{1/2} \sin(2\pi U_2)$$

ولتحويل المتغيرات من التوزيع الطبيعي القياسي الى التوزيع الطبيعي بمتوسط صفر وتباين σ^2 يتم استعمال التحويل $U = 0 + \sigma^2 Z$.

٢. توليد الاخطاء العشوائية لتتوزع توزيع طبيعي بمتوسط صفر وتباين σ^2 اي $e_i \sim N(0,\sigma^2),\ i = 1,2,...,n$، وقد تم استعمال ثلاثة مستويات من التباين لكل دالة من الدوال التي ستذكر لاحقاً وهي:

تباين عالي (High Noise) حيث:

$$\sigma = 0.5 \times \text{function Range}$$

تباين متوسط (Medium Noise) حيث:

$$\sigma = 0.25 \times \text{function Range}$$

تباين واطىء (Low Noise) حيث:

$$\sigma = 0.1 \times \text{function Range}$$

حيث ان σ الانحراف المعياري للخطأ e.

٣. توليد المتغير المعتمد Y_i ويتم من خلال النماذج المستخدمة في تجارب المحاكاة وذلك باستخدام دوال الانحدار بدلالة المتغيرين التوضيحيين اللذان تم توضيحهما في الفقرة ١ مضافاً له الاخطاء العشوائية والتي تم توليدها في الفقرة ٢ ولكل انموذج من النماذج المدروسة.

3.2 الدوال المستخدمة في تجارب المحاكاة

تتنوع الدوال بتنوع الظواهر التي تمثلها، ولا يمكن تناول جميع انماط الدوال في الوقت نفسه لما تتطلبه من الوقت والجهد والمساحة وقد حاولنا ان تكون الدوال متنوعة لتلائم واغلب الحالات وتم إستقاء هذه االدوال من بحوث منشورة وهي كالاتي:

١- الدالة الخطية التربيعية وصيغتها:

$$y = 1 + x + x^2$$

٢- الدالة الاسية وصيغتها:

$$y = \exp(2x) - 3.57$$

٣- الدالة غير الخطية وصيغتها:

$$y = 0.57 + \sin(-x)$$

٣,٣ تنفيذ تجارب المحاكاة على طرائق الانموذج الجمعي

لكل دالة من الدوال المدروسة تم عمل ما يلي:

١. توليد المتغير التوضيحي X ليتوزع توزيعا منتظماً U(0,1) مع أخطاء تتوزع توزيع طبيعي.

٢. ٢. بالنسبة لخوارزمية Back fitting تم اعتماد دالة Kernel (Epanechnikov) ذات الشكل $K(u) = (3/4)(1-u^2),\ I(|u| \le 1)$.

٣. بالنسبة لخوارزمية SIMEX تم اعتماد طريقة تقريب الاستكمال (Interpolation) ضمن طرق الاستكمال.

٣,٤ نتائج تجارب المحاكاة

تم تنفيذ برامج المحاكاة باستخدام برنامج Visual Basic أضافة لبرنامج Minitab و Matlab والتي تعد من البرامجيات القابلة للبرمجة وذو إمكانية عالية في الجوانب الرياضية والهندسية والاحصائية حيث يوظف الادوات لبرمجة متقدمة وتم استخدامه لتطبيق الجوانب النظرية وتم الحصول على النتائج لكل دالة من الدوال وكما في ادناه:

جدول (٣,١)

قيم λ لخوارزمية SIMEX لحجم عينة ٥٠ والى $\sigma = 0.1$.

Method	$\hat{Y}(\lambda_I)$	Value	$E(\hat{Y}(\lambda_I))$
SIMEX	$\hat{Y}(\lambda_1)$	0.008418	0.0083185
	$\hat{Y}(\lambda_2)$	0.009393	
	$\hat{Y}(\lambda_3)$	0.008370	
	$\hat{Y}(\lambda_4)$	0.008348	
	$\hat{Y}(\lambda_5)$	0.008328	
	$\hat{Y}(\lambda_6)$	0.008310	
	$\hat{Y}(\lambda_7)$	0.008294	
	$\hat{Y}(\lambda_8)$	0.008279	
	$\hat{Y}(\lambda_9)$	0.008265	
	$\hat{Y}(\lambda_{10})$	0.008254	
	$\hat{Y}(\lambda_{11})$	0.008244	

جدول (٣,٢)

قيم λ لخوارزمية SIMEX لحجم عينة 200 والى $\sigma = 0.1$.

Method	$\hat{Y}(\lambda_I)$	Value	$E(\hat{Y}(\lambda_I))$
SIMEX	$\hat{Y}(\lambda_1)$	0.010424	0.010603
	$\hat{Y}(\lambda_2)$	0.010454	
	$\hat{Y}(\lambda_3)$	0.010486	
	$\hat{Y}(\lambda_4)$	0.010520	
	$\hat{Y}(\lambda_5)$	0.010556	
	$\hat{Y}(\lambda_6)$	0.010594	
	$\hat{Y}(\lambda_7)$	0.010633	
	$\hat{Y}(\lambda_8)$	0.010675	
	$\hat{Y}(\lambda_9)$	0.010719	
	$\hat{Y}(\lambda_{10})$	0.010764	
	$\hat{Y}(\lambda_{11})$	0.010812	

جدول (3.3)

قيم λ لخوارزمية SIMEX لحجم عينة 400 والى $\sigma = 0.1$.

Method	$\hat{Y}(\lambda_I)$	Value	$E(\hat{Y}(\lambda_I))$
SIMEX	$\hat{Y}(\lambda_1)$	0.008834	0.008808
	$\hat{Y}(\lambda_2)$	0.008823	
	$\hat{Y}(\lambda_3)$	0.008814	
	$\hat{Y}(\lambda_4)$	0.008807	
	$\hat{Y}(\lambda_5)$	0.008802	
	$\hat{Y}(\lambda_6)$	0.008799	
	$\hat{Y}(\lambda_7)$	0.008798	
	$\hat{Y}(\lambda_8)$	0.008798	
	$\hat{Y}(\lambda_9)$	0.008801	
	$\hat{Y}(\lambda_{10})$	0.008806	
	$\hat{Y}(\lambda_{11})$	0.008812	

جدول (3.4)

قيم λ لخوارزمية SIMEX لحجم عينة ٥٠ والى $\sigma = 0.25$.

Method	$\hat{Y}(\lambda_I)$	Value	$E(\hat{Y}(\lambda_I))$
SIMEX	$\hat{Y}(\lambda_1)$	0.008834	0.008808
	$\hat{Y}(\lambda_2)$	0.008823	
	$\hat{Y}(\lambda_3)$	0.008814	
	$\hat{Y}(\lambda_4)$	0.008807	
	$\hat{Y}(\lambda_5)$	0.008802	
	$\hat{Y}(\lambda_6)$	0.008799	
	$\hat{Y}(\lambda_7)$	0.008798	
	$\hat{Y}(\lambda_8)$	0.008798	
	$\hat{Y}(\lambda_9)$	0.008801	
	$\hat{Y}(\lambda_{10})$	0.008806	
	$\hat{Y}(\lambda_{11})$	0.008812	

جدول (3.5)

قيم λ لخوارزمية SIMEX لحجم عينة ٢٠٠ والى $\sigma = 0.25$.

Method	$\hat{Y}(\lambda_I)$	Value	$E(\hat{Y}(\lambda_I))$
SIMEX	$\hat{Y}(\lambda_1)$	0.065202	0.068303
	$\hat{Y}(\lambda_2)$	0.065585	
	$\hat{Y}(\lambda_3)$	0.066046	
	$\hat{Y}(\lambda_4)$	0.066587	
	$\hat{Y}(\lambda_5)$	0.067207	
	$\hat{Y}(\lambda_6)$	0.067906	
	$\hat{Y}(\lambda_7)$	0.068685	
	$\hat{Y}(\lambda_8)$	0.069543	
	$\hat{Y}(\lambda_9)$	0.070480	
	$\hat{Y}(\lambda_{10})$	0.071496	
	$\hat{Y}(\lambda_{11})$	0.072591	

جدول (3.6)

قيم λ لخوارزمية SIMEX لحجم عينة 400 والى $\sigma = 0.25$.

Method	$\hat{Y}(\lambda_I)$	Value	$E(\hat{Y}(\lambda_I))$
SIMEX	$\hat{Y}(\lambda_1)$	0.054639	0.055785
	$\hat{Y}(\lambda_2)$	0.054644	
	$\hat{Y}(\lambda_3)$	0.054724	
	$\hat{Y}(\lambda_4)$	0.054878	
	$\hat{Y}(\lambda_5)$	0.055108	
	$\hat{Y}(\lambda_6)$	0.055412	
	$\hat{Y}(\lambda_7)$	0.055791	
	$\hat{Y}(\lambda_8)$	0.056244	
	$\hat{Y}(\lambda_9)$	0.056773	
	$\hat{Y}(\lambda_{10})$	0.057376	
	$\hat{Y}(\lambda_{11})$	0.058053	

جدول (3.7)

قيم λ لخوارزمية SIMEX لحجم عينة ٥٠ والى $\sigma = 0.5$.

Method	$\hat{Y}(\lambda_I)$	Value	$E(\hat{Y}(\lambda_I))$
SIMEX	$\hat{Y}(\lambda_1)$	0.247321	0.24999
	$\hat{Y}(\lambda_2)$	0.244838	
	$\hat{Y}(\lambda_3)$	0.243360	
	$\hat{Y}(\lambda_4)$	0.242888	
	$\hat{Y}(\lambda_5)$	0.243422	
	$\hat{Y}(\lambda_6)$	0.244961	
	$\hat{Y}(\lambda_7)$	0.247505	
	$\hat{Y}(\lambda_8)$	0.251055	
	$\hat{Y}(\lambda_9)$	0.255611	
	$\hat{Y}(\lambda_{10})$	0.261172	
	$\hat{Y}(\lambda_{11})$	0.267739	

جدول (3.8)

قيم λ لخوارزمية SIMEX لحجم عينة 200 والى $\sigma = 0.5$.

Method	$\hat{Y}(\lambda_I)$	Value	$E(\hat{Y}(\lambda_I))$
SIMEX	$\hat{Y}(\lambda_1)$	0.226730	0.25110
	$\hat{Y}(\lambda_2)$	0.227382	
	$\hat{Y}(\lambda_3)$	0.229441	
	$\hat{Y}(\lambda_4)$	0.232907	
	$\hat{Y}(\lambda_5)$	0.237780	
	$\hat{Y}(\lambda_6)$	0.244060	
	$\hat{Y}(\lambda_7)$	0.251747	
	$\hat{Y}(\lambda_8)$	0.260841	
	$\hat{Y}(\lambda_9)$	0.271342	
	$\hat{Y}(\lambda_{10})$	0.283250	
	$\hat{Y}(\lambda_{11})$	0.296565	

جدول (3.9)

قيم λ لخوارزمية SIMEX لحجم عينة 400 والى $\sigma = 0.5$.

Method	$\hat{Y}(\lambda_I)$	Value	$E(\hat{Y}(\lambda_I))$
SIMEX	$\hat{Y}(\lambda_1)$	0.228071	0.25314
	$\hat{Y}(\lambda_2)$	0.229695	
	$\hat{Y}(\lambda_3)$	0.232449	
	$\hat{Y}(\lambda_4)$	0.236332	
	$\hat{Y}(\lambda_5)$	0.241346	
	$\hat{Y}(\lambda_6)$	0.247489	
	$\hat{Y}(\lambda_7)$	0.254762	
	$\hat{Y}(\lambda_8)$	0.263164	
	$\hat{Y}(\lambda_9)$	0.272697	
	$\hat{Y}(\lambda_{10})$	0.283359	
	$\hat{Y}(\lambda_{11})$	0.295152	

جدول (3.10)

مقارنة طرائق الانحدار اللامعلمي الجمعي باستخدام معايير المقارنة لحالة الدالة الخطية التربيعية.

σ	Method	n	MSE	MGCV المقترحة
$\sigma = 0.1$	SIMEX	50	2.54169	0.185562
		200	0.8379	0.108722
		400	0.8364	0.002941
	Back fitting	50	1.00087	0.10169
		200	1.00001	0.13251
		400	1.00019	0.09824
$\sigma = 0.25$	SIMEX	50	2.93677	0.150830
		200	0.4140	0.006174
		400	0.3585	0..0.0392
	Back fitting	50	1.0301	0.041832
		200	1.0006	0.088001
		400	1.0000	0.016347
$\sigma = 0.5$	SIMEX	50	3.86160	0.205510
		200	0.0896	0.055889
		400	0.0803	0.042543
	Back fitting	50	1.00014	0.102611
		200	1.00093	0.217593
		400	1.00198	0.205452

لحالة النوع الاول من الدالة المدروسة وهي خطية تربيعية يلاحظ اخفاق معيار MSE مقارنة بالمعيار المقترح MGCV اذ يعتمد الاخير على الطريقة الشرائحية تقليص الجزاء والتي أجزت بشكل كفوء في معظم المقارنات لطرائق لامعلمية وحتى في حالة الدوال التي تعاني من تذبذبات عالية، كما اثبتت طريقة SIMEX تفوقها لجميع مستويات التباين ولحالة حجوم العينات المتوسطة والكبيرة على خوارزمية Back fitting في حين تفوقت الاخيرة في حالة حجم العينة الصغيرة.

جدول (3.11)

مقارنة طرائق الانحدار اللامعلمي الجمعي باستخدام معايير المقارنة لحالة الدالة الاسية.

σ	Method	n	MSE	MGCV المقترحة
$\sigma = 0.1$	SIMEX	50	6.2066	0.098063
		200	1.3734	0.108815
		400	0.3942	0.003811
	Back fitting	50	6.5915	0.07835
		200	6.5722	0.13231
		400	6.5546	0.05662
$\sigma = 0.25$	SIMEX	50	4.0980	0.144562
		200	2.9923	0.000707
		400	0.3102	0.037107
	Back fitting	50	5.8337	0.00874
		200	6.5768	0.01336
		400	6.6164	0.09635
$\sigma = 0.5$	SIMEX	50	8.33629	0.253307
		200	1.05523	0.043379
		400	0.20992	0.060765
	Back fitting	50	6.56072	0.042839
		200	6.71146	0.085048
		400	6.43412	0.229271

الدالة الثانية من الدوال المدروسة وهي اسيية لوحظ تفوق المعيار المقترح MGCV على معيار MSE ايضاً تفوق طريقة SIMEX لحجمي العينات المتوسطة والكبيرة، اما خوارزمية Back fitting فكانت الافضل بحجوم العينات الصغيرة اي ان الخوارزمية فشلت في تبني الدوال بحجوم العينات المتوسطة والكبيرة.

جدول (3.12)

مقارنة طرائق الانحدار اللامعلمي الجمعي باستخدام معايير المقارنة لحالة الدالة غير الخطية.

σ	Method	n	MSE	MGCV المقترحة
σ = 0.1	SIMEX	50	0.770951	0.094349
		200	0.637342	0.011600
		400	0.081234	0.012914
	Back fitting	50	0.31042	0.013800
		200	0.31932	0.010032
		400	0.30334	0.010476
σ = 0.25	SIMEX	50	0.717345	0.181930
		200	0.785834	0.088473
		400	0.051322	0.051829
	Back fitting	50	0.106171	0.114092
		200	0.102782	0.011722
		400	0. 10100	0.004296
σ = 0.5	SIMEX	50	1.000143	0.091540
		200	0.382312	0.096386
		400	0.093453	0.053963
	Back fitting	50	0.951043	0.246142
		200	0.375932	0.069273
		400	0.199605	0.019960

الدالة الثالثة وهي غير خطية لوحظ ايضا تفوق معيار MGCV واثبتت خوارزمية Back fitting الكفاءة بجميع حجوم العينات عدا حجم العينة الصغير ومستوى التباين العالي فقد كانت طريقة SIMEX الافضل.

٤. الجانب العملي

4.1 الماء أهميته ومصادر تلوثه وكيفية المحافظة عليه من التلوث:

الماء له أهمية كبيرة بحياة الكثير من الكائنات الحية بالنسبة للانسان يدخل في تركيب الخلايا بنسبة 75-95% كما يدخل في الانسجة المختلفة وعندما يتغير تركيب عناصر الماء تصبح اقل صلاحية للاستخدام ومصادر تلوث الماء هي التلوث الحراري والطبيعي، كما ان ارتفاع درجة الحرارة على النظام البيئي يؤثر على النباتات والحيوانات حيث تغير الخواص الطبيعية للماء حيث الماء الدافيء لا يحتفظ بنفس كمية الغازات التي يحتويها الماء البارد كالاوكسجين كما بارتفاع الحرارة تزداد كمية التنفس ثم تقل كمية الاوكسجين الذائب في الماء وبهذا تموت الكائنات الحية. ايضاً من مصادر تلوث الماء هو النفط كذلك المخلفات الصناعية، المخلفات البشرية والمواد المشعة، المبيدات والمخصبات الزراعية.

4.2 عينة البحث والمتغيرات المستخدمة

تم في الجانب العملي استخدام عينة من ٣١١ مفردة والتي اخذت من وزارة البيئة – قسم تلوث المياه والتي فرزت على اساس ٣ سنوات (2008,2009,2010) والى ١٢ محطة والمحطات هي (T17) جسر المثنى، (T18) جسر الائمة، (T19) جسر الشهداء، (T20) جسر الجمهورية، (T21) جسر الجادرية، (T22) مشروع ماء الرشيد، (T23) جسر ماء الزعفرانية، (T24) مشروع ماء الدورة، (D16)

نهر ديالى/ جسر ديالى الجديد، (D17) نهر ديالى /جسر ديالى القديم، (ST1) نهر المصب العام في ابي غريب، (ST2) نهر المصب العام في المحمودية. حيث تم اخذ عينة من كل محطة ومرة واحدة لكل شهر وذلك من قبل الاختصاصيين العاملين في قسم تلوث المياه. وكانت المتغيرات التي لها تأثير على نسبة تلوث المياه هي:

E.C.: القيم الموصلة الكهربائية وهي من التراكيز الايونية وقيمتها تتراوح مابين 10114-892 مايكرو موز/ سم.

T.D.S.: مجموعة المواد الصلبة الذائبة (Total Dissolved Solids) وتحديد قيمتها تختلف معاييرها اي استخدام المياه ذات الملوحة حسب المعايير الاجنبية وتقع القيمة مابين ١٠٠٠-١٥٠٠ ملغم/ لتر وان زيادتها على الحد المسموح يسبب مشاكل لصحة الانسان كما يسبب العديد من الامراض.

Ca.: ويعتبر من اهم الايونات الاساسية موجبة الشحنة والموجودة في المياه وان قيم الكالسيوم لعينات المياه تختلف في فصل الصيف عن الشتاء وتكون قيمته في الصيف مابين 1042-10 ملغم/ لتر وفي الشتاء تتراوح مابين 954-16 ملغم/ لتر.

Mg.: المغنيسيوم وياتي بعد الكالسيوم ويعتبر من الايونات الاساسية الموجبة المتواجدة في المياه وتتراوح قيمته في الصيف مابين 256-12 ملغم/ لتر اما في الشتاء يتراوح مابين 386-29 ملغم/ لتر واذا زاد او قل عن الحد المسموح تصبح المياه ملوثة وحيث ان تركيز Mg في مياه البحر تكون خمسة اضعاف تركيز الكالسيوم.

Na: الصوديوم وتتميز املاح الصوديوم بالذوبان العالي في المياه وتزداد مستويات الصوديوم في المياع الجوفية التي فيها رواسب معدن الصوديوم وان تركيز الصوديوم ماين 20-1 ملغم/ لتر وينصح للاشخاص الذين يعانون من ارتفاع ضغط الدم أو قصور القلب الاحتقاني توخي الحذر من استعمال المياه التي ترتفع فيها مستويات الصوديوم عن ٢٠٠ ملغم/ لتر.

No_3: النترات ووجد من خلال التحليلات انها تتراوح بين 290-0 ملغم/ لتر في الصيف اما في الشتاء تتراوح مابين 167-0 ملغم/ لتر وحسب المحددات الاجنبية اما حسب المحددات البيئية لنظام صيانة الانهار ٢٥ في العراق لسنة ١٩٦٧ يجب ان لا تزيد عن ١٥ وان زاد تعتبر المياه ملوثة.

C03: الكربونات ووجد ان قيمته تتراوح ما بين 13-0 ملغم/ لتر في الصيف اما في الشتاء تتراوح مابين -0 22 ملغم/ لتر وحسب المحددات الاجنبية.

Po4: الفوسفات وقيمته حسب المحددات البيئية لنظام صيانة الانهار ٢٥ لسنة ١٩٦٧ في العراق يجب ان لا تزيد عن 4-0 وان زادت تعتبر المياه ملوثة.

CI: الكلور ويعتبر من اهم المتغيرات وقيمته حسب المحددات البيئية لنظام صيانة الانهار ٢٥ لسنة ١٩٦٧ في العراق يجب ان لا يزيد عن ٢٠٠.

BOD5: يدل على الحاجة البيولوجية للاوكسجين اي وجود الاوكسجين في المياه.

ALK: يدل على قاعدية المياه.

تم اختيار المتغيرات المستقلة والتي كانت معنوية بعد تطبيق طريقة الخطوات المتسلسلة Stepwise وهي كالآتي:

١. المتغير المعتمد y_i وهو نسبة تلوث المياه وتمثل المتغير PH ويدل على حموضة المياه ويجب ان تقع قيمته بين 6.5-8.5 فاذا زادت عن القيمة المذكورة تعتبر المياه ملوثة تعتبر المياه ملوثة حسب محددات البيئية لنظام صيانة الانهار ٢٥ لسنة ١٩٦٧.

٢. المتغير المستقل الاول x_1 وهو BOD5 الذي يدل على الحاجة البيولوجية للاوكسجين اي وجود الاوكسجين في المياه وان رقم ٥ يدل على ان الفحص يتم خلال خمسة ايام واذا زاد عن قيمتة حسب المحددات البيئية لنظام صيانة الانهار ٢٥ لسنة ١٩٦٧ تكون المياه ملوثة.

٣. المتغير المستقل الثاني x_2 وهو ALK ويدل على قاعدية المياه علماً بان المتغير ليس له محدد اذا تجاوزه تعتبر المياه ملوثة في العراق ولهذا تم اعتماد محدد دولي لوزارة البيئة الامريكية حسب الموقع الالكتروني www.epa.gov/enviroed.

وبعد اختيار المتغيرات وبإستخدام المعيارين تم اعتماد أنموذج الماء من النماذج الثلاثة المدروسة والذي يعطي اقل MSE أو MGCV وللطريقتين Back fitting و SIMEX. كما في الجدول أدناه:

جدول (3.13)

مقارنة طرائق الانحدار اللامعلمي الجمعي باستخدام معايير المقارنة للدوال الثلاثة.

The Function	Method	MSE	MGCV المقترحة
الدالة الخطية التربيعية	SIMEX	2.2681	0.033057
	Back fitting	3.2642	0.426322
الدالة الاسية	SIMEX	2.3760	0.071576
	Back fitting	3.4526	0.542334
الدالة غير الخطية	SIMEX	5.7274	0.013164
	Back fitting	1.2365	0.332563

٥. الاستنتاجات

من خلال تنفيذ تجارب المحاكاة وبعد تنفيذ الطرائق على بيانات حقيقية عن تلوث المياه يمكن استنتاج ما يلي:

١. بالنسبة للمعايير لوحظ تفوق معيار MGCV المقترح على معيار MSE اذ يعتمد المعيار على طريقة تقليص الجزاء والتي اثبتت قدرتها في تبني العديد من الدوال حتى التي تعاني من ذبذبات عالية وبالتالي استطاع المعيار والذي يمكن ان يستخدم باتجاهين من التفوق من بين معايير للمقارنة بين الطرائق اللامعلمية.

٢. لحالة الدوال غير الخطية اثبتت خوارزمية Back fitting الكفاءة لمعظم مستويات التباين وحجوم العينات عدا حجم العينة الصغير والتباين العالي فكانت طريقة SIMEX هي الافضل.

٣. في حالة الدوال الاسية لوحظ تفوق طريقة SIMEX لحجمي العينات المتوسطة والكبيرة اما خوارزمية Back fitting فهي الافضل بالحجوم الصغيرة.

٤. من خلال التطبيق على بيانات تلوث المياه لوحظ ان افضل دالة تلائم انموذج الماء هي غير الخطية وللطريقتين Back fitting و SIMEX فقد كان معيار المقارنة MGCV المقترح هو الاقل.

٦. التوصيات

١. بالنظر لكون المعيار المقترح MGCV حقق كفاءة لمعظم الدوال المستخدمة وتفوق على المعيار MSE وهو اقل تحيزاً من الاخير لذا نوصي باستخدامه في حالة الطرائق اللامعلمية الجمعية.

٢. اثبتت خوارزمية Back fitting كفاءتها بحالة الدوال غير الخطية والتي لا تعاني من تذبذبات عالية لذا نوصي بتطوير الخوارزمية باستخدام الدوال الشرائحية لتعطي نتائج اكثر كفاءة مثلا لحالة الانحدار اللامعلمي الجمعي الجزئي.

٣. من خلال دراسة الجوانب النظرية للموضوع ومن خلال دراسة اسهامات الباحثين في هذا المجال لوحظ عدم تطرق اي من البحوث على مستوع القطر لحالة النماذج الجمعية التي تحتوي على تفاعلات بين المتغيرات التوضيحية لذا نوصي باستخدام هذا الجانب اي شرائح الانحدار المكيفة متعددة الانحدار MARS.

٤. بالنظر لكون مشكلة تلوث المياه أبحت مشكلة عالمية تعاني منها معظم الدول ولكون المياه هو مصدر حياة كل كائن حي اذ تعقد مؤتمرات عالمية سنوية بخصوصه لذا نوصي بالتوسع في استخدام متفيرات اخرى تتعلق بالمياه لإهمية الموضوع.

٧. المصادر

[١] قيس، ساندي، ٢٠١٢ "مقارنة مقدّرات Back fitting و SIMEX لتقدير انموذج الانحدار اللامعلمي الجمعي مع التطبيق"، رسالة ماجستير في الاحصاء كلية الادارة والاقتصاد، جامعة بغداد.

Smoothing to the state of Thin Plate Spline and the method of analysis of this type of Splines difficult, especially if he had known extent of interaction and which can not be represented easily as needs a high level of analysis and programming and this was the idea of Additive Model models and especially that there are some algorithms to overcome the effect of dimensionality problem.

Research aims to focus on the methods Nonparametric Additive Regression Methods (i.e., the case of binary variables) so that we avoid the problem of dimensionality and the algorithms used Backbiting and SIMEX, which combines simulation and interpolation, and comparison between the algorithms using the standards of existing and other proposed criterion MGCV, using many of the simulation experiments and variations and sizes of samples different, in addition to the application of algorithms on real data about water pollution and the adoption of the model that best fits the data, which gives the less comparison criteria.

[٢] علي، عمر عبد المحسنن، ٢٠٠٧، "مقارنة النماذج التجميعية المعممة باستخدام الشرائح التمهيدية عند تحليل الانحدار اللامعلمي وشبه المعلمي"، اطروحة دكتوراة في الاحصاء كلية الادارة والاقتصاد، جامعة بغداد.

[3] Buja, A., Hastic, T. & Tibshirani, R., "Linear Smoothing and additive Model", The Annals of Statistics, Vol. 17, pp. 453-510, 1989.

[4] Carrol, R.J., Maitym, H., Mamm, E. & Yu, K., "Nonparametric Additive Regression for repeatedly Measurement data", Biometrika, Vol. 96, No.2 pp. 383-398, 2009.

[5] Cook, J.R., & Stefanski, L.A., "Simulation Extrapolation estimation in parametric measurement error", American Statistical Association, Vol. 89, pp.1314-1328, 1994.

[6] Hastie, T. & Tibshirani, R.J., "Generalized Additive Model", Chapman & Hall, London, 1990.

[7] Lin, X. & Carrol, R.J., "SIMEX variance component tests in generalized linear mixed measurement error model", Biometrika, Vol. 55, No. 2, pp. 613-619.1999.

[8] Liang, H., Thurston, S.W., Ruppert, D., Apanasovich, T. & Hauser, R., "Additive Partial Linear Model with measurement Error", Biometrika, Vol. 95, pp. 667-678, 2008.

[9] Opsomer, J. D. & Ruppert, D., "Fitting a Bivariate Additive Model by Local Polynomial Regression", Annals of Statistics, Vol. 25, No. 1, pp. 168-211, 1997.

Abstract

In the absence of knowledge about the phenomenon was the experience for the first time or can not determine a causal relationship, or behavioral variables that link instead of the requirement to take the data template or form the described function phenomenon in advance are replaced with a more flexible manner so-called Nonparametric analysis.

The expansion in the Spline Smoothing Situation from Single to multiple variables showed the problem of dimensionality because we must expand in the case of Cubic Spline