مقارنة بعض طرائق التقدير المويجي لدالة الانحدار اللامعلمي عند فقدان متغير الاستجابة عشوائيا

أ.د. ظافر حسين رشيد / كلية الادارة والاقتصاد / جامعة بغداد
م.م. سعد كاظم حمزة / قسم الشؤون الادارية/ رئاسة جامعة بغداد

المستخلص

تعد مشكلة البيانات المفقودة عقبة كبيرة أمام الباحثين في عملية تحليل البيانات في مختلف المجالات ، وان هذه المشكلة متكررة الظهور في جميع مجالات الدراسات الاجتماعية والطبية والفلكية والتجارب السريرية وغيرها.

وان وجودها ضمن البيانات المراد دراستها سيؤثر بشكل سلبي على تحليلها ومن ثم يؤدي الى استنتاجات مظللة ، وان هذه الاستنتاجات ناتجة من التحيز الكبير التي تحدثه .

وعلى الرغم من كفاءة الطرائق المويجية الا انها هي الاخرى تتأثر بمشكلة فقدان البيانات ، فهي فضلا عن تأثير مشكلة الفقدان على دقة التقدير فليس بالامكان تطبيق هذه الطرائق لفقدان احد شروطها وهي حجم العينة الدايديكي (Dyadic) $n = 2^J$.

ونظراً للتأثير الكبيرة الناجم عنها فان الكثير من الباحثين ممن كرسوا بحوثهم لمعالجتها باستخدام طرائق تقليدية في معالجة البيانات المفقودة ، بينما قام الباحث باستخدام طرائق تعويض اكثر كفاءة لمعالجة البيانات المفقودة كمرحلة اولى كي تصبح البيانات جاهزة للتطبيق المويجي وقد اثبتت تجارب المحاكاة كفاءة الطرائق المقترحة على بقية الطرائق الاخرى ، كذلك تضمن البحث التصحيح التلقائي لمشكلة الحدودية عن طريق استخدام نموذج متعدد الحدود فضلاً عن استخدام قيم عتبة مختلفة ضمن التقديرات المويجي.

المصطلحات الرئيسية للبحث/ البيانات المفقودة ـ الانحدار المويجي ـ متعدد حدود موضعي – اقرب مجاور.

* البحث مستل من اطروحة دكتوراه

مجلة العلوم الاقتصادية والإدارية
العدد 90 المجلد ٢٢
الصفحات ٣٨٢ـ٤٠٦

١- المقدمة

تعد مشكلة البيانات المفقودة عقبة كبيرة أمام الباحثين في عملية تحليل البيانات في مختلف المجالات ، وفي الإحصاء الفقدان معناه ظاهرة عدم وجود قيم لبيانات بعض المتغيرات من المفترض أنها موجودة ، ولكن حدث فشل في تسجيلها لأسباب عديدة كأن تكون فشل في الاجابة عن بعض اسئلة استبانة ما في أحد المسوحات أو عطل معدات القياس كما في جهاز قياس الاشعاع الشمسي والذي يتعرض لاعطال مفاجئة بسبب الظروف الجوية .

وعلى الرغم من ان المعالجة الأولى لمشكلة البيانات المفقودة هي تجنب حدوثها الا انه في بعض الأحيان نضطر الى التعامل معها كونها قسرية الحدوث ولاسباب اهمها ان موضوع حدوثها يكون خارج سيطرة الباحث أو ان السيطرة عليها بشكل تام أمر مكلف جداً ، وعليه يجب البحث عن طرائق لمعالجة هذه البيانات كون وجود هذه المشكلة يؤثر بشكل سلبي على تحليل البيانات ومن ثم يؤدي الى استنتاجات مظللة ، وان هذه الاستنتاجات ناتجة من التحيز الكبير الذي تحدثه تلك المشكلة ، ونظراً للآثار الكبيرة التي تخلفها تلك المشكلة فان عدداً كبيراً من الباحثين كرس بحوثه ودراساته لمعالجة هذه المشكلة ، وقد شهد العقدين الاخيرين ظهور تقنيات الانحدار المويجي بشكل ملفت وفي مجالات عديدة ، وتعد هذه التقنيات اداة فعالة في تحليل البيانات وتقدير دالة الانحدار غير المعلومة ، وعلى الرغم من ان هذه التقنيات تتطلب جهداً رياضياً عالياً اثناء التطبيق الا انها بنفس الوقت تتفوق على بقية التقنيات الاخرى كالتقنيات اللبية والشرائحية كونها تملك خاصية اعظم تقليل (Minimax) وتكيفها لحالة عدم التجانس وامكانية توسيعها الى ابعاد عالية كالصور وبخوارزميات سريعة ، الا ان هذه التقنيات تعاني هي الاخرى من مشكلة وجود البيانات المفقودة لأن هذه المشكلة فضلا عن ما تسببه من تظليل في الاستنتاجات بسبب التحيز الناشئ عنها فانها تحول دون التطبيق التقديرات المويجية لانها تشترط اثناء تطبيقها تساوي المسافات بين المشاهدات (Regulary-spaced data) وحجم العينة الدايديكي (Dyadic) أي $n=2^J$ لكل J عدد صحيح .

وان هذه القيود المفروضة على التقديرات المويجية دفعت الباحثين الى التفكير لايجاد الحلول لجعل البيانات ملائمة لهذا التطبيق المويجي وذلك عن طريق معالجة البيانات المفقودة وطرائق التعويض المعروفة ، كذلك تضمن البحث استخدام نموذج الانحدار متعدد الحدود لمعالجة مشكلة الحدودية تلقائياً والتي تحدث بسبب التحويل المويجي (Wavelet Transformation) .

تم تقسيم البحث على ثمانية مباحث ، تضمن المبحث الاول المقدمة ومشكلة البحث والهدف من البحث ، اما المبحث الثاني فتضمن تعريف الانحدار المويجي وطرائق اختيار العتبة ، بينما تضمن المبحث الثالث نموذج وآلية الفقدان ، وتضمن المبحث الرابع طرائق تعويض البيانات المفقودة ، اما المبحث الخامس فقد تضمن التقدير لدالة الانحدار غير المعلومة ، بينما تضمن المبحث السادس دراسة المحاكاة ، وتضمن المبحث السابع الاستنتاجات واخيراً التوصيات ضمن المبحث الثامن .

١-١ مشكلة البحث

ان مشكلة البحث تتركز حول تأثير القيم المفقودة لمتغير الاستجابة y_i على دقة تقدير دالة الانحدار اللامعلمي غير المعلومة فضلا عن ان فقدان بعض القيم يجعل البيانات غير ملائمة لتطبيق التقدير المويجي كونه يشترط حجم العينة الدايديكي أي $n=2^J$ وتساوي المسافات بين المشاهدات .

١-٢ هدف البحث

يتناول البحث حالة انموذج انحدار لامعلمي

$$y_i = f(x_i) + e_i \quad \text{.....} \quad (1)$$

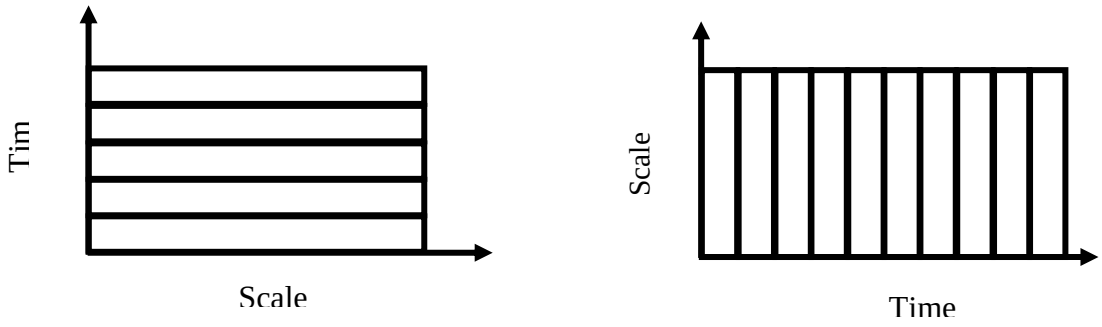
حيث ان $x_i = i/n$ متغير توضيحي ، أي أن $x \in [0,1]$ و f دالة غير معلومة و e_i يمثل التشويش (noise) و y_i متغير الاستجابة يعاني من مشكلة فقدان في بعض مشاهداتها عشوائياً .

ان الهدف من هذا البحث هو تقدير دالة الانحدار اللامعلمية $f(x_i)$ باستخدام الانحدار المويجي بواسطة ازالة التشويش ومعالجة مشكلة البيانات المفقودة في متغير الاستجابة والتصحيح التلقائي لمشكلة الحدودية وذلك عن طريق سلوك اتجاهين.

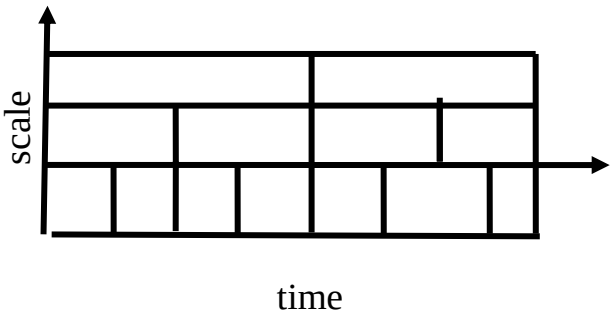
الاول معالجة مشكلة الحدودية التي تحصل بسبب التحويل المويجي عن طريق استخدام نموذج متعدد حدود من درجة قليلة، والاتجاه الثاني استخدام طرائق تعويض كفوءة للبيانات المفقودة في متغير الاستجابة كمرحلة اولى ومن ثم تقدير دالة الانحدار غير المعلومة باستخدام التقديرات المويجية .

٢- تحويل المويجات [13,19]

تعرف المويجة على انها اشارة محدودة الطول الزمني (الاستمرارية) تمتلك قيمة متوسطة مساوية للصفر ، اذ ان التحويل المويجي ظهر من اجل معالجة حالات الضعف التي تعانيها التحويلات السابقة كتحويلات فوارير وتحويلات فوارير للزمن القصير ، حيث تقوم تحويلات فوارير بنقل الاشارة من مجال الزمن الى مجال التردد وبالعكس لكن المشكلة في هذا التحويل انه يصبح غير مناسب للاشارات المختلفة (متغيرة التردد) كونه لايزودنا بمعلومات عن المحتوى الترددي خلال الزمن لذلك قدم العالم **Gabor** حلاً للمشكلة السابقة عن طريق استخدام ما يعرف بالنافذة ذات العرض الثابت باستخدام تحويل فوارير للزمن القصير (**Short Time Fourier Transformation**) بتمثيل الاشارة زمنياً وترددياً على حساب دقتها الزمنية والترددية ، اذ انه عند استخدام نافذة صغيرة يتم الحصول على دقة عالية من اجل العناصر التي تتغير بسرعة ، بينما لاتكون هذه الدقة عالية للعناصر المتغيرة ببطء ، ويحدث العكس عند استخدام نافذة كبيرة لذا تم تطوير ما يعرف بالمويجات ، حيث يمكننا هذا التحويل من تحليل الإشارة الى مجموعة من المستويات متعددة الحل (**Multiresolution**) في كل من مجالي الزمن والتردد ، و بخلاف التحويلات السابقة يستخدم التحويل المويجي نافذة متغيرة بدلاً من استخدام نافذة ثابتة ، اذ يتم تغيير عرض النافذة باستمرار للحصول على معلومات مختلفة التردد على طول المويجة . فيتم الحصول على المويجات التي تختلف تردداتها بحسب عرض النافذة المستخدم ، رياضياً يقوم التحويل المويجي على ضغط المويجة المراد معالجتها مع دالتين هما دالة المويجة الام $\psi(x)$ من اجل الحصول على مجموعة من المعاملات (**coefficients**) والتي تسمى معاملات المويجة او المعاملات التفصيلية $D(s,t)$ والثانية هي دالة القياس $\phi(x)$ وتسمى كذلك بدالة الاب للحصول على المعاملات التقريبية $A(s,t)$



شكل (1) : يبين استخدام نافذة ثابتة



شكل (2) : يبين استخدام نافذة متغيرة في التحويل المويجي

٢-1 تحويل المويجي المتقطع

ان التحويل المويجي المتقطع (Discrete wavelet Transformation) هو خوارزمية كفوءة اقترحت من قبل الباحث Mallat (١٩٨٩) لحساب معاملات المويجة لسلسلة من البيانات المشوشة من خلال تحليل الاشارة المدخلة الى حزم ترددية مختلفة وذلك عن طريق تحليلها الى ما يعرف بالمعاملات التفصيلية (Detail) $d_{j-1,k}$ وتسمى كذلك بمعاملات المويجة والمعاملات التقريبية (Approximation) وتسمى كذلك بمعاملات القياس $c_{j-1,k}$ والتي يمكن حسابه باستخدام المعادلتين الآتيتين

شكل (1) : يبين استخدام نافذة ثابتة

$$c_{j-1,k} = \sum h_n - 2k \; c_{j,n} \qquad \ldots\ldots\ldots \qquad (2)$$

$$d_{j-1,k} = \sum g_n - 2k \; c_{j,n} \qquad \ldots\ldots\ldots \qquad (3)$$

ويتم الحصول على هذه المعاملات باستخدام مجموعة من مرشحات التمرير المنخفضة والمرتفعة ولذلك سميت بشجرة (Mallats Tree) بحيث يعطي مرشح الممر المرتفع والذي يشار له بـ ($H = (hk)$) العوامل التفصيلية $d_{j-1,k}$ حيث ان j يشير الى مستوى التحليل بينما يعطي مرشح الممر المنخفض العوامل التقريبية $c_{j-1,k}$ حيث يتم التعامل في كل مرحلة من مراحل التحليل مع نصف عدد نقاط المرحلة السابقة مما يؤدي الى تسريع عملية التحليل والحصول على معاملات مساوية لعدد نقاط الاشارة المدخلة ، وبالرغم من ان كل مرشح سينتج نصف طول البيانات الأصلية الا أن الجزء الأهم من المخرجات هو الذي يتم الحصول عليه عن طريق الممر الواطئ (g) لأنه يحتوي على اغلب المعلومات المحتواة في الإشارة الأصلية وبعبارة اخرى أي يكون الاهتمام منصب عادةً على التكرارات الواطئة كونها تعطي الإشارة بشكلها الموحد ، بينما سبب إهمال مخرجات الممر العالي يعود الى انه يضم التشويشات في الاشارة والتي تكون غير مرغوب بها ومن ثم يتم استبعادها ، وبعد ذلك يتم الحصول على الإشارة الأصلية بالتجميع المتسلسل لكل العوامل الناتجة سابقاً بدءاً من آخر مرحلة تحليل [14,15,12] .

٢-2 الانحدار اللامعلمي التقليدي

في هذا المبحث سنوضح بشكل موجز طريقتين من طرائق التقدير اللامعلمي كونها تستخدم ضمن البحث وهي:

٢-٢-١ الانحدار اللبي Kernel Regression

هو ابسط اشكال الانحدار اللامعلمي ويمكن كتابته بصورة عامة على وفق الصيغة الآتية [8]

$$\hat{f}_h(x) = \frac{1}{nh}\sum_{j=1}^{n} K\left(\frac{x - x_j}{h}\right) y_j = \frac{1}{n}\sum w_j y_j \qquad \ldots\ldots \qquad (4)$$

حيث ان

$$w_j = K\left(\frac{x - x_j}{h}\right) \Big/ h \qquad \ldots\ldots \qquad (5)$$

حيث ان $K_h(x)$ هي الدالة اللبية ، و h هي عرض الحزمة .

وقد عدلت هذه الطريقة بواسطة مقدر (Nadaraya-Watson) ليكون بالصيغة التالية

$$\hat{f}_h(x) = \frac{\sum w_i y_i}{\sum w_i} \qquad \ldots\ldots \qquad (6)$$

٢-٢-٢ الانحدار متعدد الحدود الموضعي Local Polynomial Regression

ان الانحدار متعدد الحدود الموضعي (LPR) مشابه للتقدير اللبي الا انه ايجاد القيم لـ(LPR) بواسطة الانحدار الموزون الموضعي بدلاً من المتوسط المرجح موضعياً.

لنفرض ان دالة الانحدار $f(x)$ هو ممهدة كفاية لتكون مقربة (متقاربة) بواسطة توسيع تايلر [7,9]

$$f(x) \approx \sum_{j=0}^{\infty} \frac{f(j)(x_0)}{j!}(x-x_0)^j = \sum_{j=0}^{\infty} B_j (x-x_0)^j \qquad \ldots\ldots \quad (7)$$

حيث X قريبة من x_0 مقدر المربعات الصغرى يعطى بالشكل الآتي :

$$\hat{f}(x;d,h) = e_1'(X_d' W X_d)^{-1} X_d' W Y \qquad \ldots\ldots \quad (8)$$

هنا d يشير الى درجة انحدار متعدد الحدود الموضعي بينما h هي معلمة التمهيد .

و W_x مصفوفة قطرية للاوزان $W_j \equiv K_h(x_j - x) \equiv \{K[(x_j - x)/h]\}$ ، e_1 هو متجه $1\times(d+1)$ له (1) في الادخال الاول و(صفر) للبقية و k دالة لبية ، وفي حالة $d=2$ فان مصفوفة X تكون وفق الشكل الآتي :

$$x = \begin{bmatrix} 1 & (x_1 - x) & (x_1 - x)^2 \\ 1 & (x_2 - x) & (x_2 - x)^2 \\ \vdots & & \\ \vdots & & \\ 1 & (x_n - x) & (x_n - x)^2 \end{bmatrix} \qquad \ldots\ldots \quad (9)$$

٢-٣ الانحدار المويجي

يعد تقدير الانحدار المويجي من الأساليب الحديثة جداً في تقدير منحنى الانحدار والذي قدم بواسطة Donoho and Jonstone في عام (١٩٩٤) وما تزال منطقة توسعه في البحوث جارية ، وان أهم الخطوط العريضة لهذه الطريقة هو أنه عادةً ما يتم إفتراض تساوي المسافات بين النقاط $(x_1,x_2,\ldots\ldots,x_n)$ خلال الفترة [0,1] .

إذ أن : $x_i = i/n$ و أن **n** ذات حجم دايديكي* بحيث تكون بالشكل $n = 2^J$ ***J=0,1,2,…..,***

ويمكن أن تعرف تقديرات الانحدار المويجي وفق الخطوات الاتية [6]:

ليكن لدينا المشاهدات $y_i = (y_1,\ldots\ldots\ldots,y_n)'$ معطاة بالصيغة الاتية :

$$y_i = f(i/n) + \varepsilon_i \qquad \ldots\ldots \quad (10)$$

أو بصيغة المصفوفات

$$\underline{y} = \underline{f} + \underline{\varepsilon} \qquad \ldots\ldots \quad (11)$$

حيث أن $y = (y_1,.....,y_n)'$ و $\varepsilon = (\varepsilon_1,.....,\varepsilon_n)'$ يمثل التشويش ، والهدف هو تقدير الدالة f غير المعلومة حيث أن $f = (f_1,.....,f_n)'$.

١. حساب قيم معاملات المويجة w بواسطة تطبيق التحويل المويجي المتقطع على البيانات $(y_1,.....,y_n)'$ وفق الصيغة التالية

$$w = Wy \quad ..... \quad (12)$$

حيث ان W هي مصفوفة التحويل المويجي من درجة $(n \times n)$ لها علاقة بقاعدة المويجة المتعامدة التي يتم اختيارها .

٢. نعدل معاملات المويجة التي تم ايجادها من الخطوة (١) وذلك من خلال تمريرها عبر عتبة (Thresholding) ومن ثم نحسب المعاملات المعدّلة w^* .

٣. اخيراً نجد تقدير الدالة f بواسطة ايجاد معكوس التحويل المويجي المتقطع (IDWT) وفق الصيغة الآتية [5]

$$\hat{f}(x) = W^T w^* \quad ..... \quad (13)$$

* المقصود بالحجم الدايديكي هو ان حجم العينة $n = 2^J$: $J=0,1,2,...$ أي ان $n = \{(2^3 = 8), (2^4 = 16), (2^5 = 32)\}$

٢-٤ قوانين العتبة

ان الخطوة الثانية من خطوات تقدير دالة الانحدار المويجي هي ازالة التشويش الموجود في الاشارة عن طريق حد العتبة ، وباستخدام التحويل المويجي وتحديداً الخطوة الثانية منه يتم وضع عتبة ترددية مناسبة بحيث تلغي هذه العتبة معاملات التشويش وتحافظ على معاملات الاشارة الاصلية لأن معاملات التشويش تكون ذات تردد اقل من تردد معاملات الاشارة الاصلية ، ويوجد هناك نوعان من قوانين العتبة [6]

٢-٤-١ قانون قطع العتبة الناعم (Soft) :

يتم فيها إنهاء القيم ما دون العتبة الى الصفر والمحافظة على القيم الأعلى من العتبة ، وتعرف رياضياً بالعلاقة الاتية :

$$S_\lambda^s(\hat{w}_{jk}) = \begin{cases} 0 & \text{if} \quad |\hat{w}_{jk}| \le \lambda \\ \hat{w}_{jk} - \lambda & \text{if} \quad \hat{w}_{jk} > \lambda \\ \hat{w}_{jk} + \lambda & \text{if} \quad \hat{w}_{jk} < -\lambda \end{cases} \quad ..... \quad (14)$$

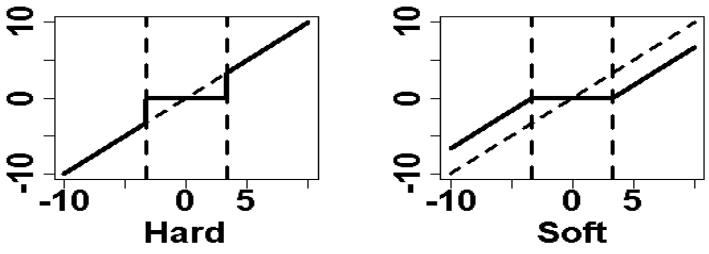
حيث أن λ هي قيمة العتبة (Thresholding value) .

٢-٤-٢ قانون قطع العتبة الصلب (Hard) :

يتم فيها تصغير القيم ما دون العتبة والمحافظة على القيم الأعلى من العتبة ويتم تعريفها رياضياً بالعلاقة التالية :

$$S_{\lambda}^{H}(\hat{w}_{jk}) = \begin{cases} 0 & \text{if} \quad |\hat{w}_{jk}| \le \lambda \\ \hat{w}_{jk} & \text{if} \quad |\hat{w}_{jk}| > \lambda \end{cases} \qquad \ldots\ldots \quad (15)$$

مع ان معادلة العتبة الناعمة أعقد من معادلة العتبة الصلبة إلا ان العديد من الباحثين اكدوا من خلال تجاربهم ان نتائج استخدام العتبة الناعمة أفضل من نتائج العتبة الصلبة [6].



شكل (3) يمثل حد العتبة الصلب والمرن

٢-٥ قيمة العتبة

كما تم الذكر ان الخطوة الثانية من خطوات التقدير باستخدام الانحدار المويجي هي حد العتبة وان الجزء المهم في هذه الخطوة هو اختيار قيمة العتبة λ ،وفيما يأتي شرح موجز لبعض الطرائق المستخدمة في اختيار قيمة العتبة

٢-٥-١ العتبة الشاملة (Universal Thresholding)

طريقة العتبة الشاملة قدمت من قبل **Donoho and Jonstone** والمعطاة على وفق الصيغة الاتية [6]

$$\lambda_{universal} = \sigma\sqrt{2\log(n)} \qquad \ldots\ldots \quad (16)$$

حيث ان :

n : عدد نقاط البيانات الكلي (مكافأة لعدد معاملات المويجي) .

σ : الانحراف المعياري لمستوى التشويش والذي يكون على الأغلب غير معلوم ويمكن استبداله بتقدير حصين هو $\hat{\sigma}$ وهذا التقدير الحصين هو وسيط الانحرافات المطلق **(Median Absolute Deviation)** لمعاملات المويجي عند المستوى الاول **(Finest)** ($j = \log(n) - 1$)

$$\hat{\sigma} = \frac{median(|w_{J-1,k} - median(w_{J-1,k})|)}{0.6745} \qquad \ldots\ldots \quad (17)$$

٢-٥-٢ العتبة (Sure Thresholding)

هذه الطريقة قدمت بواسطة **Donoho and Jonston** [6] والتي استندت في اختيار قيمة λ على تقليل تقدير المخاطرة غير المتحيزة لـ(**Stein Unbiased Risk Estimation**) (**SURE**) لكل مستوى مويجة j .

حيث اعتبر الباحثان ان معاملات المويجة عند كل مستوى على حدة كمشكلة تقدير متعدد متغيرات طبيعي مستقل ، حيث اوضح Stein انه نلك تقدير غير متحيز للمخاطرة هو

$$SURE(\lambda_j, d_{jk}) = N - 2\sum_{k=1}^{N} I\left(|d_{jk}| \le \lambda_j \right) + \sum_{k=1}^{N} \min\left(|d_{jk}| \le \lambda \right)^2 \quad \ldots\ldots \quad (18)$$

وعليه فان قيمة العتبة لـ (SURE) يمكن إيجادها من الصيغة التالية

$$\lambda_{j,SURE} = \arg\min_{0 \le \lambda \le \sqrt{2\log N}} SURE(\lambda_j, d_j k) \quad \ldots\ldots \quad (19)$$

٢-٥-٣ طريقة التقاطع الشرعية (Two Fold Cross Validation) [14,15]

ان هذه الطريقة لا يمكن تطبيقها مباشرةً على طرائق التحويل المويجي السريعة والسبب في ذلك هو اشتراط هذه التحويل أن تكون البيانات ذات حجم دايديكي أي $n=2^J$ ، ومن ثم فانها تقوم باستبعاد بعض نقاط البيانات مما يؤدي الى اختلال احد شروط تطبيق التحويل المويجي المتقطع ، ولذلك اقترح Nason طريقة (Two Fold C.V) يقوم من خلالها باستبعاد نصف البيانات حتى يضمن بقاء حجم العينة $n=2^J$ ، وسيتم توضيحها وفق الصيغة الآتية :

ليكن لدينا $y_1^0, y_2^0, \ldots, y_{n/2}^0$ تمثل نقاط البيانات الفردية و $y_1^E, y_2^E, \ldots, y_{n/2}^E$ تمثل نقاط البيانات الزوجي

وليكن $\hat{f}^{\,O}$, $\hat{f}^{\,E}$ يمثل التقديرات المويجية لنقاط البيانات الفردية والزوجية تواليا .

باستخدام حذف بيانات المؤشر الفردي ، فان البيانات الفردية المشوشة تكون وفق الصيغة الآتية :

$$\tilde{y}_i^0 = \begin{cases} \frac{1}{2}(y_{2i-1} + y_{2i+1}) & , \ i=1,2, \ldots, n/2-1 \\ \frac{1}{2}(y_{n-1} + y_i) & , \ i=n/2 \end{cases} \quad \ldots\ldots \quad (20)$$

أما الصيغة للبيانات المشوشة الزوجية تكون :

$$\tilde{y}_i^E = \begin{cases} \frac{1}{2}(y_{2i-2} + y_{2i}) & , \ i=2, \ldots, n/2 \\ \frac{1}{2}(y_n + y_2) & , i=1 \end{cases} \quad \ldots\ldots \quad (21)$$

أما التقدير النهائي لطريقة (C.V) للمخاطرة $\mu(\lambda)$ هو

$$\mu(\lambda) = \sum \left\{ \left(f_{\lambda,j}^{E}\left(\tfrac{2i}{n}\right) - \tilde{y}_i^{O} \right)^2 + \left(\hat{f}_{j,\lambda}\left(\tfrac{2i-1}{n}\right) - \tilde{y}_i^{E} \right)^2 \right\} \quad \ldots\ldots \quad (22)$$

إذا $\lambda_{n/2}$ تقلل من $\mu(\lambda)$ إذن قيمة العتبة النهائية تعطى وفق الصيغة التالية :

$$\lambda_n = \left(1 - \frac{\log 2}{\log n} \right)^{-1/2} \lambda_{n/2} \quad \ldots\ldots \quad (23)$$

٢-٦ الانحدار المويجي متعدد الحدود الموضعي Local Polynomial Wavelet Regression

طريقة الانحدار المويجي متعدد الحدود الموضعي قدمت من قبل Oh & Lee (2005) كتحسين لتعديل الحدودية في الانحدار المويجي ، اقترح باستخدام نموذج متعدد حدود موضعي من الدرجة الثانية مناسب $\hat{f}_{Lp}$ ، مقدار الانحدار المويجي متعدد الحدود الموضعي $\hat{f}_{LPWR}(x)$ يمكن ان يكتب كالآتي [18]

$$\hat{f}_{LPWR}(x) = \hat{f}_{LP}(x) + \hat{f}_{W}(x) \quad \ldots\ldots \quad (24)$$

كما موضح في Oh & Lee (2005) بأنه $\hat{f}_{LPWR}(x)$ حسبت خلال خوارزمية تكرارية مستوحاة من خوارزمية (back - fitting) لـ Has tie & Tibshirani (1990) .

وفيما يأتي خطوات مختصرة لايجاد مقدر الانحدار المويجي متعدد الحدود الموضعي $\hat{f}_{LPWR}(x)$:

١- نختار مقدر اولي $\hat{f}_0$ لـ f وليكن $\hat{f}_0^{\,*} = \hat{f}_{LPWR}$.

٢- لـ $(k = 100) : j = 1,2,\ldots\ldots,k$ تكرر الخطوات الاتية :

a. نطبق الانحدار المويجي على البواقي $y_i - \hat{f}_{LPWR}$ ونحسب $\hat{f}_w^{\,j}$.

b. نقدر $\hat{f}_{LP}^{\,j}$ بواسطة ايجاد انحدار متعدد الحدود الموضعي للـ $y_i - \hat{f}_w^{\,j}$.

٣- نتوقف اذا $\hat{f}_{LPWR} = \hat{f}_{LP}^{\,j} + \hat{f}_w^{\,j}$ يتقارب .

٣- نموذج والية الفقدان

ان النموذج وكما ذكرنا آنفاً هو نموذج انحدار لا معلمي (١) ، وبناءاً عليه فان اساس الاستدلال يبدأ بافتراض عينة عشوائية من (x_i, y_i) والتي يمكن اعادة كتابتها في حالة وجود بيانات مفقودة ، حيث ان (x_i) قد تم افتراضها تامة المشاهدة ، واما (y_i) فانها تكون معتمدة على مؤشر δ_i بحيث ان $\delta_i = 0$ اذا متغير الاستجابة (y_i) يكون مفقوداً ، وعدا ذلك $\delta_i = 1$.

$$\delta_i = \begin{cases} 0 & \text{if} \quad y_i \text{ is missing} \\ 1 & \text{if} \quad \text{o.w.} \end{cases}$$

اما آلية الفقدان فهي الجزء الذي يوضح العلاقة بين احتمالية فقدان القيمة لمتغير ما مع بقية المتغيرات في مجموعة من البيانات .

ويوجد في ادبيات الاحصاء ثلاثة انواع من آليات الفقدان شائعة الاستخدام وهي :

١- الفقدان العشوائي التام MCAR .

٢- الفقدان العشوائي MAR .

٣- الفقدان غير العشوائي NMAR .

وسيتم استخدام آلية الفقدان MAR والذي يتطلب احتمالية الآلية والتي يتم كتابتها وفق الصيغة التالية [11]:

$$p(\delta = 1 / y, x) = p(\delta = 1 / x) = p(x) \quad \ldots\ldots \quad (25)$$

---

*التقدير الاولي $\hat{f}_0$ يمكن ايجاده باستخدام دالة (supsmn) (المتاحة في R Kernsmooth Package) ، بينما معلمة التمهيد سحبت بواسطة c.v او طريقة الملئ المباشر (المتاحة عادة في Kernsmooth Package) .

٤- بعض طرائق تعويض البيانات المفقودة

٤-١ تعويض المتوسط – الوسيط Mean – Median Imputation

في هذه الطريقة يتم استبدال القيم المفقودة بالمعدل لمتغير الاستجابة ، ومن خلال استخدام المتوسط الا انه يتأثر بالقيم المتطرفة في بعض الحالات ، ومن ثم يمكن استخدامه بالوسيط .
في حالة معالجة جزء كبير من المشاهدات أي ان نسبة الفقدان عالية اقترح الباحث **Chatterjee (2009)** اضافة قيم اضافية عن طريق توليد بيانات بشكل عشوائي من التوزيع الطبيعي بمتوسط $(M = Median)$ وتباين صغير $(\sigma^2 = 0.01)$ نوع آخر من تعويض المتوسط يسمى تعويض المتوسط العشوائي والذي يعرف بـ **(zero order regression) (ZOR +)** ، والذي يمكن استخدامه لملئ المشاهدات المفقودة لمتغير الاستجابة (y_i) وببساطة فان صيغته هي عبارة عن معدل قيم (y_i) مضافاً له قيمة عشوائية مولدة من توزيع طبيعي $\bar{y} + N(0, \sigma^2)$ وبالتالي ان القيمة التقديرية لكل (y_i) مفقودة تكون على وفق الصيغة الاتية **[17]**

$$y_i = \bar{y} + N(0, \sigma^2) \quad \ldots\ldots \quad (26)$$

حيث ان (σ^2) التباين للبواقي لنموذج الانحدار ، انظر **Nittner (2003)** **[17]**

٤-٢ تعويض التمهيد اللامعلمي NonParametric Bootstrap Imputation

قدمت هذه الطريقة لأول مرة عن طريق **Efron (1979)** ومنذ ذلك الحين تتالت بحوث كثيرة واكثر تعقيداً حول طرائق التمهيد وبالخصوص حول معالجة مشكلة البيانات المفقودة وفيما يلي الخطوات الاساسية لهذه الطريقة لتقدير قيم متغير الاستجابة المفقودة **[1,2]**.

a. استبدال كل القيم المفقودة لمتغير الاستجابة بقيمة $(\bar{y})$.

b. سحب ***B*** من العينات الممهدة **(Bootstrap Samples)** المستقلة وحساب المتوسط لكل عتبة وليكن $(\bar{y}_i \quad i = 1,2,\ldots\ldots, B)$.

c. حساب متوسط التمهيد الكلي والتباين كالآتي

$$\bar{y}_b = (\sum_{i=1}^{B} \bar{y}_i)/B \quad , S_b^2 = \frac{1}{B-1}\sum_{i=1}^{B}(\bar{y}_i - \bar{y}_b)^2 \quad \ldots\ldots \quad (27)$$

d. نولد ***m*** من نقاط البيانات عشوائياً من التوزيع الطبيعي بمتوسط $(M = \bar{y}_b)$ وتباين $(\sigma^2 = S_b^2)$ لاستبدال ***m*** من المشاهدات المفقودة .

٤-٣ طريقة اقرب مجاور k- Nearest Neighbor Method

وهي طريقة تقدير لامعلمية تعتمد على التقدير اللبي بشكل اساسي الا ان الفرق ما بينها وبين التقديرات اللبية هو انه في المتغيرات اللبية ان $\hat{m}_h(x)$ تعرف لمعدل موزون لمتغيرات الاستجابة في مجاورات ثابتة حول x تحدد بواسطة الدالة اللبية وعرض الحزمة .
اما مقدرات اقرب مجاور $(k - NN)$ فاختلافها ان المجاورات مختلفة ، حيث ان هذه المجاورات تعرف من خلال متغيرات x والتي تكون مابين k اقرب مجاور لـ x ، والتي يتم حسابها بطرائق متعددة كطريقة المسافة الاقليدية ومسافة مهلونوبس وغيرها من طرائق حساب المسافة بين المجاورات ، وان هذا المقدر قدم من لدن **Stone (1977)** و **Mack (1981)** وفي مجال تقدير دالة الكثافة من قبل **Hart & Cover (1967)** .

وان الصيغة العامة لهذا المقدر في تقدير دالة الانحدار تعرف على وفق الصيغة الآتية [8,3]:

$$\hat{m}_{kNN}(x) = n^{-1}\sum_{i=1}^{n} wk_i(x)y_i \qquad \ldots\ldots \quad (28)$$

حيث ان $wk_i(x)$ تكون على وفق الصيغة الاتية:

$$w_i(x) = \frac{k\left(\frac{x_i - x}{H_n(x)}\right)\delta_i}{\sum_{i=1}^{n} k\left(\frac{x_i - x}{H_n(x)}\right)\delta_i} \qquad \ldots\ldots \quad (29)$$

إذ $k(.)$ تمثل دالة لبية مقيدة وغير سالبة وتحقق $k(u) = for\ all\ |u| > 1$ ، واما الرمز $H_n(x)$ يمثل المسافة الاقليدية (Euclidian Distance) بين x و k_n اقرب مجاور من بين قيم x_i ، حيث يتم تقدير قيم y_i المفقودة وفق الصيغة الاتية المقترحة من قبل الباحث Cheng [16,4]

$$y_i^* = \delta_i y_i + (1-\delta_i)\hat{m}_{k-NN}(x) \qquad \ldots\ldots \quad (30)$$

٥- طرائق التقدير

٥-١ طريقة الانحدار متعدد الحدود الموضعي التكراري التمهيدي

Bootstrap -Itrative Local Polynomial Wavelet :

قام الباحثان A. M. Taher & M. T. Ismail بتوظيف طريقة الباحثين Hee – Seok -oh & Thomas c. M. Lee [18] من خلال استخدام نموذج انحدار متعدد حدود موضعي هجين ، و اقترح الباحثان استخدام طريقة الانحدار اللامعلمي المقدمة من قبلهما ، وان الفكرة الأساسية لهذه الطريقة هي اعتماد تنبأ متعدد الحدود المكرر في استبدال القيم المفقودة مع ما يقاربها من قيم متوقعة من ذلك النموذج مضافاً اليها خطأ عشوائي،وتتلخص خطوات هذه الطريقة على وفق الآتي [2,1,18]:

١- نبدأ اولاً بتقدير اولي للبيانات المفقودة مستخدمين طريقة التمهيد اللامعلمي الموضحة في (٢-٤) [2,1,19] :

٢- وبعد ان اصبحت بيانات متغير الاستجابة y_i كاملة ، نجد القيم المتوقعة $\hat{f}$ باستخدام انحدار متعدد الحدود الموضعي ، ومن ثم نستبدل التقديرات السابقة للقيم المفقودة مع تلك القيم المتوقعة من النموذج المذكور

$$\hat{f}(x; p; h) = e_1'(X'_{p,x}W_xX_{p,x})^{-1}(X'_{p,x}W_xX_{p,x}) \qquad \ldots\ldots \quad (32)$$

هنا ***p*** تشير الى درجة نموذج متعدد الحدود الموضعي ، بينما ***h*** تمثل معلمة التمهيد باستخدام طريقة (Plug-In) والموضحة في الصيغة الاتية:

$$h_{AMISE} = C_1(K)\left[\frac{\sigma^2(b-a)}{Q_{22}.n}\right]^{1/5} \quad \ldots\ldots \quad (33)$$

إذ ان

$$Q_{rs} = \int_s m^{(r)}(x)m^{(s)}(x)\,f(x)dx$$

إذ ان $C_1(K)$ تمثل قيمة ثابتة معينة تعتمد على دالة اللب المستعملة ، بينما $m^{(r)}m^{(s)}$ تمثل المشتقة الثانية لدالة الانحدار في حين $f(x)$ تمثل دالة كثافة احتمالية .

أما W_x هي مصفوفة قطرية للاوزان مع

$$W_i \equiv K_h(x_j - x) \equiv \{K[(x_j - x)/h]\} \quad \ldots\ldots \quad (34)$$

e_i هي متجه من درجة $(p+1)\times 1$ اول ادخال له يساوي ١ ، اما البقية فهي اصفار ، و ***K*** هي الدالة اللبية .

٣- لـ $j = 1,2,\ldots\ldots$ تكرر الخطوة السابقة حتى نصل الى التقارب مع متوسط مربعات الخطأ للنموذج المذكور .

٤- وأخيراً فان القيم المقدرة للبيانات المفقودة النهائية من متغير الاستجابة y_i هي عبارة عن القيم التي تم الحصول عليها من آخر تكرار مضافاً لها حد الخطأ العشوائي الذي يتوزع توزيعاً طبيعياً بمتوسط صفر وتباين مساوي الى $mse^{(j)}$ ، والمقصود به هنا متوسط مربعات الخطأ من آخر تكرار لنموذج الانحدار متعدد الحدود الموضعي .

وبعد ان تم تقدير قيم متغير الاستجابة المفقودة وفق الطريقة اعلاه نطبق طريقة الباحثان **T. M. Lee & Hee-seokoh** حول الانحدار المويجي متعدد الحدود الموضعي لايجاد التقدير النهائي للدالة المراد تقديرها والموضحة في (٢-٦) .

٢-٥ الطريقة المقترحة K-Nearset Nighbor Polynomial Wavelet [14,13,3]

ان فكرة هذه الطريقة مستمدة من فكرة الباحثين **Lee, M. C. T. , Meng, L. X.** [10] إذ استخدم الباحثان طرائق تعويض تقليدية كخطوة اولى للتقدير تمثلت باستخدام طرائق المتوسط والتعويض المتعدد ، ومن ثم تطبيق طريقة الانحدار المويجي على البيانات بعد ان اصبحت تامة ، الا انه في هذا البحث تم استخدام طريقة التعويض اللامعلمي (K-NN) كخطوة اولى لتعويض البيانات المفقودة لمتغير الاستجابة y_i ومن ثم تطبيق التقديرات المويجية على البيانات التامة ، إذ تم تسمية هذه الطريقة بطريقة (NNPW) ، إذ انه بعد معالجة القيم المفقودة باستخدام الصيغة الموضحة في المعادلة (٣٠) ، إذ أن $\hat{m}_{K-NN}$ يتم الحصول عليها من المعادلة (٢٨) وبعد ذلك نطبق الخطوة الثانية من التقدير والتي تتلخص بايجاد دالة متعدد الحدود $\hat{f}p(x)$ وذلك عن طريق تقدير معالم هذا النموذج ذات الدرجة الثانية باستخدام طريقة المربعات الصغرى ، حيث ان

$$\hat{y}_i = \hat{b}_0 + \hat{b}_1 x_i + \hat{b}_2 x_i^2 \quad \ldots\ldots \quad (35)$$

إذ ان

$$\hat{y}_i = \hat{f}p(x)$$

بعد ذلك نجد البواقي عن طريق الصيغة الاتية :

$$e_i = \hat{y}_i - \hat{fp}(x) \qquad \ldots\ldots \quad (36)$$

واخيراً نطبق الانحدار المويجي التقليدي الموضح في (٢-٣) على البواقي وباستخدام قيم عتبة مختلفة للحصول على التقدير النهائي لدالة الانحدار

$$\hat{f}NNpw(x) = \hat{f}wNN(x) + \hat{f}p(x) \qquad \ldots\ldots \quad (37)$$

٦- المحاكاة

ان تطبيق ما جاء في الجانب النظري يتم عن طريق استخدام اسلوب المحاكاة (Simulation) من اجل محاكاة اكبر قدر من الحالات التي يمكن ان تواجهنا في الواقع العملي بغية الوصول الى نتائج اكثر عمومية ، وان اللجوء الى استخدام اسلوب المحاكاة كان لعدة اسباب اهمها ان هذا الاسلوب يوفر لنا اختصاراً في الوقت والكلفة بجميع اشكالها المادية والبشرية الذي تتطلبه التجارب الواقعية كالتجارب الطبية والفلكية وغيرها من التجارب التي تحتاج الى وقت كبير وتكلفة باهضة جداً .

٦-١ توليد المتغيرات

لانموذج الانحدار اللامعلمي في المعادلة (1) حيث ان $x_i = i/n$ ، $i = 1,\ldots\ldots,n$ حيث x_i ذات مسافات متساوية ضمن الفترة [0,1] وان e_i يتوزع توزيع طبيعي بمتوسط صفر وتباين ثابت σ ، $f(x_i)$ دالة الانحدار التي يراد تقديرها في ظل وجود فقدان في قيم متغير الاستجابة y ، وعليه لتوليد البيانات لـ (x_i, y_i) وفق الانموذج الانحدار اللامعلمي (١) مع دوال الاختبار الموضحة في (3-4) ، حيث ان المتغير المعتمد يتم توليده من خلال دوال الاختبار الموضحة في (3-4) مضافاً اليه حد الخطأ e_i .

ولتنفيذ تجارب المحاكاة جرى استخدام مستويات مختلفة من العوامل الآتية :

١- حجوم العينات n ، حيث تم استخدام ثلاثة حجوم للعينات وهي $2^6 = 64$ ، $2^7 = 128$ ، $2^8 = 256$ كون حجم العينة هنا يجب ان يكون $n = 2^j$ ، حيث ان J عدد صحيح موجب .
٢- نسب الاشارة الى التشويش (SNR) حيث تم استخدام اثنان من نسب التشويش SNR= 5,10 .
٣- دوال الاختبار $f(x_i)$ ، حيث جرى استخدام ثلاثة دوال اختبار مختلفة موضحة في ادناه .
٤- نسب الفقدان ، حيث تم استخدام نسبتي فقدان وهي (25% , 15%) .
٥- درجة متعدد الحدود من الدرجة الثانية أي $d = 2$.

٦-٢ دوال الاختبار [14] Test Function

وتتميز هذه الدوال كونها دوال اختبار قياسية ونموذجية ومناسبة لاستخدامها في تجارب المحاكاة كونها صممت لتعرض مجموعة من الظواهر التي غالباً ما تحدث في مجموعة البيانات المأخوذة من الواقع العملي وان هذه الدوال تكون معرفة على الفترة [0,1] وسوف نعرض تلك الدوال [6] وكالآتي :

١- دالة Doppler

$$f_1(x) = \{x(1-x)\}^{1/2} \sin\{2\pi(1+\varepsilon)/(x+\varepsilon)\}, \varepsilon = 0.05 \qquad \ldots\ldots \quad (38)$$

٢- دالة **Heavisine**

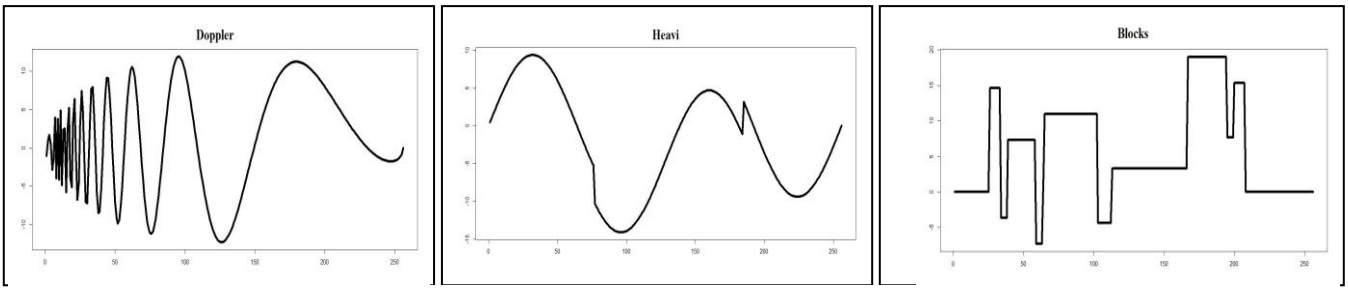
$$f_2(x) = 4\sin 4\pi x - \mathrm{sgn}(x-0.3) - \mathrm{sgn}(0.72-x) \quad \ldots\ldots \quad (39)$$

٣- دالة **Blocks**

$$f_3(x) = \sum h_j k(x - x_j), \quad k(x) = \{1 + \mathrm{sgn}(x)\}/2 \quad \ldots\ldots \quad (40)$$

$$(x_j) = (0.1, 0.13, 0.15, 0.23, 0.25, 0.40, 0.44, 0.65, 0.76, 0.78, 0.81)$$

$$(h_j) = (4, -5, -4, 5, -4.2, 2.1, 4.3, -3.1, 2.1, -4.2)$$



شكل (4) : دوال الاختبار

٦-٣ تجارب المحاكاة

هنا سيتم وصف وافي لتجارب المحاكاة بحسب دوال الاختبار المستخدمة في نموذج الانحدار اللامعلمي (1) لتوليد البيانات (x_i, y_i) ، حيث اجريت تجارب المحاكاة لتقدير الانموذج من خلال توظيف عدد من طرائق معالجة القيم المفقودة في متغير الاستجابة ولحجوم عينات ونسب اشارة مختلفة ، اذ تم توظيف طرائق المستخدمة في التقدير وهي ست طرائق تم تكرار التجربة (500) بثبات جميع العوامل عدا المتغير العشوائي والذي يعاد توليده عند تكرار كل تجربة .

في كل مرة يجري فيها توليد للبيانات ، يتم استخدام آلية الفقدان العشوائي MAR لفقدان بعض مشاهدات متغير الاستجابة عشوائية على وفق الصيغة الموضحة في المعادلة (25).

<u>اولاً : تطبيق طريقة (BILPW)</u> : ان ملخص عمل هذه الطريقة والموضحة في الخوارزمية رقم (١) يكون على وفق الخطوات الاتية :

١ـ بعد توليد كلاً من المتغير التوضيحي والمتغير المعتمد الموضحة في (1-4) و اجراء عملية الفقدان العشوائي على متغير الاستجابة وفق المعادلة (25) يتم معالجة قيم y_i المفقودة باستخدام طريقة (BILP) الموضحة في (1-3) وبذلك يتم الحصول على قيم y_i المفقودة .

٢- يتم ايجاد قيمة $\hat{flp}(x)$ عن طريق المعادلة (32) وبعد ذلك نجد التقدير النهائي للقيم المفقودة .

٣- نجد تقدير اولي $\hat{f}^0$ لـ f بحيث ان $\hat{flp} = \hat{f}^0$.

٤- نكرر الخطوات التالية لـ j من المرات حيث ان $j=1,2,.....$.

❖ نطبق التقدير المويجي على البواقي $e_i = y_i - \hat{f}lp$ للحصول على $\hat{f}_w^j$.

❖ نجد تقدير دالة متعدد الحدود الموضعي للبواقي $y_i - \hat{f}_w^j$ للحصول على $\hat{f}^j lp$ والتي تكون ثلاث تقديرات أي بعدد نوع قيم العتبة المستخدمة .

❖ اخيراً نجد التقدير النهائي لدالة الانحدار غير المعلومة باستخدام الانحدار المويجي متعدد الحدود الموضعي وفق الصيغة $\hat{f}_{LPW} = \hat{f}_{Lp}^j + \hat{f}_w^j$ من آخر تكرار ، وهنا ايضاً تكون عدد قيم الدالة المقدرة ثلاث قيم أي بعدد نوع قيم العتبة المستخدمة ، الجدير بالذكر انه تم استخدام درجة متعدد $d=2$ وتكرار (١٠٠) ، حيث تم اختيار دالة (Kernal) هي دالة (Epenchenkov) ، واما معلمة التمهيد المستخدمة في هذه الطريقة فقد تم استخدام طريقة (Plug-in dircet) الموضحة في المعادلة (٣٣) .

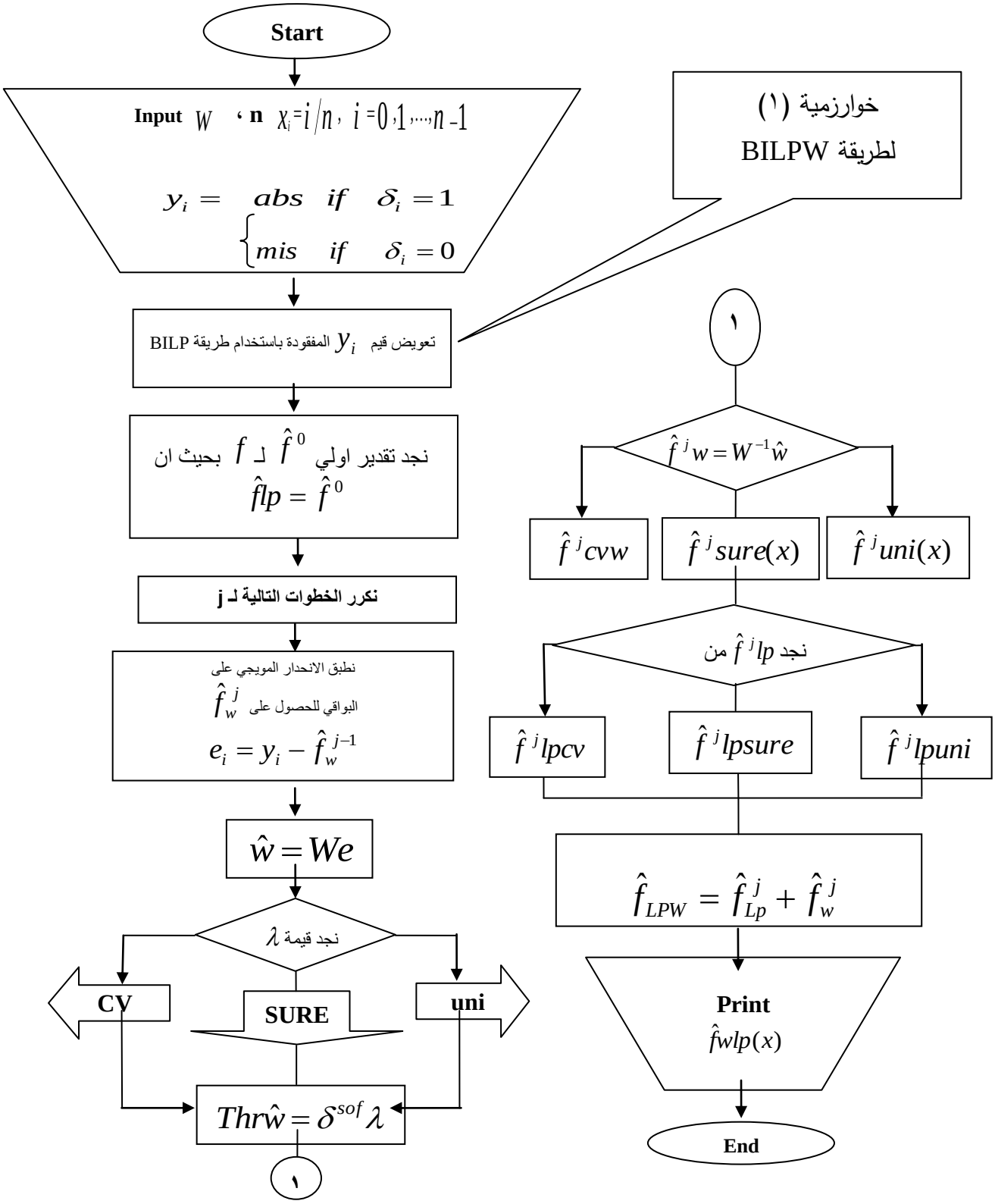
ثانياً : تطبيق طريقة (NNPW) : ان ملخص عمل هذه الطريقة والموضحة في الخوارزمية رقم (٢) يكون على وفق الخطوات الاتية :

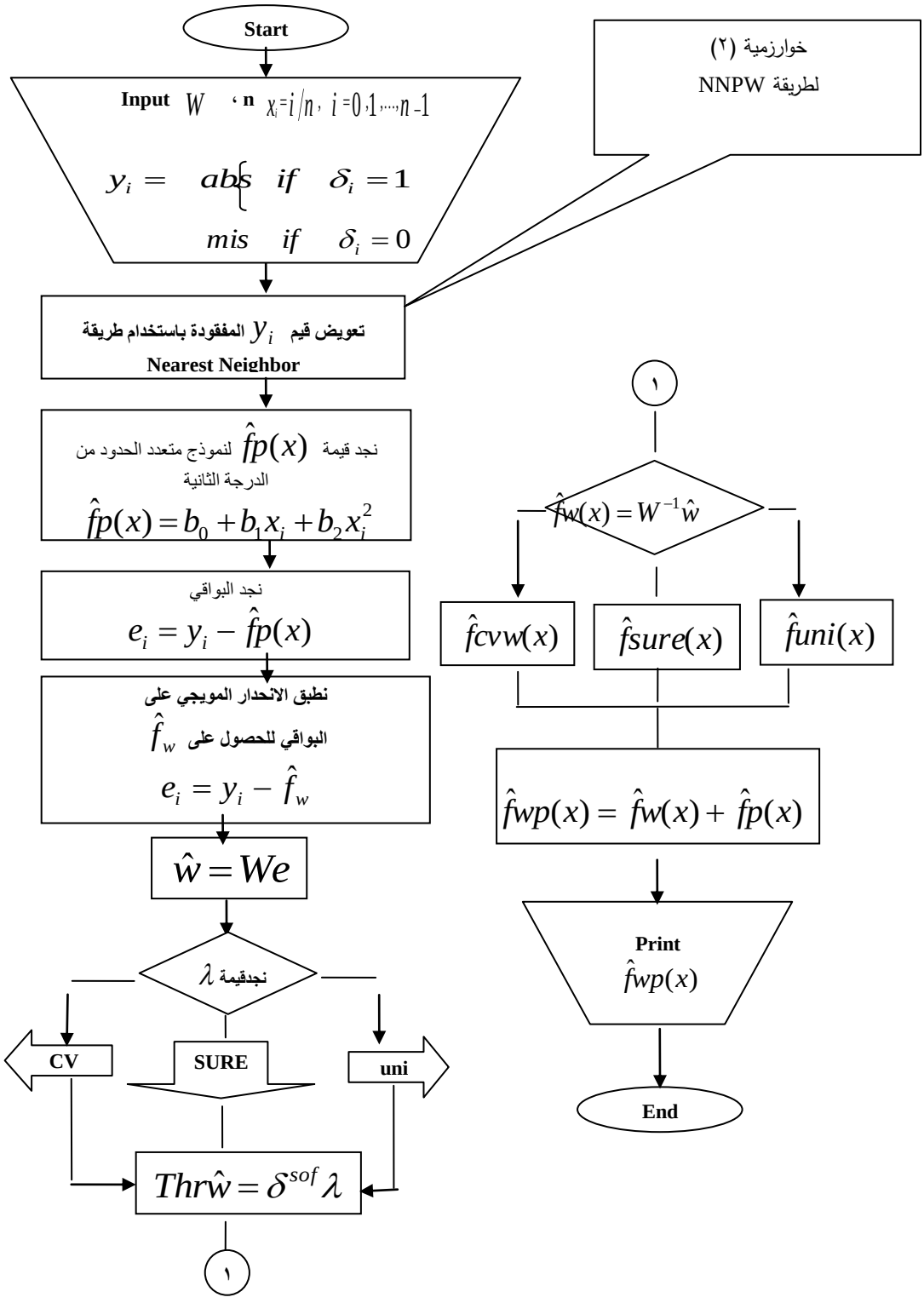
١- تطبيق طريقة تعويض الانحدار اللامعلمي باستخدام طريقة (KNN) ، حيث يتم تطبيق الصيغة الموضحة في المعادلة (٣٠) للحصول على قيم متغير الاستجابة المفقودة .

٢- نجد قيمة $\hat{f}p(x)$ لنموذج متعدد الحدود من الدرجة الثانية باستخدام المربعات الصغرى .

٣- نجد البواقي عن طريق الصيغة الموضحة في المعادلة (٣٦) .

٤- نطبق الانحدار المويجي على البواقي وخطواته نفس ماتم في الطريقة الاولى لغاية الحصول على التقدير النهائي ، الجدير بالذكر انه عند استخدام طريقة (NNPW) الموضحة في (2-3) ، تم استخدام دالة (Epenchenkov) ومسافة مهلونوبينس ، اما درجة المتعدد فهي $d=2$


Start
Input W ، n $x_i = i/n$, $i = 0,1,...,n-1$
$y_i = abs \quad if \quad \delta_i = 1$
$mis \quad if \quad \delta_i = 0$
خوارزمية (١)
لطريقة BILPW
تعويض قيم y_i المفقودة باستخدام طريقة BILP
نجد تقدير اولي $\hat{f}^0$ لـ f بحيث ان
$\hat{f}lp = \hat{f}^0$
نكرر الخطوات التالية لـ j
نطبق الانحدار المويجي على
البواقي للحصول على $\hat{f}_w^j$
$e_i = y_i - \hat{f}_w^{j-1}$
$\hat{w} = We$
نجد قيمة λ
CV
SURE
uni
$Thr\hat{w} = \delta^{sof}\lambda$
١
١
$\hat{f}^j w = W^{-1}\hat{w}$
$\hat{f}^j cvw$
$\hat{f}^j sure(x)$
$\hat{f}^j uni(x)$
نجد $\hat{f}^j lp$ من
$\hat{f}^j lpcv$
$\hat{f}^j lpsure$
$\hat{f}^j lpuni$
$\hat{f}_{LPW} = \hat{f}_{Lp}^j + \hat{f}_w^j$
Print
$\hat{f}wlp(x)$
End


Start
Input W ، n $x_i=i/n$، $i=0,1,\ldots,n-1$
$y_i = \begin{cases} abs & if \quad \delta_i = 1 \\ mis & if \quad \delta_i = 0 \end{cases}$
خوارزمية (٢)
لطريقة NNPW
تعويض قيم y_i المفقودة باستخدام طريقة
Nearest Neighbor
نجد قيمة $\hat{f}p(x)$ لنموذج متعدد الحدود من الدرجة الثانية
$\hat{f}p(x) = b_0 + b_1 x_i + b_2 x_i^2$
نجد البواقي
$e_i = y_i - \hat{f}p(x)$
نطبق الانحدار المويجي على
البواقي للحصول على $\hat{f}_w$
$e_i = y_i - \hat{f}_w$
$\hat{w} = We$
نجدقيمة λ
CV
SURE
uni
$Thr\hat{w} = \delta^{sof}\lambda$
١
١
$\hat{f}w(x) = W^{-1}\hat{w}$
$\hat{f}cvw(x)$
$\hat{f}sure(x)$
$\hat{f}uni(x)$
$\hat{f}wp(x) = \hat{f}w(x) + \hat{f}p(x)$
Print
$\hat{f}wp(x)$
End

٦-٤ تحليل النتائج

١- دالة الاختبار $f_1(x)$

عند نسبة تشويش SNR=5 نلاحظ تفوق تقدير BILPW باختلاف قيم العتبة عند حجم عينة n=64 ، بينما يتفوق تقدير NNPW عند قيمة عتبة SURE وحجم عينة n=128, 256 ، اما عند SNR=10 فنلاحظ تفوق تقدير NNPW بشكل واضح عند قيمة عتبة SURE وباختلاف حجوم العينات .
اما عند نسبة فقدان ٢٥% فنلاحظ تفوق تقدير NNPW عند قيمة عتبة SURE على جميع التقديرات وباختلاف حجم العينة ونسب التشويش .

٢- دالة الاختبار $f_2(x)$

عند نسب فقدان 25% ,١٥% ونسبة تشويش SNR=5 وباختلاف حجم العينة وقيم العتبة نلاحظ تفوق التقدير BILPW ، بينما عند نسبة تشويش SNR=10 وحجوم عينات (64, 128) فنلاحظ تفوق تقدير BILPW ، بينما عند حجم عينة ٢٥٦ وقيمة عتبة SURE فنلاحظ تفوق تقدير NNPW .

٣- دالة الاختبار $f_3(x)$

عند نسبة فقدان ١٥% ونسبة تشويش SNR=5 وحجم عينة ٦٤ نلاحظ تفوق تقدير BILPW ، بينما عند حجوم عينات (128, 256) وقيمة عتبة SURE تفوق تقدير NNPW .
اما عند نسبة تشويش SNR=10 و حجم عينة ٦٤ فنلاحظ تقارب التقديرات بشكل كبير ، اما عند حجوم عينات (128, 256) وقيمة عتبة SURE تفوق التقدير المقترح NNPW .
اما عند نسبة فقدان ٢٥% ونسبة تشويش SNR=5 وحجوم عينات (64, 128) فنلاحظ تقارب التقديرات بشكل كبير مع افضلية لتقدير BILPW ، بينما نلاحظ تفوق التقدير المقترح NNPW عند قيمة عتبة SURE وحجم عينة ٢٥٦ ، اما عند نسبة تشويش SNR=10 فنلاحظ تفوق التقدير المقترح NNPW عند قيمة عتبة SURE وباختلاف حجم العينة .

SNR (نسبة الاشارة الى التشويش) : هي مقياس يتم بواسطته المقارنة بين قيمة الاشارة وقيمة التشويش (Noise) المحمولة معها او بتعريف آخر هي النسبة بين قيمة الاشارة الى قيمة ما تحتويه من تشويش ويتم حسابها وفق الصيغة الآتية :

$$SNR = \frac{\sigma_{signal}}{\sigma_{noise}}$$

جدول (1)
معيارMSE لمقارنة التقديرات لدالة Doppler المشوّشة ونسبة فقدان ١٥%
لحجوم عينات n=256, n=128, n=64
ونسب إشارة إلى تشويش SNR=10, SNR=5

	Sure		Universal		CV	
	SNR=5	SNR=10	SNR=5	SNR=10	SNR=5	SNR=10
	n=64					
BILPW	٠.٠٠٨٥٤٩٠٤٩	0.005227427	٠.٠٠٨٥٤٩٠٤٩	٠.٠٠٥٧٠٧٧٠٨	٠.٠٠٨٦٦٠٨٠٣	٠.٠٠٥٩٥٠٩٣٣
NNPW	0.01137165	0.003476178	٠.٠١١٣٧١٦٥	0.008026567	0.01039031	0.008571250
	n=128					
BILPW	٠.٠٠٥٩٨٠٧٠٢	٠.٠٠٢٥٦٢١٧١	٠.٠٠٦٢٠١٦٩٧	٠.٠٠٤١٤٤٠٩٠	٠.٠٠٥٢٢٠٨٤٢	٠.٠٠٣٧٥١٠٤٩
NNPW	٠.٠٠٤٣٥٤٥١٩	٠.٠٠٠٧٦٧٤٦٤١	٠.٠١٢٧٨٠٢٠٣	٠.٠٠٧٨٢٦٢٥٢١	٠.٠١٢٤٥٧١٠٩	٠.٠٠٨٤٠٢٧٩٠٩
	n=256					
BILPW	0.005022379	٠.٠٠٢٠١٦٥٥٥	٠.٠٠٥١٦٣٣٩٧	٠.٠٠٤٢٥٧١٤٤	٠.٠٠٣٢٢٨٧٢٦	٠.٠٠٢٥٣٣٤٤٧
NNPW	٠.٠٠٤٥٣٩٧٤٠	٠.٠٠٠٨١٩٩٧٥	٠.٠١٢٠٤٣٧٤٧	٠.٠٠٦٦٨٧٩٤١	٠.٠٠٧٦٠٨٨٨٥	٠.٠٠٤٩٤٧٨٠٩

جدول (2) معيارMSE لمقارنة التقديرات لدالة Doppler المشوّشة ونسبة فقدان ٢٥%
لحجوم عينات n=256, n=128, n=64

	Sure		Universal		CV	
	SNR=5	SNR=10	SNR=5	SNR=10	SNR=5	SNR=10
	n=64					
BILPW	٠.٠٠٧٣٤٤٥١٠	0.007176720	٠.٠٠٧٣٤٧٨٥٧	٠.٠٠٧٢٠١٢٩٢	٠.٠٠٧٠١١٥٣٧	٠.٠٠٦٨٧٦٩٧٥
NNPW	0.001100342	0.002683086	٠.٠٠٩١٧٢٥٤٠	0.007654546	0.010084178	0.007162644
	n=128					
BILPW	٠.٠٠٧٢٤٨٢١٩	٠.٠٠٦٧١٤٧٨٤	٠.٠٠٧٢٤٨٢١٩	٠.٠٠٦٧٥٤٣٢٦	٠.٠٠٦١٩٨٥٣٣	٠.٠٠٥٤٠٩٠٧٢
NNPW	٠.٠٠٥٦٩٢١٨٤	٠.٠٠٢٠٩٣٧٤٢	٠.٠١٣٣٣٨٦٤٠	٠.٠١٥٠١٢٣٧٨	٠.٠١٠٦٦٩٩٣٩	٠.٠١٤٠٠١٤٤٢
	n=256					
BILPW	0.006098292	٠.٠٠٤٧٧٤٦١١	٠.٠٠٦١٠٧٨٨٧	٠.٠٠٥٢١٩٤٩٧	٠.٠٠٥٠٥١٥٩٦	٠.٠٠٤٥٨٧٩٧٦
NNPW	٠.٠٠١٨٠٠٣٨٦	٠.٠٠٢١٥٠٢٨٣	٠.٠١٢٤٩٤٩٠٧	٠.٠١٢٨٥٧٩٧١	٠.٠٠٩٣٢٥٧٧٩	٠.٠٠٩٤٦٨٧٥٣

ونسب إشارة إلى تشويش SNR=10

جدول (3) معيارMSE لمقارنة التقديرات لدالة Heavisine المشوّشة ونسبة فقدان ١٥%
لحجوم عينات n=256, n=128, n=64
ونسب إشارة إلى تشويش SNR=10, SNR=5

	Sure		Universal		CV	
	SNR=5	SNR=10	SNR=5	SNR=10	SNR=5	SNR=10
	n=64					
BILPW	٠.٠٠٣٠٦٦٠٩٦	0.001967682	٠.٠٠٣٠٦٦٠٩٦	٠.٠٠١٩٤٨٧١٩	٠.٠٠٢٩٤٩٣٠٤	٠.٠٠١٦١٦٨١٩
NNPW	0.007191990	0.004968865	٠.٠٠٧١٩١٩٩٠	0.004968865	0.007723249	0.005389867
	n=128					
BILPW	٠.٠٠٢٩١٦٣٥٢	٠.٠٠٢٤١٨٣٤١	٠.٠٠٤٧١٨٣٤٣	٠.٠٠٣٦٧٦٨٢٠	٠.٠٠٤٢٩٦٠٠٧	٠.٠٠٣٠٩٣٤٠١
NNPW	٠.٠٠٧٩٢٦٣٦٧	٠.٠٠٧١٨٢٢٩٥	٠.٠٠٧٩٢٦٣٦٧	٠.٠٠٧١٨٢٢٩٥	٠.٠٠٨٣٦٥٥٠١	٠.٠٠٧٥١٥٦٤٦
	n=256					
BILPW	0.003385318	٠.٠٠١٨٩٥٢٧٨	٠.٠٠٣٦٨٤٤١٤	٠.٠٠٢٣٣٢٤٩٨	٠.٠٠٢٧٣٥١٨٧	٠.٠٠١٨٦٠٠٨٥
NNPW	٠.٠٠٨٣٩٤٧١٤	٠.٠٠٠٨٢٤٥٣٧٦	٠.٠٠٨٣٩٤٧١٤	٠.٠٠٥٤٢٤٢٧٤٣	٠.٠٠٧١٣٣٦٩٧	٠.٠٠٤٣٠٠٦٣٨٣

جدول (4) معيارMSE لمقارنة التقديرات لدالة Heavisine المشوّشة ونسبة فقدان ٢٥%
لحجوم عينات n=256, n=128, n=64
ونسب إشارة إلى تشويش SNR=10, SNR=5

	Sure		Universal			
	SNR=5	SNR=10	SNR=5	SNR=10	SNR=5	SNR=10
	n=64					
BILPW	٠.٠٠٤٤٥١٩٢٥	0.003010090	٠.٠٠٤٢٨٠٥٩١	٠.٠٠٢٩٥٨٩٠٥	٠.٠٠٤٢٦١٥٧٣	٠.٠٠٢٩٧٧٠٧٤
NNPW	0.009904359	0.003827515	٠.٠٠٩٩٠٤٣٥٩	0.003827515	0.009756902	0.004040138
	n=128					
BILPW	٠.٠٠٤٤٠٠٦٩٢	٠.٠٠٤٤٩٢٦٩٧	٠.٠٠٥٢٥٦١٩١	٠.٠٠٥١٩٣٩٢٩	٠.٠٠٣٩٣٠٨٤١	٠.٠٠٤٧١٧٠٣٧
NNPW	٠.٠١٠١٥٣٨٩	٠.٠٠٩٣٤٨٧٠٠	٠.٠١٠١٥٣٨٩	٠.٠٠٩٣٤٨٧٠٠	٠.٠١٠٧٠٨٢١	٠.٠٠٩٥٠٥١١٥
	n=256					
BILPW	0.005476270	٠.٠٠٢٦١٥٩٥٧	٠.٠٠٥٥١١٧٤٢	٠.٠٠٣٢٤٠٦٩٧	٠.٠٠٣٦٥١٩٦٩	٠.٠٠٢٤١٧٢٢٦
NNPW	٠.٠١١٨٩٣٩١	٠.٠٠١٨٢٨١٠٩	٠.٠١١٨٩٣٩١	٠.٠٠٧٩٨٥٣٢٠	٠.٠١٢٠٦١٤٧	٠.٠٠٨٤١٣١٦٣

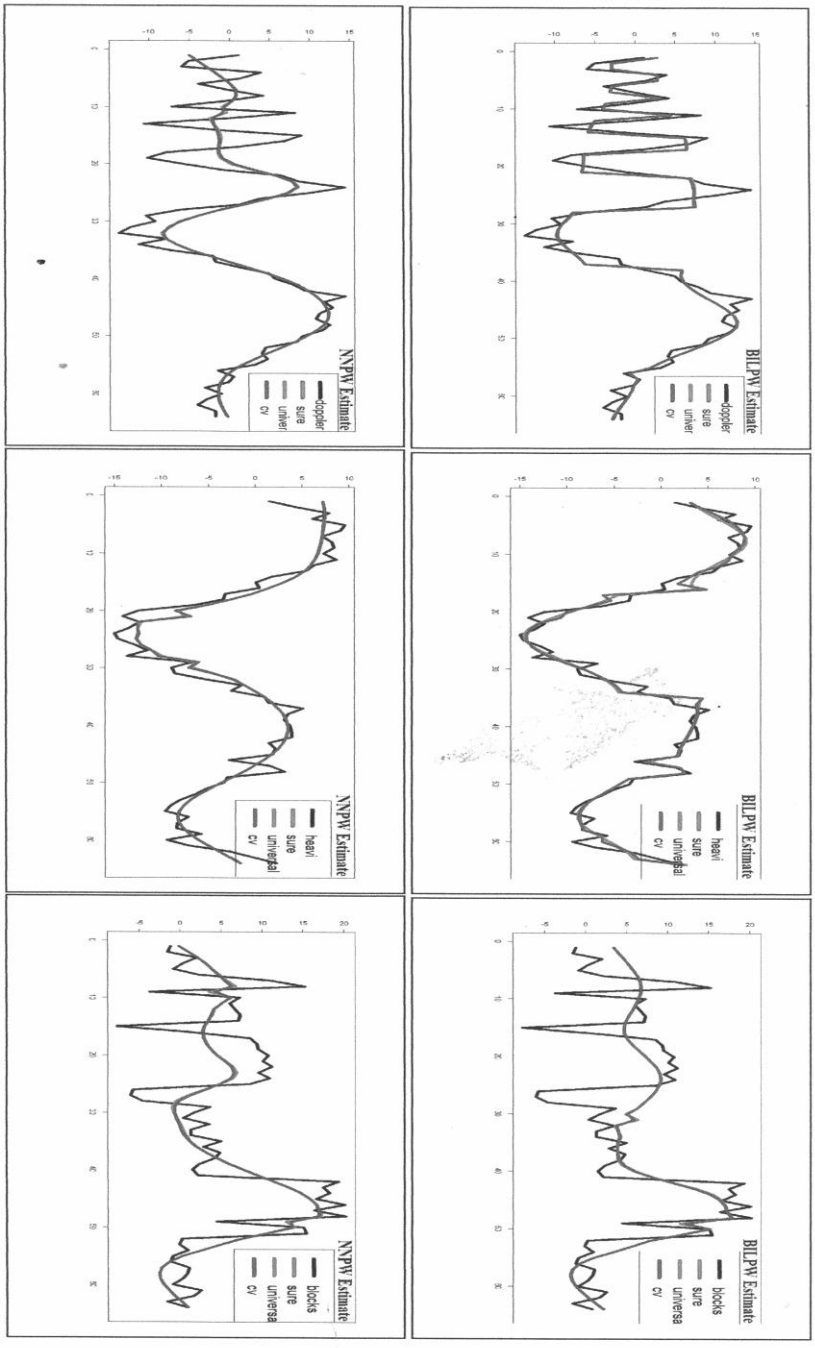
جدول (5) معيارMSE لمقارنة التقديرات لدالة Blocks المشوّشة ونسبة فقدان ١٥%
لحجوم عينات n=256, n=128, n=64
ونسب إشارة إلى تشويش SNR=10, SNR=5

	Sure		universal		CV	
	SNR=5	SNR=10	SNR=5	SNR=10	SNR=5	SNR=10
	n=64					
BILPW	٠.٠١٢٩٢٨٠٦	0.01099811	٠.٠١٢٩٣٧٤١	٠.٠١١٨٥٣٩٢	٠.٠١٣٠٨٨١٤	٠.٠١٢٦٨٠٥٩
NNPW	0.02359827	0.01694785	٠.٠٢٣٥٩٨٢٧	0.01694785	0.02422041	0.01803987
	n=128					
BILPW	٠.٠٠٧٤٧٠٢٧٩	٠.٠٠٧١٧٠٩٧٣	٠.٠٠٩٧٠٥٦٥٨	٠.٠٠٨٧٣٢٦٨٤	٠.٠٠٩٢٦٧١١٢	٠.٠٠٩٣٤٥٩٧٤
NNPW	٠.٠٠٧٢٠٠٠٨١	٠.٠٠١٠٧٤٣٢٧	٠.٠١٩٨٠٦٦٦٨	٠.٠١٢٤٣٧١٠٦	٠.٠١٤٧٢٤٨٦٥	٠.٠٠٩١٢٤٢١٧
	n=256					
BILPW	0.007693139	٠.٠٠٤٩٣٠١٤٤	٠.٠٠٨٧٩٨٦٨٥	٠.٠٠٦٩٢٤٤٧٠	٠.٠٠٦٤٠٨٣٣٦	0.005978513
NNPW	٠.٠٠٣٢٥٧٤٦٨	٠.٠٠١٢٤٤٥٩٠	٠.٠١٥٨٤١١٣٣	٠.٠١٠٠٨٧٩٣٤	٠.٠٠٨٢٩٢٠٠٥	٠.٠٠٨١٢٤٢١٣

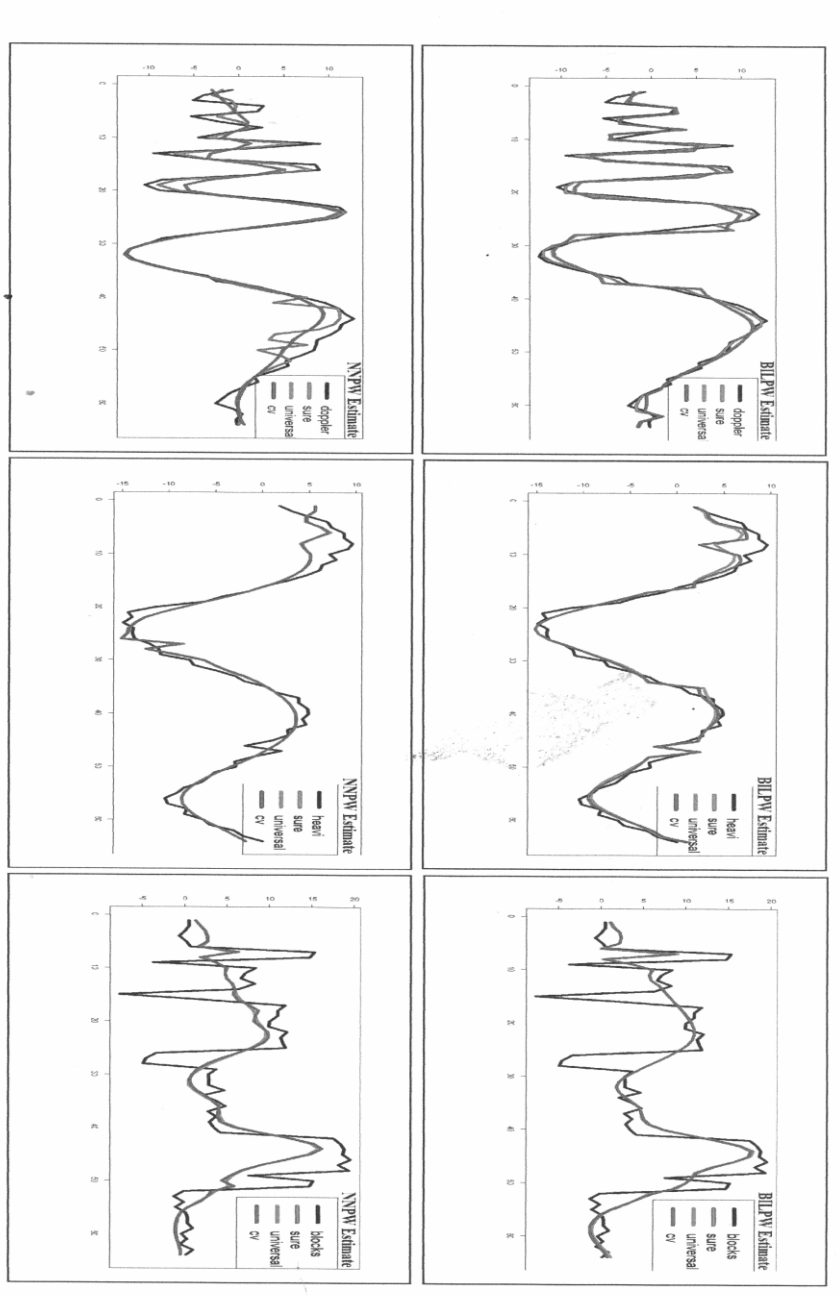
جدول (6) معيارMSE لمقارنة التقديرات لدالة Blocks المشوّشة ونسبة فقدان ٢٥%
لحجوم عينات n=256, n=128, n=64

	Sure		Universal		CV	
	SNR=5	SNR=10	SNR=5	SNR=10	SNR=5	SNR=10
	n=64					
BILPW	٠.٠١٤٨٦٥١١	0.01296007	٠.٠١٤٨٧٠٣٠	٠.٠١٢٩٠٤٨٦	٠.٠١٤٨١٠٧٥	٠.٠١٣١١٢٤٠
NNPW	0.02163477	0.002027726	٠.٠٢١٦٣٤٧٧	0.013938524	0.02023906	0.015150115
	n=128					
BILPW	٠.٠١٢٩٤٠٧٥	٠.٠١٢٢٠٣١١	٠.٠١٢٩٤٨٤٢	٠.٠١٢٨١٢٥٣	٠.٠١٣٠٧٨٧٢	٠.٠١٢٨٣٩٠٣
NNPW	٠.٠١٩٤١٧٥٠	٠.٠٠٢٠١٩٦٩١	٠.٠١٩٤١٧٥٠	٠.٠١٦٠٣٧٦١٩	٠.٠٢٠٦٠٧٤٢	٠.٠١٦٩٣٣٤٣٥
	n=256					
BILPW	0.008343347	٠.٠٠٦٦٤٢١١٣	٠.٠١٠٧٧٩٤٤٦	٠.٠٠٨٠٥٢٦٣٠	٠.٠٠٦٩٦٩٥٤٣	٠.٠٠٦٦٢١٦٨٣
NNPW	٠.٠٠٤٩٧١٢٤٩	٠.٠٠٢٢٤٤٦١٢	٠.٠٢٠١٢٤٤٨٦	٠.٠١٢١٧٢٢٦٠	٠.٠١١٧٧٤٩٢٧	٠.٠٠٨٥٩٧٦٦٥

ونسب إشارة إلى تشويش SNR=10, SNR=5



شكل (5)
نتائج تجارب المحاكاة باستخدام تقديرات الموال (Doppler, Heavi, Blocks) المشوشة عند انواع قيم عتبة مختلفة ونسبة فقدان 15% ونسبة تشويش 5 وحجم عينة 64



شكل (6)
نتائج تجارب المحاكاة باستخدام تقديرات الدالة (Doppler, Heavi, Blocks) المشوشة عند انواع قيم عتبة مختلفة ونسبة فقدان 15% ونسبة تشويش 10 وحجم عينة 64

٧- الاستنتاجات

١- تفوق التقدير BILPW باختلاف قيم العتبة ودوال الاختبار عند نسبة تشويش SNR=5 وحجم عينة ٦٤ .
٢- تفوق التقدير المقترح NNPW عند قيمة عتبة SURE ونسبة تشويش SNR=10 وباختلاف نسب الفقدان ودوال الاختبار وحجوم العينات .
٣- تفوق التقدير المقترح NNPW عند استخدام دالة Doppler ونسبة فقدان ٢٥% وقيمة عتبة SURE وباختلاف نسب التشويش .
٤- تتناقص قيمة MSE بازدياد نسب التشويش وكذلك عند زيادة حجم العينة .

٨- التوصيات

١- نوصي بدراسة طرائق التقدير المويجي بشكل اوسع ليتضمن معالجة البيانات التي تعاني من فقدان في المتغير التوضيحي وكذلك استخدام نموذج خطي عام .
٢- توظيف الطرائق المقترحة واستخدامها في معالجة حالتي الفقدان والبيانات الشاذة في آن واحد.
٣- استخدام طرائق تحويل مويجية اخرى مثل طريقة تحويل الرفع والتحويل الفاني وتوظيفها في حالة حجم العينة الاعتباطي ، ومن ثم مقارنتها مع طريقة التحويل المويجي المتقطع .

المصادر

1. Altaher, M. A. , Ismail, T. M. , (2011), "A New Method on Treating Missing Values in Polynomial Wavelet Regression", Proceedings of the Annual International Conference on Operations Research and Statistics (ORS), copyright © GSTF & ORS , ISBN : 978-981-08-8407-9 .
2. Altaher, M.A., (2012), " Local Polynomial Wavelet Regression with Missing at Random ", Applied Mathematical Sciences, Vol. 6, no. 57, 2805-2819 .
3. Boente, G. , Gonzalez, P. A. , Manteiga, G. W. , (1990), "Robust nonparametric estimation with missing data", Journal of statistical planning and inference, Vol. 139, Issues, PP. 571-592 .
4. Cheng, E. P. , (1994), "Nonparametric Estimation of Mean Functional with Data Missing at Random", Journal of the American Statistical Association, Vol. 89, No. 425 .
5. Daubechies, I. , (1992), "Ten Lectures on Wavelets", CBMS-NSF regional conference series in applied mathematics .
6. Donoho, L. D. , Johnstone, M. I. , (1994), "Ideal spatial adaptation by wavelet shrinkage", Biometrika, 81, 3, pp. 425-55 .
7. Fan,J.,Gijbiels,I.(1996)."Local polynomial modeling and its applications .london:Chapman and Hall.
8. Hardle, W. , (1990), "Applied Nonparametric Regression", Gambridg – MA : Cambridg University Press .
9. Kovac, A. , (1998), "Wavelet Thresholding for Unequally Spaced Data", PHD. Thesis, University of Bristol .
10. Lee, M. C. T. , Meng, L. X. , (2007), "Self Consistency : A General Recipe for Wavelet Estimation With Irregularly-spaced and / or Incomplete Data", arXiv : math / 0701196v1 [math. ST] .

11. Little, A. J. R. , Rubin, B. D. , (2002), "Statistical Analysis with Missing Data", John Wiley & Sons, INC.
12. Mallat, G. S. , (1989), "A Theory for Multiresolution Signal Decomposition: The Wavelet Representation ", Ieee Transactions on Pattern Analysis and Machine Intelligence. Vol. 11, No. 7.
13. Mojirsheibani, M. , (2007), "Nonparametric curve estimation with missing data : A general empirical process approach", Journal of Statistical Planning and Inference 137, 2733-2758 .
14. Nason, G. B. , (2008), "Wavelet Methods in Statistics with R", Springer.
15. Nason, G.P., (1996), " Wavelet shrinkage using cross-validation " , Journal of the Royal Statistical Society Series B. 58, 463-479 .
16. Ning, J. , Cheng, E. P. , (2012) "A Comparison Study of Nonparametric Imputation Methods", Statistical Compute, Springer, PP. 273-285.
17. Nittner, T. , (2003), "Missing at Random (MAR) in Nonparametric Regression"
18. Oh, S. H. , Lee, M. C. T. , (2005), "Hybrid Local Polynomial Wavelet Shrinkage : Wavelet Regression With Automatic Boundary Adjustment", Computational Statistics & Data Analysis 48. 809-819 .
19. Oh, S. H. , Naveau, P. , Lee, G. , (2001), "Polynomial Boundary Treatment for Wavelet Regression", Biometrika, 88, 1, pp. 291-298 .

Comparison some of methods wavelet estimation for non parametric regression function with missing response variable at random

Abstract

The problem of missing data represents a major obstacle before researchers in the process of data analysis in different fields since , this problem is a recurrent one in all fields of study including social , medical , astronomical and clinical experiments .

The presence of such a problem within the data to be studied may influence negatively on the analysis and it may lead to misleading conclusions , together with the fact that these conclusions that result from a great bias caused by that problem in spite of the efficiency of wavelet methods but they are also affected by the missing of data , in addition to the impact of the problem of miss of accuracy estimation it is not possible to apply these methods because of the miss of one of its conditions which is dyadic sample size $n = 2^J$.

According to the great impact resulted from the problem , many researchers who devoted their studies to process this problem , by using traditional methods in processing missing data , whereas the current research used imputation methods more efficient and effective to process missing data as a primary stage so that these data will be ready and available to wavelet application , as a result simulation experiment proved that the suggested methods (Nearset Nighbor Polynomial Wavelet) are more efficient and superior to other methods , this paper also includes the auto correction of boundaries problem by using local polynomial models , and using different threshold values in wavelet estimations.

Keywords / Missing data- Wavelet regression- Local polynomial- K-Nearest Neighbor