# مقارنة طرائق التعويض الأحادي عن القيمة المفقودة لأنموذج الانحدار اللامعلمي

د. قتيبه نبيل
جامعة بغداد- كلية الإدارة والاقتصاد
قسم الإحصاء

د.مناف يوسف
جامعة بغداد- كلية الإدارة والاقتصاد
قسم الإحصاء

## المستخلص

في هذا البحث سيتم دراسة أنموذج الانحدار اللامعلمي الذي يعاني فيه متغير الأستجابة من حالة فقدان (عدم استجابة) في بعض مشاهداتة وتحت أفتراض الية فقدان MCAR، إذ تم اقتراح طريقة تعويض قاعدة Kernel الأحادي اللامعلمي بدلاً عن القيمة المفقودة ومقارنة هذه الطريقة مع طريقة تعويض أقرب مجاور بأستخدام أسلوب المحاكاة والمتمثل بعدة تجارب لعدة نماذج مختلفة ولحالات مختلفة من حجوم العينة، التباين ونسب الفقدان.

## Abstract:

**In this paper, we will study non parametric model when the response variable have missing data (non response) in observations it under missing mechanisms MCAR, then we suggest Kernel-Based Non-Parametric Single-Imputation instead of missing value and compare it with Nearest Neighbor Imputation by using the simulation about some difference models and with difference cases as the sample size, variance and rate of missing data.**

## 1- مقدمة

تعد مشكلة البيانات المفقودة أو غير التامة من المشاكل الشائعة في مجال البحوث مثل استطلاع الرأي، استطلاعات دراسة التسويق، الدراسات الطبية، والعديد من التجارب العلمية. لذلك ظهرت العديد من الطرائق في التعويض عن القيم المفقودة وخاصة طرائق التعويض الأحادي Single – Imputation مثل تعويض الوسط الحسابي أو الوسيط .. الخ.

في هذا البحث سوف نقوم بدراسة أنموذج الانحدار اللامعلمي الذي يعاني فيه متغير الاستجابة $y$ من حالة فقدان (عدم استجابة) في بعض مشاهداته وتحت أفتراض الية فقدان MCAR (Missing Completely At Random) وذلك باستخدام طريقة تعويض قاعدة Kernel الأحادي اللامعلمي (KBNS) Kernel-Based Non-Parametric Single-Imputation وطريقة تعويض أقرب مجاور (NNI) Nearest Neighbor Imputation وباستخدام أسلوب المحاكاة.

## 2- طريقة تعويض قاعدة Kernel الأحادي اللامعلمي KBNS :

على فرض أن $x_i$ متغير عشوائي ذو توزيع مستقل ومتماثل i.i.d وتام المشاهدات، وأن المتغير $y_i$ يمثل متغير استجابة والمتأثر بالمتغير $X$ وعلى فرض أن الأزواج المرتبة $(x_i , y_i)$ تحقق الأنموذج التالي :[1]

$$y_i = m(x_i) + \varepsilon_i \quad \dots (1)$$

وأن المتغير $y_i$ يحتوي على بعض حالات عدم الأستجابة وأن $m(x_i)$ دالة غير معرفة. وعلى فرض أن: [3]

$$r = \sum_{i=1}^{n} \delta_i \quad \dots (2)$$

وان $m = n - r$

إذ يمثل $r$ مؤشر لمجموع حالات الأستجابة وعدم الأستجابة أو الفقدان حيث أن $\delta_i = 1$ في حالة وجود استجابة و $\delta_i = 0$ في حالة عدم الأستجابة وأن $n$ يمثل مشاهدات العينة في حين يمثل $m$ عدد حالات عدم الأستجابة للمتغير $y_i$.

وعليه يمكن الرمز لحالة الأستجابة وعدم الأستجابة بـ $S_r$ ، $S_m$ وعلى التوالي .[5] على فرض أن $y_i^* , i \in S_m$ يمثل القيم المعوضة بدل عن القيمة المفقودة أي أن :

$$y_i^* = \hat{m}_n(x_i) + \varepsilon_i^* , i \in S_m \quad \dots (3)$$

إذ يمثل $\varepsilon_i^*$ متغير الخطأ العشوائي ذو البعد $m$ ويمكن الحصول علية من الصيغة التالية :

$$m(x_i) - \hat{m}_n(x_i) \;,\; i \in S_m \qquad \dots (4)$$

أما $\hat{m}_n(x_i)$ فيتم الحصول عليه باستخدام قاعدة **Kernel** وكما يلي :[7]

$$\hat{m}_n(x) = \frac{\sum_{i=1}^{n} \delta_i \, y_i \, k\left(\frac{x - x_i}{h}\right)}{\sum_{i=1}^{n} \delta_i \, k\left(\frac{x - x_i}{h}\right) + n^{-2}} \qquad \dots (5)$$

وذلك بالاعتماد على المشاهدات التامة للأزواج المرتبـه $(x_i , y_i)$ .

وأن :

$\delta_i = 0$ إذ كان $y_i$ مفقود .

$\delta_i = 1$ إذ كان $y_i$ مشاهد .

أما $h$ فتمثل المعلمـة التمهيدية أو معلمـة عرض الحزمـة **(Bandwidth)** والتي تكون عدد موجب [7] ، أمـا $k(.)$ فتمثل دالة لبيه والتي يمكن الحصول عليها من دالة الكثافة الطبيعية القياسية او ما يطلق عليهـا بـ **Gaussian Kernel**:

$$K(z) = (2\pi)^{-\frac{1}{2}} \exp\left(-z^2/2\right) \qquad \dots (6)$$

أما فيما يخص المعلمة التمهيدية $h$ فهناك العديد من الطرائق في تقديرها ومن هذه الطرائق الصيغة التالية والتي تدعى بطريقة المصدر الطبيعي:[2]

$$h = 1.06 \cdot s \cdot n^{-\frac{1}{5}} \qquad \dots (7)$$

وهذه الصيغة تكون جيدة للبيانات التي تكون مقاربة للتوزيع الطبيعي[1].

## 3- <u>طريقة تعويض اقرب مجاور NNI</u>:

تعد هذه الطريقة من الطرائق الحديثة في التعويض عن القيم المفقودة، ولفهم آلية عمل هذه الطريقة نفرض لدينا متغير الاستجابة $y_i$ والذي يعاني من فقدان في بعض مشاهداته والمتغير التوضيحي $x_i$ الذي يكون تام المشاهدات وكما يلي:[4]

$$\underbrace{y_1, y_2, \ldots, y_{n-m}}_{\text{obs.}}, \underbrace{y_{n-m+1}, y_{n-m+2}, \ldots, y_n}_{\text{miss.}}$$

$$\underbrace{x_1, x_2, \ldots, x_{n-m}, x_{n-m+1}, x_{n-m+2}, \ldots, x_n}_{\text{obs.}} \qquad \ldots (8)$$

$$i = 1, 2, \ldots, n \quad \text{and} \quad l = n-m+1, n-m+2, \ldots, n$$

$$j = 1, 2, \ldots, n-m$$

**اذ تمثل m عدد القيم المفقودة في مشاهدات $y_i$ وان n تمثل عدد المشاهدات الكلية وعلية يتم اختيار قيمة $y_j$ ، $1 \le j \le n-m$ التي تقابل اقل فرق مطلق بين $x_j, x_l$ وحسب الصيغة التالية :**

$$\left|x_j - x_l\right| = \min_{1 \le j \le n-m} \left|x_j - x_l\right| \qquad \ldots (9)$$

**و إذا لم يكن هناك قيمة وحيدة يتم التعويض بقيمة متوسط البيانات.**

## 4- المحاكـاة

**لغرض معرفه اداء المقدرات اللامعلميه في حاله وجود مشاهدات غير تامة وبيان اختلاف كل من نسب واليه الفقدان، وتغير التباين للأخطاء وكذلك حجوم العينات فضلا عن تغير النماذج المستخدمة سوف يتم استخدام ثلاثة نماذج افتراضية وهي:**

Model I:

$$y_i = 3\cos(x_i) + 4\exp\left(-x_i^{\,2}\right) + \varepsilon_i$$

Model II:

$$y_i = \sin(2\pi x_i) + x_i^{\,2} + \varepsilon_i$$

Model III:

$$y_i = \frac{\cos(2.5\pi x_i)}{(1 + 3x_i^{\,2})}$$

**أمـا حالة الفقدان وحسب إلية الفقدان MCAR فسيتم توليدها حسب الصيغة التالية :[6]**

**MCAR:**

$$p(x) = p(\delta = 1 / X = x) = 0.9 \quad , \forall x$$

وتم افتراض قيم مختلفة لتباين الخطأ، وهي (2 ,1.5 ,1) .إما حجوم العينات المستخدمة فكانت (50, 100) ، في حين كانت نسب فقدان (10%, 20%, 30%).

والجدول الأتي يوضح قيم MSE لمقدرات دالة الانحدار اللامعلمية باختلاف النماذج المستعملة، التباينات، حجوم العينات ونسب الفقدان.

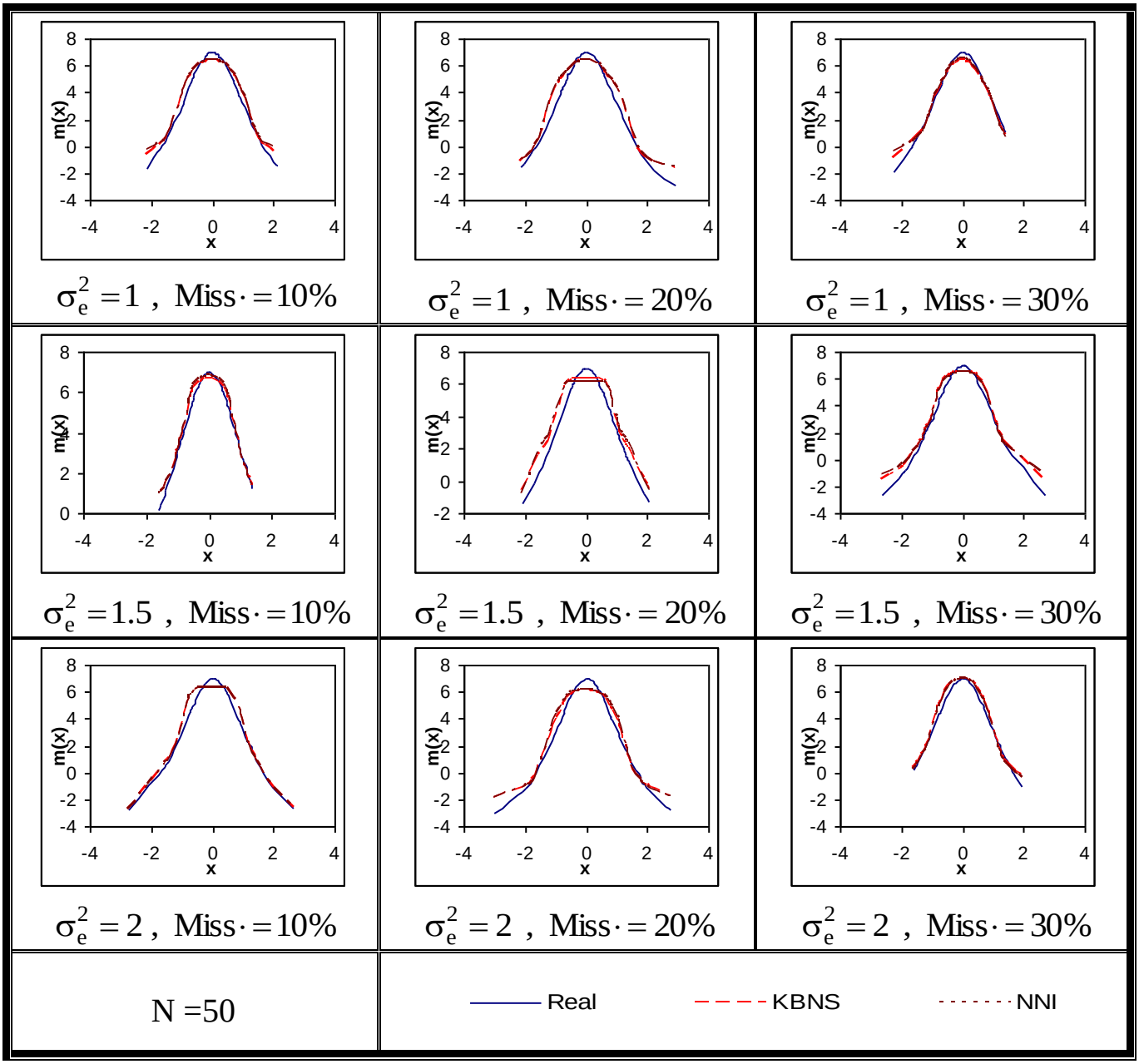
**جدول (1)**

**يشير إلى MSE لتقدير دالة الانحدار اللامعلميه $m(x)$ لجميع النماذج بالاعتماد على حجوم عينات، تباينات ونسب فقدان مختلفة**

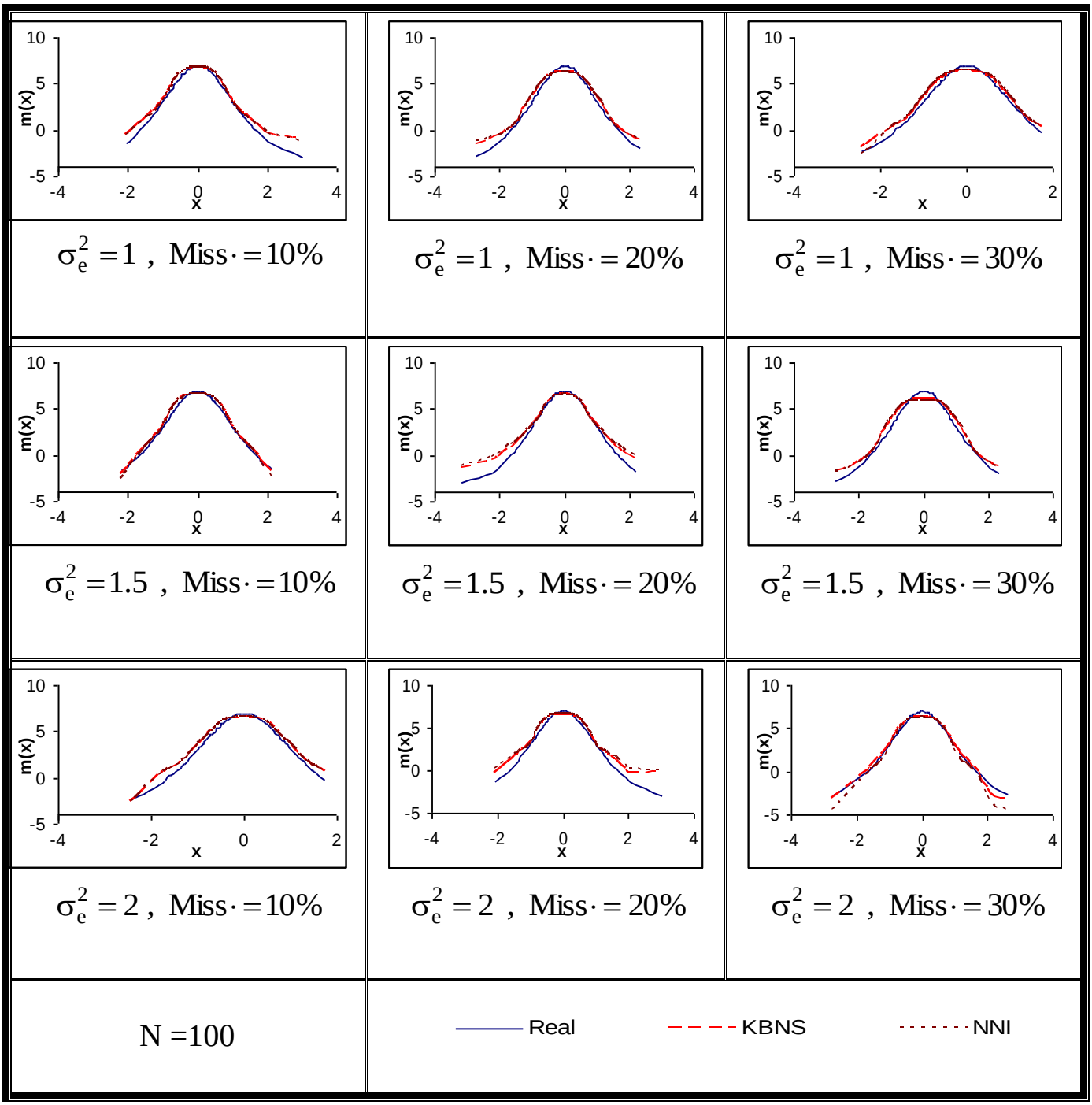
| Model | N | Methods | KBNS | | | NNI | | |
|---|---|---|---|---|---|---|---|---|
| | | $\sigma_e^2$ / Miss. | 1 | 1.5 | 2 | 1 | 1.5 | 2 |
| I | 50 | 10% | 0.2164 | 0.2709 | 0.3383 | 0.2361 | 0.3080 | 0.4054 |
| | | 20% | 0.2114 | 0.2565 | 0.3401 | 0.2548 | 0.3469 | 0.4732 |
| | | 30% | 0.2000 | 0.2473 | 0.3082 | 0.2656 | 0.3803 | 0.5462 |
| | 100 | 10% | 0.1398 | 0.1785 | 0.2242 | 0.1482 | 0.2017 | 0.2636 |
| | | 20% | 0.1351 | 0.1645 | 0.2146 | 0.1573 | 0.2052 | 0.2977 |
| | | 30% | 0.1348 | 0.1593 | 0.1919 | 0.1719 | 0.2344 | 0.3202 |
| II | 50 | 10% | 0.1523 | 0.2105 | 0.3190 | 0.1894 | 0.2631 | 0.4026 |
| | | 20% | 0.1440 | 0.2108 | 0.3008 | 0.2192 | 0.3107 | 0.4714 |
| | | 30% | 0.1381 | 0.1996 | 0.2688 | 0.2732 | 0.3864 | 0.5624 |
| | 100 | 10% | 0.1018 | 0.1380 | 0.2084 | 0.1219 | 0.1751 | 0.2595 |
| | | 20% | 0.0963 | 0.1367 | 0.1823 | 0.1355 | 0.2033 | 0.2888 |
| | | 30% | 0.0890 | 0.1252 | 0.1685 | 0.1511 | 0.2472 | 0.3366 |
| III | 50 | 10% | 0.0567 | 0.1228 | 0.2199 | 0.0761 | 0.1623 | 0.2884 |
| | | 20% | 0.0519 | 0.1083 | 0.1806 | 0.0901 | 0.1962 | 0.3265 |
| | | 30% | 0.0447 | 0.0995 | 0.1598 | 0.1074 | 0.2416 | 0.4072 |
| | 100 | 10% | 0.0301 | 0.0696 | 0.1208 | 0.0407 | 0.0934 | 0.1595 |
| | | 20% | 0.0297 | 0.0598 | 0.1040 | 0.0493 | 0.1075 | 0.1897 |
| | | 30% | 0.0252 | 0.0564 | 0.0992 | 0.0580 | 0.1417 | 0.2572 |

والإشكال الآتية توضح تأثير كل من نسب الفقدان وتغير قيم التباين على مقدرات دالة الانحدار باختلاف حجوم العينات والنماذج المستعملة.

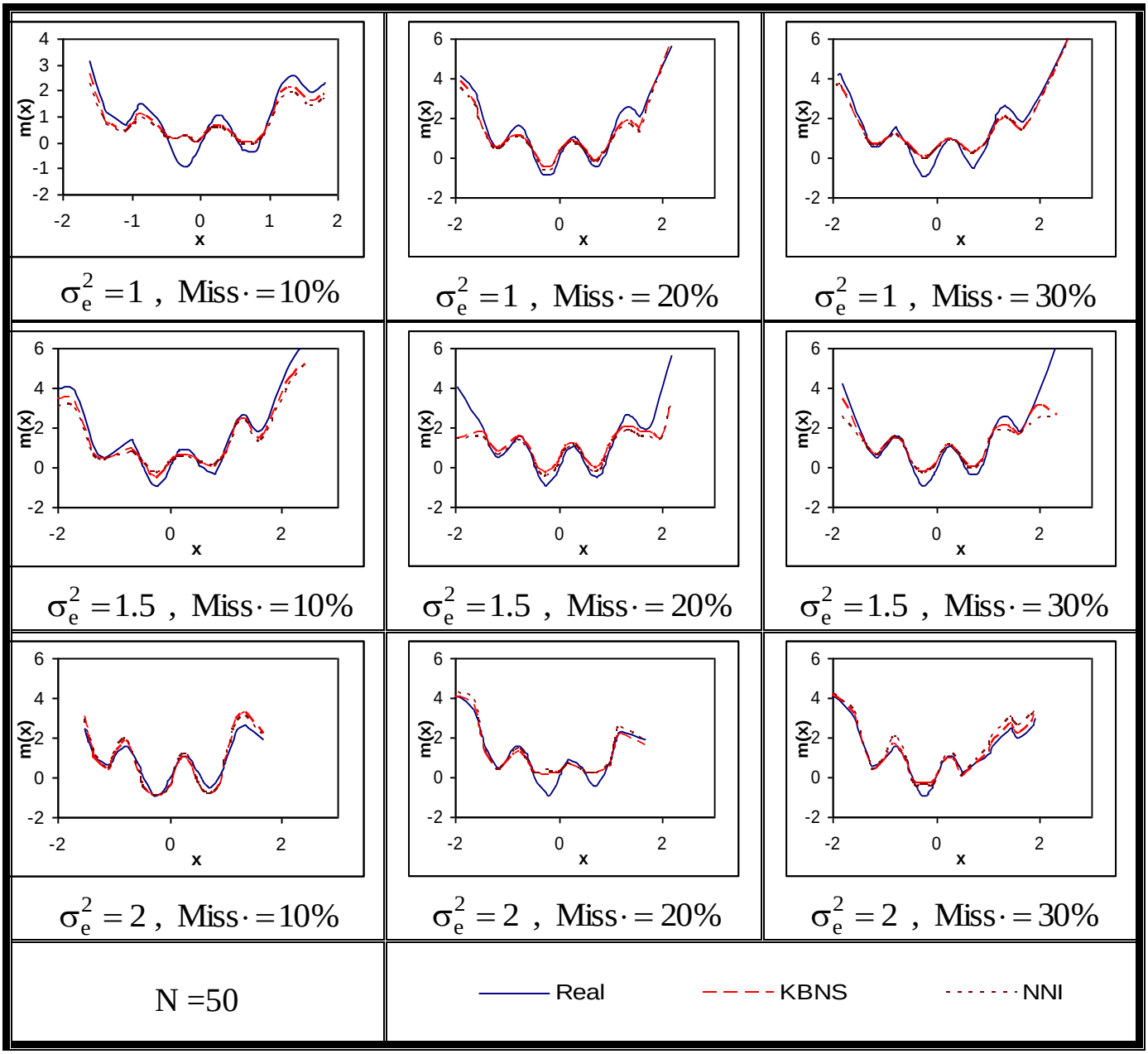
شكل (1) تأثير نسب الفقدان وتغير قيم التباين على مقدرات دالة الانحدار $m(x)$ للأنموذج الأول عندما N=50



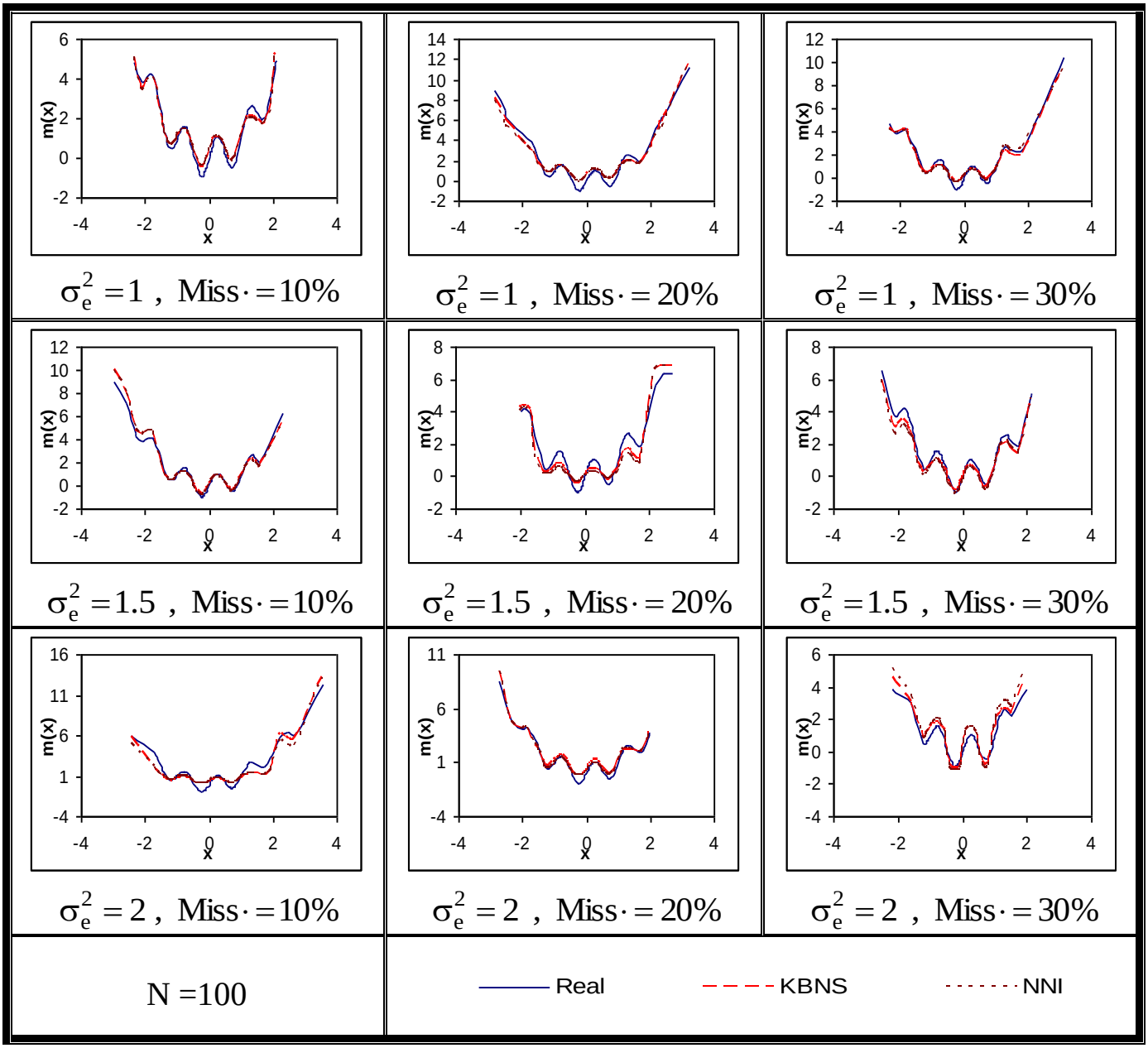
شكل (2) تأثير نسب الفقدان وتغير قيم التباين على مقدرات دالة الانحدار $m(x)$ للأنموذج الأول عندما N=100.



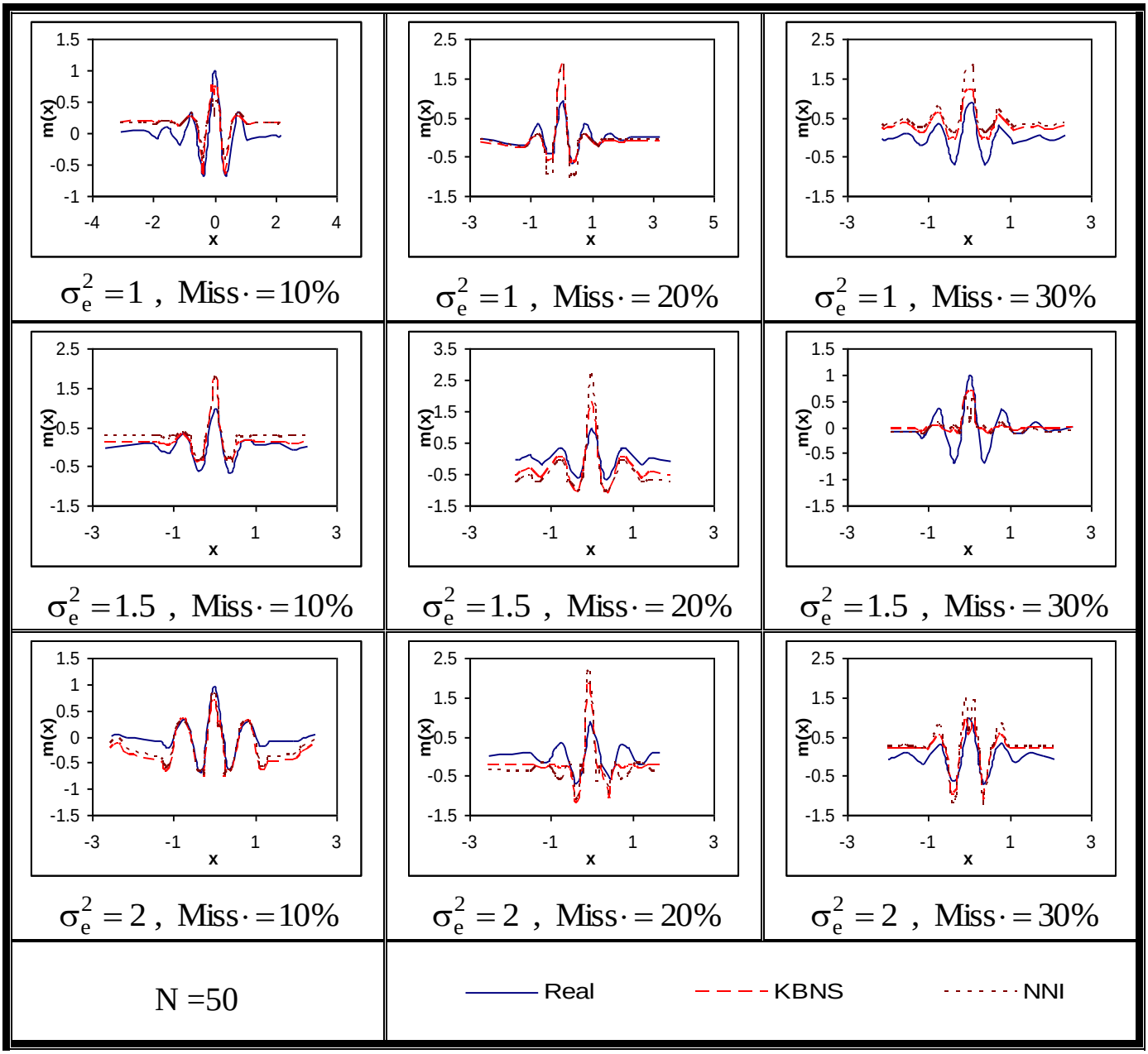
شكل (3) تأثير نسب الفقدان وتغير قيم التباين على مقدرات دالة الانحدار $m(x)$ للأنموذج الثاني
عندما N=50



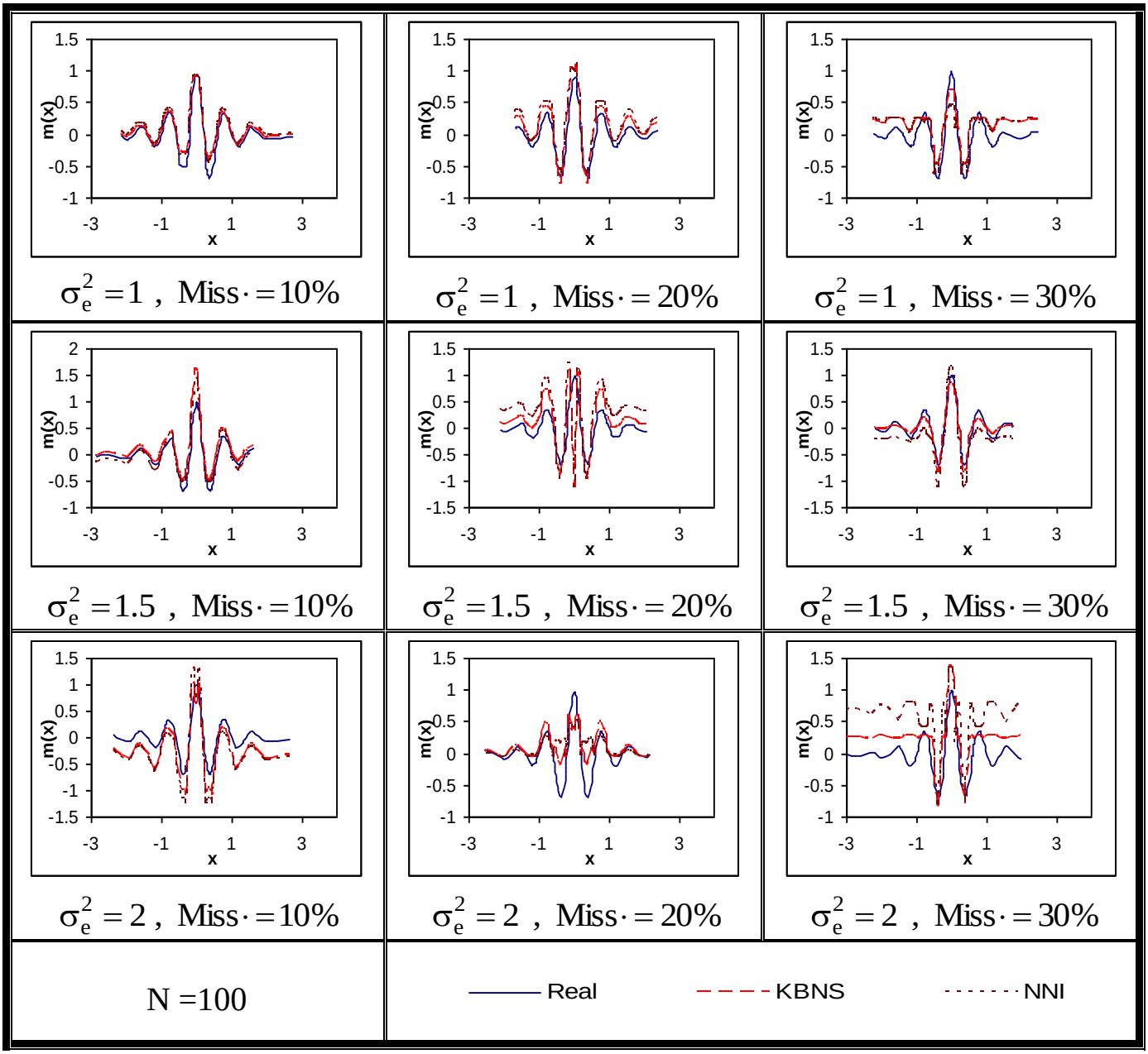
شكل (4) تأثير نسب الفقدان وتغير قيم التباين على مقدرات دالة الانحدار $m(x)$ للأنموذج الثاني عندما N=100



شكل (5) تأثير نسب الفقدان وتغير قيم التباين على مقدرات دالة الانحدار $m(x)$ للأنموذج الثالث
عندما N=50



شكل (6) تأثير نسب الفقدان وتغير قيم التباين على مقدرات دالة الانحدار $m(x)$ للأنموذج الثالث
عندما N=100



## 5 – تفسير النتائج:

يتضح من النتائج في الجدول (1) المذكور أنفا الأتي:

- مقدر NW كان المقدر الأفضل ولجميع النماذج،حجوم العينات،التباينات ونسب الفقدان.
- أشارت النتائج عند استعمال مقدرKBNS تناقص قيم MSE عند زيادة نسب الفقدان ولجميع حجوم العينات وقيم تباين(1,1.5).في حين عند قيمة تباين 2 تزايدت قيم MSE عند زيادة نسب الفقدان.
- أشارت النتائج عند استعمال مقدرKBNS ان قيم MSE تتناقص عند زيادة نسب الفقدان ولجميع حجوم العينات وقيم تباين (1,1.5). في حين عند قيمة تباين 2 تزايدت قيم MSE عند زيادة نسب الفقدان.
- أشارت النتائج عند استعمال مقدر NNI ان قيم MSE تزايدت عند زيادة نسب الفقدان ولجميع حجوم العينات وقيم التباين المستعملة.
- اشارت النتائج الى تناقص قيم MSE عند زيادة حجوم العينات ولجميع النماذج ، قيم التباينات ونسب الفقدان.
- تناقص قيم MSE لكلا مقدري دالة الانحدار عند تزايد قيم التباين ولجميع حجوم العينات،النماذج المستعملة ونسب الفقدان.
- لكل نسبه من نسب الفقدان يلاحظ تناقص قيم MSE لكلا مقدري دالة الانحدار عند تزايد قيم التباين ولجميع حجوم العينات والنماذج المستعملة.

## 6- الاستنتاجات

يلاحظ من النتائج المذكورة في المبحث السابق أفضلية مقدر KBNS المتمثل بمقدر NW ولجميع الحالات المفترضه من نماذج مفترضة، نسب فقدان، تباينات وحجوم عينات. لذا يوصى باستخدام هذا المقدر في حاله وجود بيانات مفقودة وعدم الاعتماد على مقدر NNI.

يلاحظ كذلك من الإشكال إن مقدر NW لازال يتأثر بتأثيرات الحد مما يتطلب تعديلا لهذا المقدر عند نقاط الحد.

## 7- المصادر


[1] Hardle, W. (1990). *Applied nonparametric regression*. Cambridge University press.

[2] Heumann, C.; Nittner, T.; Scheid, S. & Toutenburg, H. (2002) *Parametric and Nonparametric Regression Missing X's: A Review, Journal of the Iranian Statistical Society, 1, No.1-2, 79-110.* www. stat.uni-muenchen.de/sfb386/papers/dsp/paper286.ps.

[3] Little, R.J.A. and Rubin, D.B. (1987). *Statistical Analysis with Missing Data.* New York: John Wiley & Sons, Inc.

[4] Nittner, T. (2002) *The Additive with Missing values in the Independent Variable: Theory & Simulation, to appear in: Computational Statistics.* http://www.pms.ifi.lmu.de/research-report/index.pdf

[5] Schafer J. (1997) *Analysis of incomplete multivariate data.* 1st ed. London: Chapman and Hall.

[6] Schafer J. (2002) *Dealing with Missing Data.* Res. Lett. Math. Sci., 3, 153-160.

[7] Silverman, B.W. (1986), *Density estimation for statistics and data analysis*. London: Chapman and Hall.