تقويم بيانات العمر والنوع للتعداد العام للسكان في العراق باستعمال مقدرات Kernel Bayesian اللامعلمية

أ.م.د. قتيبة نبيل نايف القزاز الباحثة/ مروة خليل ابراهيم
قسم الإحصاء/ كلية الادارة والاقتصاد/ جامعة بغداد

المستخلص:

تعد عملية تقويم بيانات التركيب العمري والنوعي من العوامل المهمة التي تساعد اي دولة في رسم الخطط والبرامج المستقبلية، لذا تناول البحث الاخطاء التي ترافق البيانات السكانية الخاصة بتعداد العراق لعام ١٩٩٧ مستهدفاً تقويمها وتنقيحها بما يخدم الاغراض التخطيطية حيث سيتم استعمال طريقة Kernel اللامعلمية بواسطة مقدر (Nadaraya-Watson) لايجاد معادلة الانحدار التقديرية لتمهيد البيانات السكانية بالاعتماد على معلمة التمهيد (h) والتي سيتم ايجادها على وفق استعمال صيغة بيز بحالتين الاولى عندما يكون توزيع المشاهدات Lognormal Kernel والثانية عندما يكون توزيع المشاهدات Normal Kernel. ثم تتم المقارنة بين الطريقتين بواسطة دليل سكرتارية الامم المتحدة لدقة العمر والنوع وتحليل نسبتي العمر والنوع للوصول الى افضل تمهيد للبيانات. وتم التوصل الى ان البيانات الناتجة من عملية التمهيد بطريقة تقدير Kernel وحسب طريقة ايجاد معلمة التمهيد h : Bayes L-N هي الافضل وحققت اقل قيمة لدليل سكرتارية الامم المتحدة كما وانها حافظت على نسبتي العمر والنوع.

المصطلحات الرئيسية للبحث/ مقدرات Nadaraya-Watson Kernel- توزيع مقلوب گاما السابق- توزيع Lognormal Kernel اللاحق- توزيع Normal Kernel اللاحق- مقياس الامم المتحدة .

مجلة العلوم
الاقتصادية والإدارية
المجلد 20
العدد ٧6
لسنة ٢٠١٤
الصفحات ٣٨٢-٣٩٩

*البحث مستل من رسالة ماجستير

١. مقدمة:

ان لتقويم وتصحيح البيانات السكانية اهمية كبيرة للحصول على بيانات خالية من الاخطاء واستعمالها في تقدير اعداد السكان في المستقبل ومن ثم رسم الخطط والسياسات السكانية المستقبلية، إذ على الرغم من التطور السريع الذي يشهده العالم في مجال تكنولوجيا الانظمة المعلوماتية الا ان الكثير من الدول النامية ماتزال تشكو من اخطاء في بياناتها السكانية ونقص المعلومات حول تلك البيانات مما يؤدي الى اخطاء جسيمة اثناء التحليل الديموغرافي للبيانات. لذا اتجه باحثوا تلك الدول من خلال عقد دورات دولية ومحلية ومؤتمرات في البحث عن احدث الدراسات والابحاث الحديثة في مجال تمهيد البيانات السكانية وجعلها خالية وبقدر مقبول من الاخطاء.

٢. مشكلة البحث:

تتعرض بيانات التركيب العمري والنوعي لعدد من الاخطاء يمكن تقسيمها على قسمين: [2],[3],[4]

i. **اخطاء الشمول**: والتي ترجع الى قصور في العد او تكرار فيه، او الى عيوب في المراحل الادارية خاصة للفئات الخاصة مثل البدو او المقيمين في المناطق النائية مما يؤثر في عملية العد لمختلف الفئات العمرية.

ii. **اخطاء المحتوى او المضمون**: اخطاء الاجابة (وتحدث مثل هكذا اخطاء نتيجة الابلاغ الخاطئ عن العمر او تجاهل الادلاء ببيانات صحيحة)، اخطاء العدادين (تحدث بسبب قلة تدريب العدادين على طريقة طرح الاسئلة للمواطنين بدقة وسهولة) او اخطاء تجهيز البيانات (اي اثناء تفريغ البيانات وتصنيفها وتقسيمها).

ومن الجدير بالذكر عندما نذكر بيانات العمر نذكر بيانات النوع لان الاخطاء في الاعمار غالباً ما تتفاوت تبعاً للنوع، اذ تظهر بدرجة اشد في توزيع السكان بحسب النوع، اما درجة الخطأ فتختلف بحسب المراحل العمرية.

ولتلافي مثل هذه الاخطاء وبما ان العراق هو احد تلك الدول الذي تعاني بياناته السكانية منها لذا ارتأينا ان يكون الهدف من البحث هو تقويم بيانات سكان العراق لتعداد العراق عام (١٩٩٧) وتنقيحاً بما يخدم الاغراض التخطيطية المستقبلية ومن ثم اعتمادها في عدة بحوث لاحقة.

٣. طرائق تمهيد بيانات التعداد: [6],[11]

ان بيانات التعداد العام للسكان هي من أهم مصادر البيانات السكانية ويتم الاعتماد عليها بشكل كبير في كافة التحليلات الديموغرافية، لذا كان لابد من أختبار مدى دقة هذه البيانات وتحديد نوع وحجم الخطأ فيها وتصحيحها. وبسبب التطور الكبير في مجال الحواسيب من الناحيتين (software, hardware) فقد امكن تطبيق طرائق التمهيد اللامعلمية لتمهيد البيانات بسبب مرونتها واختصارها للوقت والجهد، لذا سيتم أستعمال طريقة Kernel اللامعلمية بواسطة مقدر (Nadaraya-Watson kernel estimation) لايجاد معادلة الانحدار التقديرية لتمهيد البيانات السكانية، بمعنى سيتم استعمال هذه الطريقة لتقدير m في انموذج الانحدار الظاهر في المعادلة (١) والشكل العام لانموذج الانحدار اللامعلمي هو كما يأتي:

$$Y_i = m(X_i) + \epsilon_i \qquad i = 1, \dots, n \qquad \dots (1)$$

حيث ان:

m دالة غير معلومة.

وان طريقة Nadaraya-Watson kernel estimation أُقترحت بواسطة Nadaraya و Watson في عام (١٩٦٤) وسُميت على اسمهما.

$$\hat{m}_h(x) = \frac{\sum_{i=1}^{n} K_h(x - X_i) y_i}{\sum_{i=1}^{n} K_h(x - X_i)} \qquad \dots (2)$$

$$\because \quad W_h(x - X_i) = \frac{K_h(x - X_i)}{\sum_{i=1}^{n} K_h(x - X_i)} \qquad \dots (3)$$

$$\therefore \quad \hat{m}_h(x) = \sum_{i=1}^{n} W_h(x - X_i) y_i \qquad \dots (4)$$

علماً ان:

$K_h(x - X_i)$ تمثل دالة Kernel.

h تمثل عرض الحزمة (bandwidth) او تعرف بمعلمة التمهيد (smooth parameter)، وهي عبارة عن معلمة حرة (free parameter)، حيث لها تأثير واضح في عملية التقدير لكونها تؤثر بشكل كبير في التحيز والتباين حيث بزيادة عرض الحزمة يزداد التحيز ويتناقص التباين والعكس صحيح وبالنتيجة ستؤثر في تمهيد المنحنى ومعدل اقترابه من المنحنى الحقيقي.

$$\sum_{i=1}^{n} W_h(x - X_i) = 1 \qquad \dots (5)$$

وفي اي انحدار لامعلمي التوقع الشرطي للمتغير Y بالنسبة الى المتغير X يكون بالشكل الاتي:

$$E(Y|X) = m(X) \qquad \dots (6)$$

ولاشتقاق الصيغة في المعادلة (٦)

$$E(Y|X) = \int y\, f(y|x)\, dy = \int y\, \frac{f(x,y)}{f(x)}\, dy \qquad \dots (7)$$

سنستعمل مقدر كثافة Kernel للتوزيع المشترك $f(x,y)$ و $f(x)$ حيث

$$\hat{f}(x,y) = n^{-1} h^{-2} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) K\left(\frac{y - y_i}{h}\right)$$

$$= \frac{1}{n} \sum_{i=1}^{n} K_{h_x}(x - X_i) K_{h_y}(y - y_i)$$

$$\int y\, \hat{f}(x,y) dy = \frac{1}{n} \int y \sum_{i=1}^{n} K_{h_x}(x - X_i) K_{h_y}(y - y_i)$$

$$\because \int y\, K_{h_y}(y - y_i)\, dy = y_i$$

$$\therefore \int y\,\hat{f}(x,y)dy = \frac{1}{n}\sum_{i=1}^{n} K_{h_x}(x - X_i)\,y_i \qquad \dots(8)$$

ان المعادلة (8) تمثل مقدر البسط للمعادلة (2).

$$\begin{aligned}\int \hat{f}(x,y)dy &= \frac{1}{n}\sum_{i=1}^{n} K_{h_x}(x - X_i)\int K_{h_y}(y - y_i)\,dy \\ &= \frac{1}{n}\sum_{i=1}^{n} K_{h_x}(x - X_i) \qquad \dots(9) \\ &= \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) \\ &= \hat{f}(x)\end{aligned}$$

اما المعادلة (9) فتمثل مقدر المقام للمعادلة (2)، ومن المعادلتين (8) و (9) سنحصل على مقدر Nadaraya-Watson Kernel الموضح بالمعادلة (٢).

٤. خصائص دالة Kernel: [12],[10],[8],[7],[6]

١. $K(u)$ دالة كثافة احتمالية محددة (غير سالبة) $K(u) \geq 0$ لكل (u)، بحيث $K(u): R \to R$.

٢. العزوم لدالة Kernel تحسب على وفق الصيغة الاتية:

$$M_j(K) = \int_{-\infty}^{\infty} u^j\,K(u)d(u) \qquad \dots(10)$$

3. $\int_{-\infty}^{\infty} u\,K(u)d(u) = 0 \qquad \dots(11)$

4. $C_K = R(K) = \int_{-\infty}^{\infty} K^2(u)d(u) \qquad \dots(12)$

5. $d_K = M_2(K) = \int_{-\infty}^{\infty} u^2 K(u)d(u) \qquad \dots(13)$

ومن الجدير بالملاحظة ان القيمة $\hat{f}(x)$ تكون كبيرة عندما يكون هناك عدد كبير من المشاهدات قريبة من x ، وبالعكس تكون قيمة $\hat{f}(x)$ صغيرة عندما يكون عدد قليل من المشاهدات قريبة من x .

والجدول (١) يمثل دوال Kernel [12],[10],[8],[7]

الجدول (1)
دوال Kernel

$Kernel$	$K(u)$	
Uniform (p=0)	$1/2$	$,\mathrm{I}(\lvert u\rvert \leq 1)$
Epanchnikov (p=1)	$\frac{3}{4}(1-u^2)$	$,\mathrm{I}(\lvert u\rvert \leq 1)$
Quartic or Biweight (p=2)	$\frac{15}{16}(1-u^2)^2$	$,\mathrm{I}(\lvert u\rvert \leq 1)$
Triwight (p=3)	$\frac{35}{32}(1-u^2)^3$	$,\mathrm{I}(\lvert u\rvert \leq 1)$
Gauss	$(2\pi)^{-\frac{1}{2}}exp\left(\frac{-u^2}{2}\right)$	$,\mathrm{I}(\lvert u\rvert \leq \infty)$

وهناك طرائق عدة لايجاد عرض الحزمة منها استعمال اسلوب بيز.

٥. اسلوب بيز: [8],[9],[13]

يلعب الاستدلال الاحصائي دور مهم واساسي في العديد من الدراسات الميدانية والتطبيقات العملية وان الفرع الاساسي للاستدلال الاحصائي هو التقدير الاحصائي واحدى طرائق التقدير الاحصائي هو تقدير بيز. واهم ما يميز هذا الاسلوب هو اعتبار معلمة (معالم) المجتمع تحت الدراسة متغير (متغيرات) عشوائية تتبع توزيعاً معيناً يسمى بالتوزيع الاولي (Prior)، بفرض ان الدالة $L\left(\hat{\theta},\theta\right)$ تقيس الخسارة الناتجة عن اختيار الاجراء $\hat{\theta}$ عندما تكون الحالة الطبيعية هي θ وهذه الدالة تعرف بدالة الخسارة، بعد تحديد دالة الخسارة المناسبة يكون الهدف هو اختيار المقدر الذي يجعل مخاطرة بيز Bayes risk اقل مايمكن عندما نقدر المعلمة θ بالمقدر$\hat{\theta}$، اي ان مقدر بيز $\hat{\theta}$ هو قيمة θ التي تجعل القيمة المتوقعة لدالة الخسارة بالنسبة للتوزيع اللاحق اقل مايمكن.

٥.١ التوزيعات السابقة Prior distributions:

ان اسلوب بيز يعد معلمة (معالم) التوزيع متغير عشوائي له توزيع احتمالي ويحوي على كل المعلومات حول المعلمة قبل الحصول على العينة العشوائية وهذا التوزيع يسمى (بالتوزيع السابق او التوزيع قبل المعاينة) اذ تقسم دوال الكثافة الاحتمالية السابقة بحسب طبيعة المعلومات المتوفرة على عدة انواع منها:

٥.١.١ التوزيعات السابقة المرافقة Conjugate prior distributions:

تعرف التوزيعات الاحتمالية السابقة على انها توزيعات سابقة مرافقة للتوزيع الاحتمالي الذي سحبت منه العينة العشوائية اذا كان التوزيع اللاحق ينتمي الى العائلة نفسها التي ينتمي اليها التوزيع السابق ولكن بمعلمات مختلفة.

٥.١.٢ التوزيعات السابقة غير المعلوماتية Non-informative prior distributions:

وهي عبارة عن توزيعات احتمالية تعبر عن المعرفة القليلة او شبه المعلومة لمعلمة (معالم) التوزيع قبل سحب العينة العشوائية وهذه النوعية من التوزيعات السابقة تكون مناسبة للحالات التي تكون فيها المعلومات المتوفرة لدينا حول المعلمة قبل المعاينة ليست ذات اهمية مقارنة بالمعلومات المتوقع الحصول عليها من العينة العشوائية.

٥.٢ التوزيعات اللاحقة Posterior distributions:

وهي التي تصف درجة الاعتقاد حول القيم الممكنة للمعلمة بعد الحصول على العينة، ويسمى التوزيع اللاحق بالتوزيع بعد المعاينة. ويرمز له بالرمز $\pi^*(\theta|\underline{x})$ وهو توزيع احتمالي مشروط للمعلمة θ بشرط الحصول على العينة $\underline{x}$ وبأستعمال نظرية بيز يكون وفق الصيغة التالية:

$$\pi^*(\theta|\underline{x}) = \frac{\pi(\theta)l(\theta|\underline{x})}{\int_\theta \pi(\theta)l(\theta|\underline{x})\,d\theta} \qquad \dots(14)$$

حيث $l(\theta|\underline{x})$ تمثل دالة الامكان.

اذ ان:

$$\pi^*(\theta|\underline{x}) \propto \pi(\theta)\, l(\theta|\underline{x}) \qquad \dots(15)$$

٦. اشتقاق عرض الحزمة بأستعمال صيغة بيز: [8],[9],[13]

لاشتقاق عرض الحزمة بصيغة بيز تم الاعتماد مرةً على توزيع المشاهدات Lognormal Kernel ومرة اخرى على توزيع المشاهدات Normal Kernel اما التوزيع السابق$(prior)$ يكون inverted gamma كالاتي:

***a*) توزيع المشاهدات Lognormal Kernel:**

اشتقاق عرض الحزمة h لمقدر Kernel بصيغة بيز بأستعمال توزيع Lognormal Kernel بالمعلمتين μ, σ وكالاتي:

$$K(x,\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}\,x} e^{-\frac{1}{2}\left(\frac{lnx-\mu}{\sigma}\right)^2} = \frac{1}{xh\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{lnx-lnz_j}{h}\right)^2} \qquad \dots(16)$$

ان معاملة عرض الحزمة (معلمة التمهيد h) هو كمية عشوائية لها توزيع اولي يتمثل (inverted gamma prior distribution) بالمعلمتين α و β :

$$\pi(h) = \frac{1}{\beta^\alpha \Gamma\alpha\, h^{\alpha+1}} e^{-\frac{1}{\beta h}} \qquad ,h>0 \qquad \dots(17)$$

وان:

$$\hat{f}_h(x) = \sum_{i=1}^{n} K(x, lnz_j, h) \qquad \dots(18)$$

حيث:

Z_j: المشاهدات (Observations).

دالة الكثافة الاحتمالية اللاحقة لـ (h) تكون:

$$\pi^*(h|x,\underline{z}) = \frac{f_h(x)\,\pi(h)}{\int f_h(x)\,\pi(h)\,dh} \qquad \dots(19)$$

وبما ان f_h غير معلومة، تم استعمال $\hat{f}_h(x)$ بالمعادلة (18) واصبح:

$$\hat{\pi}^*(h|x,\underline{z}) = \frac{\hat{f}_h(x)\,\pi(h)}{\int \hat{f}_h(x)\,\pi(h)\,dh} \qquad \dots(20)$$

$$\int \hat{f}_h(x)\,\pi(h)\,dh = \int_0^{\infty} \sum_{j=1}^{n} \frac{1}{x\,h\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\ln x - \ln z_j}{h}\right)^2} \cdot \frac{1}{\beta^{\alpha}\Gamma\alpha\, h^{\alpha+1}} e^{-\frac{1}{\beta h}}\,dh \qquad \dots(21)$$

نجد هنالك اختلاف بين معلمة التوزيع السابق ومعلمة توزيع Lognormal Kernel وبأستعمال تحويل الجذر التربيعي (Square root transformation) لمعلمة التوزيع السابق (h) توحد المعلمتين اي (يمتلك كِلا التوزيعان نفس قيمة المعلمة (h) اي يتم التعامل مع (h) كمتغير عشوائي له توزيع اولي)، ولايجاد التحويل:

$$\begin{aligned} f_Y(y) &= f_X\left(g^{-1}(y)\right).\left|\frac{d}{dy}g^{-1}(y)\right| \\ &= f_X(y^2).2y \\ &= \frac{1}{\beta^{\alpha}\Gamma\alpha\,(y^2)^{\alpha+1}}.e^{-\frac{1}{\beta y^2}}.2y \\ &= \frac{1}{\beta^{\alpha}\Gamma\alpha\, y^{2\alpha+1}}.e^{-\frac{1}{\beta y^2}}.2 \qquad \dots(22) \end{aligned}$$

اذا كانت $\sigma = h^2 \propto \text{Inverted Gamma}(\alpha,\beta)$ ، عندها $h = \sqrt{\sigma}$ وحسب المعادلة (16)، وبعد تعويض دالة الكثافة الاحتمالية P.d.f للتوزيع السابق حيث:

$$\pi(h) = \frac{2}{\beta^{\alpha}\Gamma\alpha\, h^{2\alpha+1}}.e^{-\frac{1}{\beta h^2}} \qquad ,h>0 \qquad \dots(23)$$

وبذلك فأن المقام في معادلة (20) يمكن كتابته بالشكل:

$$\int \hat{f}_h(x)\,\pi(h)\,dh = \int_0^{\infty} \sum_{j=1}^{n} \frac{1}{x}\frac{1}{h\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\ln x - \ln Z_j}{h}\right)^2} \cdot \frac{2}{\beta^{\alpha}\Gamma\alpha\, h^{2\alpha+1}} e^{-\frac{1}{\beta h^2}}\, dh$$

$$= \int_0^{\infty} \sum_{j=1}^{n} \frac{1}{x}\frac{1}{\sqrt{2\pi}} \frac{2}{\beta^{\alpha}\Gamma\alpha\ (h^2)^{\alpha+1}} e^{-\frac{1}{h^2}\left(\frac{1}{\beta}+\frac{1}{2}(\ln x - \ln Z_j)^2\right)}\ \ dh \qquad \ldots(24)$$

بفرض ان:

$$\beta_j^* = \left[\frac{1}{\beta} + \frac{1}{2}(\ln x - \ln Z_j)^2\right]^{-1} \qquad \ldots(25)$$

$$\alpha^* = \alpha + \frac{1}{2} \qquad \ldots(26)$$

بتغيير المتغيرات بأستبدال $h^2 = t$ نحصل على:

$$\int \hat{f}_h(x)\,\pi(h)\,dh = \int_0^{\infty} \sum_{j=1}^{n} \frac{1}{\sqrt{2\pi}}\frac{1}{x}\ e^{-\frac{1}{t\beta_j^*}} \frac{1}{\beta^{\alpha}\Gamma\alpha\, t^{\alpha+1}} \frac{1}{\sqrt{t}}\, dt$$

$$= \frac{1}{x}\sum_{j=1}^{n} \frac{(\beta_j^*)^{\alpha^*}\Gamma(\alpha^*)}{\sqrt{2\pi}\ \beta^{\alpha}\Gamma\alpha} \int_0^{\infty} \frac{1}{(\beta_j^*)^{\alpha^*}\Gamma(\alpha^*)t^{\alpha^*+1}}\ e^{-\frac{1}{t\beta_j^*}} dt \qquad \ldots(27)$$

وبذلك يكون المقام في المعادلة (20):

$$\frac{\Gamma(\alpha^*)}{x\ \Gamma\alpha} \sum_{j=1}^{n} \frac{(\beta_j^*)^{\alpha^*}}{\sqrt{2\pi}\ \beta^{\alpha}} \qquad \ldots(28)$$

وكذلك البسط في المعادلة (20) يبسط كالاتي:

$$\hat{f}_h(x)\,\pi(h) = \sum_{j=1}^{n} \frac{1}{\sqrt{2\pi}}\frac{1}{x}\frac{2}{\beta^{\alpha}\Gamma\alpha\,(h^2)^{\alpha+1}} e^{-\frac{1}{h^2\beta_j^*}} \qquad \ldots(29)$$

وبالتالي، فأن مقدر الكثافة الاحتمالية اللاحقة للـ h هو:

$$\hat{\pi}\left(h\middle|x,\underline{z}\right) = \frac{\sum_{j=1}^{n} \frac{1}{\sqrt{2\pi}}\frac{1}{x}\frac{2}{\beta^{\alpha}\Gamma\alpha\,(h^2)^{\alpha+1}} e^{-\frac{1}{h^2\beta_j^*}}}{\frac{\Gamma(\alpha^*)}{\Gamma\alpha\,(x)}\sum_{j=1}^{n}\frac{(\beta_j^*)^{\alpha^*}}{\sqrt{2\pi}\ \beta^{\alpha}}}$$

$$= \frac{\sum_{j=1}^{n} \frac{2}{(h^2)^{\alpha+1}} e^{-\frac{1}{h^2\beta_j^*}}}{\Gamma(\alpha^*)\sum_{j=1}^{n}(\beta_j^*)^{\alpha^*}} \qquad \ldots(30)$$

وتحت دالة الخسارة التربيعية فأن مقدر بيز لمعلمة h (والذي يمثل متوسط التوزيع اللاحق) يكون:

$$\tilde{h}(x) = \int_0^{\infty} h\ \hat{\pi}\left(h\middle|x,\underline{z}\right) dh$$

$$= \int_0^{\infty} h\ \frac{\sum_{j=1}^{n} \frac{2}{(h^2)^{\alpha+1}} e^{-\frac{1}{h^2\beta_j^*}}}{\Gamma(\alpha^*)\sum_{j=1}^{n}\left(\beta_j^*\right)^{\alpha^*}}\ dh$$

$$= \frac{1}{\Gamma(\alpha^*)\sum_{j=1}^{n}\left(\beta_j^*\right)^{\alpha^*}} \int_0^{\infty} h \sum_{j=1}^{n} \frac{2}{(h^2)^{\alpha+1}} e^{-\frac{1}{h^2\beta_j^*}}\ dh \qquad \ldots(31)$$

وبأستبدال $h^2 = t$ نحصل على:

$$\tilde{h}(x) = \frac{1}{\Gamma(\alpha^*)\sum_{j=1}^{n}\left(\beta_j^*\right)^{\alpha^*}} \sum_{j=1}^{n} \int_0^{\infty} \frac{1}{(t)^{\alpha+1}} e^{-\frac{1}{t\,\beta_j^*}}\ dt \qquad \ldots(32)$$

يمكننا تبسيط التكامل في المعادلة (32) كالاتي:

$$\tilde{h}(x) = \frac{\Gamma\alpha\ \sum_{j=1}^{n}\left(\beta_j^*\right)^{\alpha^*}}{\Gamma(\alpha^*)\sum_{j=1}^{n}\left(\beta_j^*\right)^{\alpha^*}} \int_0^{\infty} \frac{1}{\left(\beta_j^*\right)^{\alpha^*}\Gamma\alpha\,(t)^{\alpha+1}} e^{-\frac{1}{t\,\beta_j^*}}\ dt \qquad \ldots(33)$$

يكون عرض الحزمة h بصيغة بيز هو:

$$\tilde{h}(x) = \frac{\Gamma\alpha\ \sum_{j=1}^{n}\left(\beta_j^*\right)^{\alpha^*}}{\Gamma(\alpha^*)\sum_{j=1}^{n}\left(\beta_j^*\right)^{\alpha^*}} \qquad \ldots(34)$$

b) **توزيع المشاهدات Normal Kernel:**

وبنفس الطريقة السابقة لايجاد عرض الحزمة الامثل h بتوزيع المشاهدات (Lognormal Kernel) تم استبدال توزيع المشاهدات بـ (Normal Kernel) اي:

$$K(u) = \frac{1}{\sqrt{2\pi}} e^{\frac{-u^2}{2}}, \qquad -\infty < u < \infty \qquad \ldots(35)$$

$$u = \frac{x - X_j}{h}$$

التوزيع السابق كما ذكر بالمعادلة (٢٣) وبعد استعمال التحويل لها فأن:

$$\pi(h) = \frac{2}{\Gamma\alpha\ \beta^{\alpha} h^{2\alpha+1}} e^{-\frac{1}{\beta h^2}} \quad ,\alpha > 0$$

$$,\beta > 0$$

$$,h > 0$$

ليكون التوزيع اللاحق:

$$\hat{\pi}(h|x,\underline{z}) = \frac{\sum_{j=1}^{n}(1/h^{2\alpha+2})e^{-\frac{1}{h^2}\left(\frac{1}{\beta}+\frac{1}{2}(x-X_j)^2\right)}}{\left(\Gamma((\alpha+1)/2)/2\right)\sum_{j=1}^{n}\left(\frac{1}{\beta}+\frac{1}{2}(x-X_j)^2\right)} \quad \ldots(36)$$

وبالتالي عرض الحزمة h لتوزيع المشاهدات (Normal Kernel) بصيغة بيز تكون:

$$h^*(x) = \frac{\Gamma\alpha\ \sum_{j=1}^{n}\left(\frac{1}{\beta(x-X_j)^2}+2\right)^{\alpha}}{\sqrt{2\beta}\Gamma\left(\alpha+\frac{1}{2}\right)\sum_{j=1}^{n}\left(1/\left(\beta(x-X_j)^2+2\right)\right)^{\left(\alpha+\frac{1}{2}\right)}} \quad \ldots(37)$$

٧. دليل سكرتارية الامم المتحدة لدقة العمر والنوع: [2],[3],[4],[5],[6]

هو عبارة عن مؤشر لحساب نسبة النوع والعمر في آن واحد، ويستعمل عندما تكون بيانات العمر الرقمية غير متاحة. اذ يتطلب توفر توزيع السكان حسب فئات العمر الخمسية، والذي من شأنه المساعدة على تقليل خطأ الابلاغ عن العمر. وقد اقترحه قسم السكان بالامم المتحدة لقياس دقة بيانات العمر والنوع. ويعد من انجح المقاييس، خاصة عند مقارنة دقة البيانات بين منطقتين او تعدادين او اكثر للدولة نفسها. ويعتمد حساب الدليل على ثلاث مكونات اساسية وهي:

A : متوسط مجموع الفروق المتتالية لنسبة النوع بصرف النظر عن الاشارة (+،-).
B : متوسط مجموع انحرافات نسبة العمر عن العدد ١٠٠ بصرف النظر عن الاشارة (+،-) بالنسبة للذكور.
C : متوسط مجموع انحرافات نسبة العمر عن العدد ١٠٠ بصرف النظر عن الاشارة (+،-) بالنسبة للأناث.

ويجري حساب قيمة الدليل وفق الصيغة الرياضية التالية:

$$UN_M = 3A + B + C \quad \ldots(38)$$

وتفسر نتيجة الدليل على النحو التالي:

جدول (٢)
مستوى جودة البيانات وفق دليل سكرتارية الامم المتحدة لدقة العمر والنوع

قيمة الدليل	مستوى جودة البيانات
اقل من ٢٠	البيانات دقيقة ويمكن الاعتماد عليها
من ٢٠ الى ٤٠	البيانات تعاني من ضعف بدرجة متوسطة
اكثر من ٤٠	البيانات غير صحيحة وتشوبها اخطاء كثيرة

الجانب التطبيقي: [2],[4],[5],[6]

قبل البدء في عملية تحليل البيانات تم حساب ما يلي:

اولاً: حساب معلمة التمهيد لكلا الطريقتين اي ($h : Bayes\ N$ ، $h : Bayes\ L - N$).

ثانياً: حساب مقدر (Nadaraya Watson Kernel) وحسب المعادلة التالية:

$$\hat{P}_i = \frac{\sum_{j=1}^{m} K\left(\frac{i/m - j/m}{h}\right) \bar{P}_j}{\sum_{j=1}^{m} K\left(\frac{i/m - j/m}{h}\right)} \quad \ldots(39)$$

حيث:

m تمثل عدد الفئات العمرية.

i تمثل الفئة العمرية حيث ان $(i = 1,2,\ldots,m)$.

i/m يمثل احتمال الفئة العمرية i من عدد الفئات العمرية m بحيث i/m التوزيع المكاني المشترك (co-spatial design).

$\bar{P}_i$

وان:

$$\bar{P}_i = \frac{ni}{n} \quad \ldots(40)$$

حيث:

ni يمثل عدد المشاهدات في الفئة العمرية i.

n يمثل العدد الكلي للمشاهدات.

$\bar{P}_i$ يمثل التكرار النسبي لعدد المشاهدات بالفئة العمرية i.

$\hat{P}_i$ يمثل مقدر (Nadaraya Watson Kernel) بعد تحويل البيانات الى احتمال.

K تمثل دالة Kernel.

h تمثل معلمة التمهيد.

ووفق المعادلة (39) تم تحويل البيانات الى احتمال وبالتالي اصبحت مستمرة. إذ تم اعادة التمهيد لكل خلية اي لكل فئة عمرية.

ثالثاً: تم مقارنة البيانات قبل التمهيد مع البيانات الناتجة بعد استعمال طريقة Kernel اللامعلمية بواسطة مقدر (Nadaraya Watson Kernel) وحسب طريقة حساب معلمة التمهيد ($h : Bayes\ L - N$ ، $h : Bayes\ N$) عن طريق دليل سكرتارية الامم المتحدة لدقة العمر والنوع وتحليل نسبتي العمر والنوع.

تحليل نسبتي العمر والنوع: [2],[4],[5],[6]

لمعرفة دقة احصاءات العمر يجب ان نحسب كل من نسبة النوع لكل فئة عمرية ونسبة العمر لكل فئة عمرية وكالاتي:

$$\text{نسبة النوع} = \frac{\text{عدد الذكور}}{\text{عدد الاناث}} \times 100\ \% \quad \ldots (41)$$

$$\text{نسبة العمر (للذكور للاناث)} = \frac{\text{عدد السكان في الفئة العمرية(للذكور للاناث)}}{\text{متوسط عدد السكان في الفئة العمرية السابقة واللاحقة للفئة العمرية(للذكور للاناث)}} \times 100\ \% \quad \ldots (42)$$

ومن هاتين النسبتين تم الحكم على مدى صحة البيانات الاصلية قبل التمهيد والبيانات التي تم التوصل اليها بعد استعمال طرائق التمهيد وبالتالي الكشف عن الاخطاء التي تقع في اعداد بعض الفئات إذ عملية التحليل تكون بالصورة:

- **تحليل نسبة النوع**:

عندما ندرس تحليل نسبة النوع اولاً يجب ان نركز على ان هذه النسبة لا تتغير الا بصورة متدرجة جداً وان لكل ١٠٠ من المواليد الاناث يقابلها ١٠٥ من المواليد الذكور وغالباً ما تنحصر هذه النسبة بين ١٠٢–١٠٨ والبيانات التي تقع خارج حدود هذه النسبة يكون مشكوكاً بها وهذه النسبة تأخذ بالانخفاض التدريجي لتصل في الاعمار المتوسطة الى حوالي ١٠٠% ثم تنخفض اكثر عند فئة كبار العمر لتكون اقل من ١٠٠% لدى الذكور وذلك اما بسبب الوفاة، الهجرة او ميل النساء الى تصغير اعمارهم الحقيقية ومتى ما كانت البيانات السكانية ضمن حدود النسبة اعلاه دل ذلك على ان البيانات مقبولة ويمكن الاعتماد عليها، اما بالنسبة الى انحرافات النوع تفسر بالشكل التالي:

- **اذا كانت نسبة النوع اكبر من مئة يؤدي ذلك الى انحراف موجب اي عدد الذكور اكبر من عدد الاناث.**
- **اذا كانت نسبة النوع اقل من مئة يؤدي ذلك الى انحراف سالب اي عدد الاناث اكبر من عدد الذكور.**

- **تحليل نسبة العمر**:

لتحليل نسبة العمر في المعادلة (٤٢) يتم تقسيم الفئات العمرية الى ثلاث فئات عمرية عريضة وعلى النحو التالي:

- **فئات صغار العمر (٠ الى ١٤).**
- **فئات متوسطي العمر (١٥ الى ٦٤).**
- **فئات كبار العمر (٦٥ فأكثر).**

ومن تقسيم الفئات العريضة يمكن ان تقسيم المجتمع السكاني حيث اذا كانت نسبة صغار العمر في المجتمع اكثر من ٣٥% فان هذا المجتمع هو مجتمع فتي وكلما ازادت النسبة نقول إن المجتمع يسير في النمو إلى مجتمع شاب، اما بالنسبة الى فئة كبار العمر فاذا كانت نسبتهم في المجتمع اكثر من ١٠% هذا يعني ان المجتمع يتصف على انه مجتمع شيخوخة (هرم).

بالاضافة الى ذلك عند حساب انحرافات نسبة العمر (ذكور، اناث) عن المئة اذا كان صغيرة جداً عن المئة يعني ان البيانات السكانية (ذكور، اناث) تقع ضمن المدى المقبول، لكن في بعض الاحيان نجد ان الاناث في الفئة العمرية ٥٥ فأكثر تكون انحرافات نسبة العمر لديهم اكبر من المدى المقبول وبالتالي فأن البيانات في تلك الاعمار يكون مشكوكاً فيها وذلك بسبب ميل النساء في تلك الفئات الى اعطاء اعمار اصغر من اعمارهم الحقيقية. علماً ان النتائج التي تم الحصول عليها والموضحة بالجدولين (٣) و (٤) كانت بتطبيق برنامج الماتلاب.

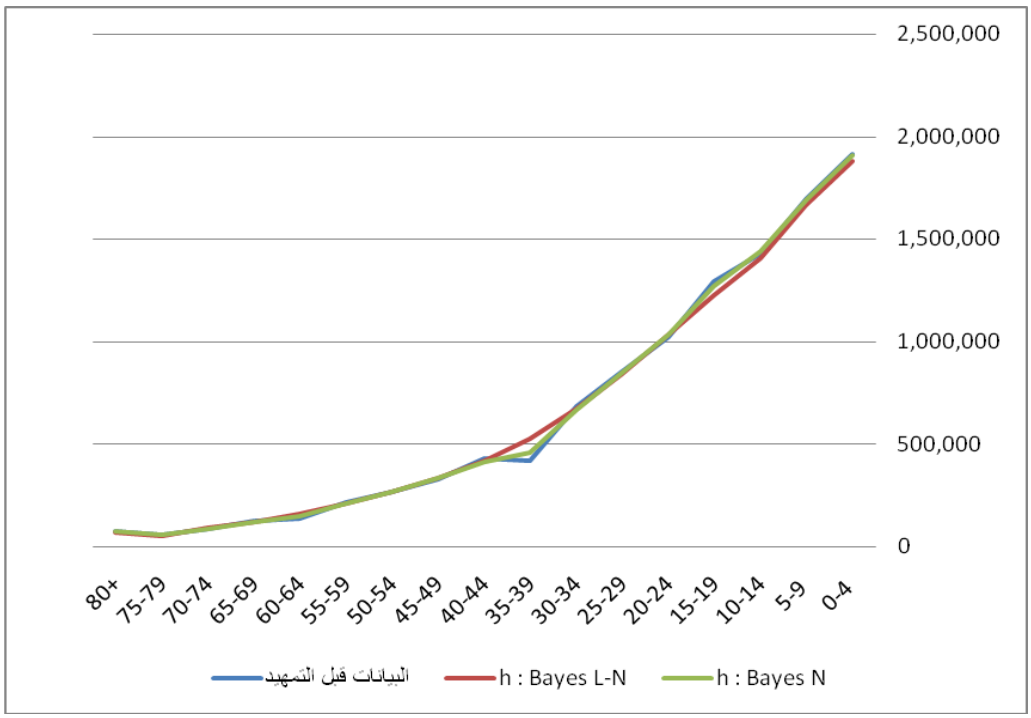
جدول (٣) تمهيد بيانات الذكور لتعداد سكان العراق عام ١٩٩٧

h : Bayes N	h : Bayes L-N	البيانات قبل التمهيد	فئات العمر
1,905,693	1,877,613	1,911,828	0-4
1,686,688	1,661,835	1,692,118	5-9
1,439,919	1,407,614	1,425,042	10-14
1,268,862	1,227,083	1,293,435	15-19
1,030,977	1,034,103	1,019,269	20-24
846,857	844,418	849,385	25-29
664,684	670,975	680,816	30-34
453,346	526,730	413,916	35-39
411,360	417,730	429,425	40-44
331,844	336,024	327,438	45-49
265,168	269,047	263,859	50-54
209,911	210,820	215,068	55-59
144,332	161,572	134,845	60-64
118,772	122,964	123,427	65-69
83,871	95,597	82,011	70-74
54,937	54,128	55,114	75-79
70,028	68,996	70,254	80+
10,987,249	10,987,249	10,987,249	المجموع

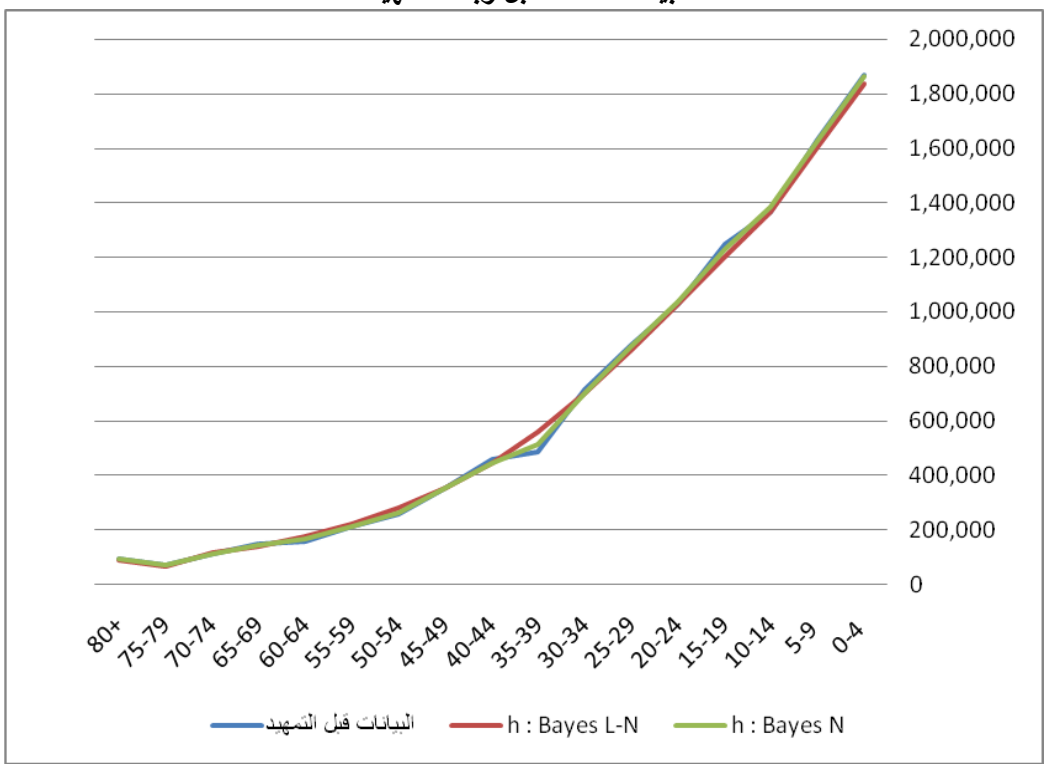
جدول (٤) تمهيد بيانات الاناث لتعداد سكان العراق عام ١٩٩٧

h : Bayes N	h : Bayes L-N	البيانات قبل التمهيد	فئات العمر
1,861,583	1,835,427	1,867,103	0-4
1,626,495	1,603,642	1,631,318	5-9
1,383,646	1,367,208	1,366,867	10-14
1,229,064	1,202,127	1,246,767	15-19
1,034,736	1,030,261	1,028,478	20-24
870,848	860,999	875,086	25-29
699,795	701,691	710,286	30-34
512,610	561,969	486,257	35-39
442,946	447,686	454,607	40-44
350,772	355,542	350,292	45-49
262,580	280,224	256,622	50-54
207,910	220,425	209,141	55-59
162,431	175,262	156,724	60-64
142,834	141,521	147,040	65-69
109,916	116,441	111,100	70-74
68,750	67,784	68,954	75-79
92,075	90,781	92,348	80+
11,058,992	11,058,992	11,058,992	المجموع

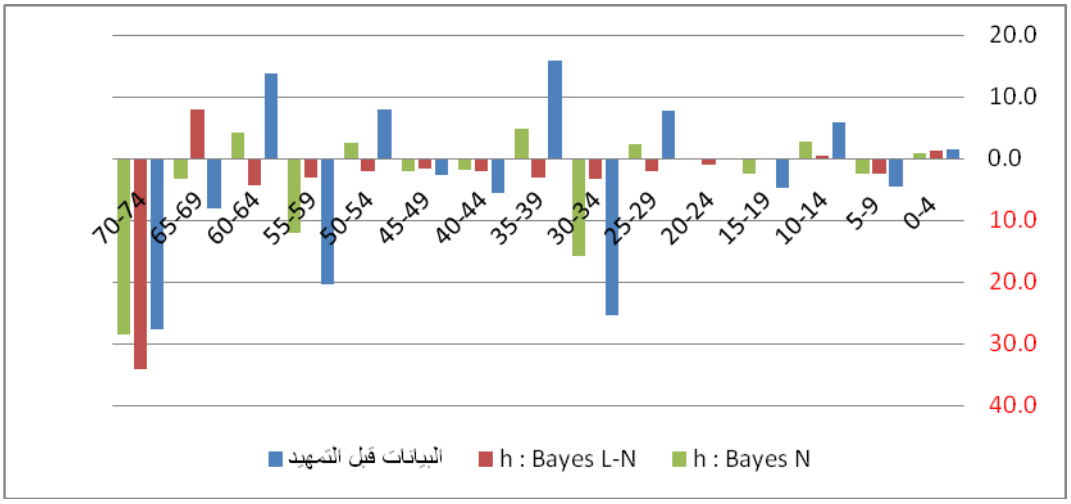
شكل (1)
بيانـات الذكور قبل وبعد التمهيد



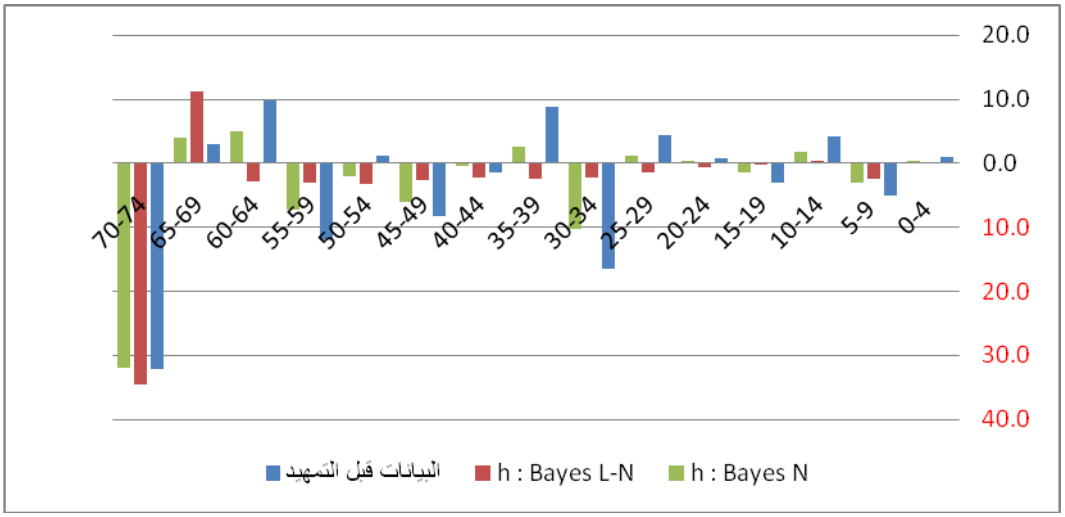
شكل (2)
بيانات الاناث قبل وبعد التمهيد



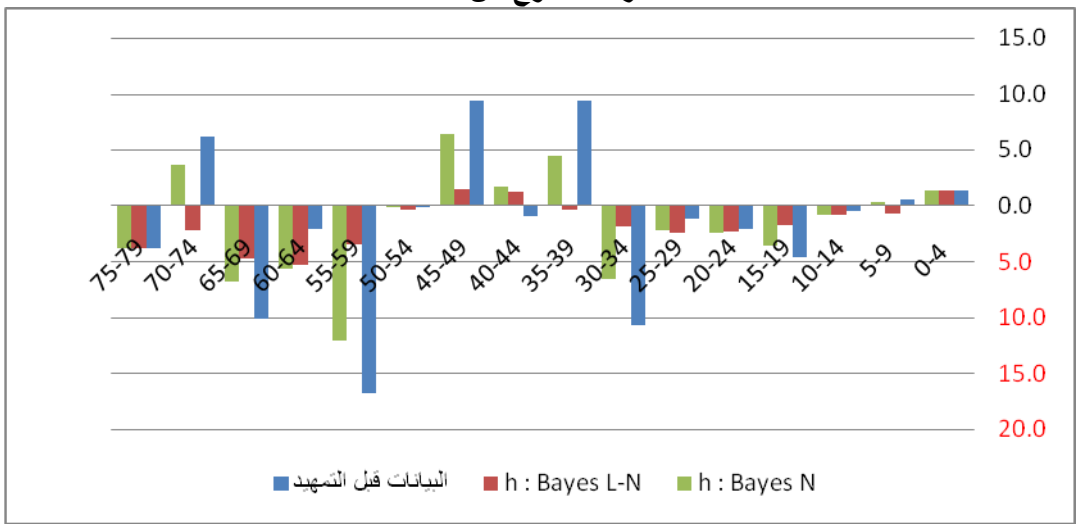
شكل (3)
انحرافات الذكور عن المئة



شكل (4)
انحرافات الاناث عن المئة



شكل (5)
انحرافات النوع عن المئة



جدول (5)
تحليل لنسبة العمر وقيمة دليل سكرتارية الامم المتحدة لدقة العمر والنوع للبيانات الاصلية والبيانات الناتجة بعد عملية التمهيد بطريقة تقدير Kernel وحسب طريقة ايجاد معلمة التمهيد

h : Bayes N	h : Bayes L-N	Real data	فئات العمر
3,767,276	3,713,040	3,778,931	0-4
3,313,183	3,265,477	3,323,436	5-9
2,823,565	2,774,822	2,791,909	10-14
9,904,024	9,753,339	9,894,276	مجموع فئة صغار العمر
22,046,241	22,046,241	22,046,241	المجموع الكلي
44.9239	44.2404	44.8797	معدل فئة صغار العمر

h : Bayes N	h : Bayes L-N	Real data	فئات العمر
261,606	264,485	270,467	65-69
193,787	212,038	193,111	70-74
123,687	121,912	124,068	75-79
162,103	159,777	162,602	80+
741,183	758,212	750,248	مجموع فئة كبار العمر
22,046,241	22,046,241	22,046,241	المجموع الكلي
3.3619	3.4392	3.4031	معدل فئة كبار العمر

22.57	15.39	32.69	دليل سكرتارية الامم المتحدة لدقة العمر والنوع

الاستنتاجات:

تم التوصل الى ان البيانات الناتجة من عملية التمهيد بطريقة تقدير Kernel وحسب طريقة ايجاد معلمة التمهيد h : Bayes L-N هي الافضل وحققت اقل قيمة لدليل سكرتارية الامم المتحدة كما وانها حافظت على نسبتي العمر والنوع.

التوصيات:

١.معالجة البيانات السكانية بالطرائق العلمية الحديثة قبل اعتمادها.

٢.المتابعة المستمرة للمؤتمرات والندوات وكذلك المجلات والمكتبات الالكترونية العلمية لمعرفة احدث الاصدارات والبحوث المنشورة في مجال تمهيد البيانات السكانية.

٣.اقامة الدورات والندوات لتوعية المواطنين بضرورة واهمية التعدادات السكانية واجراء دورات مكثفة للكادر المكلف بعملية التعداد تجنباً لحدوث الاخطاء خصوصاً بالأطراف (الارياف).

المصادر:

١. الجهاز المركزي للاحصاء وتكنولوجيا المعلومات، نتائج التعداد العام لسكان العراق عام (١٩٩٧).

٢. الجهاز المركزي للاحصاء / اليمن / تقويم بيانات التوزيع العمري والنوعي للسكان / الدراسات المعدة من بيانات تعداد ٢٠٠٤.

٣. الحميداوي، نادية علي عايد (٢٠٠٥) "الاسقاطات السكانية لمحافظة البصرة للفترة (١٩٩٧-٢٠٢٢) بأستخدام نتائج التعداد العام لسنة ١٩٩٧ في العراق"، رسالة ماجستير في الإحصاء، كلية الإدارة والاقتصاد، جامعة بغداد.

٤. الحنبلي، غيداء (٢٠٠٦) "تقييم بيانات التعداد العام للسكان عام ٢٠٠٤" منقول بواسطة منتدى الموسوعة الجغرافية http://www.4geography.com/vb.

5.Arriaga, E. E., Johnson, P. D., Jamison, E. (1994) "Population analysis with microcomputers" Presentation of techniques, Vol. I.

6. Cula, S. , HOŞGÖR, S. (2006) "APPLICATION OF KERNEL ESTIMATION METHOD FOR CORRECTION OF AGE DISTRIBUTION ERRORS IN CENSUS" NüfusbilimDergisi\Turkish Journal of Population Studies, pp. 61-71.

7. Hansen, B. E. (2009) "Lecture Notes on Nonparametrics" University of Wisconsin.

8. Kuruwita, C. (2006) "A BAYESIAN APPROACH FOR BANDWIDTH SELECTION IN KERNEL DENSITY ESTIMATION WITH CENSORED DATA" A Master's Thesis.

9. Stern, H. S. ,Gelman, A. , Carlin, J. B. , Rubin, D. B. (2003) "Bayesian Data Analysis" CHAPMAN & HALL / CRC, Texts in Statistical Science Series.

10. Turlach, B. A. (1993) "Bandwidth Selection in Kernel Density Estimation: A Review" C.O.R.E. and Institut de Statistique. Universit e Catholique de Louvain.

11. Wikipedia, the free encyclopedia, "Kernel regression, Kernel density estimation and Kernel (statistics)".

12. Zucchini, W. (2003) "APPLIED SMOOTHING TECHNIQUES Part 1: Kernel Density Estimation" http://isc.temple.edu/economics/Econ616/Kernel/ast_part1.pdf .

13. Zheng, Q. (2009) "Local Adaptive Smoothing in Kernel Regression Estimation" A Master's Thesis.

Evaluation Age and Gender for General Census of the population in Iraq by using nonparametric Bayesian Kernel Estimators

Abstract

The process of evaluating data (age and the gender structure) is one of the important factors that help any country to draw plans and programs for the future. Discussed the errors in population data for the census of Iraqi population of 1997. targeted correct and revised to serve the purposes of planning. which will be smoothing the population databy using nonparametric regression estimator (Nadaraya-Watson estimator) This estimator depends on bandwidth (h) which can be calculate it by two ways of using Bayesian method, the first when observations distribution is Lognormal Kernel and the second is when observations distribution is Normal Kernel.

then we will be compare between the result of these methods by using UN Age-Sex Accuracy Index and analysis of the Age and Gender ratios to find the method which gave the optimum smoothing for data. And we reached that the method of estimate h when observations distributed as Lognormal Kernel of Bayesian method is the best because it achieved less value of UN Age-Sex Accuracy Index.

keyword: Nadaraya-Watson Kernel estimator, inverted gamma prior distribution, Lognormal Kernel posterior distribution, Normal Kernel posterior distribution, UN Age-Sex Accuracy Index.