AL- Rafidain University College

PISSN: (1681-6870); EISSN: (2790-2293)

مجلة كلية الرافدين الجامعة للعلوم

Available online at: https://www.jrucs.iq

JRUCS

Journal of AL-Rafidain University College for Sciences

استعمال التحليل التمييزي اللبي الموقعي لتصنيف اعداد السكان في العراق حسب المحافظات

سكينه شامل جاسم منذور	شروق عبد الرضا سعيد السباح
sackineh.sh@uokerbala.edu.iq	shorouqa.@uokerbala.edu.iq
قسم الاحصاء، كلية الادارة والاقتصاد، جامعة كربلاء، كربلاء، العراق	

المستخلص

في هذا البحث تم استعمال اسلوب التحليل التمييزي الفيشر الموقعي لغرض تمييز التراكيب السكانية في محافظات العراق عن طريق اختيار مجموعات من السكان حسب المحافظات في العراق عبر مناطقهم لجغرافية وخطوطها المكانية لغرض التمييز بين الانماط المكانية للسكان ونموهم وباستعمال برنامج ماتلاب تم التوصل الى ان اسلوب التحليل التميزي الفيشر الموقعي اللبي هو الافضل من باقي اساليب التحليل التميزي وهي (التحليل التميزي الخطي – التحليل التمييزي الفيشر- التحليل التمييزي الفيشر الموقعي) , وايضا تم التوصل الى ان هنالك محافظات فيها نمو سكاني اكثر من باقي المحافظات وهي (بغداد - الانبار- البصرة- بابل- كربلاء) نسبة لمواقع المحافظات الجغرافية والنقاط المكانية للإسقاطات السكانية.

معلومات البحث

تواريخ البحث:

تاريخ تقديم البحث: 24/2/2024
تاريخ قبول البحث: 12/4/2024
تاريخ رفع البحث على الموقع: 31/12/2024

الكلمات المفتاحية:

تحليل تمييزي، اعداد السكان، تحليل التمييزي الموقعي، مواقع مشاهدات، تصنيف.

للمراسلة:

سكينه شامل جاسم منذور
sackineh.sh@uokerbala.edu.iq

DOI: https://doi.org/10.55562/jrucs.v56i1.6

1. التحليل التمييزي Discriminant Analysis

يشير التحليل التمييزي إلى تقنية إحصائية قد تحدد انتماء المجموعة استنادًا إلى مجموعة من المتنبئات القياسية التي تعتبر متغيرات مستقلة. الهدف الأساسية لهذه التقنية هي تعيين كل مشاهده (قيمة) إلى مجموعة أو فئة معينة وفقًا للسمات المستقلة للبيانات.تحليل التمييز الخطي هو تعميم لتحليل التمييز لفيشر وهو طريقة تستخدم في الإحصاء والتعرف على الأنماط وتعلم الآلة لإيجاد تركيبة خطية من الصفات التي تصنف أو تفصل صنفان أو أكثر من الأشياء أو الأحداث .

يقسم التحليل التمييزي الى صنفين :

1. التحليل التمييزي المعلمي (الخطي) تعتمد على تركيب خطي للمتغيرات .
2. التحليل التمييزي اللامعلمي (التربيعي) بينما تُستخدم دالة التمييز التربيعية عندما لا تتوفر شروط التمييز الخطي، مثل عدم تساوي مصفوفة التباين والتباين المشترك للمجموعات. يُمكن استخدام هذه الأساليب في تحليل البيانات لفهم العلاقات والاختلافات بين المجموعات المختلفة.

ستكون دالة التمييز المعلمي كالاتي:

$$x \text{ is allocated to group } g_0 \text{ if } g_0 = \underset{g \in \{1,\dots,v\}}{\operatorname{argmax}} \log(\pi_g) \; \frac{1}{2}(x \bar{x}_g)^T \sum^{-1} (x \bar{x}_g); \quad (1)$$

وبالنسبة للتحليل التمييزي اللامعلمي سيكون كالاتي:

$$x \text{ is allocated to group } j_0 \text{ if } g_0 = \underset{g \in \{1,\dots,v\}}{\operatorname{argmax}} \log(\pi_g) \; \frac{1}{2}\log|\sum_g| \; \frac{1}{2}(x \bar{x}_g)^T \sum_j^{-1} (x \bar{x}_g) \quad (2)$$

2. التحليل التمييزي لفيشر Fisher discriminant analysis

تحليل التمييز الخطي لفيشر (Fisher Discriminant Analysis) هو تقنية إحصائية تستخدم في تحليل البيانات وفصل الفئات. يهدف هذا التحليل إلى تحديد الخصائص التي تمكن من تمييز مجموعات مختلفة من البيانات. يستخدم تحليل فيشر في عدة مجالات مثل التعلم الآلي ومعالجة الصور والتعرف على الأنماط. يعتمد هذا النوع من التحليل على الاختلاف بين المجموعات والتباين داخل كل مجموعة لتحديد الخصائص التي تسهم في التمييز بينها.

3. التحليل التمييزي الفيشر الموقعي Local Fisher Discriminant Analysis

وهو اسلوب فعال لتقليل الابعاد من خلال دمج نقاط القوة لأسلوب التحليل التمييزي الفيشر والحفاظ على اسقاط المحلي (Locality Preserving Projection (LPP)) لها .تقنية خاضعة للإشراف لتقليل الأبعاد و التي تسعى الى الفصل بين الفئات (الاصناف) في المصفوفة الى اعلى حد وتقليل التشتت داخل الفئة (الصنف). ولكن يكون ادائها ضعيف عندما يكون العينات داخل الفئة (الصنف) تشكل عدة مجموعات منفصلة (اي : توزيع الفئات (الاصناف) تكون متعدد الوسائط (multimodal)) .الهدف الاساسي لطريقة LFDA هو تقدير لمستوى التشتت (الانتشار) بين و داخل الفئات (الاصناف) و بالتالي تزيد من قدرة التمييز بين الفئات و في نفس الوقت تحافظ على تركيب الموقعي داخل كل فئة (الصنف) .و بطريقة اسهل من خلال LFDA يمكن حل مشكلة القيم الذاتية المعممة وتطبيقها في اي فضاء عشوائي ذات الابعاد المنخفضة .

نفرض ان $x_i \in R^d (i = 1,2, \dots, n)$ عينات ذات البعد d و $y_i \in (i = 1,2, \dots, I)$ .n:عدد العينات و I: عدد الفئات

الغرض من هذا هو البحث على البيانات (z_i) ذات m من الابعاد من خلال مصفوفة التحويل **T** ذات البعد d×m وهي:

$$z_i = \boldsymbol{T}^T x_i. \quad (3)$$

ان $\bar{S}^{(b)}$ و $\bar{S}^{(w)}$ هما النظيران الموقعيان لمصفوفة الانتشار داخل الفئة $S^{(w)}$ و بين الفئة $S^{(b)}$على التوالي :

$$\bar{S}^{(w)} = \frac{1}{2}\sum_{i,j=1}^{n} \bar{A}_{i,j}^{(w)} (x_i - x_j)(x_i - x_j)^T \quad (4)$$

$$\bar{S}^{(b)} = \frac{1}{2}\sum_{i,j=1}^{n} \bar{A}_{i,j}^{(b)} (x_i - x_j)(x_i - x_j)^T \quad (5)$$

اذ ان :

A : مصفوفة التقارب ذات n من الابعاد وعناصرها ضمن الفترة [0,1] و التي تمثل التقارب بين x_j و x_i.

فيتم تحويل LFDA الى مصفوفة التحويل T_{LFDA} على نحو التالي :

$$T_{LFDA} = \underset{\mathrm{T} \in R^{d\times m}}{argmax} \; Tr\left((T^T \bar{S}^{(w)} T)^{-1} T^T \bar{S}^{(b)} T\right) \quad (6)$$

$$T_{LFDA} = \underset{\mathrm{T} \in R^{d\times m}}{argmax} \; Tr(T^T \bar{S}^{(b)} T) \quad (7)$$

$$\text{Subject to } \; T^T \bar{S}^{(w)} T = I \quad (8)$$

يتم تحديد T من خلال ازواج البيانات القريبة من نفس الفئة (اي المتشابه) اما بالنسبة لأزواج المختلفة (اي البعيدة) فهذا يمكن حلها بسهولة كمشكلة القيم الذاتية المعممة (Generalized).

ولحل هذه المشكلة نفرض ان $\bar{S}^{(m)}$ هي مصفوفة التشتت المختلطة المواقع (The Local Mixture Scatter Matrix) مع مصفوفة الوحدة تعرف كالاتي :

$$\bar{S}^{(m)} = \bar{S}^{(w)} + \bar{S}^{(b)} = \frac{1}{2}\sum_{i,j=1}^{n} \bar{A}_{i,j}^{(m)} (x_i - x_j)(x_i - x_j)^T \quad (9)$$

$$Tr\left((T^T \bar{S}^{(w)} T)^{-1} T^T \bar{S}^{(m)} T\right) = Tr\left((T^T \bar{S}^{(w)} T)^{-1} T^T \bar{S}^{(b)} T\right) + m \quad (10)$$

فيمكن استبدال معيار الموجود في المعادلة 7 بالمعيار $\bar{S}^{(m)}$ و يعطى تحليلاً لحل T_{LFDA} اي ان :

$$T_{LFDA} = \left(\bar{\emptyset}_1 \middle| \bar{\emptyset}_2 \middle| \dots \middle| \bar{\emptyset}_m\right) \quad (11)$$

اذ ان $\{\bar{\emptyset}_i\}_{i=1}^{d}$ هي المتجهات الذاتية المعممة المرتبطة بالقيم الذاتية $\bar{\lambda}_1 \geq \bar{\lambda}_2 \geq \cdots \geq \bar{\lambda}_d$ لمشكلة القيم الذاتية المعممة وهي :

$$\bar{S}^{(m)} \bar{\emptyset} = \bar{\lambda} \bar{S}^{(w)} \bar{\emptyset} \quad (12)$$

ويمكن التعبير عنها بشكل مصفوفة كالاتي :-

$$\bar{S}^{(m)} = X \bar{L}^{(m)} X^T \quad (13)$$

اذ ان :

$$\bar{L}^{(m)} = \bar{A}^{(m)} - \bar{D}^{(m)} \quad (14)$$

و ان $\bar{D}^{(m)}$ مصفوفة قطرية ذات n من الابعاد عناصرها تكون $\bar{D}_{i,i}^{(m)} = \sum_{j=1}^{n} \bar{A}_{i,j}^{(m)}$.

و بالمثل $\bar{S}^{(w)}$ يمكن التعبير عنها من خلال :-

$$\bar{S}^{(w)} = X\bar{L}^{(w)}X^T \quad (15)$$

اذ ان: $\bar{L}^{(w)} = \bar{A}^{(w)} - \bar{D}^{(w)}$ ، $\bar{D}_{i,i}^{(w)} = \sum_{j=1}^{n} \bar{A}_{i,j}^{(w)}$ وبالتالي تم معالجة المشكلة للقيم الذاتية المعممة في المعادلة (12) .

4. التحليل التمييزي الفيشر الموقعي اللبي Kernel Local Fisher Discriminant Analysis

يقوم اسلوب KLFDA بأداء يغير من اسلوب LFDA غير الخطي، وهو يستخدم اسلوب kernel لتقليل الأبعاد. وقد وجد أن KLFDA ينتج تحولًا مدمجًا يزيد من فصل بين الفئات ويحافظ على البنية المحلية (الموقعي) (Local) داخل الفئات في الفضاء ذو الأبعاد المنخفضة. ونظرًا لاستخدامه لحل مشكلة القيمة الذاتية المعممة، يمكن تطوير KLFDA لمجموعات بيانات عالية الأبعاد وذات موثوقية حسابية. ان التحليل التمييز الخطي يبحث عن الاتجاهات النقاط البيانات البعيدة عن بعضها البعض في الفئات المختلفة .

نفرض ان لدينا مجموعة من العينات $X=\{x_1, x_2, \dots, x_i\}$ وان $x_i \in R^n$ ينتمون الى مجموعة خطية C ذات الميزات $Z_i = a^T x_i$

و تكون دالة الهدف كالاتي :-

$$a_{opt} = argmax \frac{a^T S_b a}{a^T S_w a} \quad (16)$$

اذ ان :-

$$S_b = \sum_{k=1}^{c} I_k \, (\mu^{(k)} - \mu)(\mu^{(k)} - \mu)^T \quad (17)$$

$$S_w = \sum_{k=1}^{c} \sum_{i=1}^{j_k} \, (x_i{}^{(k)} - \mu^{(k)})(x_i{}^{(k)} - \mu^{(k)})^T \quad (18)$$

اذ ان :

μ : يمثل جميع المتوسطات المتجهات العينات .

I_k : عدد العينات في الفئة (الصنف) k-th.

$\mu^{(k)}$: معدل متجه k للفئة k-th.

$x_i{}^{(k)}$: عينة i للفئة k-th.

S_w : مصفوفة الانتشار داخل الفئة .

S_b : مصفوفة الانتشار بين الفئة .

و لحل المشكلة الذاتية هي المتجهات الذاتية الامثل $\boldsymbol{a}s$ المتناظرة مع القيم الذاتية غير صفرية اذ ان :

$$S_b \boldsymbol{a} = \lambda S_w \boldsymbol{a} \quad (19)$$

يتم انشاء رسمين بيانيين بين و داخل الفئة اي (G_b ، G_w) لكل نقاط x_i و يمكن تقسيم مجموعة $N(x_i)$ للجار الاقرب للصنف k الى مجموعتين فرعيتين و تكون كالاتي :

$N_w(x_i)$: تحتوي الجار الذي يشارك نفس التسمية x_i .

$N_b(x_i)$: تحتوي على الجار الذي لها تسميات مختلفة .

نفترض ان W_w و W_b هما مصفوفتان الاوزان لـ G_w و G_b على التوالي ،فيعرفان كالاتي :-

$$W_{b,ij} = \begin{cases} 1 & if\ x_i \in N_b(x_j) or \quad x_j \in N_b(x_i) \\ 0 & otherwise \end{cases}$$
$$W_{w,ij} = \begin{cases} 1 & if\ x_i \in N_w(x_j) or \quad x_j \in N_w(x_i) \\ 0 & otherwise \end{cases} \quad (20)$$

فيتم الرسم البياني لحل المشكلة لـداخل و بين الفئة في مساحة منخفض الابعاد و التي يجب ان تكون نقاط G_w قريبة مع البعض قدر الامكان . في حين تضل نقاط G_b بعيدة .

تعرف SS_b و SS_w بقياس جودة الخرائط و تكون كالاتي :-

$$SS_b = \sum_{ij} (Z_i - Z_j)^2 W_{b,ij} = 2\boldsymbol{a}^T X(D_b - W_b)X^T \boldsymbol{a} = 2\boldsymbol{a}^T X L_b X^T \boldsymbol{a} \quad (21)$$

$$SS_w = \sum_{ij}(Z_i - Z_j)^2 W_{w,ij} = 2\boldsymbol{a}^T X(D_w - W_w)X^T\boldsymbol{a} = 2\boldsymbol{a}^T X L_w X^T\boldsymbol{a} \quad (22)$$

اذ ان :D_w : مصفوفة قطرية والتي مدخلاتها عبارة عن اعمده او (الصفوف نظراً لتماثل) بمجموع قيم W_w اي ان :

$D_{w,ii} = \sum_j W_{w,ij}$و $L_w = D_w - W_w$ و هي عبارة عن مصفوفة لابلاس

وبالمثل $L_b = D_b - W_b$

والمعيار المناسب لاختيار افضل خريطة هو تحسين من اداء دالة الهدف و هي :

$$a_{opt} = argmax\ \frac{SS_b}{SS_w} = argmax\ \frac{\boldsymbol{a}^T X L_b X^T\boldsymbol{a}}{\boldsymbol{a}^T X L_W X^T\boldsymbol{a}} \quad (23)$$

المتجه المثلى هي المتجهات الذاتية التي تعطي اعلى قيمة ذاتية لمشكلة القيمة الذاتية المعممة.

$$XL_bX^T\boldsymbol{a} = \lambda XL_WX^T\boldsymbol{a} \quad (24)$$

بتوسيع نسخه الخطية LFDA الى RKHS(Reproducing Kernel Hilbert Spaces) عالي الأبعاد يتم اختيار Z_i غير الخطي :

$$Z_i = f(x_i) = P_\emptyset^T \emptyset(x_i) = \sum_{j=1}^m \alpha_j \emptyset(x_j)^T \emptyset(x_i) = \sum_{j=1}^m \alpha_j k(x_j, x_i) \quad (25)$$

اذ ان :$\emptyset$:دالة موقعية غير خطية nonlinear mapping function, $k(x_j, x_i)$: اللبية Mercer Kernel, $P_\emptyset$

:التحويل الاسقاطي Projecting Transformation

يمكن كتابة المعادلة (25) بالشكل الاتي :

$$\alpha_{opt} = argmax \frac{\alpha^T K L_b K \alpha}{\alpha^T K L_w K \alpha} \quad (26)$$

و بعد ادخال اللبية المتجه المثلى هي المتجهات الذاتية التي تعطي اعلى قيمة ذاتية لمشكلة القيمة الذاتية المعممة اي ان :

$$KL_bK\alpha = \lambda KL_wK\alpha \quad (27)$$

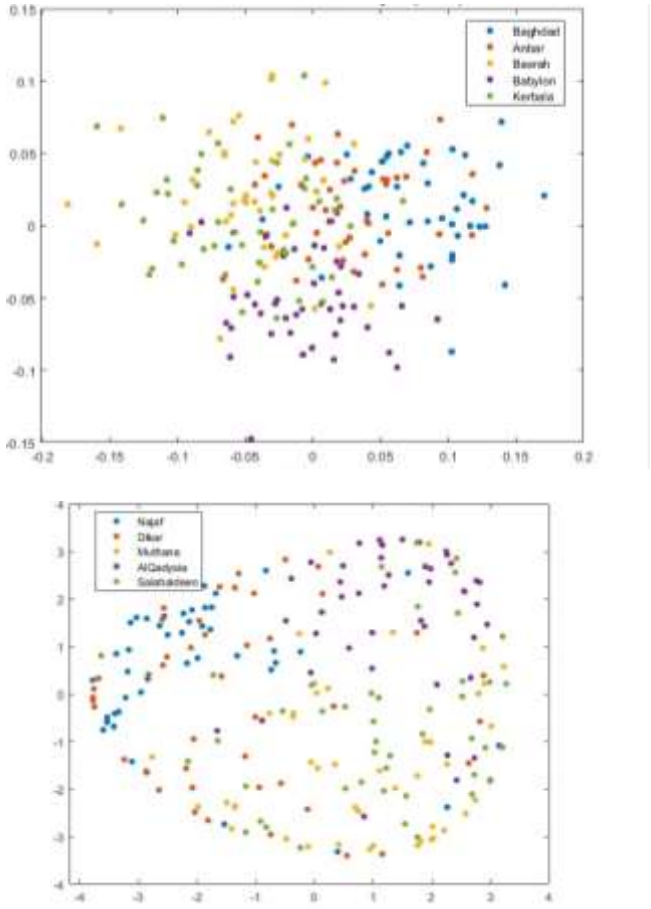
5. الجانب التطبيقي

تم اخذ بيانات تمثل اعداد السكان في العراق للمحافظات (بغداد- البصرة- بابل – الناصرية- كربلاء- صلاح الدين- القادسية-المثنى- الانبار-ذي قار) وقد واجها صعوبة في الحصول على البيانات ولكن تم تجميع البيانات من مواقع وزارة التخطيط العراقية وموقع الجهاز المركزي للإحصاء وبعض السنوات قمنا بتقديرها بواقع (45) مشاهدة من سنة (1978) ولغاية (2022) لغرض تطبيق اساليب التحليل التمييزي وهي (الخطي - الفيشر- الفيشر الموقعي –الفيشرالموقعي اللبي) وتم مقارنة الاساليب باستعمال معيار دقة التصيف وكانت النتائج كما الجدول (1) الآتي:

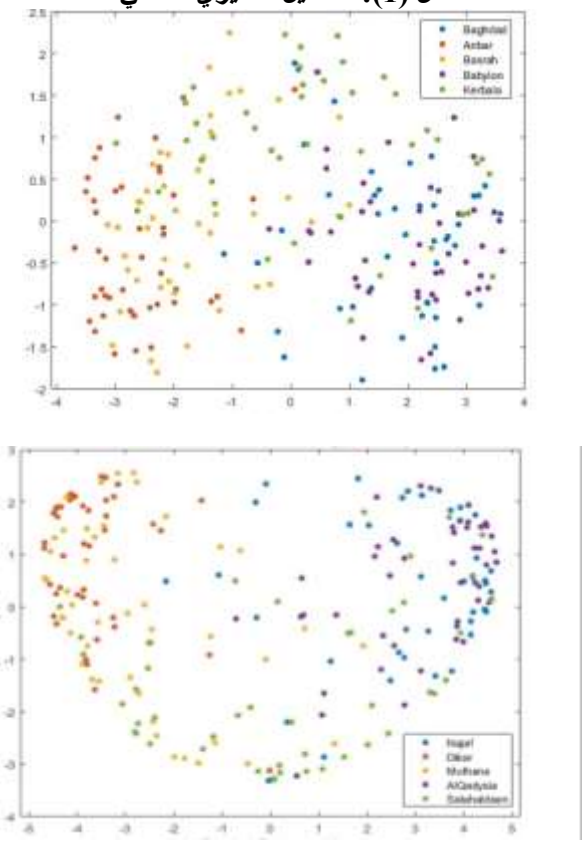
جدول (1): مقايس الدقة والتصنيف لكل اسلوب من اساليب التحليل التمييزي

اسلوب التحليل التمييزي	مقياس الدقة
التحليل التمييزي الخطي	0.42233
التحليل التمييزي الفيشر	0.71564
التحليل التمييزي الفيشر الموقعي	0.89668
التحليل التمييزي الفيشر الموقعي اللبي	0.98997

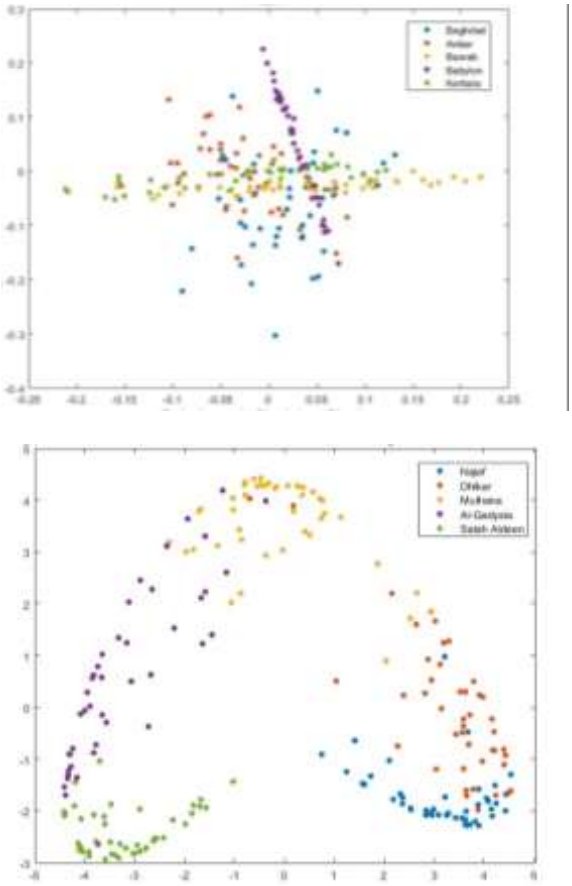
اذ نلاحظ من جدول (1) ان التحليل التميزي الفيشر الموقعي اللبي كان له اعلى مقايس دقة تصنيف بلغ (0.98997) وان التحليل التميزي الخطي له اقل معيار دقة تصنيف بلغ (0.42233) .



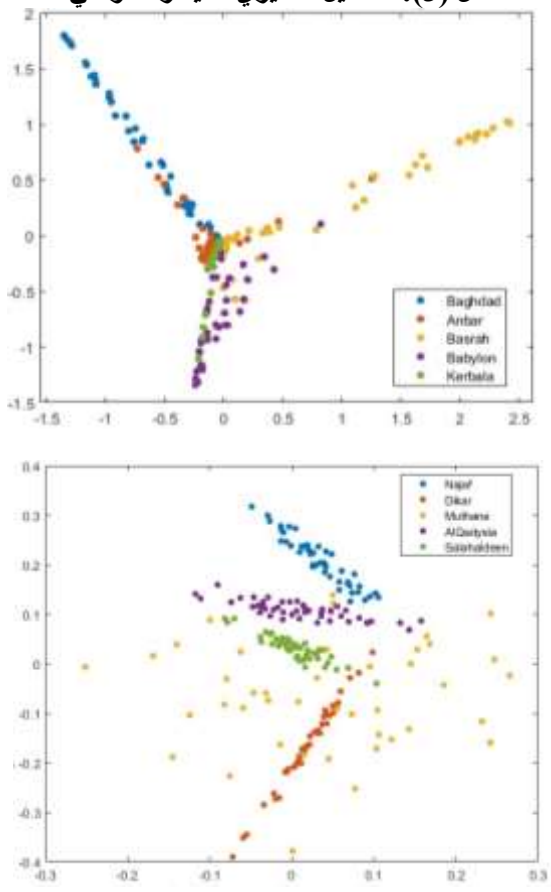
شكل (1): التحليل التميزي الخطي



شكل (2): التحليل التميزي الفيشر



شكل (3): التحليل التميزي الفيشر الموقعي



شكل (4): التحليل التميزي الفيشر الموقعي اللبي

نلاحظ من الجدول (1) والاشكال (1) الى (4) ان التحليل التمييزي الفيشر الموقعي اللبي كان هو الافضل في تصنيف اعداد السكان في العراق حسب المحافظات الذي بين بان المحافظات (بغداد – انبار – بصرة- بابل- كربلاء) حسب مواقع الجغرافية للمحافظة اكثر تكتل واكثر تقارب مما يدل على ان هذه المحافظات تعاني من كثافة سكانية عالية. اما المحافظات (النجف- ذي قار- المثنى – القادسية – صلاح الدين) اظهر التحليل التميزي لها تباعد اكثر للكتل السكانية مما يشير ان هذه المحافظات ذات كثافة سكانية واطئة مقارنة بالمحافظات في الجزء الاول من التحليل التمييزي.

6. الاستنتاجات

من خلال ما تم التوصل اليه من نتائج تم التوصل الى الاستنتاجات الآتية :

1. لمواقع المشاهدات (المحافظات) اثر كبير في تقدير اللب التحليل التمييزي .
2. استعمال اللبية في التحليل التمييزي زاد من دقة الاسلوب.
3. اسلوب التحليل التميزي الفيشر الموقعي اللبي هو الافضل من باقي اساليب التحليل التميزي.
4. المحافظات (بغداد – انبار – بصرة- بابل- كربلاء) حسب مواقع الجغرافية للمحافظة اكثر تكتل واكثر تقارب مما يدل على ان هذه المحافظات تعاني من كثافة سكانية عالية.
5. المحافظات (النجف- ذي قار- المثنى – القادسية – صلاح الدين) اظهر التحليل التميزي لها تباعد اكثر للكتل السكانية مما يشير ان هذه المحافظات ذات كثافة سكانية واطئة.

المصادر

[1] سكينة شامل جاسم، (2015) .، "مقارنة بعض اساليب تقدير المعلمة التمهيدية لدوال اللب متعددة المتغيرات وتوظيفها في الدوال التمييزية مع تطبيق عملي"، رسالة ماجستير ، كلية الادارة والاقتصاد، جامعة بغداد .

[1] Chang.S, Cheng. Y, Wei . C, Jen .M, (2012)," kernel local fisher discriminant analysis based manifold regularized SVM model for financial distress predictions" , Expert Systems with Applications , journal homepage: www.elsevier.com/locate/es

[2] Ghojogh.B , Karray. F, Crowley.M , (2022) , " Fisher and kernel fisher discriminant analysis Tutorial", arXiv: 1906.09436.

[3] Gramacki, A. (2018). Nonparametric Kernel Density Estimation and Its Computational Aspects. Springer.

[4] Li, C. and Wang, B. (2014) Fisher Linear Discriminant Analysis. CCIS Northeastern University..

[5] Wen .G, Markey.M, (2024), " the use of the kernel local fisher discriminant analysis for the channelization of the Hotelling model observer", Medical Imaging 2015: Image Perception, Observer Performance, and Technology Assessment, 94160Y (17 March 2015); doi: 10.1117/12.2081956.

[6] Zhang, Q., & Suganthan, P. N. (2019). Dimensionality reduction using kernel discriminant analysis with a generalization of pairwise scatter matrix. Neurocomputing, 325, 71-81

Journal of AL-Rafidain University College for Sciences (2024); Issue 56; pp. 61-68

AL- Rafidain University College

PISSN: (1681-6870); EISSN: (2790-2293)

Journal of AL-Rafidain University College for Sciences

Available online at: https://www.jrucs.iq

JRUCS

Journal of AL-Rafidain University College for Sciences

Using Kernel Local Fisher Discriminant Analysis to Classify Population in Iraq According to Government

Sakina S. Jassim Mandour	Shorouk A. Saeed Al-Sabbah
sackineh.sh@uokerbala.edu.iq	shorouqa.@uokerbala.edu.iq

Department of Statistics, Faculty of Administration and Economics, Kerbala University, Kerbala, Iraq

Article Information

Article History:
Received: February, 24, 2024
Accepted: April, 12, 2024
Available Online: December, 31, 2024

Keywords:
Discriminant analysis, Kernel estimator, population numbers, estimator, locations of observations, classification

Correspondence:
Sakina S. Jassim Mandour
sackineh.sh@uokerbala.edu.iq

Abstract

In this paper , the method of spatial discriminant analysis was used for the purpose of distinguishing population structures in the governorates of Iraq by selecting groups of residents according to the governorates in Iraq across their geographical regions and spatial lines for the purpose of distinguishing between the spatial patterns of the population and their growth. Using the Matlab program, it was concluded that the method of spatial discriminant analysis Fisher's on-site discriminant analysis is better than the rest of the discriminant analysis methods, which are (Fisher's linear discriminant analysis - core discriminant analysis - on-site discriminant analysis), and it was also found that there are governorates with more population growth than the rest of the governorates, which are (Baghdad - Anbar - Basra - Babylon - Karbala) Relative to the geographical locations of the governorates and the spatial points of the population projections.

DOI: https://doi.org/10.55562/jrucs.v56i1.6