3-0: تمهيد

تعتبر الطرق الإستكشافيه الإبتدائيه ذات فائده في فهم الطبيعه المعقده لعلاقـة المتغيـرات المتعدده .ويعتبر اسلوب فحص البيانات من اجل التوصل الي هيكل للتجمعـات "الطبيعيـه" مـن الأساليب الإستكشافيه الهامه فهي تساعدنا في تحديد الأبعاد والتعرف على القيم الشاذه بالإضافه الى اقتراح روابط او فروض تصف العلاقة بين المتغيرات.

لا يضع تحليل التجميع أي فروض تخص عدد المجموعات أو هياكلها مما يجعله اسـلوب بدائي لوصف التجمعات .وتتم عملية التجميع بناء على أوجه التماثل أو الإختلاف وتعـد مقـاييس التماثل أو البيانات في حساب التماثل مدخلات لإسلوب تحليل التجميع.

كثيراً ما نهتم بتجميع عدد من الوحدات في مجموعات متجانسة داخلياً في ضوء مشاهدات ذات بعد P مأخوذة من هذه الوحدات في المتغيرات $x_1, x_2, \ldots, x_p$ مثلاً تجميـع أشـخاص حسـب قراءات أخذت منهم إلى مجموعات قد تمثل أجناس أو أعراف مختلفة . كذلك قد يهمنا تجميع عـدد من المتغيرات في مجموعات قد ترتبط المتغيرات في كل منها بعامل معين.

3-1 المبحث الأول: مقاييس البيانات الرقمية Measures for numerical data

الهدف الأساسي من تحليل التجميع هو اكتشاف التجمعات الطبيعيه للمفردات أو المتغيرات لذا يجب علينا ان نتوصل أولاً الى مقياس كمي لقياس التوافق "التماثل" بين المفردات.

في أغلب الأحيان يمكننا تجميع المفردات بمجرد النظر في اشكال انتشارها "ذات بعدين او ثلاثه" وذلك على الرغم من عدم وجود تعريف دقيق لمفهوم التجميع الطبيعي .

وللإستفاده من المقدره العقليه على تجميع الأشياء المتشابهه فقد تم التوصل حديثآ الى اساليب بيانيه عديدة لرسم المشاهدات ذات الأبعاد المتعدده في بعدين فقط.

3-1-1: المسافة Distance

عند محاولة تجميع وحدات لابد أن يكون لدينا أولاً مقياس للمسافة بين أي وحدتين مثلاً A و B. وهناك عدة مقاييس للمسافة منها ، المسافة الإقليدية Euclidian Distance والإنحراف Deviance وهو مربع المسافة الإقليدية ومقياس مهالونبس للمسافة Mahalanobis Distance

والذي يسمي أيضاً المسافة الإحصائية ،مسافة مانهاتن Manhattan Distance ، المسافة القصوى Maximum Distance ، مسافة منكوفسكي Minkowski Distance ، المسافة المتوسطة Average Distance. فكبر المسافة يشير إلى وجود عدد كبير من الأزواج غير المتماثلة.[1]

وفيما يلي نتناول هذه المقاييس بإيجاز :

3-1-2: المسافة الإقليدية Euclidean Distance

المسافة الاقليدية معروفة لنا جميعاً فإذا نظرنا إلى النقطة $P=(x_1,x_2)$ في المستوى فإنه وفقاً لنظرية فيثاغورث يمكن إيجاد المسافة الخطية بينها وبين نقطة الأصل $0=(0,0)$ ويرمز لهذه المسافة بالرمز $d=(0,p)$ كما يلي:

$$d=(0,p)=\sqrt{x_1^2+x_2^2}\ldots\ldots(1-3)$$

وعموماً إذا كان لدينا نقطة في فراغ ذي p من الأبعاد ، أي إذا كان لدينا النقطة $p=(x_1,x_2,\ldots,x_p)$ فإن المسافة بينها وبين نقطة الأصل $0=(0,0,\ldots,0)$ هي :

$$d(0,p)=(\sqrt{x_1^2+x_2^2+\ldots+x_p^2}$$

ويمكن تعميم هذا المفهوم لقياس المسافة بين وحدتين A وB فاذا كانت $\underline{X}_A$ و$\underline{X}_B$ مشاهدات ذات بعد P (فيP متغير) على الوحدتين A وB فإن المسافة الإقليدية بينهما تعرف:[2]

$$\bar{d}_{\Delta}=\sqrt{(\underline{X}_A-\underline{X}_B)'(\underline{X}_A-\underline{X}_B)}\ldots\ldots(2-3)$$

3-1-3: الانحراف Deviance

يعرف الإنحراف للمسافة بين وحدتين A وB كما يلي:

---

1. ريشلرد جونسون ، ديرن وشرن ، التحليل الإحصائي للمتغيرات المتعددة من الوجهة التطبيقية ، تعريب . عبدالمرضي حامد عزام - دار المريخ للنشر، المملكة العربية السعودية – 1418هـ - 1998م. ص (855).

2- عزة أحمد – تجميع مستشفيات ولاية الخرطوم – بحث تكميلي لنيل درجة الماجستير في الإحصاء –جامعة النيلين – يناير 2009م ص(11)

$$(\underline{X}_A - \underline{X}_B)'(\underline{X}_A - \underline{X}_B) = \text{الإنحراف}^{3} \ldots\ldots\ldots\ldots (3-3)$$

أي أنه مربع المسافة الإقليدية.

3-1-4: مسافة مهالانوبس Mahalanobis Distance

يتم إيجاد هذه المسافة عن طريق العلاقة :

$$d_{mah} = \sqrt{(X-Y)\Sigma^{-1}(X-Y)'} \ldots\ldots\ldots\ldots (4-3)$$

Σهي عبارة عن مصفوفة التباينات والتغايرات.[4]

3-1-5: المسافة القصوى : Maximum Distance

ويطلق على هذه المسافة أيضاً المسافة الجزئية (sup distance) إذ نوجد أقصى مسافة بين المتغيرات فمثلاً إذا كان لدينا المتغيرين X,Y في d من الأبعاد المسافة القصوي بينهما توجد كما يلي:[5]

$$d_{\max}(x,y) = \max_{1\le k\le d}\left|x_j - y_j\right| \ldots\ldots\ldots\ldots (5-3)$$

3-1-6: مسافة منكوفسكي: Minkowski Distance

المسافة الإقليدية ومسافة مانهاتن والمسافة القصوى عبارة عن ثلاث حالات خاصة من مسافة منكوفسكي وهذه المسافة تعرف بـ. :

$$d_{\min}(x,y) = (\sum_{j=1}^{d}\left|x_j - y_j\right|^{r})^{\frac{1}{r}},,,, r \ge 1 \ldots\ldots\ldots\ldots (6-3)$$

عندما:

3-عزة أحمد – مرجع سايق ص(12)

[4] - ملرتن تي ويلز- وأخرون – تجميع البيانات – سلسلة إصدارات الجمعية الأمريكية الأمريكية (فرجينيا)– جمعية الصناعات والتطبيقات الرياضية(بينسيلفيتيا) – الولايات المتحدة الأمريكية – سبتمبرر-2007م. ص(72)

[5] - ملرتن تي ويلز- وأخرون – مرجع سابق ص (72)

$$r = 1,2 and \infty$$

تنتج المسافة الإقليدية مسافة مانهاتن والمسافة القصوى على التوالي.

3−1−7: المسافة المتوسطة Average Distance:

هذه المسافة يمكن إيجادها من المسافة الإقليدية كالآتي:

أفرض أن لدينا المتغيرين x,y ذات بعد d المسافة المتوسطة تصبح [6]:

$$d_{ave}(x,y) = \left(\frac{1}{d_j} \sum \left(x_j - y_j \right) \right)^{\frac{1}{2}} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(7-3)$$

[6] - مارتن تي ويلز - وأخرون – مرجع سابق ص (74)

3-2 طرق التجميع والدمج Clustering Method and Amalgamation

وفي هذا المبحث سنتناول بإيجاز المفاهيم الأساسية المرتبطة بتجميع وحدات في ضوء مشاهدات $\underline{X}'=\left[x_1, x_2, \ldots, x_p\right]$ مأخوذة منها.

توجد ثلاث طرق مختلفة يمكن استخدامها لعنقدة أو تجميع البيانات وهي:

1-التحليل العنقودي بواسطة الـk متوسط : يستخدم عند معرفة عدد العناقيد المطلوبة حيث تدخل قبل اجراء التحليل.

2- التحليل العنقودي الهرمي: يستخدم عندما يكون عدد البيانات صغير.

3- التحليل العنقودي ذو الخطوتين : يستخدم عندما يكون عدد البيانات كبير جداً أوعندما تكون البيانات مختلطة ما بين المتغيرات المتصلة والمتغيرات الوصفية.

وسوف يقتصر الباحث على التحليل العنقودي الهرمي والتحليل العنقودي بواسطة الـ .K متوسط.

3-2-1الدمج: Amalgamation

مثلما تتطلب عملية التجميع تحديد مقياس للمسافة نحتاج أيضاً لتحديد طريقة لدمج العناقيد. وهناك ثلاثة طرق للدمج هي طريقة الربط المفرد وطريقة الربط الكامل وطريقة الربط المتوسط والتي يتم تناولها بإيجاز كما يلي:

3-2-2 طريقة الربط المفرد: Single Linkage Method

تعتبر واحدة من أسهل طرق التحيل العنقودي – عرفت بواسطة فلوريك Florek (1951م)- والربط المفرد يعرف بعدة تعريفات منها الجار الأقرب ، طريقة الحد الأدنى توظف أقرب الجارين لدراسة او قياس الاختلاف بين مجموعتين، ويظهر كما مبين:

الشكل(3-1) الربط المفرد

المصدر : اعداد الدارس باستخدام برنامج MS-Word

في هذه الحالة ومن الشكل (3−1) يمكن ان نفترض ان a هي اقرب وحدة في المجموعة الاولي الي المجموعة الثانية وb هي اقرب وحدة في المجموعة الثانية الي المجموعة الاولي عليه اعتماد الماسفة بين a و b لقياس الاختلاف بين المجموعتين يسمي الربط المنفرد

والصيغة الرياضية لهذه الطريقة تكون كما يلي:

أفرض أن لدينا $C_i, C_j \& C_k$ عبارة عن ثلاث مجموعات فالمسافة D بين C_k و $C_i \cup C_j$ يمكن الحصول عليها من صيغة لانس – ويليامز Lance – Williams كالآتي[7]:

$$
\begin{aligned}
&D\left(C_k, C_i \cup C_J\right) \\
&= \frac{1}{2} D\left(C_k, C_i\right) + \frac{1}{2} D\left(C_K, C_j\right) - \frac{1}{2}\left|D\left(C_k, C_i\right) - D\left(C_k, C_j\right)\right| \\
&= \min\left\{D\left(C_k, C_i\right), D\left(C_k, C_j\right)\right\} \quad \ldots\ldots (8-3)
\end{aligned}
$$

3−2−3: طريقة الربط الكامل Complete Link Method

هذه الطريقة أكثر حذراً إذ تعتبر المسافة بين أي مجموعتين هي المسافة بين أبعد وحدتين فيهما كما يلي:

الشكل(3-2) الربط الكامل[8]

[7] - مارتن تي ويلز وآخرون- تجميع البيانات – مرجع سابق ص (118)

[8] - د. ريتشارد جونسون ، ديرن وشرن، مرجع سابق ص (866)

المصدر : اعداد الدارس باستخدام برنامج MS-Word

فلنفرض ان a هي ابعد وحدة في المجموعة الاولي من المجموعة الثانية و b هي ابعد وحدة في المجموعة الثانية من المجموعة الاولي عليه يتم استخدام هاتين الوحدتين لقياس الاختلاف بين المجموعتين.

والصيغة الرياضية لهذه الطريقة تكون كما يلي:

أفرض أن لدينا $C_i, C_j \& C_k$ عبارة عن ثلاث مجموعات فالمسافة D بين C_k و $C_i \cup C_j$ يمكن الحصول عليها من صيغة لانس – ويليامز Lance – Williams كالآتي:[9]

$$\begin{aligned} & D(C_k, C_i \cup C_J) \\ & = \frac{1}{2} D(C_k, C_i) + \frac{1}{2} D(C_K, C_j) + \frac{1}{2} \left| D(C_k, C_i) - D(C_k, C_j) \right| \\ & = \max \{ D(C_k, C_i), D(C_k, C_j) \} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (9-3) \end{aligned}$$

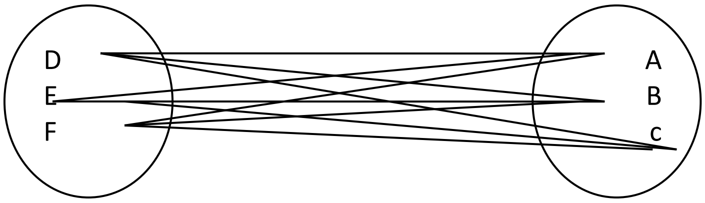
4-2-3 :طريقة الربط المتوسط :Average Linkage Method

طريقة الربط المتوسط تنظر للمسافة بين مجموعتين على أنها متوسط المسافة بين جميع الأزواج التي ينتمي أحد عناصرها إلى إحدى المجموعتين بينما ينتمي العنصر الآخر إلى المجموعة الأخرى ، كما يلي:

الشكل (3-3) الربط المتوسط[10]

[9] - مارتن تي ويلز وآخرون- تجميع البيانات – مرجع سابق ص (121).

[10] - د. ريتشارد جونسون ، ديرن وشرن، مرجع سابق ص (866).



المصدر : اعداد الدارس باستخدام برنامج MS-Word

في هذه الحالة اذا افترضنا (A,B,C) تنتمي للمجموعة الاولي و (D,E,F) تنتمي الي المجموعة الثانية حيث يتم حساب المسافات بين الوحدات في المجموعة الأولي مع المجموعة الثانية لإيجاد متوسط المسافة لاستخدامها في قياس الاختلاف بين المجموعتين

والصيغة الرياضية لهذه الطريقة تكون كما يلي:

أفرض أن لدينا $C_i, C_j \,\&\, C_k$ عبارة عن ثلاث مجموعات فالمسافة D بين C_k و $C_i \cup C_j$ يمكن الحصول عليها من صيغة لانس – ويليامز Lance – Williams كالآتي:[11]

$$D(C_k, C_i \cup C_J) = \frac{|C_i|}{|C_i| + |C_j|} D(C_k, C_i) + \frac{|C_j|}{|C_i| + |C_j|} D(C_k, C_j) \ldots\ldots (10-3)$$

أفرض أن C, C' عبارة عن مجموعتين غير خاليتين عليه يمكن إيجاد المسافة بطريقة الربط المتوسط الآتي:

$$D(C, C') = \frac{1}{|C||C'|} \sum_{x \in c, y \in c'} d(x, y) \ldots\ldots (.11-3)$$

أفرض أن C_1, C_2, C_3 عبارة عن ثلاث مجموعات غير خالية عليه أفرض أن :

$$D(C_i, C_j) = \frac{1}{n_i n_j} \sum (C_i, C_j), 1 \le i \le j \le 3 \ldots\ldots (12-3)$$

حيث:

[11] -مارتن تي ويلز وآخرون- تجميع البيانات – مرجع سابق ص (123)

$n_i = |C| \, \& \, n_j = |C_j|$ والمجموع الكلي لمسافات المجموعات $\sum(C_i, C_j)$ لـ . C_i, C_j

هذا يعني أن:

$$\sum(C_i, C_j) = \sum_{x \in C_i y \in C_j} d(x, y)$$

ومن المعادلات (3−11) و(3−12) نستنتج:

$$\begin{aligned}
&D(C_k, C_i \cup C_g) \\
&= \frac{n_2}{n_2 + n_3} D(C_1, C_2) + \frac{n_3}{n_2 + n_3} D(C_1, C_3) \\
&= \frac{n_2}{n_2 + n_3} \cdot \frac{1}{n_1 n_2} \sum(C_1, C_2) + \frac{n_3}{n_2 + n_3} \cdot \frac{1}{n_1 n_2} \sum(C_1, C_3) \\
&= \frac{1}{n_1(n_2 + n_3)} \sum(C_1, C_2 \cup C_3)
\end{aligned}$$

بعد ذلك:

$$\sum(C_1, C_2) + \sum(C_1, C_3) = \sum(C_1, C_2 \cup C_3)$$

وهذا يتحقق أيضاً للمعادلة (3−11)

هناك عدة طرق للتجميع توفرها النظرية الإحصائية ونتناول فيما يلي بشكل موجز إثنين من أهم هذه الطرق وهي طريقة التجميع المتدرجة وطريقة الـ(K) متوسط واللتين سيتم تطبيقهما في التجميع وسنفترض أن المطلوب هو تجميع وحدات وليس متغيرات وهي الحالة التي تهمنا

3−2−5 طرق التجميع المتدرجة: Clustering Methods

هناك نوعان من طرق التجميع المتدرجة :

1− طريقة الفصل المتدرجة Divisive hierarchical methods ،حيث نبدأ بمجموعة تتضمن جميع العناصر ثم تقسم هذه المجموعات إلى مجموعتين فرعيتين بحيث تكون العناصر الموجودة في مجموعة منها "بعيدة" عن العناصر الموجودة في المجموعة الأخرى ، يتم بعد ذلك تقسيم كل

من هاتين المجموعتين إلى مجموعات فرعية غير مماثلة ، نستمر في ذلك حتى يكون لدينا عدداً من المجموعات الفرعية مساوياً لعدد العناصر ، أي حتى يكوِّن كل عنصر مجموعة بنفسه.[12]

2- طريقة التجميع المتدرجة Agglomerative Hierarchical Methods وهي إحدى طرق الإدماج "linkage methods" المتتالية التي تلائم تجميع المفردات وكما تلائم تجميع المتغيرات وهذا لا يتحقق بالنسبة لجميع طرق الإدماج المتدرجة الأخرى ، وفيما يلي خطوات تنفيذ الطرق التجميع المتدرجة لإدماج المجموعات عند وجود N من العناصر (المفردات أو المتغيرات) وتسمى هذه الخطوات بالطريقة العامة:[13]

1- نبدأ بعدد N من المجموعات ، كل مجموعة بها عنصر واحد ونوجد مصفوفة المسافة (أو مصفوفة قيم معامل التماثل المستخدم) $D = d_{ik}$ وهي مصفوفة متماثلة أبعادها $(N \times N)$

2- نبحث في مصفوفة المسافة عن أقرب زوج من المجموعات (الزوج الأكثر تماثلاً). نفترض أن d_{uv} تشير الي المسافة بين الزوج الأكثر تماثلاً U,V.

3- ندمج المجموعة U مع المجموعة V، ونستخدم الرمز $(U\ V)$ للاشارة الى المجموعة الجديدة ، ونعدل من عناصر مصفوفة المسافة على النحو التالي:

(a) نحذف الصفوف والأعمدة المناظرة للمجموعتين U، V.

(b) نضف صفاً وعموداً جديدين يعطيان المسافة بين المجموعة الجديدة$(U\ V)$ والمجموعات الأخرى

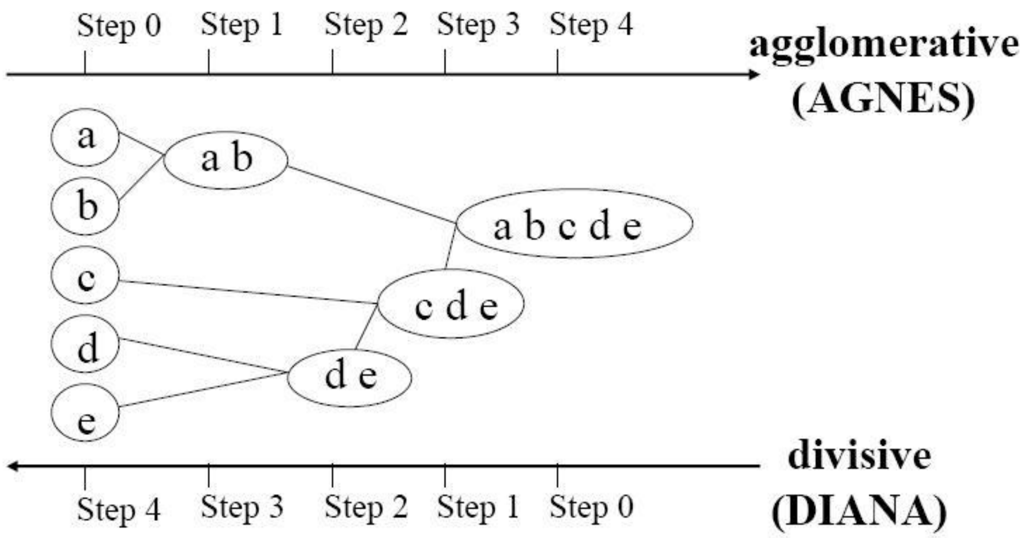
4-نكرر الخطوتين (a) ، (b) عدد $N-1$ من المرات (في النهاية تتجمع جميع العناصر في مجموعة واحدة) نقوم بتسجيل هوية المجموعات التي أدمجت والمسافات (أو قيم معامل التماثل المستخدم) التي تم عندها الإدماج.

[12] www.norusis .com /pdf/ spc-v13 pdf.

[13] - د. ريتشارد جونسون ، ديرن وشرن، مرجع سابق ص (867).

ويمكـن عـرض نتـائج هـاتين الطـريقتين بيانيـاً فـي فـراغ ذي بعـدين فـي شـكل بيـاني يعـرف باسـم" الدندوجرام "Dendogram ".

الشكل (3-4) طريقة التجميع المتدرجة وطريقة الفصل المتدرج[14]



1. **المصدر :** فرح عبدالله محمد ,(2010) ،"**تصنيف الولايات السودانيه ذات الخصائص الديمغرافية المتشابهة بإستخدام التحليل العنقودي للعام 2002 م**" , جامعة السودان للعلوم والتكنولوجيا.

نلاحظ في الشكل اعلاه في حالة التجميع المتدرج في الخطوة Step0 اعلي الشكل ان كل مفردة تعتبر عنقود لوحدها وعند اجراء عملية التجميع ننتقل الي الخطوة Step1 لنجد ان العنقودين a,b تم دمجها في عنقود واحد نسبة لوجود تقارب بينهما اما البقية (c,d,e) تظل في عناقيد لوحدها ، وفي الخطوة Step2 نلاحظ ان العنقودين (d,e) تم جمعهما معا في عنقود واحد نسبة لدرجة التقارب بينهما والبقية (a,b) في عنقود و (c) في عنقود لوحدها ، في الخطوة Step3 تم دمج العنقود (c) مع العنقود (d,e) ليصير لدينا عنقودين فقط هما (c,d,e) و (a,d) اما في الخطوة Step4 تم تجميع اخر عنقودين من الخطوة Step3 في عنقود واحد كبير يجمع جميع العناقيد في الخطوة Step0 في عنقود واحد وبذلك تكون طريقة التجميع المتدرج وصد وصلت الي آخر خطوة.

---

[14] كي –بينق زاهنق – وسائل التحليل العنقودي في البيانات البحثية -رسالة دكتوراة -جامعة ماكوري –سدني – أستراليا 2007م.

اما في خطوة الفصل المتدرج اسفل الشكل ففي الخطوة Step0 نجد ان هذه الطريقة تعتبر كل العناصر عبارة عن عنقود واحد كبيرة وعند الانتقال للخطوة الاولي في الطريقة Step1 نجد انه تم فصل العنصرين (a,b) عن العنقود الكبير لتصبح عنقود لوحدها والبقية تصبح (c,d,e) وفي عند الانتقال الي الخطوة Step2 نجد ان العنصر (c) انفصل عن العنقود الكبير واصبح عدد العناقيد 3 وهي (a,b) و (c) و (d,e) وفي الخطوة Step3 نجد ان العنقود (d,e) قد انفصل الي عنقودين هما (d) و (e) فاصبح عدد العناقيد اربعة وهي (a,b) و (c) و (d) و (e) وفي الخطوة الختامية لطريقة الفصل المتدرج تكون كل العناصر قد اصبحت في عناقيد كل علي حده.

ونلاحظ ان كلا الطريقتين تعملان عكس بعضهما الاولي تجمع بالتدريج والاخري تفصل بالتدريج

6-2-3 مصفوفة القرابة[15] Proximity Matrix:

هي مصفوفة متماثلة مدخلاتها عبارة عن مربع المسافة الإقليدية لكل زوج من الحالات الجدول رقم (3-1) يمثل نموذج لمصفوفة القرابة ، ونلاحظ أن قيم القطر الرئيسي (0) مما يعني أنه لاتوجد فروق بين الحالة ونفسها وكذلك القيمة (2.6) تعني أن المسافة بين (E) و(D) هي عبارة عن أصغر قيمة في المصفوفة مما يعني أنهما متماثلين بدرجة عالية ، وأيضاً نجد القيمة (12.8) تعني أن المسافة بين (F) و(G) هي أكبر المسافات الإقليدية مما يعني أنهما على درجة عالية من عدم التماثل.

1-د. محفوظ جودة ، التحليل الإحصائي المتقدم باساخدام **SPSS** دار وائل للطباعة والنشر- عمان الأردن الطبعة الأولى 2008م ص(100)

جدول (3-1) مصفوفة القرابة

samples	A	B	C	D	E	F	G
A	0	7.5	10.4	6.4	4.7	6.6	8.8
B	7.5	0	12.4	9.3	5.6	9.2	8.8
C	10.4	12.4	0	5.4	3.7	9.1	10.6
D	6.4	9.3	5.4	0	2.6	5.4	10.7
E	4.7	5.4	3.7	2.6	0	4.7	5.4
F	6.6	9.2	9.8	5.7	4.8	0	12.8
G	8.8	8.8	10.6	10.7	5.4	12.8	0

المصدر :المرجع نفسه، ص[19]

3-2-7: شكل الألواح الجليدية: The Icicle Plot

هو عبارة عن شكل يبين الطريقة التي تتم بها دمج الحالات في كل خطوة من خطوات التحليل العنقودي ، ويسمي شكل الالواح الجليدية لأن لها نفس شكل الألواح الجليدية المتشكلة من مساقط الماء.[16] ولكي نتمكن من فهمه لابد أن نقرأه من أسفل إلى أعلى.

2-www.norusis .com /pdf/ spc-v13 pdf

شكل (3-5) الالواح الجليدية[17]

Case

			5			**8**		**6**		**7**		**3**		**2**		**4**		**1**	
عدد العناقيد	**1**	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
	2	x	x	x	x	x	x	x	x	x	x	x	x	x	x		x	x	x
	3	x	x	x	x		x	x	x	x	x	x	x	x	x		x	x	x
	4	x	x	x	x		x		x	x	x	x	x	x	x		x	x	x
	5	x	x	x	x		x		x	x	x	x	x		x		x	x	x
	6	x	x	x	x		x		x		x	x	x		x		x	x	x
	7	x		x	x		x		x		x	x	x		x		x	x	x
	8	x		x	x		x		x		x	x	x		x		x		x

المصدر :المرجع نفسه ، ص[20]

3-2-8: خطوات التجميع[18] :The Agglomeration Schedule

شكل الألواح الجليدية لا يمكّن من معرفة أصغر مسافة إقليدية بين الحالتين أو الحالات التي تم دمجها في عنقود لذلك علينا اللجوء لما يسمى بخطوات التجميع جدول (3-3) فنجد في هذه الخطوة عمود يسمى المعامل Coefficient به قيم المسافات الإقليدية قيم (التماثل) التي أستخدمت لتكوين العناقيد .أما عمود (Stage Cluster First appears) يبين لنا الخطوة التي يظهر فيها العنقود لأول مرة .

[17] www.noursis.com/pdf/spc/-v13pd-

[18] - محفوظ جودة ، التحليل الإحصائي المتقدم باساخدام SPSS مرجع سابق ص (100)

جدول (3–2) خطوات التجميع

المرحلة	الانضمام للعناقيد		المعاملات	الظهور اول مرة		المرحلة
	العنقود 1	العنقود 2		العنقود 1	العنقود2	التالية
1	4	5	2.566	0	0	4
2	1	11	2.694	0	0	5
3	3	10	3.903	0	0	4
4	3	4	4.023	3	1	6
5	1	8	4.279	2	0	11
6	3	15	4.336	4	0	13
7	2	9	5.205	0	0	14

المصدر : المرجع نفسه، ص[21]

3–2–9: الديندوجرام The Dendrogram

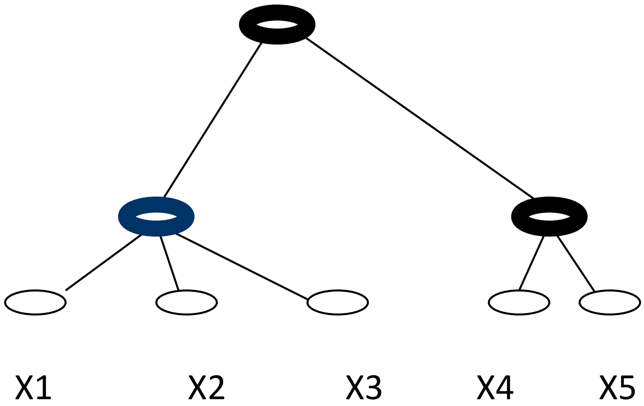
هو عبارة عن شكل بياني يوضح المسافات التي تم فيها تكوين العناقيد ويظهر ذلك في الشكل رقم (3-7) ويقرأ من الشمال إلى اليمين الخطوط العمودية تشير إلى العناقيد المدمجة فموقع الخط يشير إلى المسافة التي تمت فيها عملية الدمج وتكوين العناقيد ، فأول عمود رأسي يقابل أصغر مسافة إقليدية ولكي نفهم شكل الديندوجرام لابد لنا ان نعرف مايسمى بالـ n-tree . :

أفرض ان $D=(x_1,x_2,...,x_n)$ وأن J مجموعة جزئية من D تحقق الشروط التالية:

1. J∈D
2. Empty set $\Phi \in$ J
3. For all i=1,2,3,...,n $\{x_i\} \in J$
4. if A, B ∈B ∩ J, then A $\in \{\Phi, A, B\}$.

ا لـ n-tree في الشكل (3-6)العقد الطرفية المرسومة بدائرة فقط من غير تظليل تبين البيانات بصورة مفردة ،أما العقد المرسومة بدوائر مظللة المجموعات أو العناقيد.

الشكل (3-6) الشجرة ن – 5-tree[19]



X1 الي X5 = الوحدات التي تتكون منها العناقيد

المصدر : المرجع نفسه ، ص [22]

من ذلك نستطيع أن نصف شكل الديندوجرام بأنه شجرة تعرف با لـ.(5-tree) يحقق الشرط التالي:

$$h(A) \leq h(B) \Leftrightarrow A \subseteq B.....................................(13-3)$$

لكل مجموعة جزئية من البيانات A وB بشرط $A \cap B \neq \Phi$

عندما h(A) &h(B) تعرف ا لارتفاعات لـ.(A,B) على التوالى .

من الشكل أعلاه أفرض أن (h_{ij}) عبارة عن إرتفاع زوج البيانات(X_i,X_j) فالقيمة الأصغر لـ.(h_{ij}) تمثل أكبر معامل تماثل أو تشابه بين(X_i,X_j) وكذلك أكبر قيمة لـ.(h_{ij}) تمثل أصغر قيمة لمعامل التماثل أو تشابه بين((X_i,X_j)).

ا لإرتفاع في الديندوجرام يحقق الشرط التالي:

$$h_{ij} \leq \max\{h_{ik}, h_{jk}\} \forall j,k \in \{1,2,...,n\}......................(14-3)$$

ويمكن عرض شكل الديندوجرام رياضياً بالمعادلة التالية:

---

[19] - مارتن تي ويلز وآخرون- تجميع البيانات – مرجع سابق ص (112)

$$C:[(0,\infty)\rightarrow E(D)]$$

التي تحقق:

$$c(h)\subseteq c(h^{/})\rightarrow if\ \leq h^{/}$$

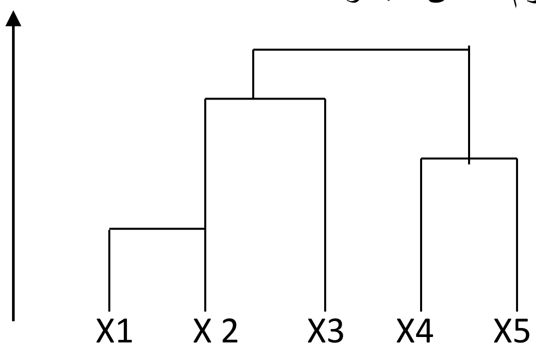
c (h) is eventually in D×D

$c(h+\delta)=c(h)$ for some small δ >0

عندما يكون لدينا المجموعة D و E(D) عبارة

عن مجموعة من العلاقات المتكافئة في D ويمكن تمثيل ذلك بالشكل التالي:

شكل (3-7) ديندوجرام لخمس مجموعات [20]



المصدر : المرجع نفسه ، ص [23]

X1 الي X5 = الوحدات التي تتكون منها العناقيد

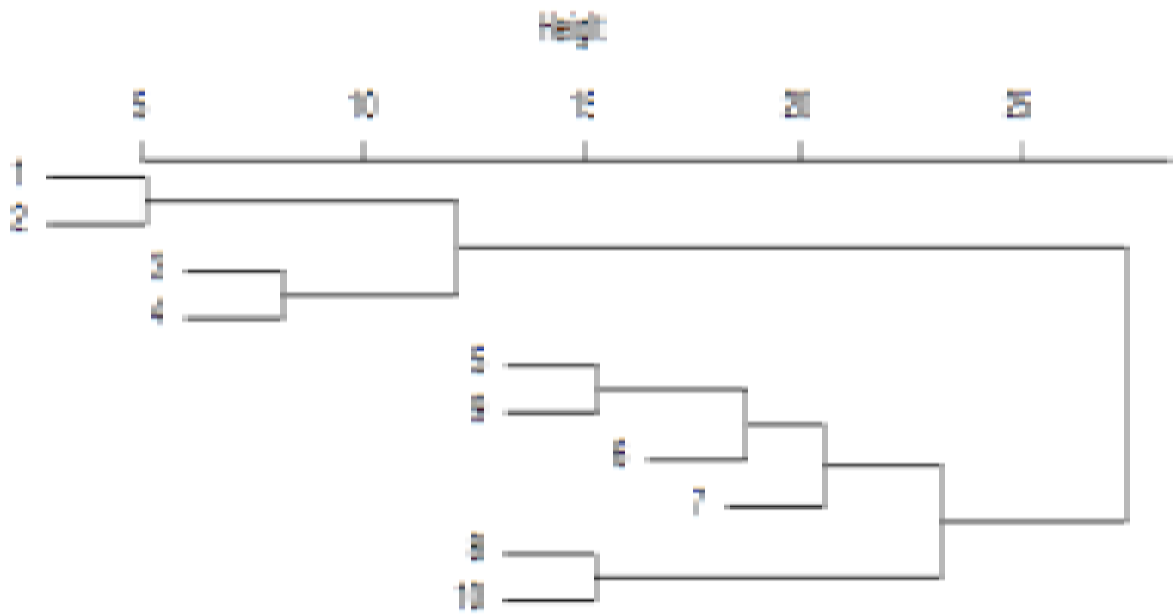
المعادلة (c) تحوي المعلومات في شكل الديندوجرام ويمكن بيان ذلك من الشكل (3–8) كالآتي:

$$c(h)=\begin{cases}\{(i,i):i=1,2,3,4,5\} if\ 0\leq h<1,\\ \{(i,i):i=3,4,5\}\cup\{(i,j):i,j=1,2\} if\ 1\leq h<2\\ \{(3,3)\}\cup\{(i,j):i,j=1,2\}\cup\{(i,j):i,j=4,5\} if\ 2\leq h<3 \\ \{(i,j):i,j=4,5\}\cup\{(i,j):i,j=1,2,3\} if\ 3\leq h<4\\ \{(i,j):i,j=1,2,3,4,5\} if\ 4\leq h\end{cases} ....(15-3)$$

شكل (3-8) الديندوجرام[21]

[20] - مارتن تي ويلز وآخرون- تجميع البيانات – مرجع سابق ص (112)

www.uga.edu/strata/software/pdf/cluster[21]



المصدر : المرجع نفسه، ص [24]

3-2-10:طريقة K من المتوسطات

هي إحدى طرق التجميع غير المتدرجة Nonhierarchical Clustering Methods لتجميع المفردات وليس المتغيرات في K من المتوسطات، إقترح MacQueen عام 1976م اسم K من المتوسطات(K-means) لوصف طريقته وتعتبر من أبسط طرق العنقدة وأكثرها شيوعاً وذات كفاءة عالية ، وتستخدم ما يسمى النقاط المركزية (Centroids) التي تضع كل مفردة في المجموعة التي يكون وسطها الحسابي أقرب لها ، وتتكون هذه الطريقة في أبسط صورها من ثلاث خطوات[22] هي:

1- تقسيم المفردات إلى K من المجموعات الأولية.

2- ضع كل مفردة من المفردات الموجودة في المجموعة التي يكون وسطها الحسابي أقرب ما يكون لها (عادة ماتستخدم المسافة الإقليدية لحساب المسافة سواء تم ذلك باستخدام المشاهدات الفعلية

22 - www.ranger.uta.edu/chqding/paper/k-means.com.

أو المشاهدات المعيارية) نعيد حساب الوسط الحسابي للمجموعة التي أضيفت إليها المفردة الجديدة وللمجموعة التي فقدت منها المفردة.

3- نكرر الخطوة الثانية إلى أن تتوقف عملية توزيع المفردات على المجموعات ، وبدلاً من أن نبدأ في الخطوة الأولي بتجزئة جميع المفردات إلى K من المتوسطات من المجموعات الأولية ، فإنه يمكننا البدء بتحديد k, (نقاط الأساس) ثم نقوم بتنفيذ الخطوة الثانية.

ويعتمد التوزيع النهائي للمفردات على المجموعات إلى حد ما ، على التقسيم الأولى المستخدم أو على الاختيار الأولى لنقاط الأساس .

توجد بعض الحجج القوية التي تؤيد عدم تحديد عدد المجموعات K مسبقاً[23] ومن بينها:

1- إذا وقعت نقطتين أو أكثر من نقاط الأساس ، دون تعمد ، في مجموعة واحدة، فإنه من الصعب التمييز بين المجموعات الناجمة عن ذلك.

2- قد يؤدي وجود مشاهدة شاذة في الحصول على مجموعة واحدة على الأقل متناثرة المفردات.

3- وحتى إذا تكون المجتمع من Kمن المجموعات ، فإن طريقة المعاينة المستخدمة يمكن أن تؤدي إلى عدم احتواء العينة على بيانات من المجموعة الأصغر. في هذه الحالة، يؤدي توزيع البيانات على K من المجموعات إلى وجود مجموعات ليس لها معني.

المثال التالي يوضح طريقة التجميع بواسطة الـ .(K) متوسط:

أفرض أننا نستطيع قياس المتغيرين X_2، X_1 لكل مفردة من المفردات كما في الجدول التالي:

جدول (3-3) قيم المتغيرينX_1، X_2

[23]د. ريتشارد جونسون ، ديرن وشرن، مرجع سابق ص (893).

المفردة	**X1**	**X2**
A	5	3
B	-1	1
C	1	-2
D	-3	-2

المصدر : (فرح عبدالله ، [2010])

الهدف هو تقسيم هذه المفردات لمجموعتين أي K=2 بحيث يكون قرب المفردات الموجودة في أي مجموعة من بعضها البعض أكبر من قربها من مفردات المجموعات الأخرى.

- نقوم بتقسيم المفردات عشوائياً لمجموعتين(AB) والمجموعة (CD).
- نقوم بحساب إحداثيات المركز $(\overline{X}_1, \overline{X}_2)$ لكل مجموعة كالآتي:

جدول (3–4) احداثيات مركز X_1 ,X_2

المجموعة	$\overline{X}_1$	$\overline{X}2$
AB	2	2
CD	−1	−2

المصدر : المرجع نفسه ، ص [26]

نقوم بحساب المسافة الإقليدية بين كل مفردة وبين إحداثيات نقطتي المركز بين كل مفردة وبين إحداثيات نقطتي المركز ونعيد توزيع كل مفردة على المجموعة الأقرب لها ، وإذا انتقلت مفردة ما من مجموعة إلى أخرى يجب تعديل نقاط المركز المناظرة قبل مواصلة العمل كالآتي:

$$d^2(A.(AB)) = (5-2)^2 + (3-2)^2 = 10$$
$$d^2(A,(CD)) = (5+1)^2 + (3+2)^2 = 61$$

حيث أن المفردة A أقرب إلى المجموعة (AB) منها للمجموعة (CD) فتبقي في مجموعتها.

وبمواصلة حساب المسافات المربعة :

$$d^2(B.(AB)) = (-1-2)^2 + (1-2)^2 = 10$$
$$d^2(B,(CD)) = (-1+1)^2 + (1+2)^2 = 9$$

وبالتالي نقوم بوضع المفردة B في المجموعة (CD) لنحصل على المجموعة (BCD) كما نحصل على نقاط المركز الجديدة التالية:

جدول (3−5) احداثيات المركز الجديدة لـ X_1, X_2.

المجموعة	$\overline{X}_1$	$\overline{X}2$
A	5	3
CDB	−1	−1

المصدر :المرجع نفسه ، ص [27]

ومرة أخرى نقوم بفحص كل مفردة لمعرفة إمكانية إعادة توزيع المفردات وبحساب المسافات الإقليدية نحصل على مايلي:

جدول (3−6) توزيع المجموعات النهائي[24]

المجموعة	**A**	**B**	**C**	**D**
A	0	40	41	89
CDB	52	4	5	5

المصدر : المرجع نفسه ، ص [27]

من ذلك يكون كل مفردة قد تم وضعها في المجموعة التي يكون مركزها أقرب ما يكون لها بالتالي تتوقف العملية أي أننا نحصل في النهاية على المجموعتين A و(BCD).

[24] ريتشارد جونسون ، ديرن وشرن، مرجع سابق ص (887)