الفصل الخامس: نماذج التصنيف " Modèle de classification "

تمثل نماذج (تقنيات) التصنيف جزئا من الاحصاء الاستكشافي المتعدد الابعاد. حيث تهدف الى توضيح هيكل مجموعة من المعطيات الكبيرة الحجم, والتي تسمح بصياغة الفرضيات التي يتم اختبارها في مراحل لاحقة. تتميز هذه الطرق عن طرق التصنيف التي لها غرض التفسير او التنبؤ.

V-1 تقديم الطريقة

فيما يلي سوف نتطرق الى التعريف بالطريقة , الهدف منها وكذا بعض المفاهيم الاساسية

V-1 -1 التعريف بالطريقة

تجدر الاشارة الى ان الادبيات التي تناولت طرق التصنيف متعددة , يمكن الاشارة الى أعمال Jain et Dubes عام 1988, Everitt عام 1993 , Mirkinعام 1996, Han et al. في 2001و Ghosh في 2002. كما يمكن ان نجد مجموعة هامة جدا من الاعمال الحديثة الخاصة بالتقنيات التصنيف في ابحاث كل من Becher و غيرهم (2000), Han و Kamber (2001), Berkhin (2002) , Andritsos (2002) و Portier (2003).

هناك العديد من المصطلحات المستعملة في الادبيات الدالة على تقنية التصنيف (Classification) , من بينها: التصنيف الاتوماتيكي او الالي (Classification automatique) , التحليل العنقودي (Analyse typologique) , التصنيف الرقمي (Taxonomie numérique) وهذا في البيولوجيا والاحياء, علم تصنيف الامراض (Nosologie) في الطب , التقسيم في نظرية التمثيل البياني (Partition . dans la théorie des graphes)

المصطلح الانجليزي والذي يعبر عن تقنية التصنيف هو بمعنى التجميع "Clustering" (التصنيف الغير خاضع للإشراف = Non supervised classification). أما المصطلح الانجليزي " Classification" (Supervised classification) فيستعمل في حالة وجود اقسام تم تعريفها من قيل ويتعلق الامر حينها بتصنيف مفردة جديدة في احد هذه الاقسام مع ضمان حد ادنى من خطأ التصنيف (Nakache et Confais, 2003).

V-1 -2 ميادين التطبيق

تقنيات التصنيف تستعمل في مجالات عدة. نذكر منها على سبيل المثال:

- في ميدان الطب : نجد أن مسؤول عن مصلحة استشفائية يبحث عن تجميع المرضى في مجموعات جزئية مختلفة من أجل تحديد السلوك العلاجي الذي يجب أخذه أمام نوع معين من المرضى.
- في ميدان التسويق : من أجل تسويق أول منتوج , فان المسؤول عن مصلحة التسويق للمؤسسة يمكنه أن يبحث عن تشكيل مجموعات من المدن المتماثلة بالنضر الى عدة معايير و يختار من اجل كل مجموعة المدينة النموذجية المستعملة كسوق اختباري.

- ميدان السياسة : كمترشح في الانتخابات المحلية يحدد برنامجه الانتخابي بالمقارنة بالبرامج الانتخابية الاخرى بالاعتماد على مختلف الخصائص او الامتيازات.
- ميدان الاشهار : تتمثل في مسالة تكييف حملة اعلانية مع مجموعة متجانسة من الافراد (متصفحي مجلة معينة , ...).
- ميدان المناخ: كمثال يمكن تصنيف بعض المدن الفرنسية من خلال 12 متغيرة او سلوك والتي تمثل متوسط درجات الحرارة خلال 12 شهر وذلك لمدة 30 سنة .

مجالات التطبيق متعددة والتي جاءت في مختلف الادبيات التي تناولت تعدين البيانات " Data mining" والتي تناولت جوانب عدة :

- تحليل المعطيات المكانية « Analyse de données spatiales",
- الوراثة القياسية "Génétique quantitative»,
- المعلوماتية الحيوية " Bio-Informatique", ...

في جميع هذه الحالات فانه هذه التقنية تهدف الى تبسيط حقيقة معقدة والتي من اجلها يكون أي تصنيف قبلي لا يفرض على الفور.

V-1 -3 الهدف من طرق التصنيف:

تتمثل في البحث عن تجميع الافراد (تصنيف تصاعدي) المشكلة لمجموع ما (او تجزئة هذه المجموعة في حد ذاتها (تصنيف تنازلي)) والتي تسلك سلوكا مماثلا وفقا لمجموعة من المتغيرات ، معايير, إلى أقسام متجانسة. حيث أن هذه العملية ينتج عنها تمثيل بياني يعرف باسم مخطط الشقوق أو" شجرة الدندروغرام = Dendrogramme " أو شجرة التصنيف الهرمي.

كما تهدف ايضا الى استكشاف البيانات, تقليصها في النهاية , تأكيدها سواء عن طريق قبول او رفض مجموع الفروض المناقشة, تصرف الافراد وفقا لطبيعة المجموعة التي يتواجدون فيها . تجميع الافراد في أقسام يمكن ان يولد فرضيات يمكن اختبارها

طرق التصنيف لمجموعة من الافراد يمكن تقسيمها الى نوعين اساسين : التصنيف حسب التقسيم " Classification par partition" او "Partitionnement" و التصنيف الهرمي. يستعمل النوع الاول عدد معين من الفئات في البداية , حيث يرمي هذا النوع على العموم الى تقسيم المعطيات لمجموعة ما الى K فئة مختلفة. وبهذا يكون كل فرد منتمي الى الفئة باعتبار المسافة الاقرب او معيار التشابه.

الصنف الثاني من الطرق والذي ينتج سلسلة من الاقسام المتداخلة من الاقل حجم الى الاكثر حجم , فانه يؤدي الى نتائج في شكل مخطط الشقوق أو" شجرة الدندروغرام = Dendrogramme " أو شجرة التصنيف الهرمي. والتي تسمح بمرئية نظام الاقسام المرتبة عن طريق التضمين.

في هذا الفصل سوف نقوم بتسليط الضوء فقط على طرق التصنيف الهرمي التي تعتمد على مفهوم المسافة او معيار التشابه وطرق التصنيف الغير هرمي والتي تتبنى مبدأ العشوائية في الاختيار المبدئي لعدد الفئات غير أن كلتا الطريقتين تعتمد على خوارزمية تراكمية معينة " Algorithme agglomératif" .

2-V طريقة التصنيف الهرمي "La classification hiérarchique (CH)"

تعتبر طريقة التصنيف الهرمي من اشهر طرق الاحصاء الوصفية، المتعددة الابعاد .والتي تستعمل كثيرا على غرار طرق التحليل العاملي ACP، AFC ، حيث تستخدم على نطاق واسع خصوصا في البيولوجيا وعلوم البيئة والاحياء وغيرها. وهذا من اجل توزيع عناصر مجموعة أم A في شكل أقسام أو أفواج. بحيث كل فوج يكون متجانس قدر الامكان (تجانس بين عناصر الفوج الواحد) ، كما يجب على الفروع أن تكون مختلفة او متنافرة قدر الامكان عن بعضها البعض. وهذا انطلاقا من جداول البيانات المستطيلة أو ذات المدخلين أي الجداول التي تعتمد عليها طرق التحليل العاملي ACP و AFC ، والتي تحوي بيانات كمية أو بيانات وصفية. وبالتالي يمكن اعتبار هذه الطريقة كدعامة للنتائج المحصل عليها في طرق التحليل العاملي.

في الغالب ما نكون غير راضيين عن هذه التقسيمات الحاصلة عن هذه الطريقة، لذلك نسعى إلى إنشاء تسلسل هرمي لمجموع الافراد المشكلة للمجموعة وذلك وفقا لفرض مجموعة من القيود المختلفة.

1-2-V بعض المفاهيم الاساسية " Notions de bases"

- **التصنيف الهرمي:** ويعني به انه إذا كان لعنيا مجموعة العناصر A.فإن التصنيف الهرمي لهذه المجموعة هو مجموعة الاجزاء ذات الاربع خصائص:
  - الجزء الفارغ (المجموعة الخالية) جزء منه.
  - المجموعة A نفسها هي جزء منه.
  - الاجزاء الصغيرة او المخفضة الى عنصر واحد هي جزء منه.
  - اذا كان X و Y جزء من A فإن X و **Y** منفصلين ، أو X يحتوي Y ،أو Y يحتوي X .
- **شجرة التصنيف:** هي رسم بياني متجذر بحيث : الاوراق هي الاجزاء التي تحتوي على عنصر واحد، والتي توجد دائما في التسلسل الهرمي، الجذر هو المجموعة الكاملة والتي تكون دائما في التسلسل الهرمي حيث ان الاوراق تمثل القمة. كل جزء له سلف واحد، باستثناء الجذر الذي ليس له أي سلف. والا سنجد جزأين متداخلين وهذا غير موجود في التسلسل الهرمي.

يكون التسلسل الهرمي جيدا او قيم ، إذا تمكنا من ربط كل جزء ‘‘ Partie ’’ بقيمة عددية تحقق لنا التعريف التالي:

$$X \subseteq Y \Leftrightarrow f(x) \leq f(y)$$

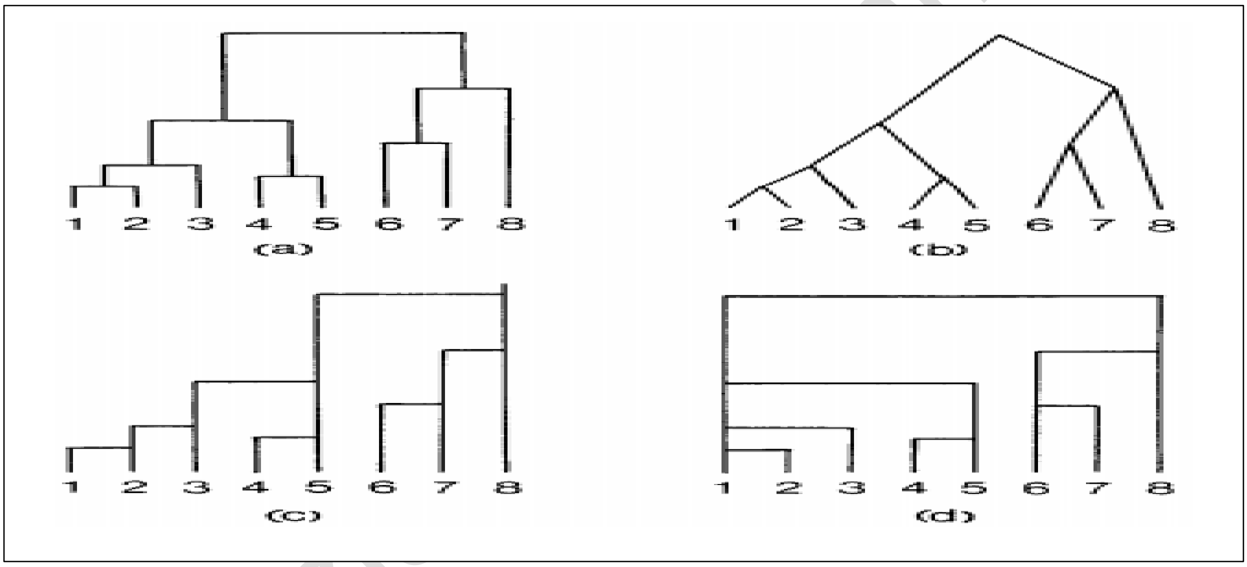
حيث تضع هذه القيمة الاوراق في الاسفل والجذر في الاعلى.

- تجدر الاشارة إلى انه يمكن الحصول على عدد كبير من التمثيلات المتجذرة (شجرة التصنيف) لذلك نلجأ الى فرض بعض القيود وكذا استخدام بعض المؤشرات من اجل الحصول على تمثيل قيم يجمع بين الافراد الاكثر تجانسا الى الاقل تجانسا.

مثال: لتكن لدينا المجموعة A فإن العناصر {1.2.3.4.5.6.7.8} =A فإنه يمكن الحصول على الشجرة التالية والتي تمثل تسلسلا هرميا قيما مشكلا من مجموع الاجزاء:

{{Ø، {1} ، {2}، {3} ، {4} ، {5}، {6}، {7}، {8}, a = {1،2}, b= {1،2،3} ، c = {4،5}، d = {1،2،3،4},E = {6،7}، f = {6،7،8} , g= {1،2،3،4،5،6،7،8}}.

يمكن الحصول على اشكال مختلفة لشجرة التصنيف مثل :



- **المسافة بين المفردات "Distance entre individus"**

تعتمد طريقة التصنيف الهرمي اساسا على مفهوم المسافة بين الافراد. حيث يمكن تقدير عدم التجانس لجزء (يحوي مجموعة من افراد المجموعة E) اعتمادا على المسافة بين المفردات المحتوات داخل ذلك الجزء(القسم). كما يمكن تقدير التفرقة (Dissimilarité) ما بين جزئين مختلفين بالاعتماد على المسافة بين فردين ينتميان كل منهما على قسم مختلف.

هناك العديد من الطرق لتقدير المسافة بين الافراد. على غرار المسافة الجينية (تستخدم في علم الوراثة)، المسافة البيئية (تستخدم في علم البيئة)، المسافة المورفومترية .....وغيرها من المسافات الاخرى. ولكن لكي تكون أكثر دقة فمن الافضل تحديد الاختلافات والتحدث عنها بشكل عام. فإذا كان لدينا «n» مفردة مشكلة للمجموعة A ،فإنه يمكن الحصول على مصفوفة التفرقة ’’ Matrice de dissimilarité ‘‘ والتي تتميز

بكونها مصفوفة مربعة (m،m). جميع قيمها اكبر او يساوي 0 . ومتناظرة بالنسبة للقطر المساوي الى 0 . أي انها تحقق مجموعة الشروط التالية:

$$1 \le i \le n \qquad\qquad \Rightarrow \quad d_{ii} = 0$$
$$1 \le i \le n \quad 1 \le j \le n \quad \Rightarrow \quad d_{ii} \ge 0$$
$$1 \le i \le n \quad 1 \le j \le n \quad \Rightarrow \quad d_{ij} = d_{ij}$$

هذه التفرقة(او الاختلاف) هو متري Métrique او قياسي بحيث :

$$1 \le i \le n \quad 1 \le j \le n \quad 1 \le k \le n \quad \Rightarrow \quad d_{ij} \le d_{ik} + d_{kj}$$

حيث يمكن التعبير عنه بالمسافة الاقليدية او مربع المسافة الاقليدية « Distance euclidienne » . وهو من اكثر المسافات استعمالا. وهي ممثلة في مسافة هندسية لفضاء متعدد الابعاد ويعطي بالعلاقة المتتالية:

$$1 \le i \le n \quad 1 \le l \le n \quad \Rightarrow \quad d(i,l) = \sqrt{\sum_{j=1}^{p} (x_{ij} - x_{lj})^2}$$

حيث ان$d(i,l)$ هي المسافة الاقليدية بين المفردتينن i و l.

يمكننا دائما تقريب التفرقة عن طريق المسافة الإقليدية. ولهذا ما يهمنا الان هو العلاقة بين المسافات بين العناصر والتسلسل الهرمي للأجزاء (الاقسام). حيث يحدد التسلسل الهرمي اولا المسافة بين الافراد التي لها خصائص ملحوظة.

V-2-2 مبدأ التصنيف التسلسلي الهرمي (كيف يمكن تشكيل k قسم)

الجزء الاهم والاكبر من طرق التصنيف التسلسلي ينتهج نهج الخوارزمية وليس التقنيات الرياضية المعقدة: لذلك فالنتائج المحصلة في نهاية السلسلة هي نتيجة عمليات بسيطة ومتكررة .حيث تجمع الافراد الاكثر قربا يستلزم العديد من الشروط القبلية من اجل اجراء تقدير او قياس للتقارب بين كل زوج من الافراد. هذه الشروط يمكن تخصيصها في النقاط التالية:

- الحاجة الى تقدير المسافة بين جميع الافراد وبالتالي الحصول على مصفوفة التفرقة.
- الحاجة الى تحديد معيار للمقارنة وهذا من اجل تجميع الافراد.
- الحاجة الى طريقة تصنيف وبالتالي تحديد خوارزمية قبلية.
- **التصنيف التسلسلي الهرمي وخوارزمية التصنيف**

يمكن ان نميز بين نوعين من طرق التصنيف التسلسلي: التصاعدي والتنازلي. حيث تهتم الاولى بإنشاء جزء عن طريق تجميع جزئين (قسمين) موجودين اصلا . أما التنازلي فتقوم على تقسيم (تفكيك) على العكس جزء (قسم) موجودة اصلا من اجل انشاء مجموعتين جزئيتين جديدتين او اكثر.

من أجل التجميع بحيث ان يكون وفقا لمعيار او مؤشر معين، في البداية من الطبيعي تجميع المفردتين الاكثر قربا وفقا لمعنى التفرقة عند الانطلاق. ولكن مباشرة بعد هذه العملية يمكن اعادة تجميع سواء مفردتين، او مفردة مع جزء معين او بعد ذلك بقليل قسمين معا. وهذا وفقا للخوارزمية التالية:

المرحلة 1: n فوج (قسم)

تقوم الخوارزمية ببحث المفردتين الاكثر قربا وفقا لمفهوم المسافة الاقليدية وهذا من اجل تكوين الزوج الاول ،والذي يحرص على تعظيم التباين الداخلي (ما بين المجموعات) حيث تقوم الخوارزمية بتعويض المفردتين عن طريق الزوج لتقلص بذلك عدد المفردات الى(n-1) .

المرحلة 2: (n-1) قسم

تقوم الخوارزمية بإعادة حساب المسافات بين جميع الافراد. مرة اخرى تقوم الخوارزمية بتجميع المفردتين الاكثر قربا من خلال تكوين زوجين جديدين مما يسبب زيادة جديدة في الجمود داخل المجموعة الى جانب انخفاض في عدد الافراد.

المرحلة 3 : (n -2) قسم , حيث تقوم بتكرار الاجراء السابق.

المرحلة n: 1 قسم

باعتبارها اخر خطوة من الخوارزمية، هذه المرة يكون الجمود الكلي محتوى في الجمود داخل المجموعة ، حيث لم يعد هناك أي جمود بين المجموعات.

-تستمر الخوارزمية بطريقة تصاعدية حتى يتم الحصول على مجموعة واحدة فقط. في الاخير ،فان طريقة التسلسل الهرمي تقوم بإنتاج هرم للأفراد، حيث يتم تجميعها تدريجيا في مجموعات اكبر واكبر بحيث انه:

-**في القاعدة**: مخطط الشجرة (شجرة المجموعات)، بشكل كل فرد بمفرده مجموعة.

-**في القمة**: ينتمي جميع الافراد الى نفس المجموعة.

ان الانتقال من مستوى واحد من التسلسل الهرمي الى المستوى الموالي يتمثل في دمج المجموعتين الاكثر تشابها، فالفكرة العامة هي تقليل التباين داخل المجموعة (الجمود الذاتي = $Inertie_{\ intra-classe}$) وتعظيم التباين بين المجموعات (الجمود الخارجي = $Inertie_{\ inter-classe}$).

تعتبر خوارزمية CHA بسيطة جدا، حيث يتمثل بداية في الوضع الابتدائي الاولي ، حيث أن كل فرد من E يمثل فوجا في حد ذاته، وهذا يعني n فوج ككل . نتيجة لهذا فإن لدينا الشروط الاولية. وفقا لمبدأ العطالة (الجمود) هي كالاتي:

$$Inertie_{\ intra-classe} = 0 \quad , \ Inertie_{\ inter-classe} = Inertie_{\ Totale}$$

-**مؤشرات التصنيف**

هناك العديد من المؤشرات المستعملة في طرق التصنيف من أهمها:

- مؤشر ادنى مسافة = Saut minimum (نجد ايضا: القفزة الدنيا، الاقرب جار, الارتباط البسيط).
- مؤشر التجميع لأقصى مسافة = Lien complet (الارتباط الكامل، ابعد جار).
- مؤشر التجميع باعتبار المسافة المتوسطة = Distance moyenne
- مؤشر وارد Ward.

العمل بهذه المؤشرات يكون دائما وفقا للخوارزمية CHA المذكورة سلفا والذي يمكن من الحصول على شجرات تصنيف خاصة بكل مؤشر.

- **CHA وفقا لمؤشر saut minimum**

حيث يعرف المسافة بين جزأين بواسطة العلاقة التالية:

$$D^2(Y,X) = \min D^2(y,x) \quad / \, x \in X \text{ و } y \in Y$$

حيث يتم الجمع في كل خطوة من خطوات CHA بين الجزئين اللذان يمثلان اكثر تقارب أي اننا نأخذ أصغر قيمة للمسافة المتاحة (جدول المسافة). هذه القيمة تتعلق بالجزئيين X وY اللذان يتم جمعهما، (كل جزء يحمل في البداية عنصر وحيد فقط وبالتالي المسافة المقدرة بين هذين الجزئيين تقدر بالمسافة بين هذين العنصرين. يتم الاحتفاظ بالمسافات الاخرى ما عدا التي تكون منحصرة بين هذا الجزء المكون والعناصر الاخرى فيتم حسابها وفقا ل:

$$\begin{aligned} D^2(A \cup B, X) &= \min\left(D^2(A,X), D^2(B,X)\right) \\ &= \frac{1}{2} D^2(A,X) + \frac{1}{2} D^2(B,X) - \frac{1}{2} |D^2(A,X) - D^2(B,X)| \end{aligned}$$

مثال: تم جمع نتائج 6 أطفال في العاشرة من العمر في 6 اختبارات فرعية لاختبار (الدرجات من 0 الى 5).حيث كانت المتغيرات الملحوظة هي: PUZ (تجميع الاشياء). CUB (مكعب ROHS)، CAL (الحساب الذهني)،MEM (الذاكرة الفورية للأرقام)، COM (فهم الجمل) ، VOC (المفردات) حيث كان البرتوكول المرصود كمايلي:

T0	**CUB**	**PUZ**	**CAL**	**MEM**	**COM**	**VOC**
I1	4,00	3,00	3,00	2,00	2,00	1,00
I2	2,00	,00	1,00	3,00	1,00	1,00
I3	1,00	2,00	1,00	4,00	3,00	3,00
I4	,00	1,00	,00	3,00	1,00	,00
I5	2,00	,00	1,00	3,00	1,00	,00
I6	4,00	2,00	4,00	2,00	1,00	2,00

• من اجل عمل CHA فإن أول خطوة نقوم بها هي حساب مصفوفة التفرقة وهذا بحساب المسافة بين مجموع الافراد مثنى مثنى مقدرة بمربع المسافة الاقليدية:

Matrice de proximité

Observation	**Carré de la distance Euclidienne**					
	1	2	3	4	5	6
1	,000	19,000	23,000	32,000	20,000	4,000
2	19,000	,000	14,000	7,000	1,000	19,000
3	23,000	14,000	,000	17,000	19,000	27,000
4	32,000	7,000	17,000	,000	6,000	38,000
5	20,000	1,000	19,000	6,000	,000	22,000
6	4,000	19,000	27,000	38,000	22,000	,000

Ceci est une matrice de dissimilarité

حيث:

$$d^2(I_1, I_2) = \sum_{j=1}^{6} (x_{I_1 j} - x_{I_2 j})^2$$
$$= (4-2)^2 + (3-0)^2 + (3-1)^2 + (2-3)^2 + (2-1)^2 + (1-1)^2 = 19$$

أما باقي المسافات فتحسب بنفس الطريقة.

• وفقا للخوارزمية فإنه في البداية لدينا 6 أقسام (n =6) باعتبار أن كل عنصر يشكل قسم لوحده. ولهذا من اجل تشكيل الزوج الاول فإننا نختار لذلك وكخطوة ثابتة في كل مرحلة من مراحل الخوارزمية وكذا من اجل جميع المؤشرات المستخدمة – اقرب عنصرين وذلك باعتبار اصغر مسافة ثم بعد ذلك نحتفظ بالمسافات الاخرى التي لا تتغير (بين مختلف الازواج الاخرى)، أما الباقي (التي تتغير وتكون محصورة بين هذا الزوج المشكل والعناصر الاخرى) يتم حسابها وفقا لمؤشر ادنى مسافة. وعليه يحصل: من خلال جدول التفرقة نلاحظ ان اقرب عنصرين هو I_2, I_5 . وعليه نقوم بجمعهما للحصول على الزوج الاول والذي نسيمة I_7 .

$$\begin{array}{c|ccccc} & I1 & I7 & I2 & I4 & I6 \\ \hline I1 & & & & & \\ I7 & & & & & \\ I2 & & & & & \\ I4 & & & & & \\ I6 & & & & & \end{array}$$

$$\begin{array}{cc} & \begin{array}{ccccc} I1 & I7 & I2 & I4 & I6 \end{array} \\ \begin{array}{c} I1 \\ I7 \\ I2 \\ I4 \\ I6 \end{array} & \begin{pmatrix} 0 & 19 & 23 & 32 & 4 \\ & 0 & 10 & 6 & 19 \\ & & 0 & 17 & 27 \\ & & & 0 & 42 \\ & & & & 0 \end{pmatrix} \end{array}$$

• نكرر نفس العملية في الخطوة الثانية من الخوارزمية حيث أن في هذه المرة لدينا 5 أقسام ($n-1= 6-1$). حيث من خلال مصفوفة التفرقة الجديدة نلاحظ ان 4 هو اقل مسافة وهي محصورة بين العنصرين I_1 و I_6 وعليه يتم جمعهما من اجل تشكيل الزوج الثاني والذي نسميه I_8.

$$\begin{array}{c c} & \begin{array}{cccc} I8 & I7 & I3 & I4 \end{array} \\ \begin{array}{c} I8 \\ I7 \\ I3 \\ I4 \end{array} & \begin{pmatrix} 0 & 19 & 23 & 32 \\ & 0 & 10 & 6 \\ & & 0 & 17 \\ & & & 0 \end{pmatrix} \end{array}$$

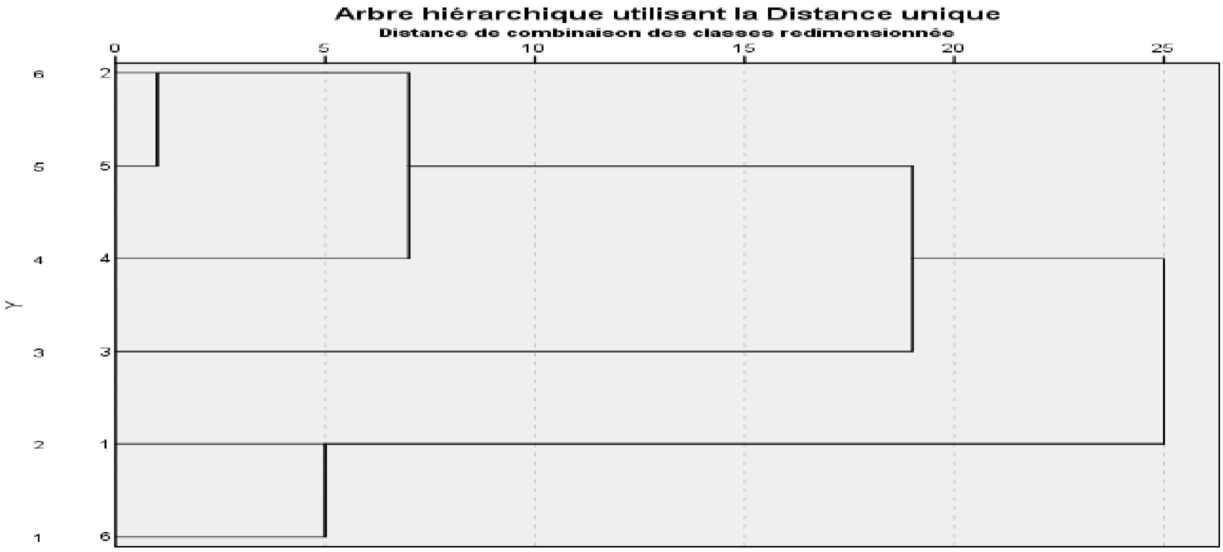
نقوم بنفس الاجراءات للحصول على الجدول T_3 . ثم تنتقل للخطوة الثالثة حيث تبقى لدينا 4 اقسام. في هذه المرة نقوم بتجميع الجزء I_7 المشكل من I_2 و I_5 مع العنصر I_4 وليكن الزوج I_9 فيحصل على:

$$\begin{array}{c c} & \begin{array}{ccc} I8 & I9 & I3 \end{array} \\ \begin{array}{c} I8 \\ I7 \\ I3 \end{array} & \begin{pmatrix} 0 & 19 & 23 \\ & 0 & 10 \\ & & 0 \end{pmatrix} \end{array}$$

نفس الاجراءات بالنسبة للخطوة 4 حيث تبقى لنا ($n-3=3$) 3 عناصر فقط. نجمع الجزء المشكل I_9 مع العنصر I_3 لنحصل على جدول يحوي عنصرين فقط من I_{10} و I_8

$$\begin{array}{c c} & \begin{array}{cc} I8 & I9 \end{array} \\ \begin{array}{c} I8 \\ I7 \end{array} & \begin{pmatrix} 0 & 19 \\ & 0 \end{pmatrix} \end{array}$$

نقوم بتجميع آخر للعنصرين من أجل الحصول على مجموعة واحدة كاملة تشمل جميع العناصر الستة. ثم نقوم برسم شجرة التصنيف(دندروغرام) الموافقة لمؤشر الخطوة الدنيا. فنحصل على:



- **مؤشر أبعد جار (المسافة الاقصى)، الرابط الكامل Lien complet**

او ما يعرف عادة بالتجميع حسب القطر، حيث نشتق المسافة بين جزئيين او زوجين من المسافة بين المفردات بالعلاقة التالية:

$$\mathrm{D}^2(\mathrm{Y},\mathrm{X}) = \max \ \mathrm{D}^2(\mathrm{y},\mathrm{x}) \quad / \ \mathrm{x} \in \mathrm{X} \text{ و } \mathrm{y} \in \mathrm{Y} \text{ -}$$

حيث يتم الاحتفاظ بالمسافات الاخرى، أما ما يتم حسابه ، فيكون باتباع للمسافة المحينة وفقا ل:

$$D^2(A \cup \mathrm{B}, \mathrm{X}) = \max\left(\mathrm{D}^2(A,X), D^2(B,X)\right)$$
$$= \frac{1}{2}D^2(A,X) + \frac{1}{2}D^2(B,X) + \frac{1}{2}\left|D^2(A,X) - D^2(B,X)\right|$$

مثال : انطلاقا من مصفوفة التفرقة في المثال السابق نقوم بتجميع العناصر وفقا لمعيار أبعد جار حيث سوف ننتهج نفس الخوارزمية مع نفس الاجراءات.

-تكون اول خطوة يجمع العنصرين I_2, I_5 لنحصل على الزوج I_7 وعليه يكون بشكل الجدول كالتالي:

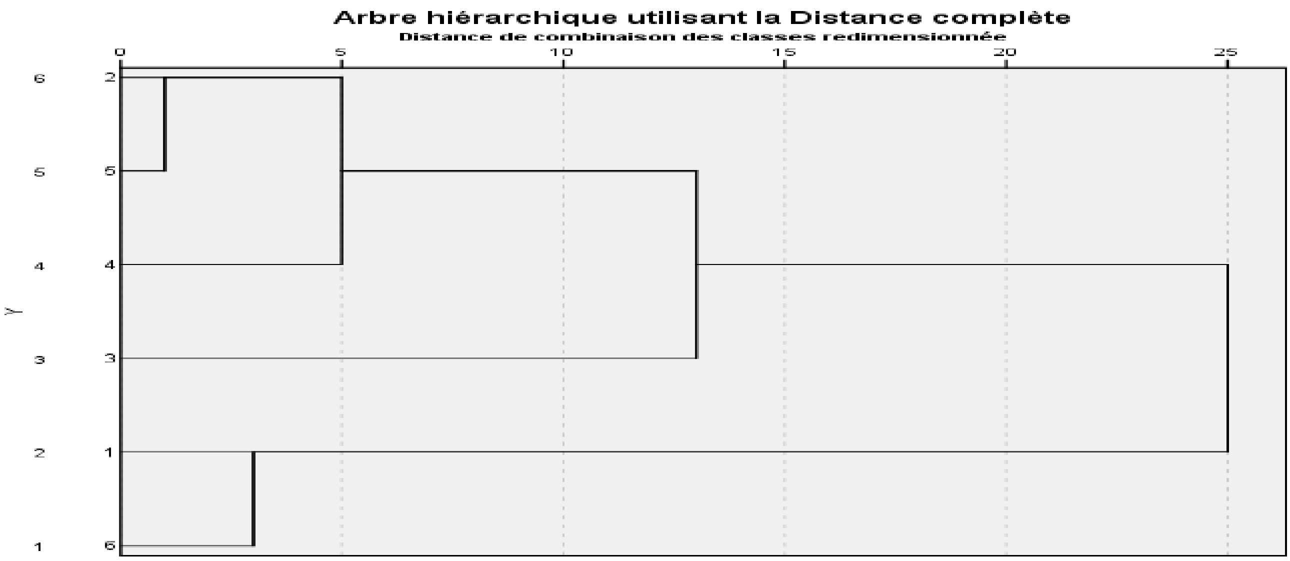
$$\begin{array}{cc} & \begin{array}{ccccc} I1 & I7 & I3 & I4 & I6 \end{array} \\ \begin{array}{c} I1 \\ I7 \\ I3 \\ I4 \\ I6 \end{array} & \begin{pmatrix} 0 & 20 & 23 & 32 & 4 \\ & 0 & 19 & 7 & 22 \\ & & 0 & 17 & 27 \\ & & & 0 & 42 \\ & & & & 0 \end{pmatrix} \end{array}$$

$$D^2(I_2 \cup I_5, I_3) = \max\left(\mathrm{D}^2(I_2, I_3), D^2(I_5, I_3)\right) = 19$$

بتكرارا نفس الخطوات نتحصل على مجموع المراحل التالية:

$$\begin{array}{c} I8 \\ I7 \\ I3 \\ I4 \end{array}\begin{pmatrix} 0 & 19 & 23 & 32 \\ & 0 & 10 & 6 \\ & & 0 & 17 \\ & & & 0 \end{pmatrix} \qquad \begin{array}{c} I8 \\ I7 \\ I3 \end{array}\begin{pmatrix} 0 & 19 & 23 \\ & 0 & 10 \\ & & 0 \end{pmatrix} \qquad \begin{array}{c} I8 \\ I7 \end{array}\begin{pmatrix} 0 & 19 \\ & 0 \end{pmatrix}$$

- بعد آخر خطوة والتي تقوم بتجميع آخر جزئين نقوم برسم شجرة التصنيف وفقا لمؤشر أبعد جار.



- **مؤشر المسافة المتوسطة:** او متوسط المجموعة او ما تعرف بطريقة المجموعة الزوجية غير المرجحة (UGPMA).حيث تعرف المسافة بين جزئين بالعلاقة التالية:

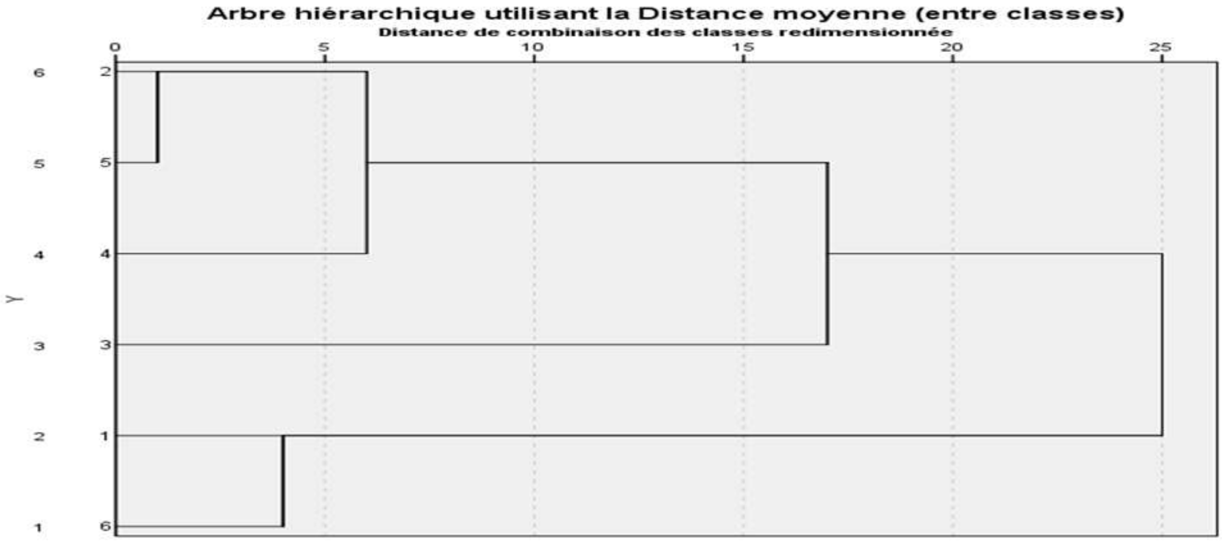
$$D^2(Y,X) = \min D^2(y,x) \quad / \; x \in X \text{ و } y \in Y$$

طريقة CHA وفقا لهذا المؤشر تحتفظ بجميع الخصائص الاخرى. حيث تحسب المسافات الجديدة باستعمال المسافة المحينة:

$$D^2(A \cup B, X) = \frac{n_A D^2(A,X) + n_B D^2(B,X)}{n_A + n_B}$$

ثم نقوم برسم الشجرة وفقا لمؤشر المسافة المتوسطة

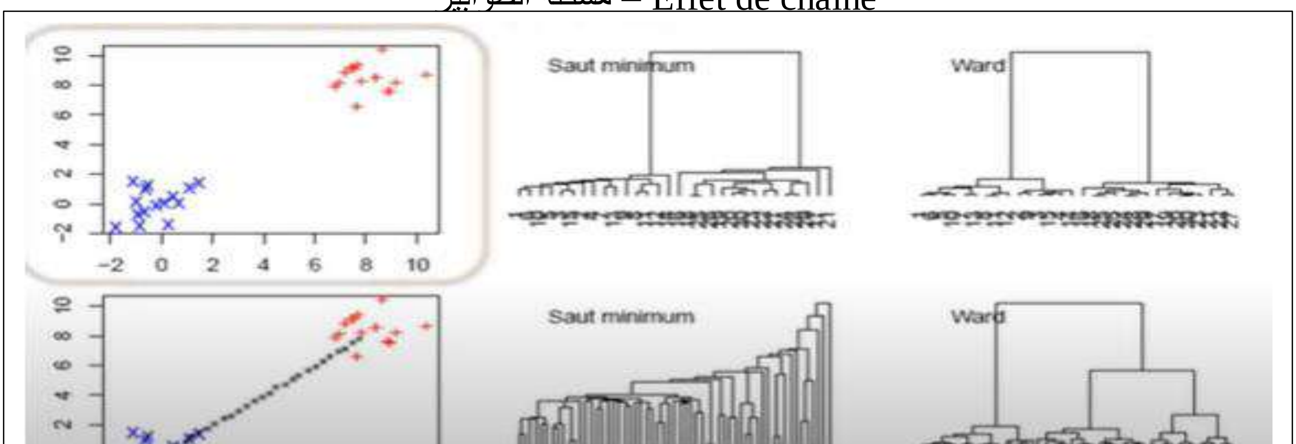
<u>مثال:</u> باتباع نفس الخوارزمية يمكن الحصول على الشجرة الدندروغرام التالية وهذا باستعمال المؤشر المسافة المتوسطة:



- **مؤشر وارد "Ward "**

العمل بالمؤشرات السابقة وفقا للخوارزمية التصنيف الهرمي يمكن من الحصول على نتائج متشابهة نوعا ما (شجرات تصنيف متماثلة)غير أنها ليست بالمرضية لأنها تشوه قليلا المجموعة الاصلية خصوصا في حالات الطوابير (effet de chaine) أي عندما يكون هناك تتابع بين عناصر المجموعة الكلية والتي تنحدر بين جزءين مختلفين، وهذا لأنها تقوم على اساس تجميع العناصر الواحد تلو الاخر. حيث انه لا يمكن الفصل بين هذه المجموعات المشكلة بعد التجميع. لذلك نلجأ في أغلب الحالات الى مؤشر أكثر دقة والذي يعرف بمؤشر وارد.

Effet de chaine = مشكلة الطوابير



من البديهي، فإن أحسن طريقة لإنشاء مجموعات متجانسة فيما بينها اي داخل المجموعة الواحدة ومختلفة من مجموعة لأخرى، هو صنع حزم للأفراد وفقا للمسافة الفاصلة بينهم. وهذا يعني تقليل المسافات الفاصلة بين مختلف الافراد(العناصر) داخل الحزمة الواحدة (العبوة) مع زيادة المسافات الفاصلة بين الافراد الذين ينتمون الى حزم مختلفة. بعبارة اخرى فانه سوف يعتمد في ذلك على مبدأ العطالة (Inertie) او التباين الكلي (variation totale) .

حيث يعتمد مؤشر Ward على هذا المفهوم. باعتباره مقياس لمدى تشتت المفردات الاحصائية حول مركز الثقل ''G''(centre de gravite) للمجموعة الكلية E . او حول مركز الثقل للمجموعة التي ينتمون اليها (العطالة داخل المجموعة = inter intra- groupe).

حيث انه كلما انخفض التباين ، كلما زاد تجمع الافراد حول مركز الثقل المعتبر. وباعتبار ان التباين الكلي او العطالة الكلية هو ثابت وذلك بفضل نظرية Huygens :

التباين الكلي = التباين داخل المجموعات + التباين بين المجموعات

باعتبار ان عناصر المجموعة" E " لا تتغير ، لذلك فان التباين داخل المجموعات هو الذي يتغير عند كل قسم من K قسم ثابت وفقا للتقسيمات المعتبرة n عنصر من k قسم.

ولهذا فان الهدف وراء كل هذا هو ايجاد احسن تجميع ممكن ، وهو الترجمة الرياضية بمعنى ان فردين من قسم واحد اكثر تقاربا ب التغيير داخل المجموعة ولو ان الفردين من قسمين مختلفين متباعدين بالتغير ما بين المجموعات وعليه باعتبار ثبات التباين الكلي فان التقليل من I_{intra} يؤدي بالضرورة الى تعظيم قيمة I_{inter}. لهذا يمكن الاعتماد على احد المقدرين اما بتعظيم I_{inter} او تقليل من I_{intra} . وكل منهما يمكن من الحصول على مجموعة الافراد الاكثر تجانسا فيما بينهما.

-طريقة عمل مؤشر " Ward"

تنطلق من تجميع اين يكون لدينا قسم واحد فقط حيث ان هذا القسم يحوي عنصر واحد فقط وبالتالي :

$$1 \text{ classe} = 1 \text{ individu} \Longrightarrow inertie_{\ inter} = 1$$

و هذا باعتبار انه لا يوجد تباين داخل ذلك القسم فان التصنيف امثلي. نقوم بتجميع القسم A والقسم B كخطوة ثانية ، مما يؤدي الى تناقص في قيمة التباين بين الاقسام.

$$\text{Inertie (A) + Inertie (B) = Inertie (A}\cup\text{B) - } \frac{n_a n_b}{n_a+n_b}\, d^2(a,b)$$

حيث ان : $n_a; n_b$ هما عدد العناصر داخل القسمين B وA على الترتيب ;
$d^2(a,b)$ هما المسافات بين مراكز ثقل القسمين B وa .

ومنه يمكن تقريب تباين اتحاد القسمين B وA الى مجموع تباين كل مجموعة عن طريق التقليل من كمية $\frac{n_a n_b}{n_a+n_b}\, d^2(a,b)$ والتي تحتوي مجموعة من اثقال المفردات وكذا المسافة مربعة.

هذه الكمية $\frac{n_a n_b}{n_a+n_b}$ تسمح بتجميع الاشياء للأفراد ذات الاوزان الحقيقية وبالتالي فانها تتجنب تأثير التتابع المتسلسل, اما الكمية الثانية فتسمح بتجميع الاقسام التي تكون مراكز ثقلها قريبة من بعضها البعض حيث يمكن حساب مركز ثقل مشترك للمجموعتين القسمين B و A بالعلاقة :

$$g_{AB} = \frac{n_a g_a + n_b g_b}{n_a + n_b}$$

وعليه فاذا اردنا ان نقوم بتجميع القسمين B وA فإننا نقوم بحساب المؤشر والذي يستعمل وفقا للعلاقة:

$$\mathrm{D}^2(\mathrm{A} \cup \mathrm{B}, \mathrm{X}) = \frac{(\mathrm{n_A} + \mathrm{n_X})\mathrm{D}^2(\mathrm{A}, \mathrm{X}) + (\mathrm{n_B} + \mathrm{n_X})\mathrm{D}^2(\mathrm{B}, \mathrm{X}) - \mathrm{n_X}\, \mathrm{D}^2(\mathrm{B}, \mathrm{A})}{\mathrm{n_A} + \mathrm{n_B} + \mathrm{n_X}}$$

مثال: بالعودة للمثال السابق وبالضبط الى المصفوفة المسلفات قم بعمل تصنيف تسلسلي باستعمال مؤشر WARD.

الحل: من اجل الاجابة نقوم أولا بحساب مصفوفة النقل انطلاقا من مصفوفة المسافة السابقة وهذا باستعمال العبارة:

$$\mathrm{g_{AB}} = \frac{\mathrm{n_a g_a + n_b g_b}}{\mathrm{n_a + n_b}}$$

وعليه تكون مصفوفة مراكز الثقل:

$$\begin{array}{c} \\ I1 \\ I2 \\ I3 \\ I4 \\ I5 \\ I6 \end{array}
\begin{array}{c} \begin{array}{cccccc} I1 & I2 & I3 & I4 & I5 & I6 \end{array} \\
\begin{pmatrix} 0 & 9{,}5 & 11{,}5 & 16 & 10 & 2 \\ & 0 & 5 & 3{,}5 & 0{,}5 & 9{,}5 \\ & & 0 & 8{,}5 & 9{,}5 & 13{,}5 \\ & & & 0 & 3 & 21 \\ & & & & 0 & 11 \\ & & & & & 0 \end{pmatrix} \end{array}$$

باستعمال المؤشر نكون قد تحصلنا على مجموع النتائج التالية:

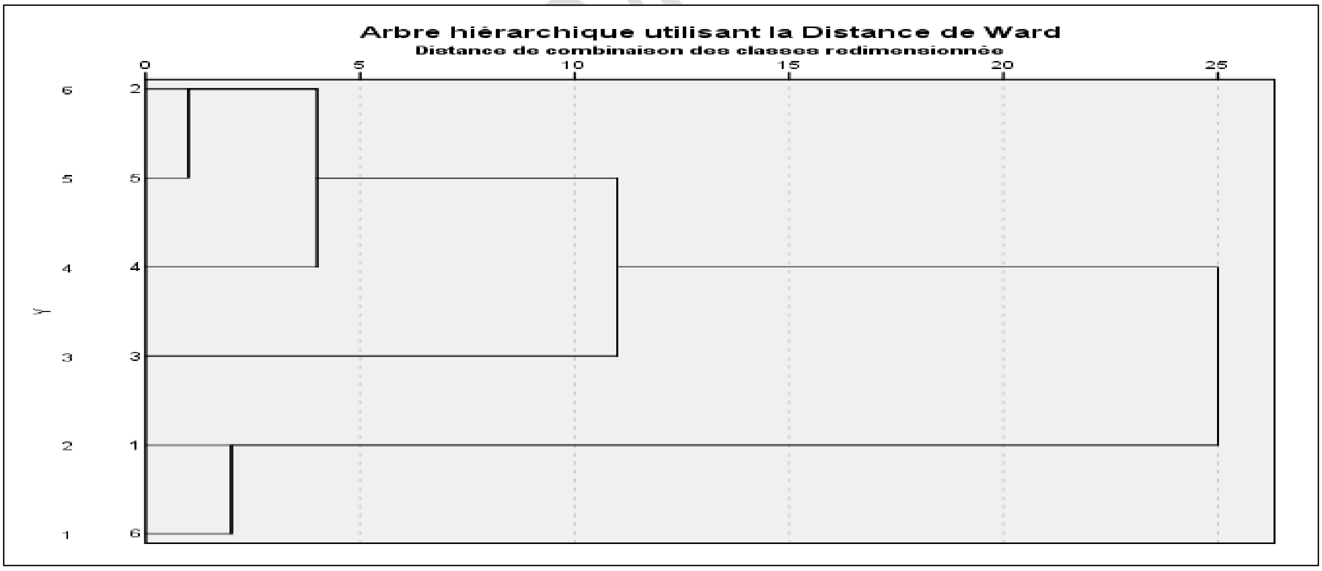
$$\begin{array}{c} I1 \\ I7 \\ I3 \\ I4 \\ I6 \end{array} \begin{pmatrix} 0 & 12{,}84 & 11{,}5 & 16 & 2 \\ & 0 & 9{,}5 & 4{,}17 & 13{,}5 \\ & & 0 & 8{,}5 & 13{,}5 \\ & & & 0 & 21 \\ & & & & 0 \end{pmatrix} \qquad \begin{array}{c} I8 \\ I7 \\ I3 \\ I4 \end{array} \begin{pmatrix} 0 & 15{,}4 & 16 & 24 \\ & 0 & 9{,}5 & 4{,}17 \\ & & 0 & 8{,}5 \\ & & & 0 \end{pmatrix}$$

$$\begin{array}{c} I8 \\ I9 \\ I3 \end{array} \begin{pmatrix} 0 & 25{,}05 & 16 \\ & 0 & 10;33 \\ & & 0 \end{pmatrix} \qquad \begin{array}{c} I8 \\ I10 \end{array} \begin{pmatrix} 0 & 25{,}43 \\ & 0 \end{pmatrix}$$

حيث:

$$12,84 = d^2(I_7, I_1) = D^2(I_2 \cup I_5, I_1) = \frac{(n_{I_2}+n_{I_1})D^2(I_2,I_1)+(n_{I_5}+n_{I_1})D^2(I_5,I_1) - n_{I_1} D^2(I_2,I_5)}{n_{I_1}+n_{I_2}+n_{I_5}}$$

- يمكن القول ان مجموع النتائج المحصل عليها بواسطة المؤشرات المختلفة، هي نتائج مهمة، غير انها ليست أمثليه بالمقارنة بالنتيجة الممكن الحصول عليها، بواسطة مؤشر وارد. كون هذا الاخير يمكن من الحصول على مجموعات منفصلة تماما واكثر تجانسا.



3-V التصنيف غير الهرمي (Classification non hiérarchique)

التصنيف أو التقسيم غير الهرمي، يرمي إلى تفكيك مجموعة جميع الأفراد إلى m مجموعة منفصلة أو الى فئات تكافؤ؛ بحيث العدد m من الفئات ثابت.
النتيجة التي يتم الحصول عليها هي تقسيم مجموعة الأفراد أو مجموعة من الأجزاء أو فئات المجموعة I الأولى من الأفراد بحيث أن:

- أي فئة من الفئات ليست فارغة؛
- الفئتان المختلفتان منفصلتان؛
- كل فرد ينتمي إلى فئة.

تسمى هذه الخوارزمية "التجميع حول المراكز المتغيرة". هناك نسخة مختلفة قليلاً، تُعرف باسم "النوى او السحب الديناميكية"، وهي تمثيل كل مجموعة ليس من خلال مركزها، ولكن من خلال مجموعة من النقاط (الأساسية) يتم اختيارها عشوائيًا داخل كل مجموعة. ثم نحسب المسافة "المتوسطة" بين كل ملاحظة وهذه النوى وننتقل إلى المهمة.

يمكن القول أن طريقة التصنيف الغير هرمي هي طريقة هدفها تقسيم الملاحظات إلى K قسم حيث تنتمي كل ملاحظة إلى القسم مع المتوسط الاقرب. يمكن ان نقتبس طريقتين معروفتين تقومان على مبدأ k-Means:

- طرق المركز المتنقل (Méthodes de centres mobiles).
- طرق النوى الديناميكية (Méthodes des nuées dynamiques).

- **طريقة المركز المتنقل**

تتمثل هذه الطريقة في إنشاء قسم من K فئات عن طريق اختيار k فرد اولي، ويتم اختيار الفئات عشوائيًا من مجموعة الأفراد. بعد هذا الاختيار، نقوم بإرسال كل فرد إلى أقرب مركز عن طريق إنشاء K فئة، وسيتم استبدال مراكز الفئات بمراكز الثقل وسيتم إنشاء فئات جديدة وفقًا لنفس المبدأ.

بشكل عام ، يكون القسم الذي تم الحصول عليه هو الأمثل محليًا لأنه يعتمد على الاختيار الأولي للمراكز. لذلك، تتنوع النتائج بعد تنفيذ كل عمليتين للخوارزمية بشكل كبير.

- **طريقة النوى الديناميكية**

في هذه الحالة، فإن المشكلة المطروحة هي البحث عن قسم من K فئة (k ثابت) لمجموعة من n فرد. إنها خوارزمية تكرارية.

ليكن I مجتمع من الأفراد، هذا المجتمع يمكن تمثيله خلال R ويشكل سحابة من n نقطة. نبحث عن تشكيل K فئة من i. يتم تمثيل كل فئة بمركزها، والذي يسمى أيضًا النواة، المكون من مجموعة فرعية صغيرة من الفئة التي تقلل من معيار الاختلاف.

تجدر الاشارة الى كلا الصنفين السابقين يعتمد على ما يعرف ب خوارزمية "K-Means"

- **خوارزمية K-Means :**

K-Means هي خوارزمية تجميع غير هرمية غير خاضعة للرقابة. يتيح تجميع ملاحظات مجموعة البيانات في مجموعات K المتميزة. وبالتالي، سيتم العثور على بيانات مماثلة في نفس المجموعة. علاوة على ذلك ، لا

يمكن العثور على ملاحظة إلا في مجموعة واحدة في كل مرة (عضوية حصرية). لذلك لا يمكن أن تنتمي نفس الملاحظة إلى مجموعتين مختلفتين.

يمكن تلخيص هذه الخوارزمية فيما يلي :

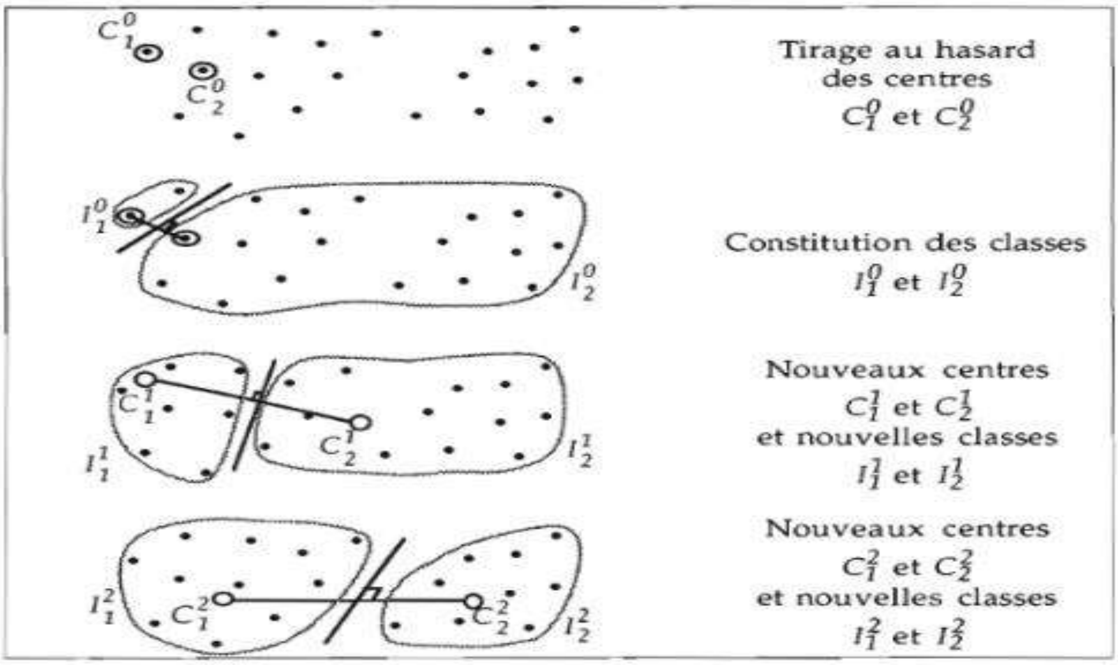
- **المدخلات :** في الاول تكون لدينا المصفوفة X والمشكلة من (n مشاهدة , p متغيرة), نقوم بتحديد او وضع K مركز فئات ابتدائي والذي يرمز له بالرمز G_K. وهنا يكون مبدأ العشوائية في اختيار K مفردة أو بالمقابل K متوسط محسوب من خلال تقسيم عشوائي للأفراد الى K قسم.
- **التكرارات (المراحل) حتى التقارب**

توزيع: هنا نقول بإرسال كل فرد الى القسم الذي يكون قريبا الى مركزه حيث نقوم في كل مرة بتحيين مراكز الاقسام من اجل كل مفردة تم معالجتها .

تمثيل: إعادة حساب مراكز الاقسام من خلال الأفراد الذين تم ارسالهم او الملتحقين بالأقسام الملائمة. هذه العملية تكون وفق الخاصية الاساسية القائمة على مبدا العطالة حيث في كل مرحلة تنخفض فيها العطالة بين الفئات .

- **المخرجات:** قسم من الأفراد يتميز ب K مركز للفئات G_K

كما يمكن تجسيد هذه الخوارزمية في المخطط التالي :



Lebart et al., 1995 ; page 149.

- **ايجابيات وسلبيات الطريقة :**

الايجابيات :

- قابلية التوسع (Scalabilité): القدرة على التعامل مع قواعد البيانات الكبيرة جدًا. يجب فقط حفظ اشعة المتوسطات في الذاكرة الرئيسية.

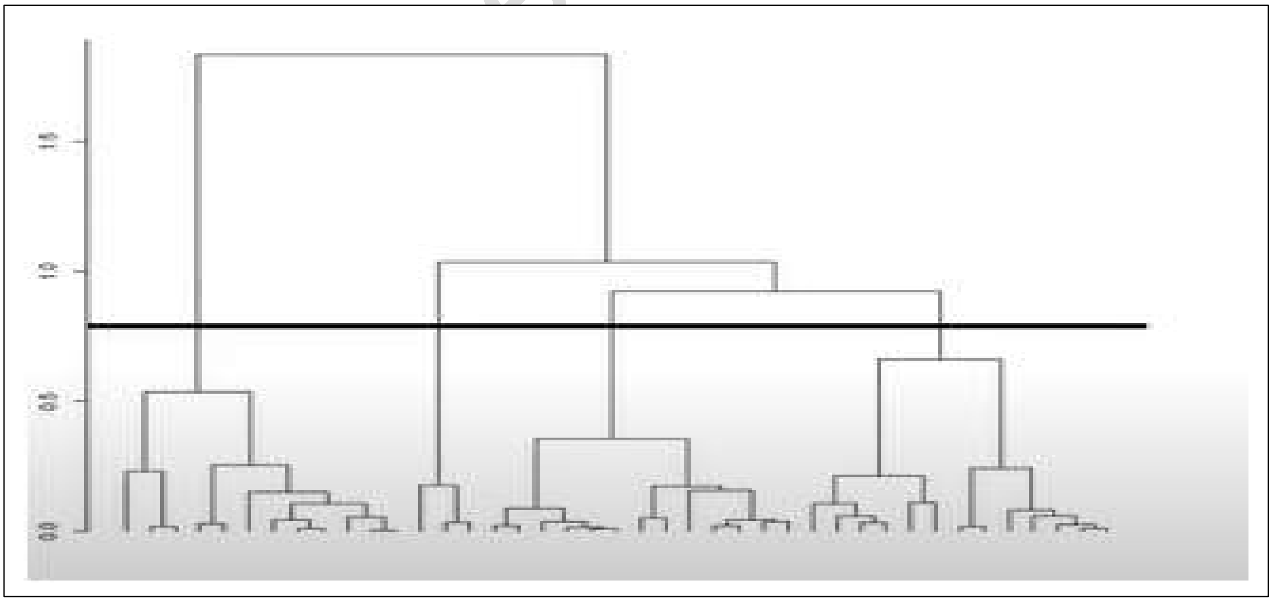
- التعقيد الخطي (Complexité linéaire) فيما يتعلق بعدد الملاحظات (عدم حساب المسافات مثنى مثنى للأفراد، راجع CAH).

السلبيات:

- لكن البطء على كل حال بسبب الحاجة لتمرير الملاحظات عدة مرات.
- ينتج عن التحسين حد أدنى محلي من الجمود الذاتي داخل الفئة W.
- يعتمد الحل على الاختيار الأولي لمراكز الصف و بالتالي تكريس مبدأ العشوائية.
- قد يعتمد الحل على ترتيب الأفراد (MacQueen)[1].

- **تقطيع الشجرة "Ttroncature de l'arbre"**

كل شجرة تصنيف تنتهي بالقطع. هذا الاخير هو الذي يسمح بتحديد عدد الاقسام الطبيعية الممكن الحصول عليها خلال تجميع معين(مجموعة) (مجتمع النباتات، مجتمع الاحياء......). ولهذا فان الوقوف عند مستوى قطع معين هو من أساسيات التصنيف الهرمي (niveau de coupure).



[1] تتمثل في خلط الأفراد بشكل عشوائي قبل تمريرهم حتى لا تعتمد على تنظيم غير خاضع لرقابة الملاحظات.

ولهذا وجب وضع مجموعة من المعايير والتي تضمن الى حد كبير من جودة الصورة (التمثيل الجيد = qualité de l'image) وبالتالي امكانية التجميع مع الحصول على مجموعات او اقسام منفصلة تماما، بحيث ان كل قسم يكون اكثر تجانسا مما يسمح من التقليل من ضياع المعلومة جراء التجميع. ويمكن تلخيص اهم هذه المعايير فيمايلي :

1- **التغير أو التباين "variabilité"** حيث انه من خلال تفكيك التباين الكلي الى تباين داخل المجموعات واخر خارج المجموعات يمكن ان تقارن بينهما كمالي: $0 \leq \frac{\text{Inertie}_{\text{inter-classe}}}{\text{Inertie}_{\text{intra-classe}}} \leq 1$

حيث يمكن ان نميز بين :

$$\frac{\text{Inertie}_{\text{inter-classe}}}{\text{Inertie}_{\text{totale}}} = 0 \quad \Leftrightarrow$$ هنا لا نستطيع التصنيف أساسا

$$\frac{\text{Inertie}_{\text{inter-classe}}}{\text{Inertie}_{\text{totale}}} = 1 \quad \Leftrightarrow$$ التصنيف مثالي

2- **عدد المفردات وعدد المجموعات "Nombre"**

هذا المعيار مرتبط بالأول حيث انه اذا كان هناك عدد كبير من الاقسام فانه من السهل رؤية اقسام متجانسة، اما اذا كان هناك عدد قليل من الاقسام فان التغير داخل المجموعات يكون كبير مما ينقص من تجانس هذه المجموعات. وهذا يعني ان وجود عدد كبير من المفردات لا يعني وجود عدد كبير من التقسيمات والعكس. غير انه في الواقع فوجود عدد كبير من المفردات ينتج عنه عدد معتبر من التجمعات ووجود عدد قليل من الافراد ينتج عنه عدد قليل من المجموعات.

3- **اختبار مؤشر التجميع**: راينا انه باختلاف المؤشر فإننا نحصل على نتائج مختلفة حيث وفقا لطريقة التجميع يمكن الحصول على نتائج مهمه ولكنها غير مرضية فمؤشر Ward سمح بإعطاء تقسيم مثالي الى درجة كبيرة.

4- **وفقا لمنحنى القيم الذاتية:**

يمكن حساب قيمة المعلومة الضائعة عند الانتقال من مستوى تجميع (مرحلة الى مرحلة اخرى)الى اخر أي من n قسم الى (n-1) حتى غاية الوصول الى مجموعة واحدة. وعليه بقلب هذه السيرورة يمكن تحديد مستوى قطع معين وهذا عن طريق معرفة حجم المعلومة التي يمكن تحديد مستوى قطع معين وهذا عن طريق معرفة حجم المعلومة التي يمكن ربحها عند مستوى قطع معين عند الانتقال من مجموعة الى مجموعتين فاكثر وهذا عن طريق معرفة النسبة :

$$\% \text{ النسبة} = \frac{\text{Inertie}_{\text{inter-classe}}}{\text{Inertie}_{\text{totale}}}$$

حيث يمكن ان نتوقف عند المستوى الذي تكون عنده حجم المعلومة التي يمكن ربحها غير معتبرة أي اقل بكثير من 10%.

5- وفقا لطول الاغصان "Le long des branche "

نعلم ان الاغصان هي التي تحمل الاوراق. وعليه فانه من الظاهر انه كلما كان القطع عند مستوى معين من طول هذه الاغصان، كلما كان التقسيم جيد والنتائج جد مرضية مما يسمح بالحصول على تقسيمات اكثر تجانسا أي زاد حجم التباين بين المجموعات.

كما تجدر الاشارة الى ان عملية القطع ترتبط ايضا ب:

- نوع الدراسة وكذا نوع المعطيات وطبيعتها.
- مدى مقروئية الفئات المحصل عليها ودلالتها من الناحية العملية والعلمية.

هناك ايضا خوارزمية تعرف باسم K-maens تعطي لطريقة التصنيف الطابع الاوتوماتيكي خصوصا اثناء القطع، غير اننا لم نسلط عليها الضوء وهذا لأنها تعتمد في مبدئها على العشوائية في تحديد التقسيمات التي يمكن الحصول عليها ثم مبدأ التكرار في العملية عدة مرات.

3-2-V <u>التصنيف التسلسلي الهرمي للمتغيرات الكيفية:</u>

يمكن القيام بتطبيق هذا النوع من الطرق على المتغيرات للكيفية وهذا باتباع هذين الاستراتيجيتين:

- جعل هذه المتغيرات كمية.
- نقوم بعمل ACM مع الحفاظ بالبعد الاول والثاني فقط أي اهمال المحاور الاخرى.
- نقوم بعمل CHA انطلاقا من المركبات الاساسية لطريقة التحليل العاملي للتوفيقات.
- استعمال مقاييس ملائمة للمعطيات للكيفية j : مؤشر التفرقة ، مؤشر Jaccard الى اخره.

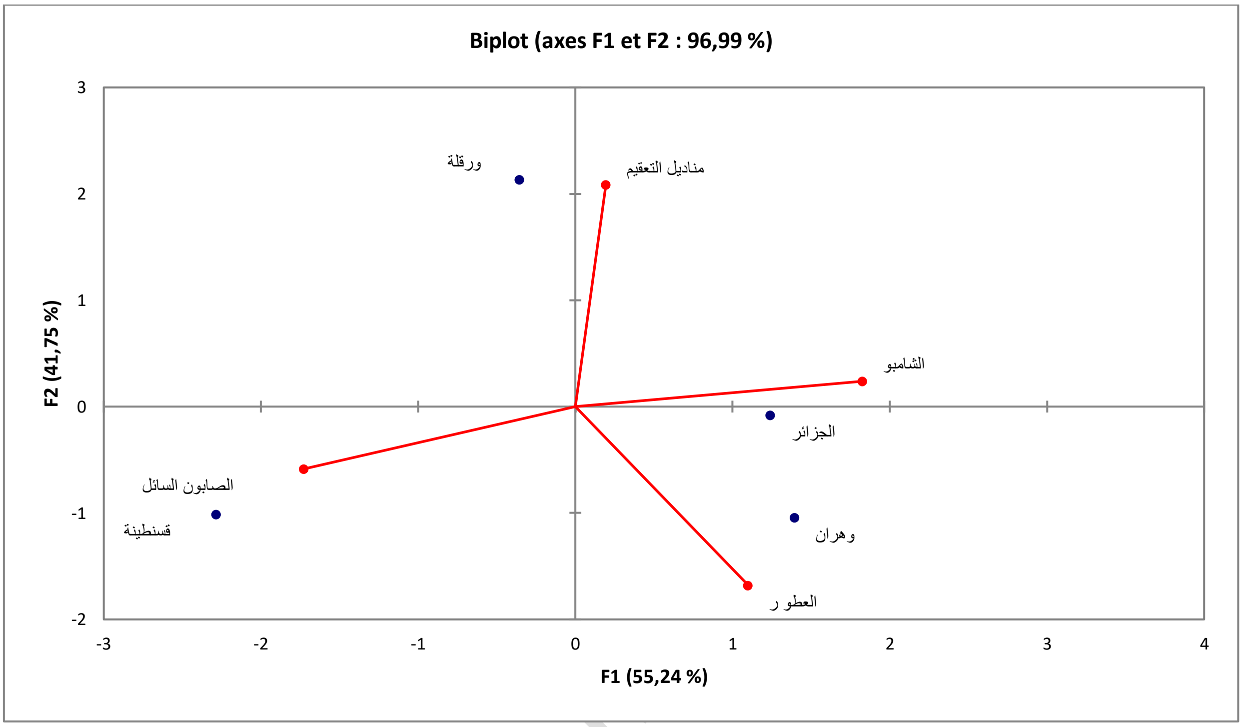
4-V دراسة حالة مع التطبيق

الجدول التالي يبين مبيعات احدى المؤسسات الناشئة و الناشطة في مجال ادوات التجميل والعناية بالبشرة وهذا بعد مرور شهرين عن بداية النشاط , حيث قامت بترويج منتجاتها في المدن الكبرى في الجزائر كبداية فقط قبل عملية التوسع:

	الصابون السائل لليدين	**العطو ر**	**الشامبو**	**مناديل التعقيم**
الجزائر	500	1000	1300	400
قسنطينة	1600	700	600	100
وهران	800	1500	1400	300

1200	1100	200	900	**ورقلة**

حيث اعطت طريقة تحليل المركبات الاساسية النتيجة التالية :



المطلوب :

1– علق على النتيجة ؟

2- قم بتنصيف هذه المدن باستعمال طريقة التصنيف التسلسلي التصاعدي وفقا لجميع المؤشرات التي درسناها؟ -ماذا تلاحظ؟

3- قارن تلك النتائج المحصلة بنتيجة ACP ؟

الحل:

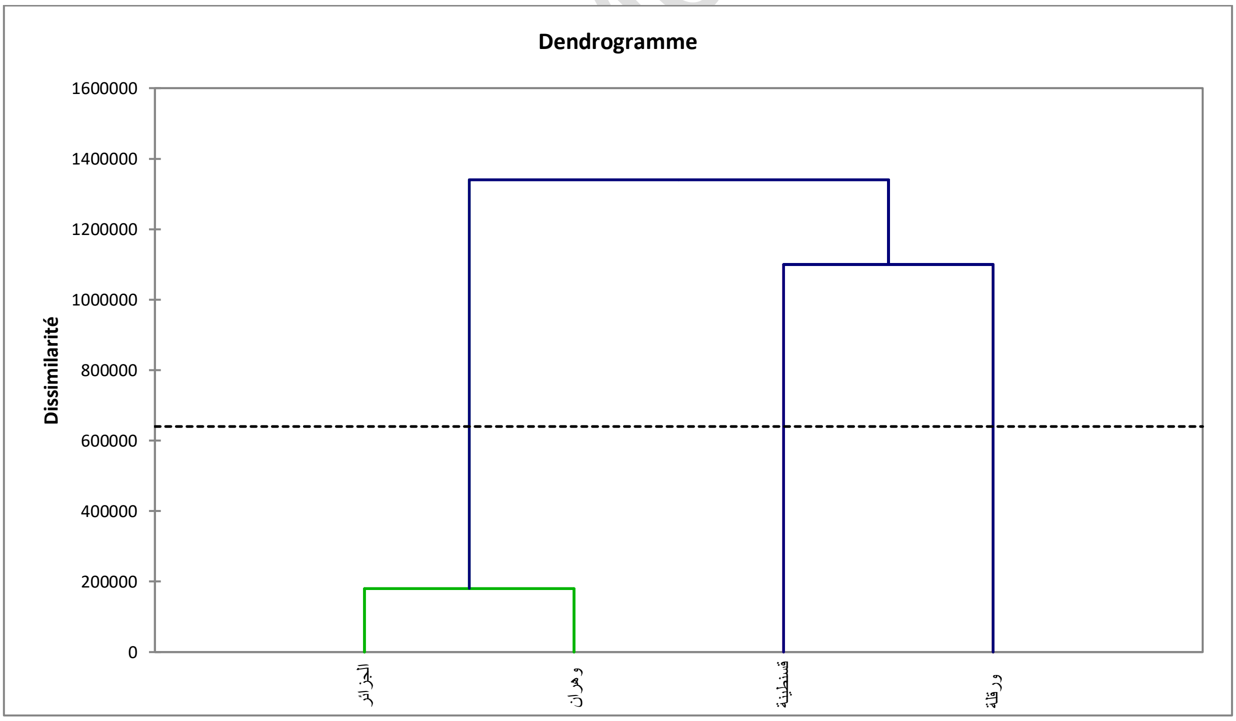
-1- من خلال المخطط يظهر لنا ان نسبة المعلومة المفسرة هي %96 من اجمالي المعلومة وهي نسبة كبيرة جدا بمعنى ان هناك ضياع قليل للمعلومة وهذا خلال المخطط العاملي الاول حيث تم الاحتفاظ فقط بالعاملين الاول والثاني بنسبة 55,24 % و%41,75 على الترتيب. كما يلاحظ ايضا ان معظم المفردات جاءت مرتبطة بالمحور الاول (الجزائر , وهران وقسنطينة) ولذلك يمكن ان نعتبر هذا المحور بمحور التناظر الكلي حيث يناظر بين المحموعة الاولى التي تضم كل من الجزائر ووهران والمجموعة الثانية التي تضم قسنطينة وهذا من وجهة نظر المتغيرتين الشامبو والصابون السائل حيث نجد ان اعلى نسبة مبيعات للمنتج الاول (الشامبو) حققت في المجموعة الاولى ونسبة قليلة في المجموعة الثانية والعكس بالنسبة للصابون السائل. اما بالنسبة للمفردة الرابعة والتي تمثل المجموعة الثالثة فجاء ارتباطها قويا موجبا بالنسبة للمحور الثاني والذي يمكن اعتباره كمحور للتناظر الجزئي, حيث ان نسبة المبيعات كانت مرتفعة بالنسبة

للمنتوج مناديل التعقيم في مدينة ورقلة وعلى العكس من ذلك بالنسبة لمنتوج العطور. وعليه يمكن القول ان هناك ثلاث مجموعات مختلفة من وجهة نظر المتغيرات باعتبار طريقة هي طريقة تصنيف بامتياز ACP.

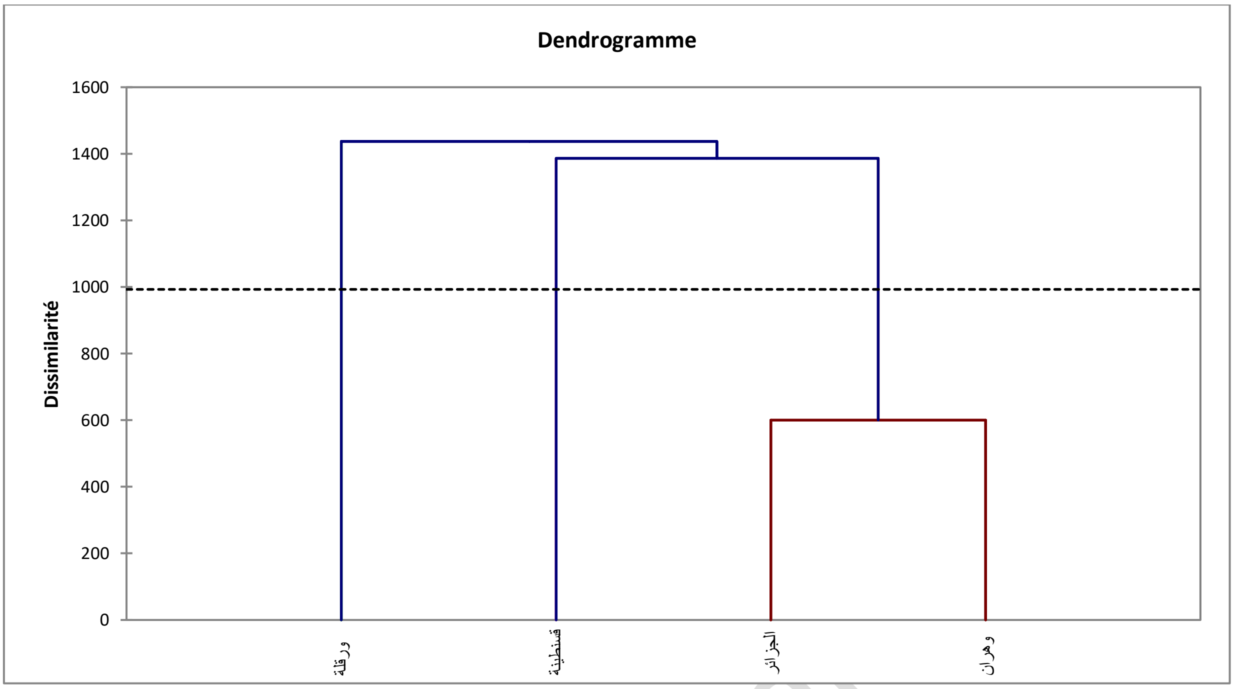
2- تصنيف المدن من حيث المنتجات الاكثر مبيعا باستخدام اسلوب التصنيف التسلسلي التصاعدي:

Matrice de proximité (Distance euclidienne)				
	الجزائر	قسنطينة	وهران	ورقلة
الجزائر	0	1371,131	600,000	1216,553
قسنطينة	1371,131	0	1400,000	1483,240
وهران	600,000	1400,000	0	1612,452
ورقلة	1216,553	1483,240	1612,452	0

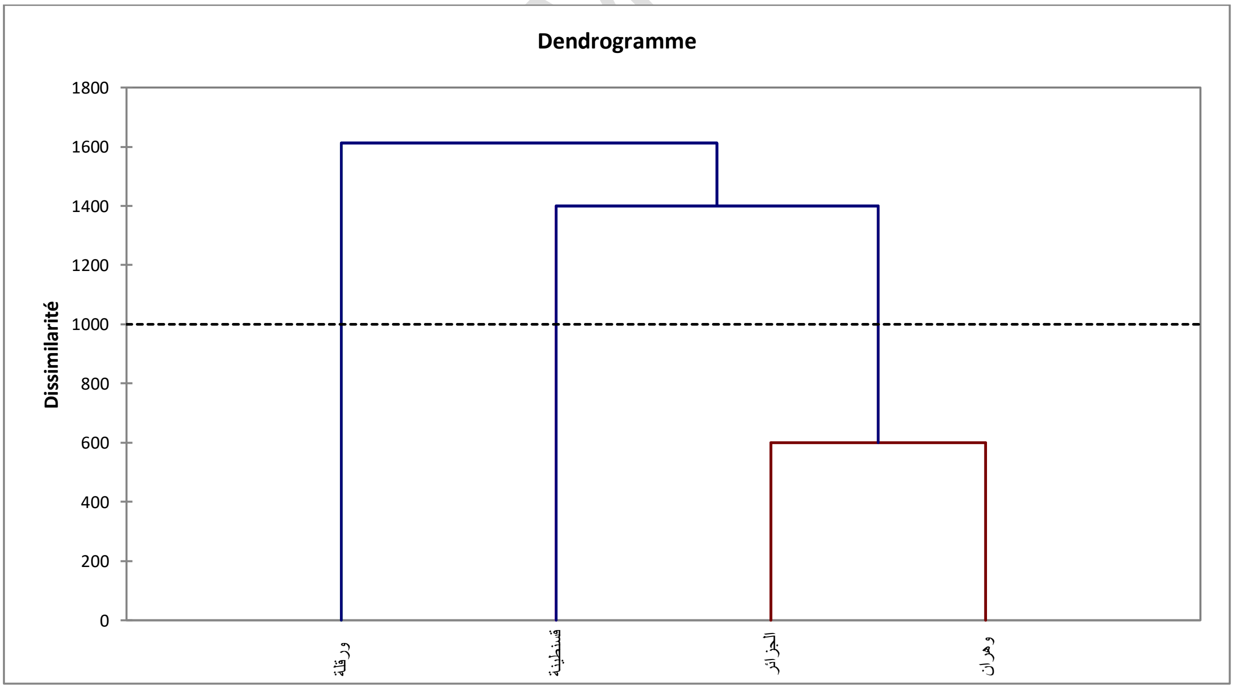
-التجميع وفقا لمؤشر وارد



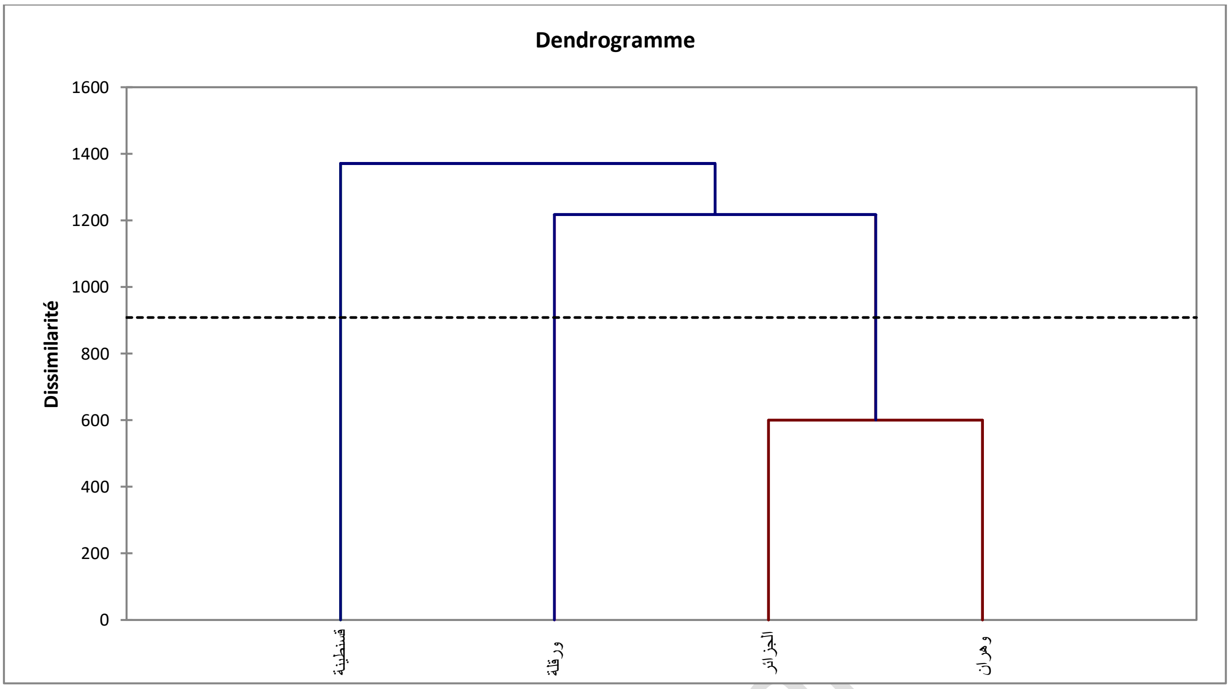
-التجميع وفقا لمؤشر المسافة المتوسطة



-التجميع وفقا لمؤشر المسافة القصوى



-التجميع وفقا لمؤشر المسافة الدنيا



-نلاحظ ان جميع الطرق اعطت نتائج متشابهة الى حد كبير كما انها جاءت مرضية .

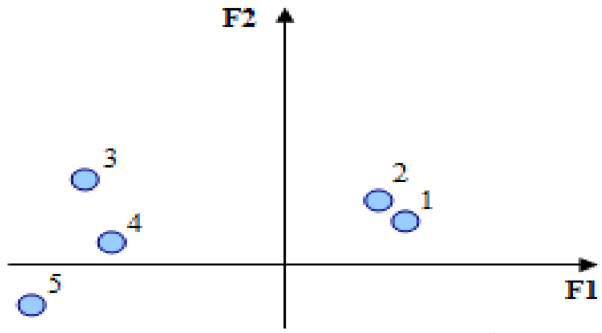
-من خلال مقارنة نتائج الطريقتين نلاحظ ان كلاهما سمح بتنصيف الافراد الى ثلاث مجموعات حيث جاءت نتيجة اسلوب التصنيف مدعمة لنتيجة ACP .

لو اعتمدنا على القيم الذاتية فإننا نجد ان الانتقال من مجموعتين الى ثلاث مجموعات يمكن ان يكسبنا 41,748 % + 55 ,241% وهذا جيد اما الانتقال من ثلاث مجموعات الى اربعة فيمكن ان يضيف لنا 3,011 في المئة فقط وهي نسبة ضعيفة جدا يمكن ان توقعنا في الخطأ.

	F1	F2	F3
Valeur propre	2,210	1,670	0,120
Variabilité (%)	55,241	41,748	3,011
% cumulé	55,241	96,989	100,000

<u>تمرين أخر لمتغيرين تم الحصول عليها عن طريق AFC</u>

ليكن خمس مفردات موزعة في فضاء متعدد الابعاد وفقا لمجموعة من المتغيرات ، حيث كان اسقاطهم في المخطط العاملي خلال المركبتين الاساسيتين الاولى والثانية فقط (F1،F2) باستعمال طريقة , AFC حيث كانت النتائج كالتالي:

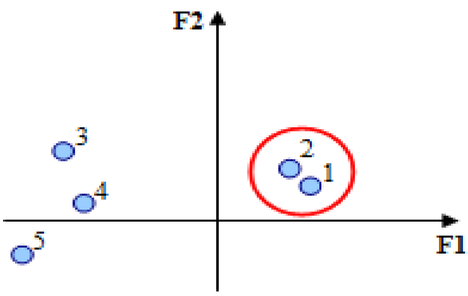


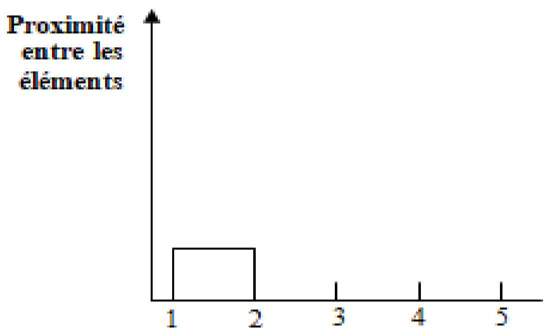
-قم بتجميع الافراد باستعمال مؤشرات WARD ثم قم برسم شجرة التصنيف.؟

<u>**ملاحظة**</u>: مقدار المسافة غير مهم مقارنة بنوعية النتائج الممكن الحصول عليها وكذا طريقة التجميع.

الخطوة1: 5 مفردات، 5 أقسام:

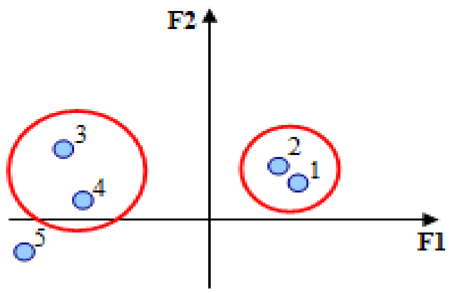
نقوم بتشكيل مصفوفة المسافات بين مختلف المفردات ، ثم نقوم بتجميع المفردتين الاكثر قربا.

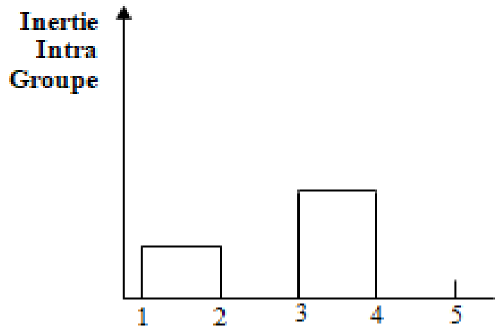




الخطوة 02: 05 مفردات، 4 أقسام

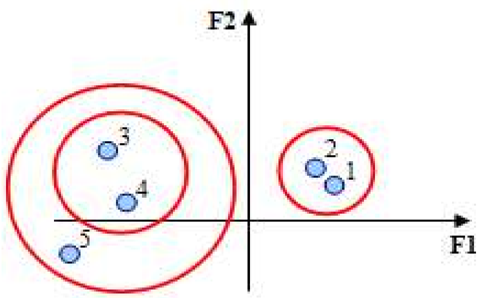
باستعمال معيار تجميع وليكن معيار WARD والذي يبحث عن تصغير حجم التباين بين الاقسام (وهذا يقودنا الى تعظيم التباين بين الاقسام، لان التباين الكلي ثابت ومساوي الى مجموع التباين داخل الاقسام وكذا بين الاقسام)، نقوم بتقدير المسافة بين قسم وكذا العناصر الفردية (المفردات). وعليه نقوم بجمع المفردين 1و2 في قسم ثم نقوم بمقارنة المسافات بين هذا القسم والمفردات الثلاثة المتبقية (3، 4 ،5)، ثم من جديد، نقوم بجمع المفردتين الاكثر قربا.

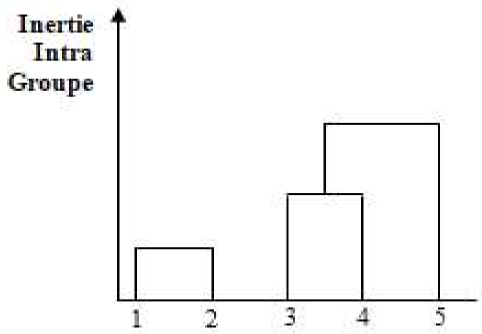




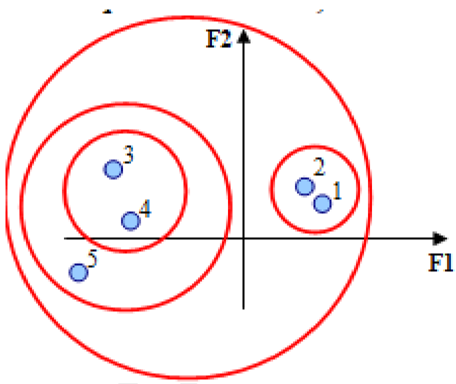
الخطوة 03: 05 مفردات، 3 أقسام

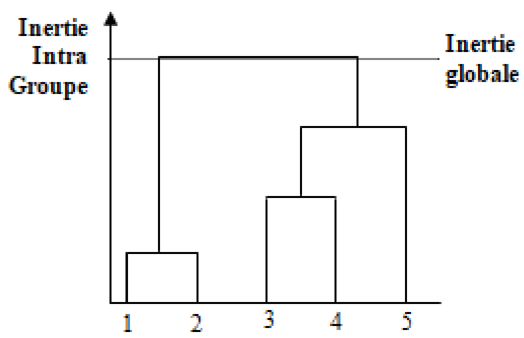
نقوم بإعادة نفس الاجراءات السابقة ، باتباع نفس المعيار التجميع WARD وكذا المسافة الاقليدية.



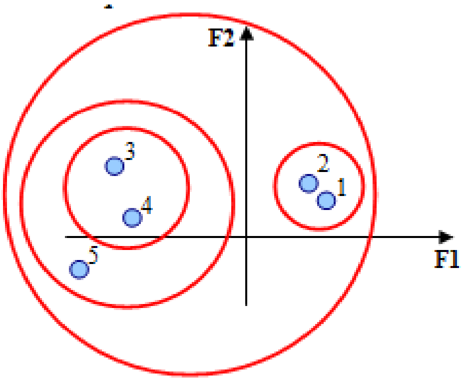


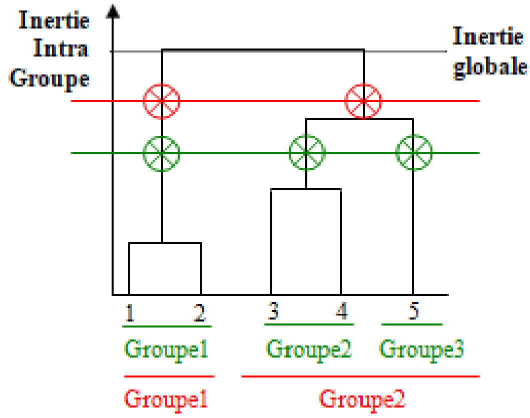
الخطوة 04: 05 مفردات، 2 أقسام





الخطوة الاخيرة: 05 مفردات، قسم وحيد , يتم عند هذه الخطوة اختيار مستوى قطع معين وهذا بالاعتماد على المعايير السابقة.





قائمة المراجع

- **HUSSON. F,** « Classification ascendante hiérarchique (CAH) », Laboratoire de mathématiques appliquées - Agrocampus Rennes, France, P43/ URL / math.agrocampus-ouest.fr/infoglueDeliverLive/digitalAssets/100457_AnaDo_CLASSIF_cours_slides.pdf .

- **CARPENTIER F-G.,** 2013/2014, « Analyse multidimensionnelle des données - Master 2ème année - Psychologie Sociale des Représentations », Réf : PSR92C – (polycopié et fichiers de données utilisés) / URL / http://geai.univ-brest.fr/~carpenti/

- **RAKOTOMALALA.R,** « Méthode de centres mobiles – Classification par partitions- Les méthodes de réallocation » ; Université Lumière Lyon 2, France , 31 pages . URL /http://tutoriels-data-mining.blogspot.fr/

- **ALVIN.C. R, WILEY. J & SONS** 2002, « Methods of Multivariate Analysis, Second Edition », Canada, P270.