محاضرات و تطبيقات حول التحليل العنقودي

التصنيفي

-مقدمة

1-التحليل العنقودي

1.1-تعريف و أهداف التحليل العنقودي

2.1 - مراحل التحليل العنقودي

2- التحليل العنقودي باستخدام SPSS

1.2-: التحليل الهيكلي

2.2-: التحليل الغير هيكلي

-الخاتمة

مقدمة

- التحليل العنقودي هو أسلوب إحصائي يُستخدم لتقسيم البيانات إلى مجموعات متجانسة تُعرف بـ"العناقيد" بناءً على تشابه الخصائص بين العناصر. يُعتبر هذا التحليل مفيدًا في العديد من المجالات مثل التسويق، الطب، والعلوم الاجتماعية، حيث يساعد في كشف الأنماط وتقسيم البيانات بطرق تسهّل فهمها واتخاذ القرارات.

تعريف التحليل العنقودي

يعتبر من الأساليب المهمة في تحليل البيانات، إذ يستخدم لغرض تصنيف ودراسة تجمعات البيانات أو المشاهدات بغية الوصول إلى وصف دقيق يستخدم هذا التحليل لتجميع المفردات بشكل عناقيد بالاعتماد على مقدار التشابه بينها، وتتم العنقدة بشكل متسلسل (الطريقة العنقدية)أو الأسلوب غير المتسلسل (الطريقة غير العنقدية) .

-الهدف

-لهدف من التحليل العنقودي هو تكوين مجموعات مفرداتها متقاربة أكثر ما يمكن و مجموعات مختلفة عن بعضها أكثر ما يمكن.

مراحل التحليل العنقودي

المرحلة الأولى: اختيار معايير التصنيف

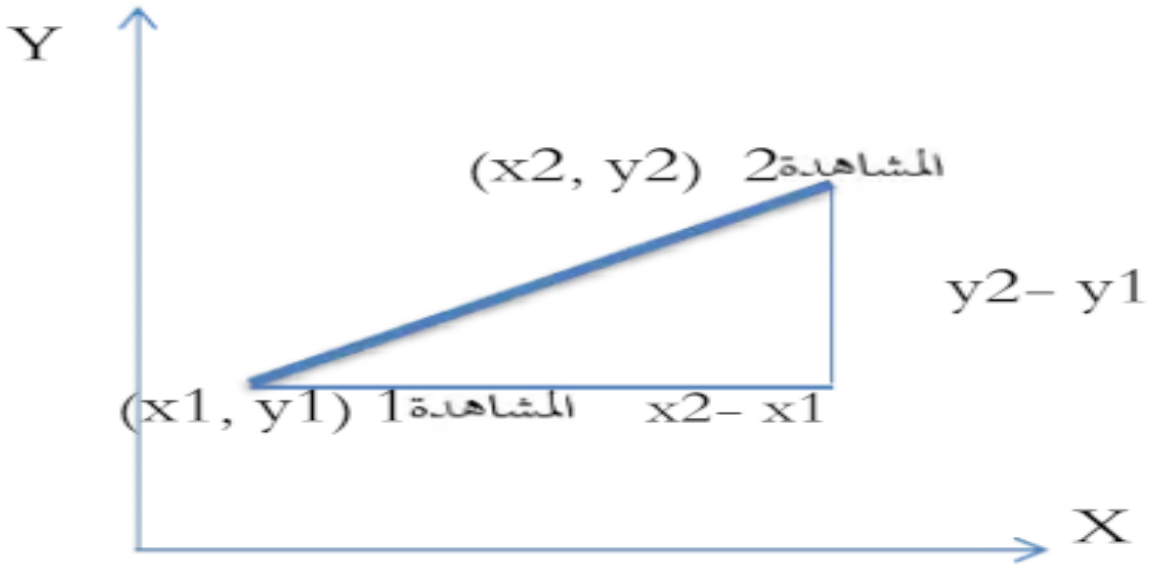
الكمية والجودة ومكان وطريقة الشراء

المرحلة الثانية حساب البعد

على خلاف التحليل العاميلي الذي يستعمل مصفوفة الارتباط لتجميع المتغيرات يقوم التحليل العنقودي في الأساس على حيتم التركيز على تحديد المتغيرات الأساسية لتبسيط الدراسة الإحصائية وتكوين مجموعات مميزة، وفي حال وجود عدد كبير منها يُفضل استخدام تقنية التحليل إلى مكونات أساسية ثم استعمال المتغيرات المجمعة لتصنيف العينة، مثل تصنيف الزبائن بناءً على ساب البعد، الذي تقوم من خلاله بتحديد مدى تقارب أو تباعد المفردات عن بعضها لتكوين مجموعات متجانسة ويختلف حساب البعد حسب نوع المتغيرات المستعملة حيث تشير قيمة البعد الكبيرة إلى التباعد بين المفردات، في حين تشير القيمة الصغيرة إلى التقارب بينهما

أـ المتغيرات الكمية

عند وجود هذه المتغيرات يوجد الكثير من الطرق الحساب البعد ولعل أكثرها استعمالا البعد الإقليدي ، الذي يقوم على حساب الجذر التربيعي المجموع مربعات للفوارق بين قيم لكل المتغيرات.



$$D = \sqrt{\sum_{k=1}^{p}(xik - xjk)^2}$$

مثال

ليكن لدينا طول و وزن الطالبين التاليين:

	الطول	الوزن
الطالب 1	192	86
الطالب 2	187	76

بما أن المتغيرين كميين إذا يمكن استعمال البعد الإقليدي ذو الصيغة التالية:

ومنه:

$$D1,2 = \sqrt{(192 - 187)^2 + (86 - 76)^2}$$

$$D1,2 = 11,18$$

كما يمكن استعمال معاملات أخرى لحساب البعد مثل

- معامل الإرتباط لبيرسون: يستعمل لحساب البعد إذا كانت المتغيرات تتغير في نفس الاتجاه و في اتجاه متعاكس.
- بعد تشيبيشب : يحسب القيمة المطلقة الأعظمية لقيم المتغيرات المعنية بالتصنيف.
- بعد نينكاوسكي : وهو الجذر النوني المجموع الفوارق المطلقة بين القيم للمفردات الإحصائية بالقوة النونية.

ب- المتغيرات الاسمية

بالنسبة للمتغيرات الاسمية نستعمل مصطلح مؤشر التماثل (الاختلاف) كمؤشر JACCARD مثلا

$$Sij = \frac{a}{a+c}$$

مثال:

لتكن لدينا البيانات التي تم جمعها عن ثلاثة أشخاص كالاتي:

الجدول رقم (13) : مثال بيانات اسمية

	المستوى التعليمي			الجنس		العمر		
	ابتدائي	ثانوي	جامعي	ذكر	أنثى	15-35	35-55	55-75
الشخص1	1	0	0	1	0	1	0	0
الشخص2	0	1	0	0	1	1	0	0
الشخص3	0	0	1	1	0	0	1	0

و بالتالي يمكن حساب مؤشر Jaccard للوحدات:

أ- 1 و 2: $S1,2= \frac{1}{1+4} = \frac{1}{5}$

ب- 2 و 3: $S2,3= 0$

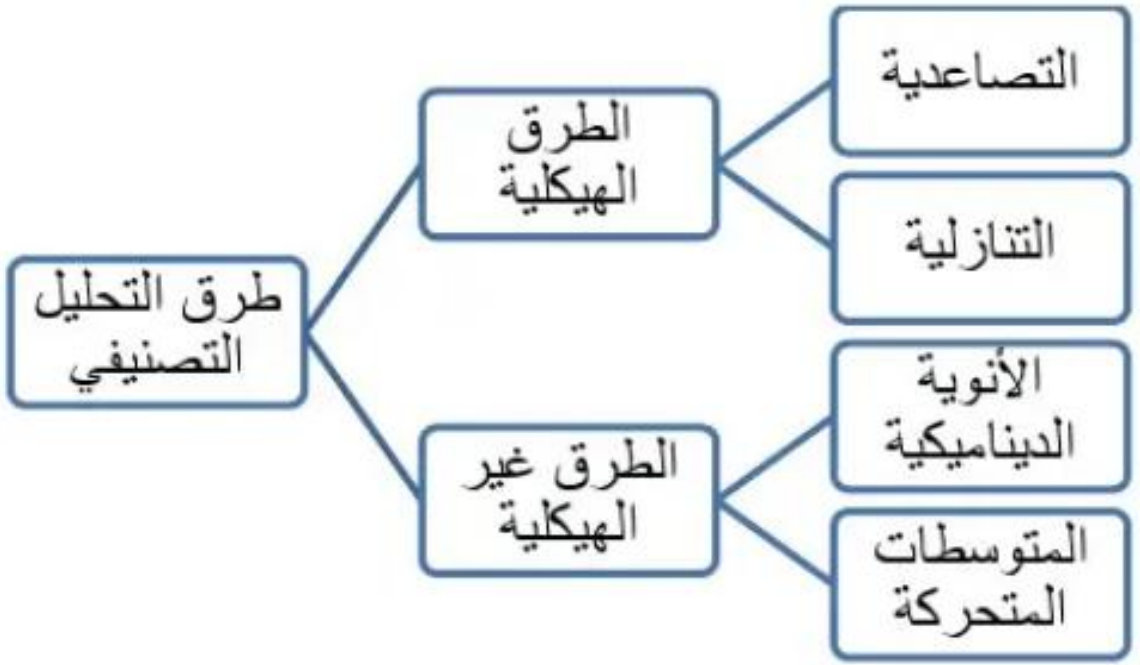
ت- 1 و 3: $S1,3= \frac{1}{1+4} = \frac{1}{5}$

- و يمكن استعمال المؤشرات التالية:
- – **بعد سوكال و ميشنر** : يمثل النسبة بين المفردات المتقاربة و المجموع الكلي للقيم.
- – **مؤشر روجرس وتنيموتو** : هذا المؤشر يستعمل مؤشر مضاعف الأهمية للمفردات غير المتطابقة.
- – بالإضافة إلى **بعد سوكال** و **سنيث**.
- انطلاقا من حساب الأبعاد بين كل زوج من مفردات العينة يتم تكوين مصفوفة البعد كالأتي:

n	3	2	1	
البعد1، n	البعد1، 3	البعد1، 2	0	1
البعد2، n	البعد2، 3	0	البعد2، 1	2
البعد3، n	0	البعد3، 2	البعد3، 1	3
..... 0	....	.....	.....	.....
0	البعد3، n	البعد2، n	البعد 1، n	n

المرحلة الثالثة: اختيار خوارزمية التحليل التصنيفي

يمكن تصنيف المفردات الإحصائية بطريقتين: **الطريقة الهيكلية**، التي تعتمد على تكوين مجموعات في شكل شجرة، إما تنازليًا من مجموعة كلية إلى مجموعات فردية، أو تصاعديًا من مفردات فردية إلى مجموعة كلية، و**الطريقة غير الهيكلية**.



أ- الطرق الهيكلية:

سنحاول توضيح الطريقة هيكلية التصاعدية من خلال استعمال مثال عبارة عن خمسة دول تحاول تصنيفها في مجموعات وفقا لعدد سكانها و مساحتها كما يوضح الجدول:

الدول	المساحة	الكثافة السكانية
1	8280	1624
2	41308	2795
3	26013	1320
4	31582	1600
5	27208	2795

Matrice de proximité

Observation	Distance euclidienne				
	1	2	3	4	5
1	,000	33048,752	17735,606	23302,012	18964,188
2	33048,752	,000	15365,958	9799,138	14100,000
3	17735,606	15365,958	,000	5576,035	1898,328
4	23302,012	9799,138	5576,035	,000	4534,303
5	18964,188	14100,000	1898,328	4534,303	,000

Ceci est une matrice de dissimilarité

المصدر: مخرجات SPSS22

أ- الطرق الغير هيكلية

تتميز هذه الطرق بقدرتها على معالجة العديد من المفردات الإحصائية واختيار عدد المجموعات بناءً على معيار محدد، مع توصية بالبدء بعينة صغيرة لضمان دقة التحديد.

- طريقة المربعات المتحركة: تقوم هذه الطريقة على فصل عدد من المفردات الإحصائية في nمجموعة مختارة، و يتم اختيار بصفة عشوائية أو من اختيار الباحث عنصر من هذه المجموعة يمثلها، والتي يبدأ من خلالها تجمع المفردات الإحصائية حولها انطلاقا من العناصر الأقرب، و بمجرد نسب كل العناصر إلى المجموعات يتعين حساب متوسط المجموعات النهائي.

— طريقة الأنوية الديناميكية:

ديناميكية أعم من طريقة المربعات المتحركة، حيث تُعبّر عن كل مجموعة بنواة من العناصر بدل عنصر واحد. تُحدث مراكز المجموعات مع كل إعادة نسب عنصر جديد، مما يجعل التجميع أبسط وإمكانية الوصول إلى الحل الأمثل أسرع. تُستخدم بكثرة لقدرتها على التعامل مع عينات تفوق 100 مفردة إحصائية، مع افتراض وجود فرقة إحصائية تمثل المجموعة.

- وصف المجموعات : يتم وصف المجموعات انطلاقا من المتغيرات المستعملة للتصنيف، حيث يمكن التعرف على ملمح الوحدات المكونة للمجموعات انطلاقا من المتوسط، التباين و الانحراف المعياري. أي اعطاء تسمية لكل مجموعة.

المرحلة الخامسة: التأكد من النتيجة

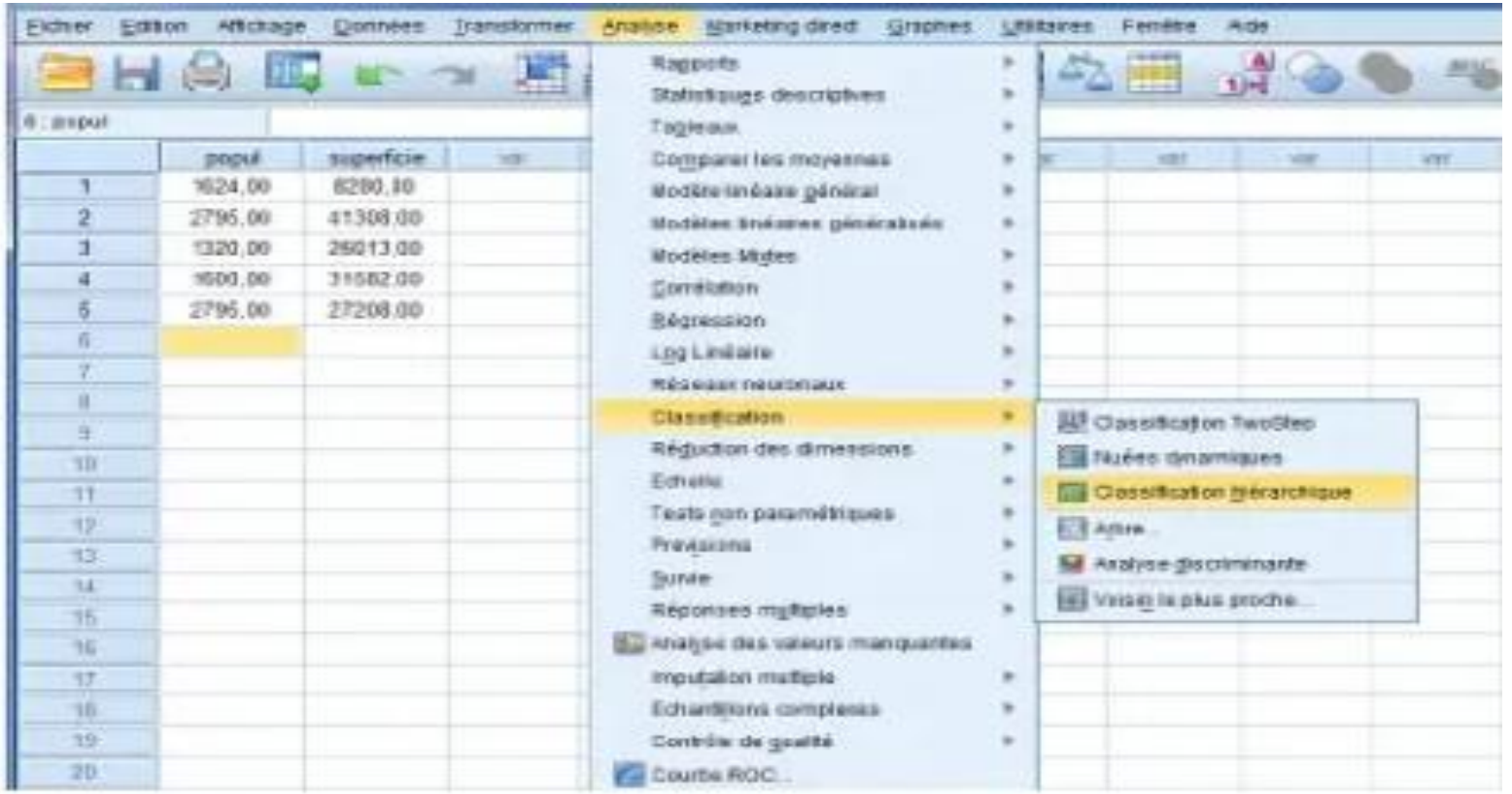
اختبار النتيجة من خلال الاختبارات الاحصائية يعد أمر صعب نوعا ما ولذا يتم اللجوء إلى بعض الطرق الاخرى

- مقارنة التصنيف المتوصل إليه بالتصنيف الطبيعي (الحقيقي) إذا وجد.
- تقسيم العينة إلى اثنين و تطبيق نفس الخوارزمية على الجزئين و مقارنة النتيجتين و في حالة التعارض يتم رفض النتيجة.
- نلاحظ توزيعات المجموعات إذا كانت متطابقة يلغى التصنيف

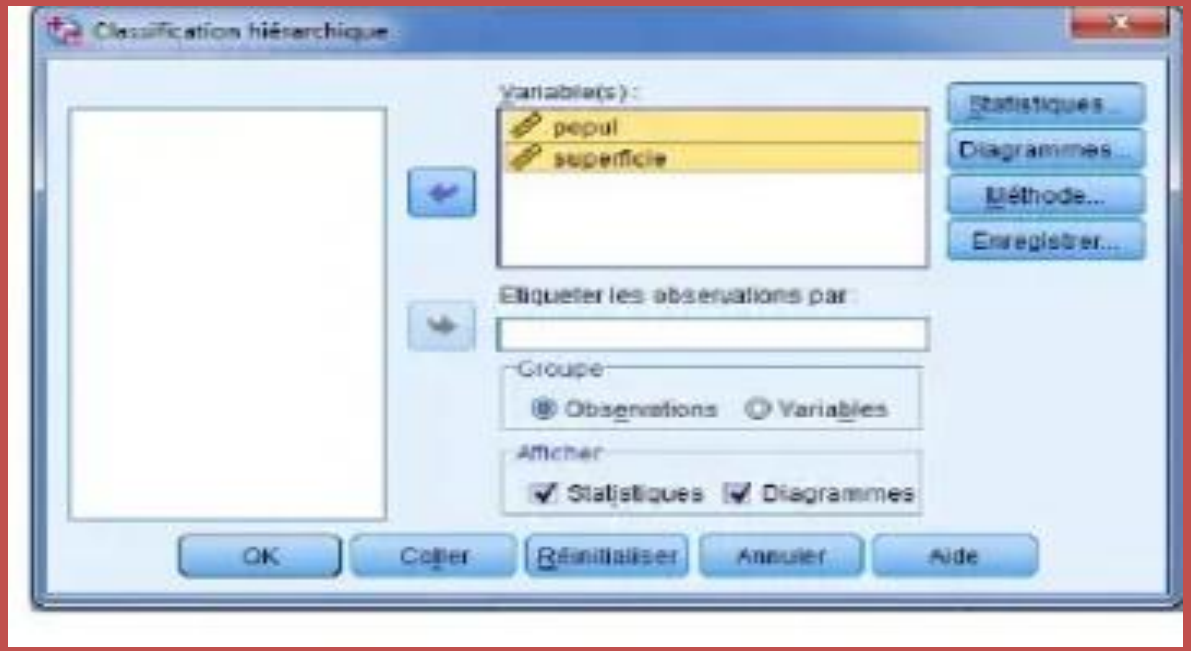
- التحليل العنقودي باستخدام SPSS

للتمكن من فهم اجراءات تطبيق التحليل العنقودي الهيكلي SPSS نستعمل المثال السابق وفقا للمرحلة المبينة في الشكل التالي

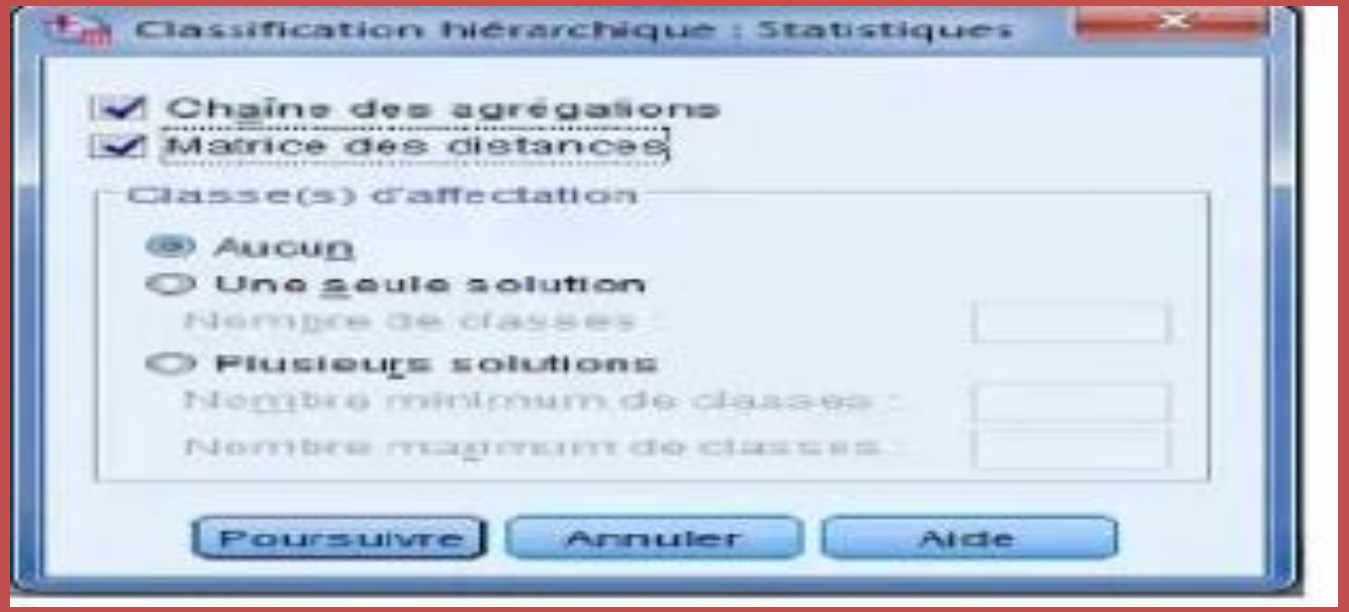
أ – التحليل الهيكلي :



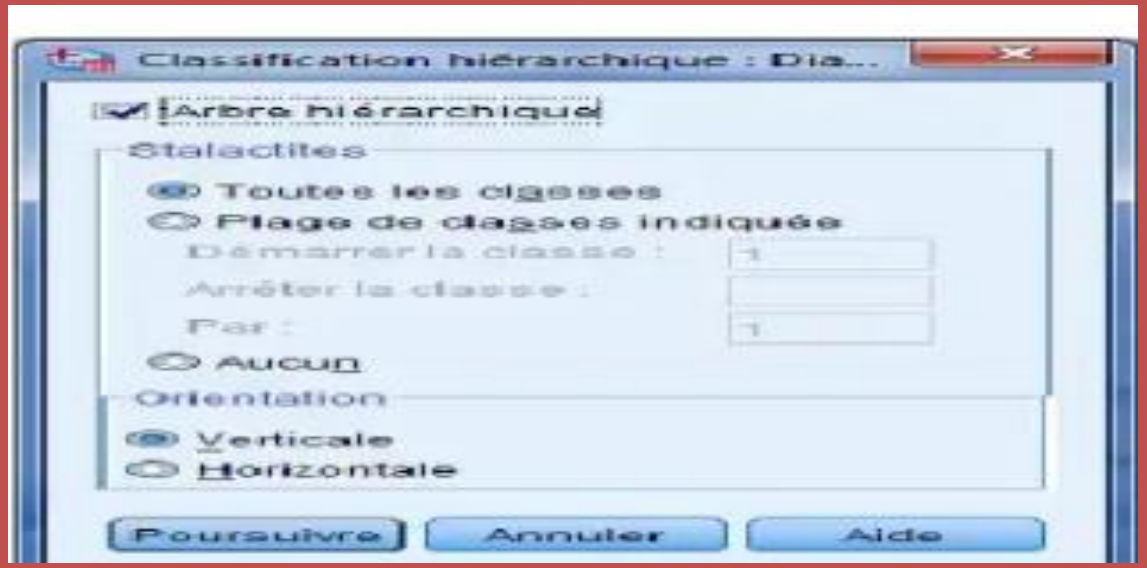
فيظهر صندوق الحوار التالي



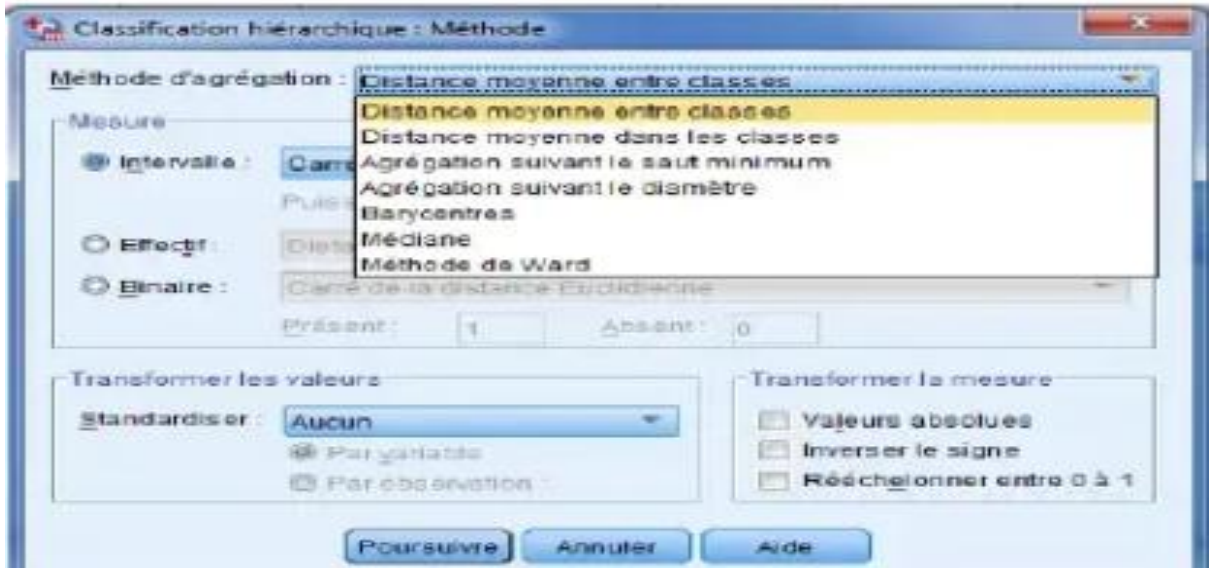
بعد إدخال المتغيرات التي تعتبر أساسا للتصنيف نضغط على STATISTIQUESللحصول على مصفوفة الأبعاد وكذا خوارزمية تجميع الفئات (المجموعات) كالاتي:



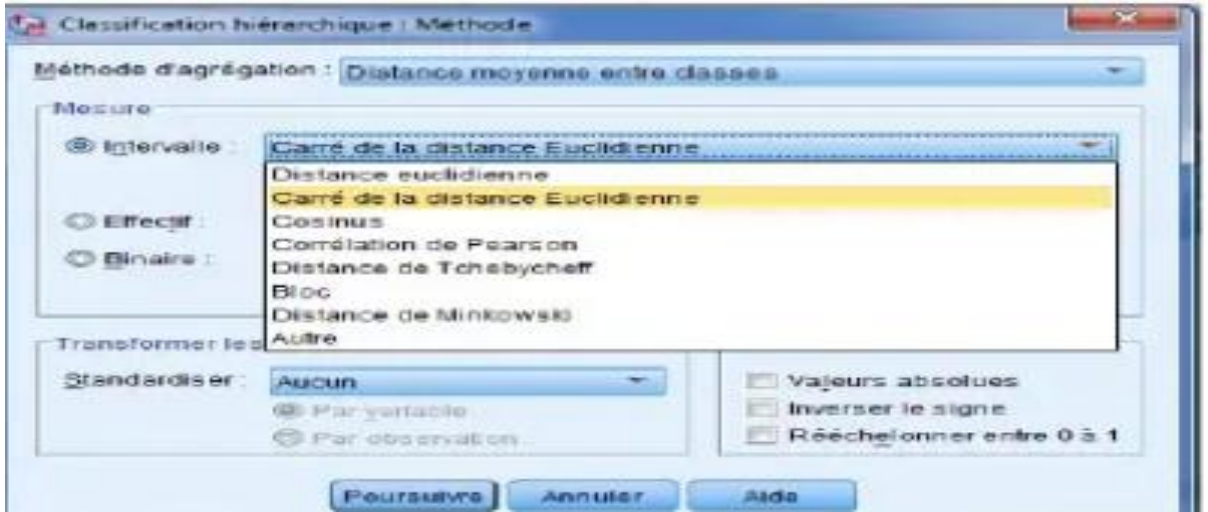
و من خلال DIAGRAMMESيمكن الحصول على الشكل الممكلي (الشحرة) مجميع المفردات الإحصائية للمعينة المدروسة في مجموعات، ويمكن الإختبار هذا بين أن يكون التمثيل أفقي أو عمودي كما يظهر في هدا الصندوق

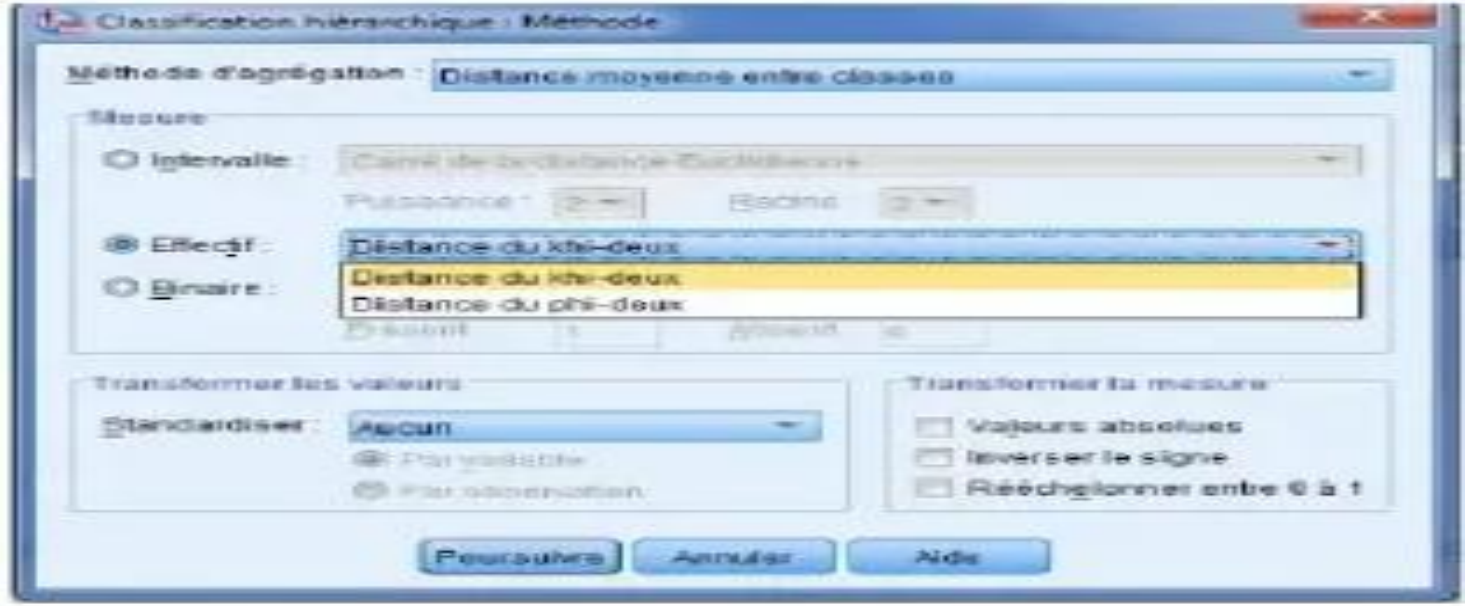


بينما من خلال الضغط على MÉTHODE يمكن الاختيار بين طرق التجميع الموضحة في صندوق الحوار الموالي:

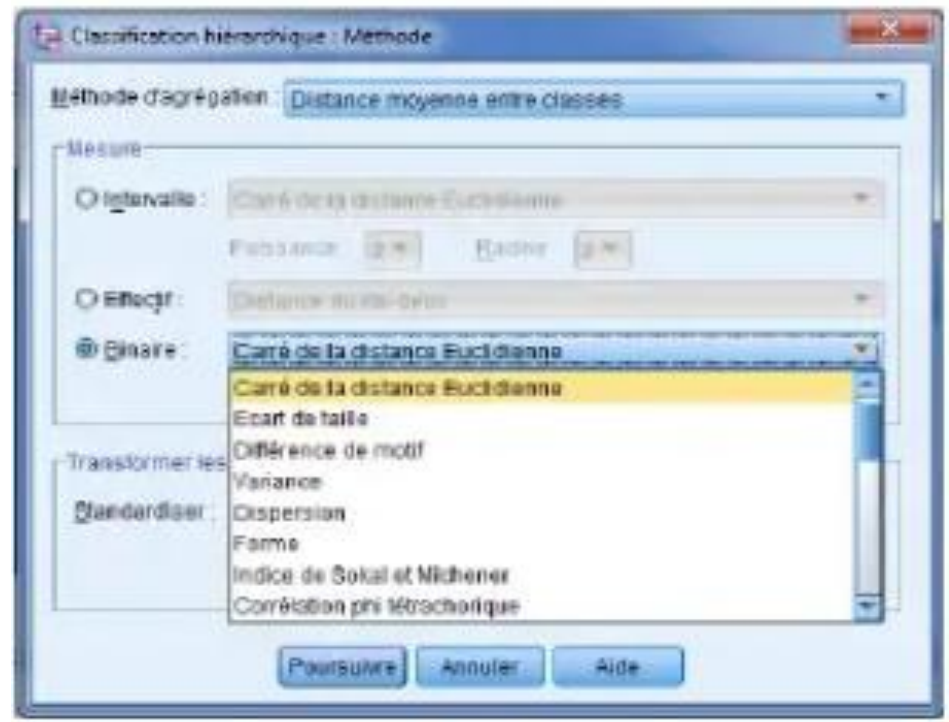


و كذا يمكن اختيار أسلوب قياس البعد بين المتغيرات و هذا حسب طبيعة هذه المتغيرات:





و منه نحصل على النتائج التالية :



1- مصفوفة البعد:

مصفوفة الأبعاد و التي ذكرنا سابقا أنها متناظرة و تقيس البعد بين كل زوج من مفردات العينة كما يوضح الجدول، وكما ذكرنا سابقا هذه القيم تختلف باختلاف نوع المتغيرات المستعملة (كمية، إسمية، تكرارات)

و طريقة قياس البعد المستعمل:

Observation	Distance euclidienne				
	1	2	3	4	5
1	,000	33048,752	17735,606	23302,012	18964,188
2	33048,752	,000	15365,958	9799,138	14100,000
3	17735,606	15365,958	,000	5576,035	1898,328
4	23302,012	9799,138	5576,035	,000	4534,303
5	18964,188	14100,000	1898,328	4534,303	,000

Ceci est une matrice de dissimilarité

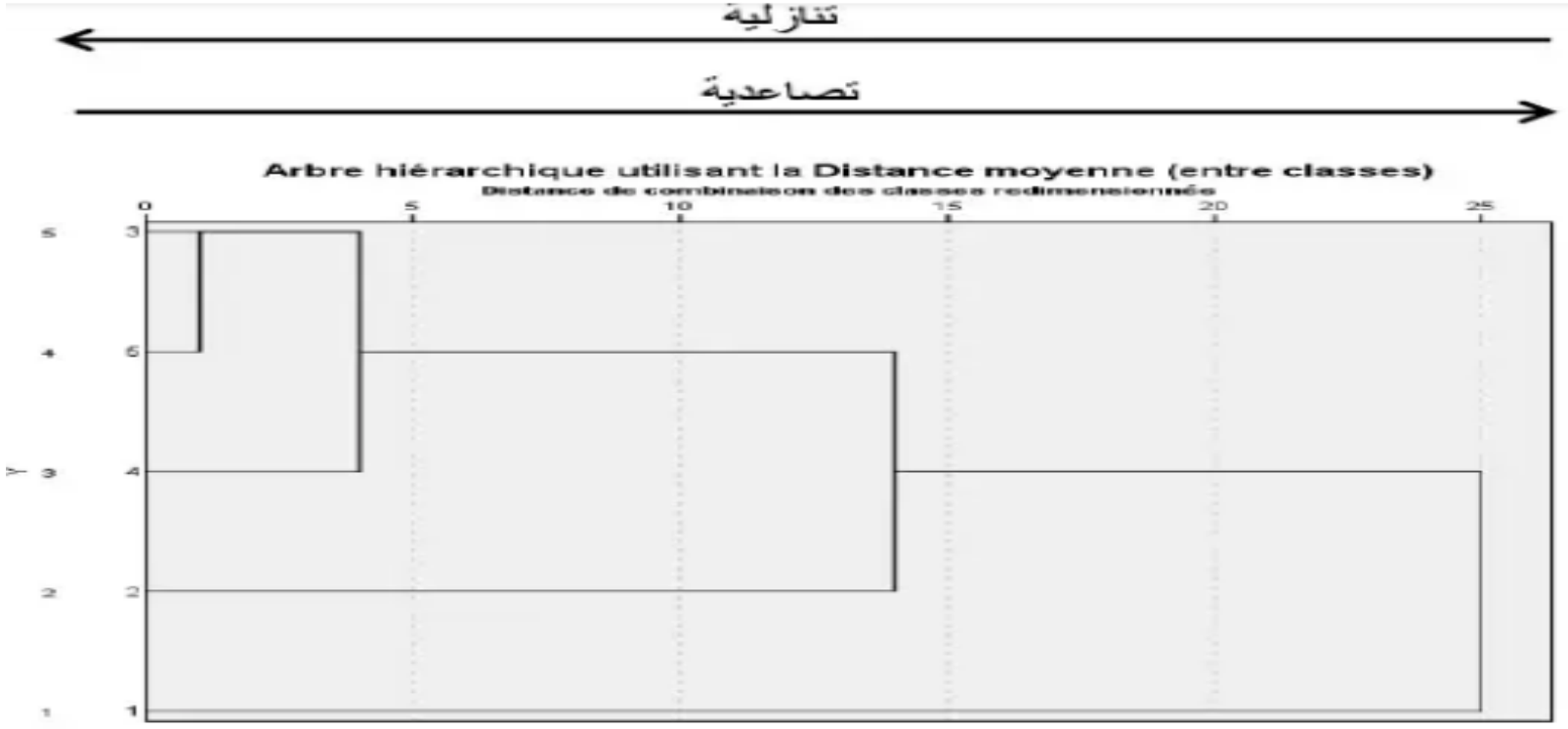
المصدر: مخرجات SPSS22

و باستعمال هذا الجدول يتم تجميع المفردات الاحصائية في مجموعات انطلاقا من البعد الأقل إلى البعد الذي يليه إذا كانت الطريقة تصاعدية، أما إذا كانت تنازلية فمن البعد الأكبر إلى البعد الذي يليه.

و الجدول التالي يوضح المراحل المتبعة للتجميع حيث يتم قراءة الجدول أفقيا فمثلا نقوم بوضع العنصر 3 و 5 في نفس المجموعة بعدها نذهب للمرحلة الموالية المشار إليها في اخر خانة و هي السطر الثاني في هذا المثال و نضم 4 إلى المجموعة السابقة. وتواصل العملية حتى تكوين جميع المجموعات.

Etape	Regroupement de classes		Coefficients	Etape d'apparition de la classe		Etape suivante
	Classe 1	Classe 2		Classe 1	Classe 2	
1	3	5	1898,328	0	0	2
2	3	4	5055,169	1	0	3
3	2	3	13088,365	0	2	4
4	1	2	23262,640	0	3	0

المصدر: مخرجات SPSS22



المصدر: مخرجات SPSS22

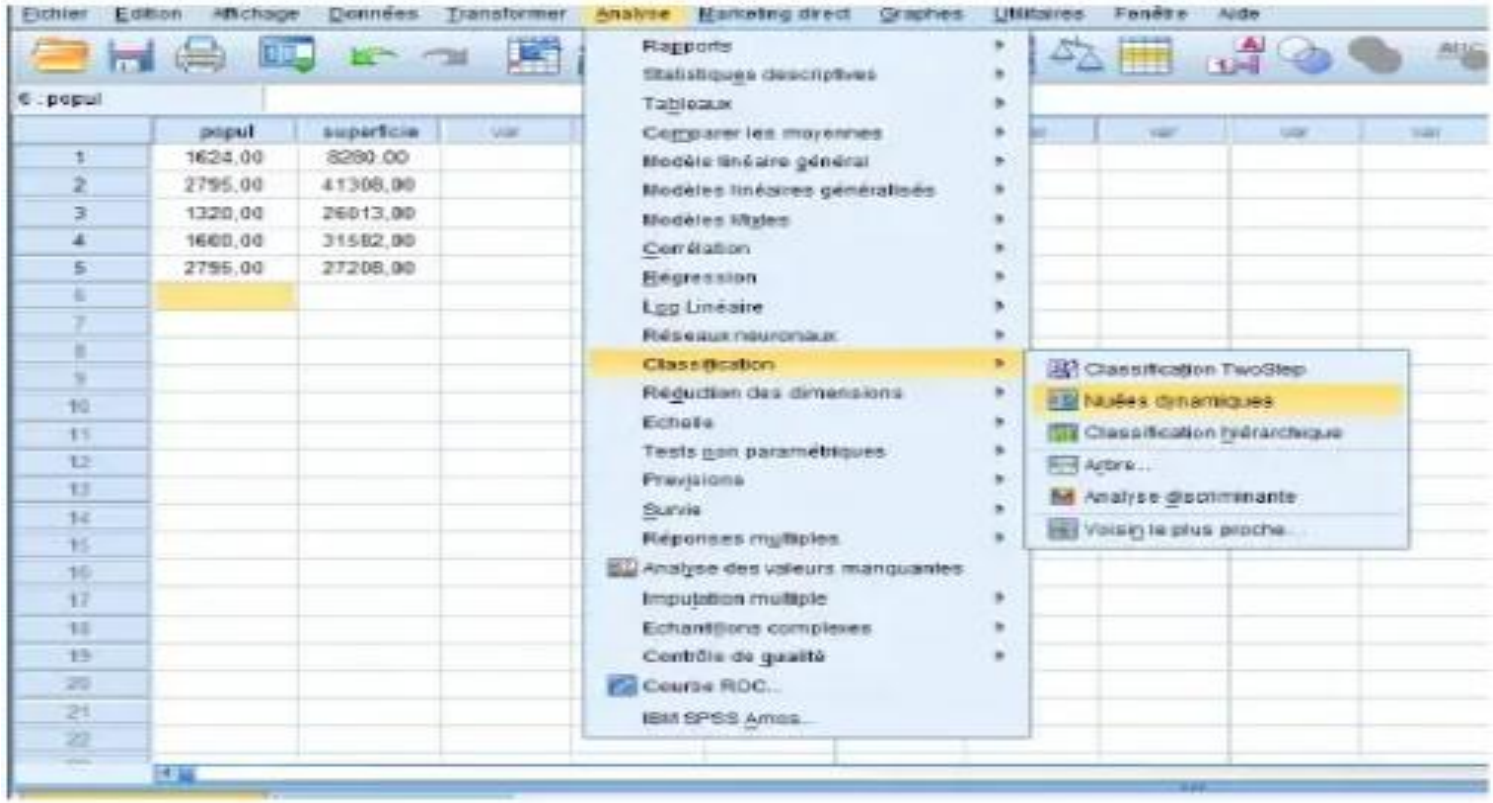
المجموعات	المجموعة الأولى	المجموعة الثانية	المجموعة الثالثة
	[4، 3، 5]	[2]	[1]
	$\bar{X}$	$\bar{X}$	$\bar{X}$
الكثافة السكانية	1905	2795	1624
المساحة	28267.66	41308	8280

و باستعمال هذا الجدول يتم تجميع المفردات الاحصائية في مجموعات انطلاقا من البعد الأقل إلى البعد الذي يليه إذا كانت الطريقة تصاعدية، أما إذا كانت تنازلية فمن البعد الأكبر إلى البعد الذي يليه.

و الجدول التالي يوضح المراحل المتبعة للتجميع حيث يتم قراءة الجدول أفقيا فمثلا نقوم بوضع العنصر 3 و 5 في نفس المجموعة بعدها نذهب للمرحلة الموالية المشار إليها في اخر خانة و هي السطر الثاني في هذا المثال و نضم 4 إلى المجموعة السابقة. وتواصل العملية حتى تكوين جميع المجموعات.

ب- التحليل العنقودي الغير هيكلي

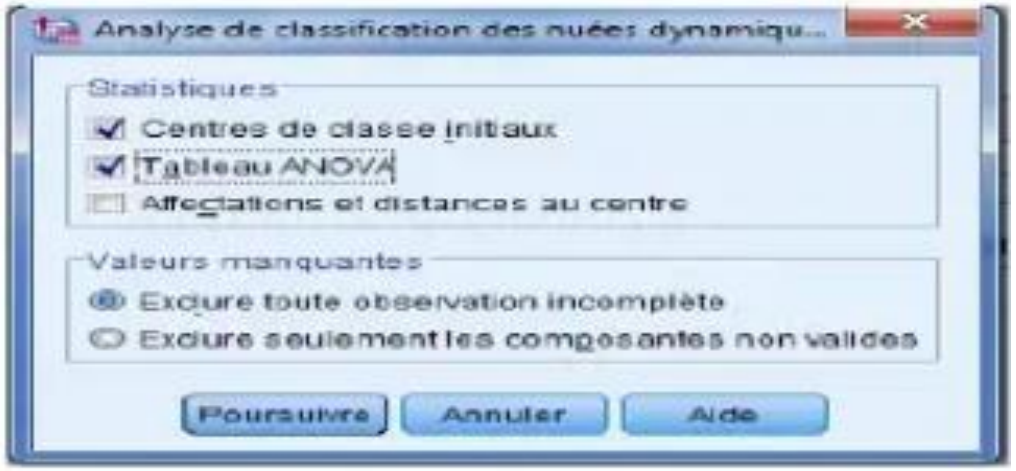
- يتم اما من خلال الأنوية الديناميكية أو المتوسطات المتحركة كما يلي :



بعد إدخال المتغيرات التي تم اختيارها كأساس للتجميع كما يلي :

يظهر صندوق الحوار التالي الذي يتم من خلاله حفظ المجموعات التي نسبت إليها كل مفردة إحصائية، و بعد كل مفردة إحصائية عن مركز الفئة (المجموعة) المنتمية إليها.

أما من خلال صندوق الحوار التالي فيمكن الحصول على جدول التباين الأحادي الذي سنأتي على شرح استعماله في التحليل التصنيفي عند تحليل النتائج، ويمكننا أيضا من الحصول على مراكز الفئات الإبتدائية قبل بداية نسب المفردات الإحصائية إليها.



و منه نحصل على النتائج التالية

	Classe		
	1	2	3
popul	1624,00	2795,00	1320,00
superficie	8280,00	41308,00	26013,00

المصدر: مخرجات SPSS22

لجدول السابق يمثل مراكز الفئات الابتدائية التي أشرنا سابقا أنها تعين من طرف الباحث أو يتم اختيارها جزافا.
أما الجدول الموالي فهو يمثل سيرورة عملية التغيير التي تحدث للمجموعات نتيجة نسب متغيرات جديدة لها، فتلاحظ مثلا أن المجموعة الأولى والثانية لم يحدث فيها اي تغيير و بالتالي فهي تأخذ قيمة الصفر.

Itération	Changements dans les centres de classes		
	1	2	3
1	,000	,000	2329,323
2	,000	,000	,000

المصدر: مخرجات SPSS22

و هذا ما يأكده الجدول الموالي حيث نلاحظ تطابق بين متوسط الفئات الإبتدائي والنهائي للمجموعتين الأولى و الثانية ذلك أنهما ضمتا نفس العنصرين فقط. في حين أن المجموعة الثالثة حدث فيها تغييرات لأنها ضمت عناصر جديدة.

	Classe		
	1	2	3
popul	1624,00	2795,00	1905,00
superficie	8280,00	41308,00	28267,67

المصدر: مخرجات SPSS22

أما من خلال الجدول الموالي و الذي يمثل تحليل التباين أحادي البعد فإنه يتم تحديد أي المتغيرات (متغيرين في هذه الحالة الأكثر قدرة على الفصل بين المجموعات.

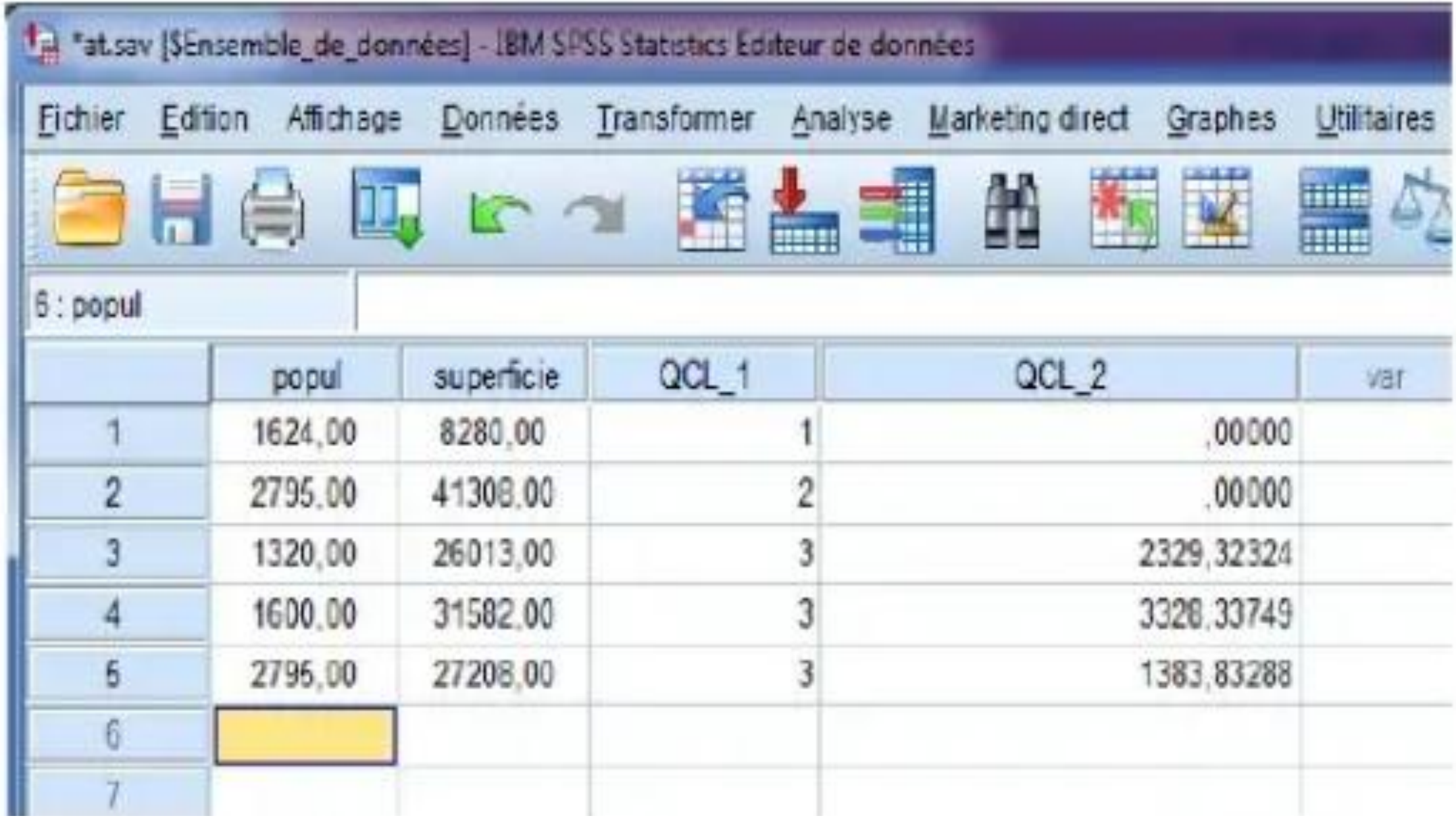
	Classe		Erreur		F	Signification
	Moyenne des carrés	ddl	Moyenne des carrés	ddl		
الكثافة السكانية	398442,400	2	613675,000	2	,649	,606
المساحة	279952012,067	2	8595610,333	2	32,569	,030

Nombre d'observations dans chaque classe

Classe	1	1,000
	2	1,000
	3	3,000
Valides		5,000
Manquantes		,000

المصدر: مخرجات SPSS22

أما الجدول التالي فيمثل عدد المفردات الإحصائية في كل مجموعة إذ نلاحظ أن المجموعتين الأولى و الثانية تتكون كل منهما من مفردة إحصائية واحدة، أما المجموعة الإحصائية الثالثة فتتكون من ثلاث مفردات إحصائية. أي مجموع خمسة مفردات إحصائية.



	popul	superficie	QCL_1	QCL_2	var
1	1624,00	8280,00	1	,00000	
2	2795,00	41308,00	2	,00000	
3	1320,00	26013,00	3	2329,32324	
4	1600,00	31582,00	3	3328,33749	
5	2795,00	27208,00	3	1383,83288	
6					
7					

خاتمة

يُعد التحليل العنقودي أداة فعّالة لتصنيف البيانات واكتشاف الأنماط المخفية فيها، خاصة عند استخدام برامج متقدمة مثل SPSS. ومع ذلك، يبقى تحديد العدد الأمثل للعناقيد من أبرز التحديات التي تواجه الباحثين، حيث يتطلب موازنة بين دقة النتائج وقابلية تفسيرها. لذا، فإن استخدام أساليب دقيقة ومعايير واضحة لتحديد العناقيد يُعزز من فعالية هذا التحليل ويسهم في الحصول على رؤى إحصائية دقيقة وذات قيمة.