Journal of Al Azhar University-Gaza *(Natural Sciences)*, **2021, 23** : **39-66**

استخدام التحليل العنقودي الهرمي وغير الهرمي في تصنيف القوى العاملة في فلسطين (دراسة تطبيقية مقارنة)

Classification of labor force in Palestine using hierarchical and non-hierarchical cluster analysis "Comparative Applied Study"

مؤمن محمد الحنجوري **محمد بسام حمد**

جامعة الأزهر – غزة

moamin2000@hotmail.com

Received 18/08/2021 *Accepted* 5/12/2021

الملخص:

في هذا البحث تم استخدام طريقتي التحليل العنقودي الهرمي وغير الهرمي في تصنيف القوى العاملة في فلسطين حسب المحافظات الفلسطينية، وقد تم التوصل إلى أن هناك تقارباً في تصنيف القوى العاملة في فلسطين بين ست محافظات والتي شكلت العنقود الأول وهي(بيت لحم ،جنين ،رام الله ،قلقيلية ،القدس ،أريحا) وأيضاً كان هناك تقارب في تصنيف القوى العاملة بين خمس محافظات، والتي شكلت العنقود الثاني وهي(الخليل ،طولكرم ،نابلس ،سلفيت ،طوباس) وأيضاً كان هناك تقارب في تصنيف القوى العاملة بين خمس محافظات، والتي شكلت العنقود الثالث وهي(شمال غزة ،مدينة غزة ،دير البلح ،خان يونس ،رفح)،وقد تم إجراء مقارنة بين طريقتي التحليل العنقودي الهرمي وغير الهرمي وهي:

(Hierarchical Cluster Analysis , K-Means Clustering ,K-Mediods Clustering)

وكانت الأفضلية لطريقة Hierarchical Cluster Analysis، وتم عمل مقارنة لاختيار أفضل طريقة للربط من خلال التجميع العنقودي الهرمي:

(Average linkage method, Complete linkage method, Ward linkage method , Single linkage method)

وكانت الأفضلية لطريقة الربط الهرمية Ward، وهي التي تم استخدامها في التصنيف .

الكلمات المفتاحية: التحليل العنقودي الهرمي وغير الهرمي_ طريقة الر بط الهرمي_ خوارزمية(k-means)

Abstract:

this research uses two methods of hierarchical and non-hierarchical cluster analysis to classify the labor force in Palestine according to the Palestinian governorates,

The research has been concluded that there is convergence at the classification of the labor force in Palestine between six governorates, which formed the first cluster is (Bethlehem, Jenin, Ramallah, Qalqilya, Jerusalem, Jericho) and also there is convergence at the classification of the labor force among the five governorates that formed the second cluster is (Hebron, Tulkarm, Nablus, Salfit, Tubas) and there is also a convergence at the classification of the labor force among the five governorates, which formed the third cluster is (North Gaza, Gaza City, Deir Al-Balah, Khan Yunis, Rafah), and a comparison is made between The two methods of hierarchical and non-hierarchical cluster analysis, namely:

(Hierarchical Cluster Analysis , K-means Clustering ,K-mediods Clustering)

The preference goes for a Hierarchical Cluster Analysis method, a comparison is made to choose the best method of linking through hierarchical cluster grouping: Average linkage method, Complete linkage method, Ward linkage method). (Single linkage method,

The preference goes to the hierarchical linking method, Ward, which used in the classification.

Keywords: hierarchical and non-hierarchical cluster analysis - hierarchical association method - k-means algorithm

المقدمة:

يعتمد أسلوب التحليل الإحصائي على نواة المشكلة محل القياس والتحليل ونوع البيانات المتوافرة علما بأنَّ أسلوب التحليل الذي يستخدم لدراسة مشكلة معينة، قد لا يكون مناسباً لدراسة مشكلة أخرى، وذلك لاختلاف طبيعة المشكلة ونوع البيانات.

ونظراً لتعدد أهداف البحوث والدراسات وتعدد متغيراتها الاقتصادية والاجتماعية، فقد يتم تحديد الأساليب الإحصائية، التي تتفق مع طبيعة البيانات وتحقق أهداف البحوث والدراسات والتي من أهمها أسلوب التحليل الإحصائي متعدد المتغيرات " Multivariate statistical analysis" والذي يتضمن مجموعة من الأساليب أهمها:

١ –أسلوب تحليل المكونات الرئيسية Principal components Analysis .

٢ –أسلوب التحليل العاملي Factor analysis.

٣ –أسلوب التحليل العنقودي Cluster Analysis.

وسيتم من خلال هذا البحث التركيز على أسلوب التحليل العنقودي (Cluster Analysis) على اعتبار أن هناك الكثير من الدراسات والبحوث تناولت أسلوب المكونات الرئيسية، وأسلوب التحليل العاملي، في دراسة تصنيف أفراد القوى العاملة في فلسطين، وندرة الدراسات التي تناولت أسلوب التحليل العنقودي في تصنيف أفراد القوى العاملة في فلسطين حسب المحافظات الفلسطينية،(الجاعوني،2001).

يستخدم التحليل العنقودي لتحديد المجموعات الفرعية المختلفة للمشاهدات خارج مجموعة من المشاهدات بناءً على مجموعة من الخصائص. في ضوء التحليل العنقودي، من الممكن علاج الأفراد المتباينين بطريقة مناسبة من خلال أخذ اختلافهم في الاعتبار. سيؤدي ذلك إلى تعزيز دقة وكفاءة نماذج التقدير والتنبؤ. تهدف هذه الدراسة إلى تقييم أداء التسلسل الهرمي وغير الهرمي لدراسة أوجه التشابه / الاختلافات الإقليمية في سوق العمل الفلسطيني والتي لا يمكن ربطها بتحليل وصفي بسيط من الظواهر المرتبطة. يجب أن نضع معايير للمقارنة المكانية للأراضي الخاضعة للتحليل من أجل تطوير السياسات العامة على الصعيدين المركزي والإقليمي. يجب أن تكون هذه السياسات هي الأنسب لمحاربة مشاكل البطالة المرتبطة بها. جدير بالذكر أن توليد السياسات العامة للتوظيف، والتي تستخدم لمحاربة ظاهرة البطالة المستمرة، استحقت اهتمامًا خاصًا في الاقتصاد الفلسطيني على مدار العقود الماضية(الجهاز المركزي للإحصاء الفلسطيني، 2018)، ومع ذلك، لا يُعرف الكثير عن صورة الأفراد العاطلين عن العمل المسجلين في مناطق السلطة الوطنية الفلسطينية. هذه المعرفة لها أهمية حاسمة في تطوير سياسات سوق العمل العامة التي تستهدف تحديداً ملامح البطالة الإقليمية.

يسعى هذا البحث إلى تقييم أداء طريقة تحليل العنقودي الهرمي وغير الهرمي، بحيث يتم تجميع الوحدات الإقليمية في فئات، وفقًا لأوجه التشابه التي لوحظت من خلال مجموعة من المتغيرات التوضيحية المقدمة. الهدف من ذلك هو اكتشاف وجود التجانس بين المناطق المختلفة بناءً على طريقة إحصائية متعددة المتغيرات –منهجية تحليل المجموعات. تسمح هذه الطريقة بالحصول على تجزئة للأراضي حسب المناطق التي تتميز بملف تعريف يحدد "السكان"، في المتوسط ، عدد السكان المسجلين العاطلين عن العمل، سيكون الغرض منه تحديد أنماط الاستقرار (أو التطور) لمتوسط ملامح السكان المسجلين في جميع المناطق الستة عشر التي تشكل مناطق السلطة الوطنية الفلسطينية. للوصول إلى الهدف المذكور، سيتم أخذ مجموعة من المتغيرات التي أعلنتها المؤسسة العامة التي تدير سجلات الأفراد العاطلين عن العمل –الجهاز المركزي للإحصاء الفلسطيني (PCBS)، في الاعتبار.

مشكلة البحث:

تتعرض القوى العاملة عادة للعديد من المشكلات الرئيسية والمشتركة بين العديد من المجتمعات والبلدان، حيث تكثر مشكلات القوى العاملة في دول العالم الثالث، وفي المناطق الفقيرة من العالم، في حين تقل مشكلات هذه الفئة كلما تحسنت أحوال الدولة وأوضاعها المادية، تكمن المشكلة في دراسة أوجه التشابه و الاختلافات الإقليمية في الأراضي الفلسطينية وفقًا لأوجه التشابه التي لوحظت من خلال مجموعة من المتغيرات التوضيحية التي قدمت صورة محددة لأفراد القوى العاملة المسجلة باستخدام التحليل العنقودي الهرمي وغير الهرمي، ويحاول البحث الإجابة على الأسئلة التالية، والتي تعبر عن المشكلات قيد البحث والمشار إليها في الأسطر السابقة:

- ما هو حجم أفراد القوى العاملة في السوق العمل الفلسطيني؟
- هل يوجد تشابه او اختلاف بين أفراد القوى العاملة بالنسبة للمحافظات الفلسطينية؟

أهداف البحث:

يهدف هذا البحث إلى استخدام التحليل العنقودي الهرمي وغير الهرمي في تصنيف المناطق الفلسطينية إلى مجموعات متجانسة وفقًا لملف تعريف محدد للبطالة المسجلة، بالإضافة لتحقيق الأهداف التالية:

- تصنيف البيانات وتحديد العلاقة بين العناصر من حيث التشابه أو الاختلاف بناءً على القياسات ذات الصلة.
- تقييم أداء طريقتي التحليل العنقودي الهرمي وغير الهرمي.
- مقارنة بين طريقتي التحليل العنقودي الهرمي وغير الهرمي.

الإطار النظري:

التحليل العنقودي الهرميHierarchical Clustering Analysis

لا تتطلب طريقة التحليل العنقودي الهرمي المعرفة المسبقة بعدد العناقيد التي سيتم تصنيف الحالات على أساسها ،التحليل العنقودي الهرمي يناسب العينات الصغيرة نسبيا(جودة،2007) ،في التصنيف الهرمي لا يتم تقسيم البيانات إلى عناقيد في خطوة واحدة ،ونحتاج بدلاً من ذلك إلى مراحل متتالية.

يمكن تقسيم الأسلوب الهرمي إلى الأساليب التجميعية والتي تبدأ بسلسلة من الاندماجات المتتالية من n من الوحدات والتي تتحول إلى عناقيد، ثم الأساليب التقسيمية والتي تقسم n من الوحدات بصورة متتالية إلى تقسيمات دقيقة. ويمكن اعتبار هذين النوعين من الأساليب الهرمية على أنها محاولات للعثور على العنقدة الأكثر كفاءة ،ذلك في كل مرحلة من مراحل التقسيم المتقدم أو تجميع البيانات، وعندما يتم استخدامها فإن الانقسامات والاندماجات التي تتم تصبح نهائية، بما أن

الأسلوب الهرمي التجميعي يخفض البيانات في نهاية المطاف إلى عنقود واحد يضم كل الوحدات ،بينما الأساليب التقسيمية في الختام تجزئ المجموعة الكاملة من البيانات إلى n من المجموعات كل مجموعة تضم وحدة واحدة ،فإن الباحث يحتاج أن يقرر في أي مرحلة من مراحل التحليل يرغب في التوقف من أجل تعريف العناقيد وتحديد عددها. هناك عدة طرق للتجميع توفرها النظرية الإحصائية ونتناول فيما يلي بشكل موجز اثنين من أهم هذه الطرق وهي طريقة التجميع الهرمية والطريقة غير الهرمية واللتين سيتم تطبيقهما في التجميع وسنفترض أن المطلوب هو تجميع وحدات وليس متغيرات وهي الحالة التي تهمنا(Feinstein,1996; Aldenderfer et al.,1984).

طرق التجميع الهرمية:Clustering Methods

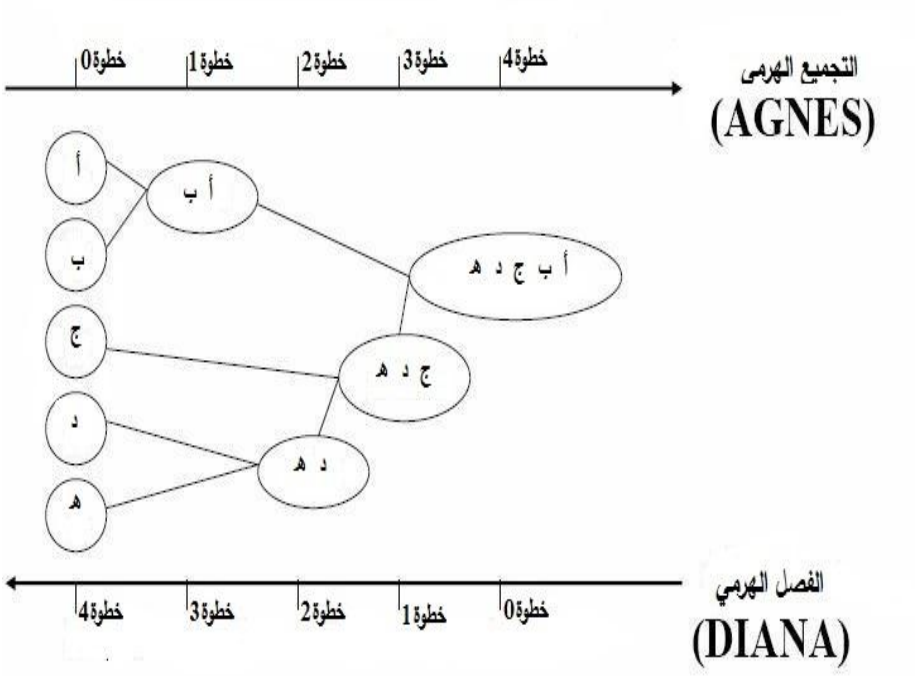
هناك نوعان من طرق التجميع الهرمية:

1. **طريقة الفصل الهرميةDivisive 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 عددٌ من المجموعات الفرعية مساوياً لعدد العناصر، أي حتى يكون كل عنصر مجموعة بنفسه، وفي هذا النوع من التحليل جميع العناصر تتجمع في عنقود واحد وبعد ذلك يتم تصنيف العناصر في عناقيد أصغر فأصغر(نامق،2010).
2. **طريقة التكتل الهرميةAgglomerative Hierarchical Methods**، وهي إحدى طرق الربط "linkage methods"المتتالية التي تلائم تجميع المفردات وكما تلائم تجميع المتغيرات وهذا لا يتحقق بالنسبة لجميع طرق الربط الهرمية الأخرى، وفيما يلي خطوات تنفيذ طرق التجميع الهرمية لربط المجموعات عند وجود N من العناصر (المفردات أو المتغيرات) وتسمى هذه الخطوات بالطريقة العامة(جونسون،1998):
   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 ×N).
   2) نبحث في مصفوفة المسافة عن أقرب زوج من المجموعات (الزوج الأكثر تماثلاً) نفترض أن d_{AB} تشير إلى المسافة بين الزوج الأكثر تماثلاً A,B.
   3) ندمج المجموعة A مع المجموعة B، ونستخدم الرمز (AB) للإشارة إلى المجموعة الجديدة، ونعدل من عناصر مصفوفة المسافة على النحو التالي:
      i. نحذف الصفوف والأعمدة المناظرة للمجموعتين B,A.

ii. نضـف صـفاً وعمـوداً جديـدين يعطيـان المسـافة بـين المجموعـة الجديـدة (AB) والمجموعات الأخرى

4) نكرر الخطوتين (i)، (ii) 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.

ويمكن عرض نتائج هاتين الطريقتين بيانياً في فراغ ذي بعدين في شكل بياني يعرف باسم الديندوجرام Dendogram، كما في شكل (1).



شكل (1): طرق التجميع الهرمية

المصدر: عبد الله (2010)

الدمج Amalgamation :

مثلما تتطلب عملية التجميع تحديد مقياس للمسافة نحتاج أيضا لتحديد طريقة لدمج العناقيد. وهناك عدة طرق للدمج هي طريقة الربط المفرد وطريقة الربط الكامل وطريقة الربط المتوسط والطريقة المركزية والوسيط والطريقة الهرمية والتي يتم تناولها بإيجاز كما يلي:

- **طريقة الربط المفردSingle Linkage Method :**

تُعَدُ هذه الطريقة من أبسط طرق التحليل العنقودى ومن أقدمها حيث يمكن استعمالها مع كل مقاييس التشابه ومقاييس المسافات وتعتبر واحدة من أسهل طرق التحليل العنقودي (الشكرجي،2008) والربط المفرد يعرف بعدة تعريفات منها الجار الأقرب ،طريقة الحد الأدنى توظف أقرب الجارين لدراسة أو قياس الاختلاف بين مجموعتين (جونسون،1998)

في هذه الحالة يمكن أن نفترض أن a هي أقرب وحدة في المجموعة الأولى إلى المجموعة الثانية و b هي أقرب وحدة في المجموعة الثانية إلى المجموعة الاولى وعليه يتم اعتماد المسافة بين b، a لقياس الاختلاف بين المجموعتين وذلك يسمى الربط المفرد، والصيغة الرياضية لهذه الطريقة تكون كما يلي:

افرض أن لدينا C_i U C_j&C_kعبارة عن ثلاث مجموعات فالمسافة D بين C_kو C_i U C_jيمكن الحصول عليها من صيغة لانس – ويليامز Lance – Williams كالأتي(مارتن وليز ،واخرون،2007):

$$D(C_k, C_i \cup C_j) = min\{D(C_k, C_i), D(C_k, C_j)\} \dots\dots\dots\dots\dots\dots (1)$$

- **طريقة الربط الكامل Complete Link Method**

هذه الطريقة أكثر حذراً إذ تعتبر المسافة بين أي مجموعتين هي المسافة بين أبعد وحدتين فيهما (جونسون،1998)، فلنفرض أن a هي أبعد وحدة في المجموعة الأولي من المجموعة الثانية و b هي أبعد وحدة في المجموعة الثانية من المجموعة الأولي، عليه يتم استخدام هاتين الوحدتين لقياس الاختلاف بين المجموعتين. والصيغة الرياضية لهذه الطريقة تكون كما يلي:

افرض أن لدينا C_i U C_j&C_kعبارة عن ثلاث مجموعات فالمسافة D بينC_kو C_i U C_jيمكن الحصول عليها من صيغة لانس–ويليامز Lance– Williams كالاتي(مارتن وليز وآخرون،2007):

$$D(C_k, C_i \cup C_j) = max\{D(C_k, C_i), D(C_k, C_j)\} \dots\dots\dots\dots\dots\dots (2)$$

- **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،(جونسون،1998)، في هذه الحالة إذا افترضنا (A,B,C) تنتمي للمجموعة الأولى و (D,E,F) تنتمي إلى المجموعة الثانية حيث يتم حساب المسافات بين الوحدات في المجموعة الأولى مع المجموعة الثانية لإيجاد متوسط المسافة لاستخدامها في قياس الاختلاف بين المجموعتين، والصيغة الرياضية لهذه الطريقة تكون كما يلي:

افرض أن لدينا C_k&C_j,C_iعبارة عن ثلاث مجموعات فالمسافة D بينC_kو C_j U C_i,يمكن الحصـول عليهـا مـن صـيغة لانـس–ويليـامز Williams –Lance كالاتي(مـارتن وليـز وآخرون،2007):

$$D(C_k, C_i \cup C_j) = \frac{|C_i|}{|C_i| + |C_j|} D(C_k, C_i) + \frac{|C_j|}{|C_i| + |C_j|} D(C_k, C_j) \dots\dots\dots\dots\dots (3)$$

افـرض ان C, C' عبـارة عـن مجمـوعتين غيـر خـاليتين عليـه يمكـن ايجـاد المسـافة بطريقـة الـربط المتوسط كالاتي:

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

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

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

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

هذا يعني أن:

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

ومن المعادلات نستنتج :

$$D(C_k, C_i \cup C_g)$$
$$= \frac{n_2}{n_{2+n_3}} D(C_1, C_2) + \frac{n_3}{n_{2+n_3}} D(C_1, C_3)$$
$$= \frac{n_2}{n_{2+n_3}} \cdot \frac{1}{n_1 n_2} \sum (C_1, C_2) + \frac{n_3}{n_{2+n_3}} \cdot \frac{1}{n_1 n_2} \sum (C_1, C_3)$$
$$= \frac{1}{n_1 (n_2 + n_3)} \sum (C_1, C_2 \cup C_3) \dots\dots\dots\dots\dots\dots\dots\dots (4)$$

بعد ذلك:

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

- **طريقة المركز centroid**

وتتلخص هذه الطريقة بحساب المتوسط العام عن طريق جمع حاصل ضرب متوسط كل مجموعة بعدد مفرداتها وقسمتها على عدد المفردات الكلي فلنفرض أن A هي المجموعة الاولي و B هي المجموعة الثانية.

$$D(A,B) = d(\overline{y_A}, \overline{y_B})$$

$$\bar{Y}_A = \sum_{I=1}^{NA} \frac{Y_i}{n_A}$$

$$\bar{Y}_{AB} = \frac{n_A \bar{Y}_A + n_B \bar{Y}_B}{n_A + n_B}$$

- **الوسيط *Median*** (رشيد وآخرون ،2011)

تستخدم هذه الطريقة في حالة كون عدد مفردات أحد العناقيد أكبر من الأخرى وفي هذه الحالة عند استخدام طريقة الربط المركزية، فإن مركز العنقود الجديد يميل الى العنقود ذي المفردات الأكبر ولتفادي هذه المشكلة نستخدم الوسيط بدلاً من الوسط الحسابي الموزون لحساب مركز العنقود الجديد وفق الصيغة الآتية:

$$M_{AB} = \frac{(\bar{Y}_A + \bar{Y}_B)}{2}$$

وعندها يتم الربط بين أي عنقودين لهما أصغر مسافة بين وسطيهما في كل مرحلة.

- **الطريقة الهرمية Ward Method** (رشيد وآخرون ،2011):

تعتمد على استخدام مربع المسافات داخل كل عنقود ومربع المسافات بين العناقيد والمعبر عنهما بالصيغ الاتية لكون أن AB هو العنقود الناتج من ربط العنقودين A وB:

$$SSE_A = \sum_{i=1}^{n_A} (Y_i - \bar{Y}_A)'(Y_i - \bar{Y}_A)$$

$$SSE_B = \sum_{i=1}^{n_B} (Y_i - \bar{Y}_B)'(Y_i - \bar{Y}_B)$$

$$SSE_{AB} = \sum_{i=1}^{n_{AB}} (Y_i - \bar{Y}_{AB})'(Y_i - \bar{Y}_{AB})$$

ويتم ربط أي عنقودين بحيث يقلل الزيادة في مربع المسافات (SSE) ويعبر عن مقدار تلك الزيادة كالاتي:

$$I_{AB} = SSE_{AB} - (SSE_A + SSE_B) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(5)$$

و لإيجاد التشابه أو عدم التشابه بين كل زوج من العناصر في مجموعة البيانات، فلابد من حساب مقياس المسافة والذي يعتبر من أحد خطوات التحليل العنقودي الهرمي ،إن حساب التشابه بين العناصر يمكن أن يقاس بمقدار الاختلاف بين العناصر وذلك حسب العلاقة التالية:

التشابه = (1 − عدم التشابه)

يـتم المقارنـة بـين مؤشـرات الاخـتلاف للبيانـات ومـن ثـم اسـتخدام ارتباطـات الترتيـب (Rankindex) بـين مؤشـرات المسـافة (Euclidean, Gower, Manhattan ,Brary, Kulkulas).

التحليل العنقودي غير الهرمي Nonhierarchical Cluster Analysis :

تبـدأ عمليـة تصـنيف التحليـل العنقـودي غيـر الهرمـي بتجزئـة مجموعـة مـن البيانـات إلـى عناقيد جزئية وذلك بتحديد فئة من هذه العناقيد غير المتداخلة التي تحتوي على علاقات غير هرمية فيما بينهما.

لا تملك طرائق العنقدة غير الهرمية أشكال تشبه الشجرة حيث تشكل العناقيد باتباع العنقدة التجزيئية (partitional clustering) للحصول على تجزئة مفردة لمجموعة من العناصر إلى عناقيد بالاعتماد على الأمثلة التكرارية للدالة المعيارية (دالة الهدف) التي تعكس القبول او التوافق بين عناصر البيانات وعملية تجزئتها.

تعتمد طرائق التجزئة التي تستخدم مربع الخطأ (Square error) الى تحديد عدد العناقيد (K) وتمثيل كل عنقود بنموذج أولي، ومن ثم العمل على تصغير دالة الهدف التي تمثل مجموع كل عناصر البيانات لمربع المسافة بين العناصر والنموذج الأولي للعنقود المحدد لها ، وغالبا ما تمثل النماذج الأولية هذه مراكز العناقيد(Takane,1995; Bentler,2004). ومن طرائق التحليل العنقودي غير الهرمي خوارزمية (K– Means)، وخوارزمية (K–Medoid) وخوارزمية (Isodata).

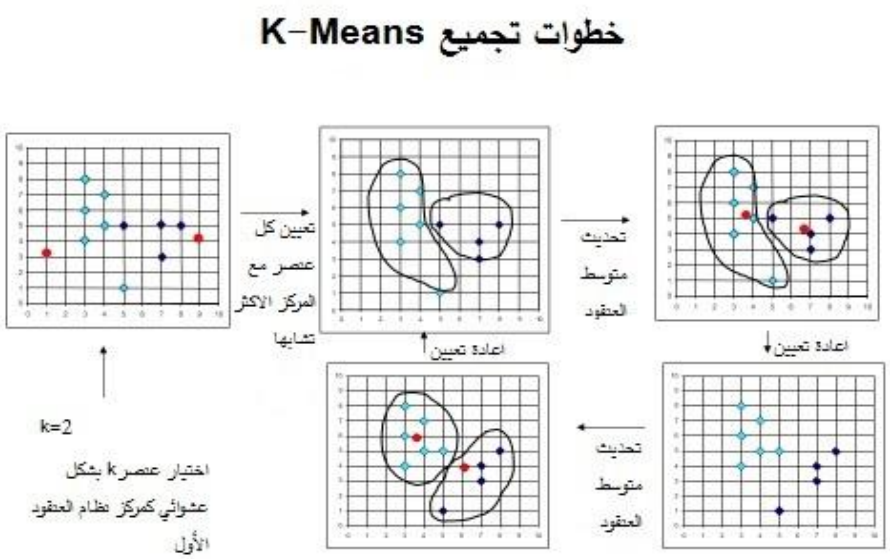
سنتطرق في هذا التحليل الى خوارزمية (K –Mediod،K–Means) في التطبيق للمقارنة مع التحليل العنقودي الهرمي.

طريقة خوارزميةk means من المتوسطات k–means Clustering Method

أول من استخدم مصطلح ال K–Means كان ل(MacQueen,J) سنة 1967 وتعد K–mean واحدة من أكثر خوارزميات التجميع شيوعًا(MacQueen,1967)، وتصنف خوارزمية K –Means العناصر إلى عدد محدد مسبقا من العناقيد وهو K عنقود وتتم عملية اختيار المراكز العنقودية في هذه الخوارزمية بشكل عشوائي،ويفضل أن تكون هذه المراكز بعيدة عن بعضها البعض قدر الإمكان،تؤثر نقطة البدء العشوائية على فعالية عملية التجميع و النتائج، وتعتمد عملية المقاربة المعنقدة على قيم المراكز الأولية بشكل رئيسي(L.Rousseeuw., 2010; Dunham,2003).

تهدف طريقة (k –means) إلى تجميع العناصر المتماثلة مع بعضها البعض اعتمادا على خصائصها في k عنقود، وتتم عملية العنقدة من خلال تقليل المسافات بين العناصر ومركز العنقود(shen,2007).

وتقوم طريقة (k –means) على أساس تصنيف الحالات (المفردات) في مجموعات متجانسة من حيث الخصائص والصفات، وذلك باستخدام خوارزميات يمكنها معالجة عدد كبير من الحالات، وتسمى هذه الطريقة بطريقة التحليل العنقودي السريع لأنها تقوم بعملية التحليل والتصنيف في وقت قصير نسبيا،بفرض أن $X = (X_{1p}, X_{2p}, \ldots, X_{np})$تمثل مصفوفة مجموعة من البيانات التي يتم تصنيفها إلى K من العناقيد وتمثل هذه العناقيد موجهات (nx1) ممكن اختيارها عشوائياً من عناصر البيانات الكلية أو تمثل أول العينات K أو تمثل العينات المستندة على بعض المعلومات السابقة. ويتم تثبيت معيار التوقف (termination toleranceϵ) لطريقة العنقدة مثلاً 0.01 و 0.001وهكذا، كما في شكل (2) يبين خطوات (k –Means)(Kamber and Han.,2001) .



شكل (2)

المصدر :(Kamber and Han.,2001)

حساب عدد العناقيد في K–Means :

1- تحويل البيانات الموجودة في المتغيرات إلى قيم معيارية إذا كانت المتغيرات مقاسة بوحدات مختلفة ،توليد مصفوفة التجزئة $\lfloor u_{ij} \rfloor$=U لكل (i=1,2,…..,k ,j= 1,….,n) توليداً قياسياً منتظماً.

2- تحديد عدد العناقيد المطلوب إجراء التصنيف على أساسه، واختيار مراكز العناقيد.

3- ربط العناصر بالعناقيد الأكثر تشابهاً والتي تكون أقرب مركز متوسط لقيم العنقود مع إعادة حساب متوسط قيم العنقود المستقبل للعنصر الجديد والعنقود الذي فقد العنصر، وذلك باستخدام مقياس المسافة الملائم ،بحيث يتم حساب المسافة بين نقطة التقاء كل زوج من العناصر.

4- إعادة تحديد أعضاء العنقود $\lfloor u_{ij} \rfloor$ =U لتصغير مربع الخطأ بين عناصر البيانات ومراكز العناقيد الحالية.

5- إعادة تكرار الخطوات السابقة حتى الوصول إلى التقارب المتمثل بحالة عدم التغير في مراكز العناقيد.

طريقة عنقدة K_ MedoidsK_Medoids Clustering Method

تعمل طريقة (K_Medoid) المقترحة من قبل(Kaufman –Rousseeuw) عام1987 على إيجاد مجموعة غير متداخلة من العناقيد من خلال تجزئة فضاء المسافة الى K من العناقيد، بحيث يملك كل عنقود عناصر التمثيل أو العناصر النموذجية التي يتم اختيارها وتعيينها مركزيا من مجموعة البيانات، والتي يطلق عليها ب(Medoids)،وأكثر طريقة شائعة هي طريقة التجزئة حول Medoids (Partitioning Around Medoids Algorithm) أو ما يطلق عليها اختصارا (PAM) وتستخدم عندما تكون البيانات كبيرة جدا(Takane,1995).

تتمثل طريقة (K –Medoids) بتجزئة مجموعة البيانات من X الى K من العناقيد، وهي تشبه طريقة (K –Means) من حيث استخدام المدخلات والمخرجات الأساسية ،والاختلاف الرئيسي بينهما يتمثل في حساب مراكز العناقيد ، حيث يمثل مركز العنقود الجديد اقرب نقطة بيانات (عنصر) الى متوسط عناصر العناقيد (البابا،2014).

صلاحية العنقود Cluster Validation:

تستخدم لمقارنة أداء أساليب العنقدة والتي يمكن استخدامها لمقارنة خوارزميات التجميع المتعددة في وقت واحد ،ويوجد عدة مقاييس لقياس صلاحية العنقود منها التحقق الداخلي والتحقق من الاستقرار.

الجانب التطبيقي:

تم الحصول على بيانات هذا البحث من الجهاز المركزي للإحصاء الفلسطيني 2018 (مسح شامل للأفراد القوى العاملة في فلسطين 2018). في هذا الجزء من البحث سيتم تطبيق أسلوبي التحليل العنقودي الهرمي وغير الهرمي على البيانات التي تم وصفها في هذا البحث حسب المحافظات الفلسطينية الستة عشر وبهدف إيجاد تجمعات من المحافظات تكون متجانسة فيما بينها من أجل التصنيف والمقارنة في آن واحد للقوى العاملة بما يحقق المتطلبات الوطنية الراهنة والمستقبلية. وعلى هذا الأساس تم تصنيف 22 متغير تحتوي على متغيرات وصفية ومتغيرات كمية منها (العمر، الحالة التعليمية، بلد التخرج، نوع الشغل، الجنس، المهنة) حسب المحافظات الفلسطينية والبالغ عددها 16 محافظة حتى العام 2018 وقد بلغ حجم أفراد القوى العاملة 90994 فرد منهم 45754 ذكر و45240 أنثى، كما في جدول (1) يوضح بعض الاحصاءات الوصفية لمتغيرات البحث الكمية المستخدمة في هذا البحث، وفي جدول (2) يوضح بعض الاحصاءات الوصفية لمتغيرات البحث المستخدمة في هذا البحث.

جدول(1): الاحصاءات الوصفية لمتغيرات البحث الكمية

	معدل الأجر اليومي بالشيكل	العمر	عدد السنوات الدراسية
المتوسط	118.889	31.95	9.92
الوسيط	100.000	27.00	10.00
الانحراف المعياري	85.6193	17.825	4.064
التباين	7330.663	317.730	16.519
الحد الادنى	1.7	10	0
الحد الاعلى	1658.4	98	31

جدول(2): الاحصاءات الوصفية لمتغيرات البحث الوصفية

		التكرار	النسبة
الالتحاق بالنظام التعليمي	ملتحق حاليا بالنظام التعليمي	28786	31.6
	التحق بالتعليم وترك	30364	33.4
	التحق بالتعليم وتخرج	29107	32.0
	لم يلتحق بالتعليم ابدا	2737	3.0
	المجموع	90994	100.0
الحالة التعليمية	أمي	2757	3.03

	ملم	10438	11.47
	ابتدائي	18569	20.41
	اعدادي	28693	31.53
	ثانوي	15804	17.37
	دبلوم متوسط	4187	4.60
	بكالوريوس	9732	10.70
	دبلوم عالي	55	0.06
	ماجستير	627	0.69
	دكتوراه	132	0.15
	المجموع	90994	100.0
حالة اللجوء	لاجئ مسجل	39819	43.8
	لاجئ غير مسجل	262	0.3
	ليس لاجئ	50913	56.0
	المجموع	90994	100.0
المحافظة	جنين	6302	6.9
	طوباس	3611	4.0
	طولكرم	4409	4.8
	نابلس	7248	8.0
	قلقيلية	3740	4.1
	سلفيت	2470	2.7
	رام الله	2622	2.9
	اريحا	2795	3.1
	القدس	5935	6.5
	بيت لحم	5233	5.8
	الخليل	10130	11.1
	شمال غزة	7415	8.1
	مدينة غزة	9967	11.0
	دير البلح	6557	7.2
	خان يونس	6928	7.6
	رفح	5632	6.2
	المجموع	90994	100.0

استخدام التحليل العنقودي الهرمي وغير الهرمي في تصنيف القوى العاملة في فلسطين

مكان العمل	الضفة الغربية	15950	17.5
	قطاع غزة	8907	9.8
	إسرائيل والمستوطنات	3900	4.3
	خارج البلاد	78	0.086
	المجموع	28835	31.7

التحليل العنقودي الهرمي Hierarchical Clustering Analysis:

يجب أن يكون لمؤشر المسافة الجيد للتحليل العنقودي تشابه ترتيب مرتفع ،كما هو موضح في جدول (3).

جدول (3): مقاييس الاختلاف *Dissimilarity indices*

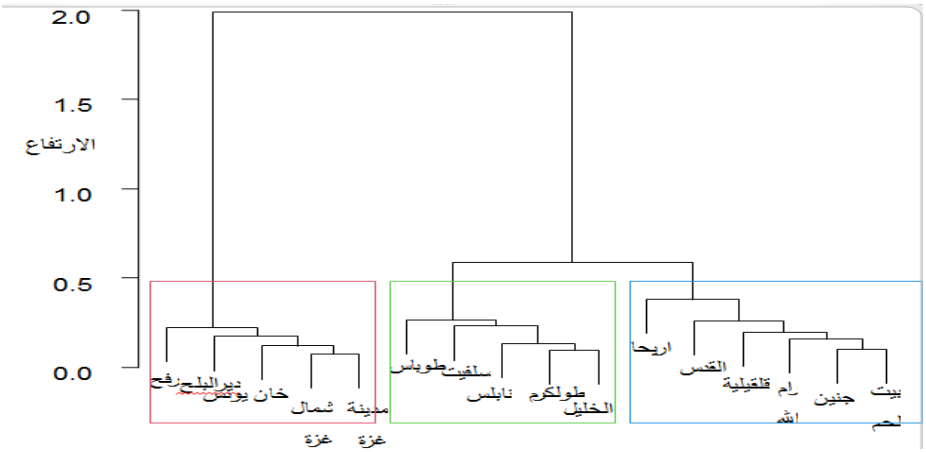
Distance	Manhattan	Euclidean	Gower	Brary	Kulkulas
Rank index Coefficients	0.8754697	0.8139955	0.9656439	0.4194218	0.5581504

ومن خلال جدول (3) يعتبر مقياس المسافة Gower من أفضل مقاييس الاختلاف.

ويستخدم مؤشر Gower لمعالجة بيانات المتحولات النوعية أو المختلطة دون تحويلها إلى متحولات ثنائية.

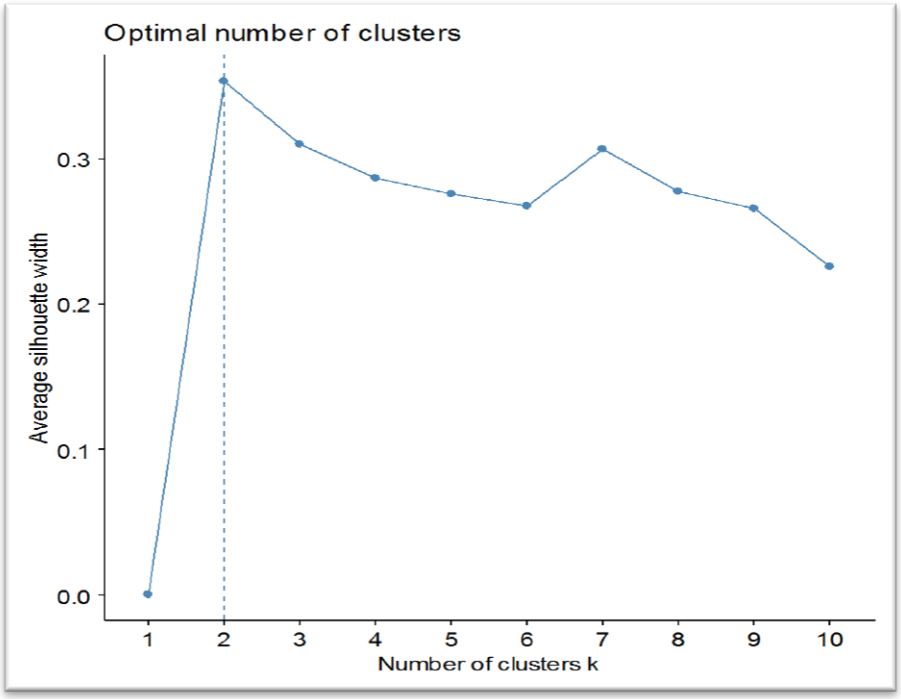
حساب عدد العناقيد في التحليل العنقودي الهرمي:

يعد تحديد العدد الأمثل للعناقيد في مجموعة البيانات قضية أساسية في تقسيم التكتل ،والذي يتطلب على الباحث تحديد عدد العناقيد k المراد إنشائها ، ويعتمد على الطريقة المستخدمة لقياس أوجه التشابه والمعلمات المستخدمة في التقسيم.



شكل (3): طريقة الربط الهرمي للمحافظات الفلسطينية

نلاحظ من شكل (3) أن كل عنصر من العناصر يتجمع مع نفسه لذا لدينا 16 مشاهدة اذا انتقلنا خطوات من الأسفل إلى الأعلى ،يمكننا أن نلاحظ أن (جنين) و (بيت لحم) في مجموعة واحدة بارتفاع = 0.1 حسب برنامج R، المجموعة الأولى تتكتل في عناقيد أخرى بارتفاع =0.2،في المجموعة الثانية نلاحظ أن نابلس موجودة في مجموعة واحدة ،و(الخليل) و(طولكرم) في مجموعة أخرى ،تتجمع هاتان المجموعتان في عناقيد أخرى بارتفاع =0.2 ، وهكذا في المجموعة الثالثة ترتبط العناصر مع بعضها، لتصبح في عنقود واحد ،يتم استخدام شكل(3) لمعرفة عدد المشاهدات في كل عنقود ،نرغب في حل حيث لا يوجد الكثير من العناقيد مع عدد قليل من المشاهدات ،لأنه قد يجعل من الصعب تفسير نتائجنا ،يبدو التوزيع بين العناقيد جيدًا ، وهناك عدة طرق لتحديد العدد الأمثل للعناقيد سنأخذ طريقة واحدة وهي متوسط الصورة الظلية (Average silhouette) :



شكل (4):متوسط الصورة الظلية لتحديد العدد الأمثل للعناقيد

تقيس هذه الطريقة جودة التكتل أي أنها تحدد مدى جودة كل عنصر داخل مجموعته. يشير ارتفاع (Average silhouette) إلى تعنقد جيد.

يحسب أسلوب متوسط الصورة الظلية (Average silhouette) من المشاهدات لقيم مختلفة من عدد العناقيد K ،العدد الأمثل للعناقيد هو 2 باعتباره موقع الحد الاقصى ،كما هو

موضح في شكل(4)وبذلك يكون العنقود الذي يزيد من متوسط الصورة الظلية(Average silhouette)عبر مجموعة من القيم المحتملة لـ([1990] Kaufman and Rousseeuw) k . ويمكن حسابها على النحو التالي:

1. حساب خوارزمية التجميع (مثل k-Means) لقيم k المختلفة. على سبيل المثال ، من خلال تغيير k من 1 إلى 10 عناقيد.
2. لكل k ،يتم القيام بحساب متوسط الصورة الظلية(Average silhouette) للمشاهدات.
3. رسم منحنى متوسط الحجم حسب عدد العناقيدk.
4. يعتبر موقع الحد الأقصى هو العدد المناسب للعناقيد.

جدول التجميع Agglomeration Schedule:

جدول (4) يمثل جدول التجميع و يوضح المراحل السابقة واللاحقة لربط المفردات والمتغيرات المتقاربة حسب المسافات بينهما.

جدول (4): جدول التجميع للعناقيد

جدول التجميع						
المرحلة	العناقيد المجمعة		المعاملات	الظهور الاولي للعنقود		المرحلة التالية
	العنقود 1	العنقود 2		العنقود 1	العنقود 2	
1	(شمال غزة)12	(مدينة غزة)13	.488	0	0	2
2	(شمال غزة)12	(دير البلح)14	.958	1	0	5
3	(طولكرم)3	(الخليل)11	1.428	0	0	4
4	(طولكرم)3	(سلفيت)6	1.857	3	0	8
5	(شمال غزة)12	(خانيونس)15	2.259	2	0	13
6	(رام الله)7	(القدس)9	2.632	0	0	11
7	(طوباس)2	(نابلس)4	3.000	0	0	8
8	(طوباس)2	0طولكرم)3	3.326	7	4	14
9	(قلقيلية)5	(رفح)16	3.594	0	0	13
10	(اريحا)8	(بيت لحم)10	3.823	0	0	12
11	(جنين)1	(رام الله)7	3.998	0	6	12
12	(حنين)1	(اريحا)8	3.711	11	10	14
13	(قلقيلية)5	(شمال غزة)12	3.401	9	5	15
14	(جنين)1	(طوباس)2	2.210	12	8	15
15	(جنين)1	(قلقيلية)5	-.473-	14	13	0

يعرض جدول(4) خطوات التجميع فيمكن تحديد المفردات أو المجموعات التي يتم ربطها معا في كل خطوة من خطوات التحليل ،الذي يعرض الكائنات أو العناقيد مجتمعة في كل مرحلة (العمود الثاني والثالث) المسافات التي يحدث عندها هذا الدمج،على سبيل المثال ،في المرحلة الأولى يتم دمج محافظتي شمال غزة ومدينة غزة على مسافة 0.488 بشكل تصاعدًي ،المجموعة الناتجة هي تمت تسميتها كما هو مشار إليها بواسطة أول محافظة متضمنة في هذا الدمج وهي المحافظة رقم 12 (شمال غزة) وهكذا إلى آخر الجدول ،ومن الجدول(4) تم الاعتماد على أكبر قيمة لمعامل الارتباط(Coefficient) بين المجاميع وقد بلغت 3.998 في المرحلة 11 والتي تشمل المجموعة الأولى (جنين) والمجموعة السابعة (رام الله)حيث العلاقة بينهم طرديه وقوية ،وأقل علاقة في المرحلة 15 بين المجموعة الأولى(جنين)والمجموعة الخامسة(قلقيلية) وقد بلغت(0.473 −) وهي علاقة ضعيفة.

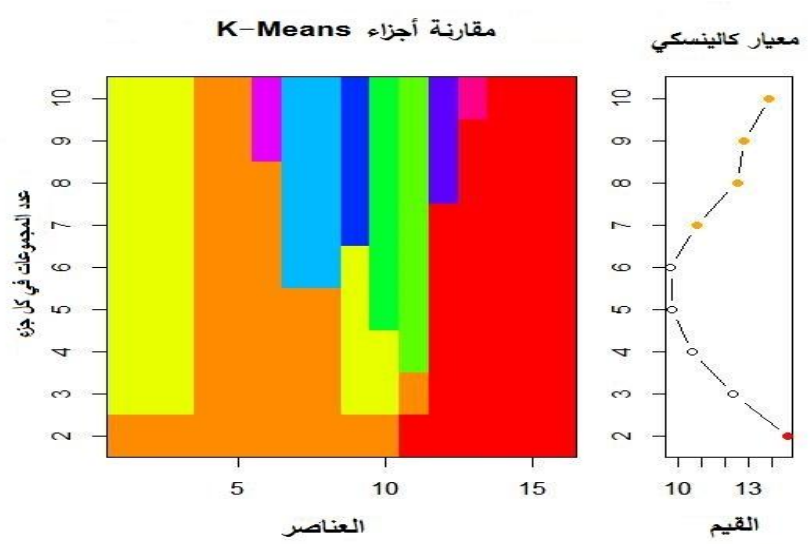
التحليل العنقودي غير الهرمي:

خوارزمية k−Means :

تعتبر هذه الطريقة من الطرق البسيطة للغاية حيث تتكون من خطوتين ،يتم تهيئة البيانات عن طريق اختيار عشوائي لمراكز العنقود ، على سبيل المثال اختيار عشوائي للعناصر في مجموعة البيانات أو القيم العشوائية داخل النطاق لكل متغير ، ثم يتم تكرار الخطوتين التاليتين:

1. حساب مسافة العنصر إلى جميع مراكز العنقود وتعيين العنصر إلى أقرب مركز ثم قم بذلك لجميع العناصر.
2. استبدال مراكز العنقود عن طريق جميع العناصر المخصصة لها.

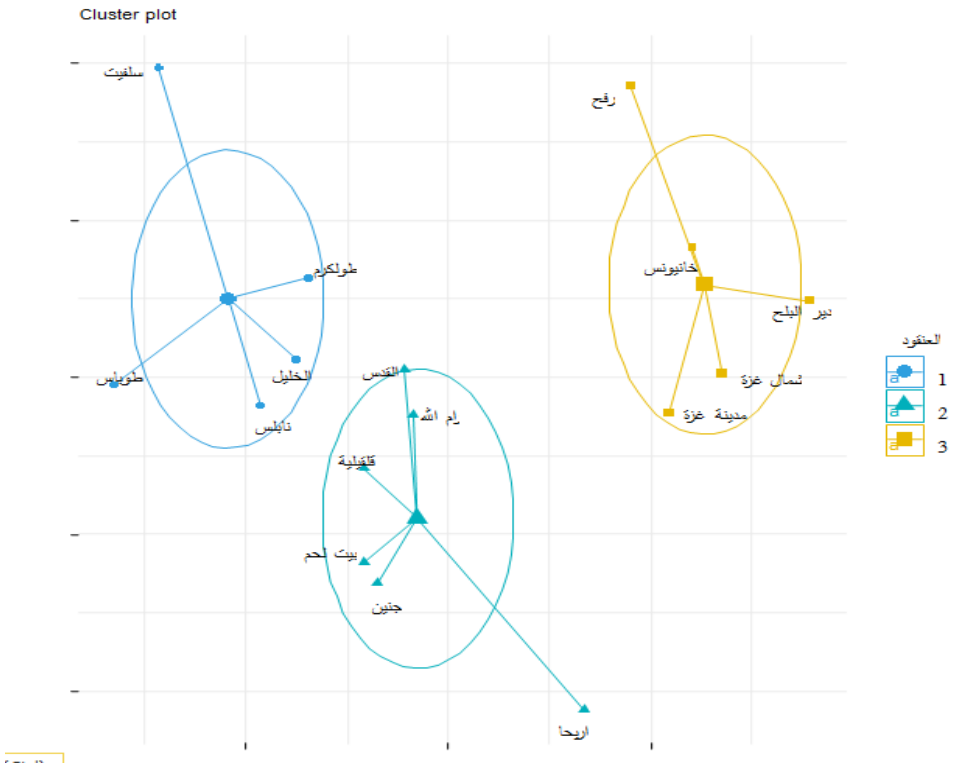
يوجد عدة معايير لحساب عدد العناقيد لخوارزمية k−Means سنستخدم أحد المعايير لحساب عدد العناقيد وهو معيار كالينكسي Calinsky(CH) لتشخيص عدد العناقيد التي تتناسب مع البيانات.



شكل (5): طريقة معيار كالينكسي

تمثل عدد العناقيد التي تتناسب مع بيانات المحافظات ،حيث يوجد عدة ألوان لهذا الشكل والتي تمثل مراكز العناقيد ومن شكل(5) نلاحظ أن النقطة الحمراء تمثل العدد الصحيح للعناقيد وهي 2 كما هو موضح في شكل(5).

يتضح من خلال شكل (6) أن المخطط **(CLUSPLOT)** يحتوي على ثلاثة عناقيد وأن العناقيد الثلاثة تحتوي على 5 و 5 و 6 محافظة، ونلاحظ هناك خطوط خارج الشكل الدائري في المجموعة الأولى والثانية والثالثة ،وهذا يدل على بعدها من مركز العنقود يمكن تصنيفها محافظات شاذة لا تتمتع بنفس خصائص المحافظات الاخرى القريبة من مركز العنقود .

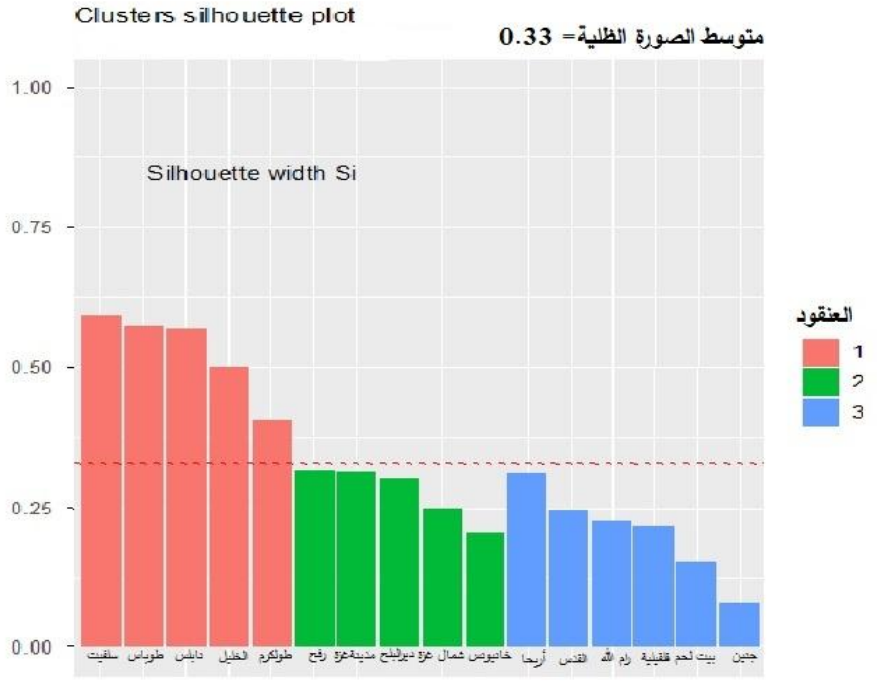


شكل (6):مخطط *Clust plot* العنقودي

ومن جدول(5) وشكل (7) يتضح أنه تم تقسيم المحافظات إلى ثلاثة عناقيد حيث إن العنقود الأول يضم 5 محافظات ومتوسط عرض الصورة الظلية لهذا العنقود يساوي0.53 والعنقود الثاني يضم 5 محافظات ومتوسط عرض الصورة الظلية لهذا العنقود يساوي0.27 والعنقود الثالث يضم 6 محافظات ومتوسط عرض الصورة الظلية لهذا العنقود يساوي0.20،يتضح من خلال شكل(7) وجدول(5) أن العنقود الأول متجمع جيدًا مقارنة مع العناقيد الأخرى، أما العنقود الثالث يعتبر أضعف متجمع مقارنة مع العناقيد الأخرى، لأنه كلما كان متوسط الصورة الظلية للعنقود قريب من الواحد يعتبر عنقود متجمع جيدا واذا كان قريب من −1 يعتبر عنقود متجمع ضعيف .

جدول (5): متوسط عرض الصورة الظلية للعنقود

Cluster العنقود	Size الحجم	sil.width. ave متوسط عرض الصورة الظلية
1	5	0.53
2	5	0.27
3	6	0.20



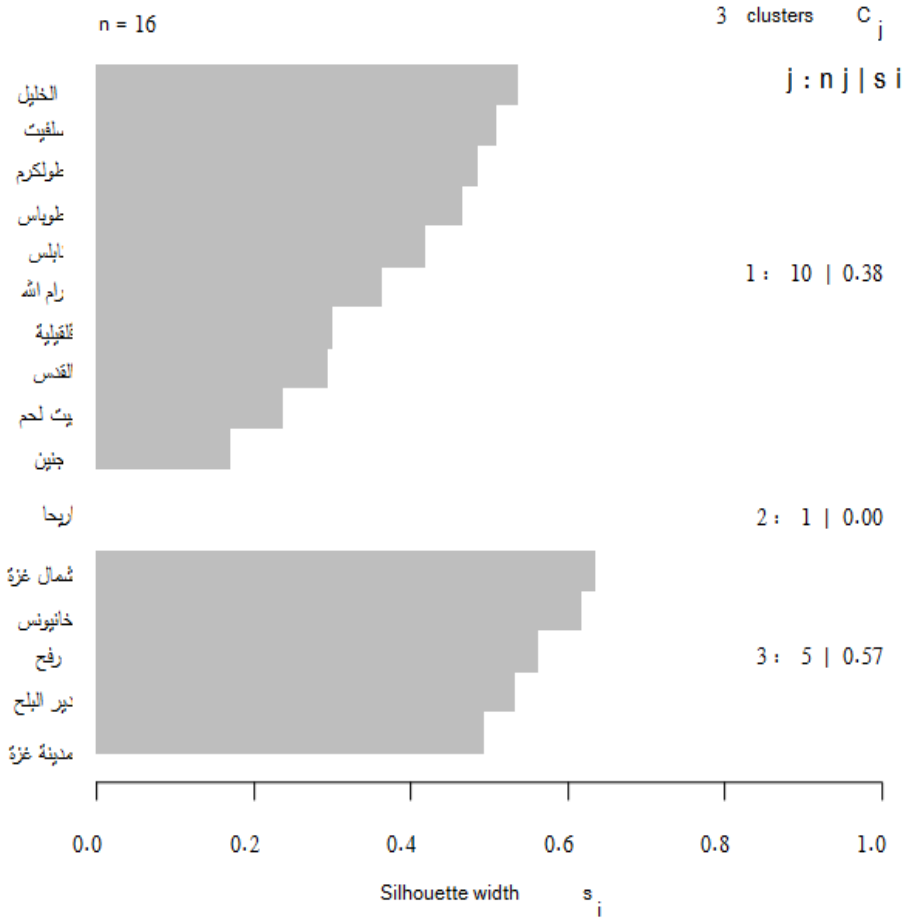
شكل (7):متوسط الصورة الظلية للعناقيد ***Clusters silhouette plot***

حيث أظهر شكل (7) المتوسط الكلي لعرض الصورة الظلية (**Average silhouette width)** لمجموعة البيانات هذه يساوي 0.33 وهذا يدلل على أن المتوسط الكلي لعرض الصورة الظلية متجمع ضعيف نوعا ما .

خوارزمية k-Medoid:

من خلال شكل(8) وجدول (6) يتضح أنه تم تقسيم المحافظات إلى ثلاثة عناقيد حيث ضم العنقود الأول عشر محافظات ومتوسط عرض الصورة الظلية لهذا العنقود يساوي 0.38 والعنقود الثاني ضم محافظة واحدة ومتوسط عرض الصورة الظلية لهذا العنقود يساوي 0 والعنقود الثالث ضم خمس

محافظات ومتوسط عرض الصورة الظلية لهذا العنقود يساوي 0.57 حيث تُظهر رسم الصورة الظلية أن عدد المحافظات هو ست عشرة محافظة وأن عدد العناقيد ثلاثة، ومتوسط عرض الصورة الظلية الكلي (Average silhouette width) لمجموعة البيانات هذه تساوي 0.41 ، وهذا يدل على أن المتوسط الكلي لعرض الصورة الظلية متجمع جيد نوعا ما.



شكل (8):متوسط عرض الصورة الظلية للعناقيد حسب خوارزمية PAM

جدول (6): متوسط عرض الصورة الظلية للعنقود

Cluster العنقود	Size الحجم	ave .sil.width متوسط عرض الصورة الظلية
1	10	0.38
2	1	0.0
3	5	0.57

التحقق الداخلي Internal Validation:

وهو طريقة تستخدم لتقييم نتائج العنقدة من حيث الكميات التي تم الحصول عليها من مجموعة البيانات نفسها حيث يقيس جودة التعنقد.

سنستخدم بعض المقاييس لقياس صلاحية العنقود الداخلي وهي:

- Connectivity الاتصال
- Silhouette Width الصورة الظلية
- Dunn Index مؤشر دان حيث يتم قياس هذا المؤشر $D = \frac{min.separation}{max.diameter}$

جدول (7):مقاييس التحقق لمقارنة خوارزميات التجميع

(Score) Number of clusters=3	Validation Measures مقاييس التحقق	Clustering method طرق التعنقد
8.8944	Connectivity	Hierarchical (الهرمي)
0.3487	Silhouette	
0.5331	Dunn	
8.8990	Connectivity	K means **(غير الهرمي)**
0.3490	Silhouette	
0.5335	Dunn	
15.4214	Connectivity	Pam (غير الهرمي)
0.2777	Silhouette	
0.3349	Dunn	

من خلال الجدول (7) يتضح أن التجميع الهرمي(Hierarchical) مع ثلاثة عناقيد يؤدي أفضل أداء مع (Dunn ، Silhouette، Connectivity).

وبالتالي يجب تقليل مؤشر Connectivity ،بينما يجب تكبير كل من مؤشر Dunn و Silhouette Width. وهكذا يبدو أن التجميع الهرمي يتفوق على خوارزميات التجميع الأخرى تحت كل مقاييس التحقق.

التحقق من الاستقرار(الثبات) Stability Validation:

التحقق من الاستقرار يتضمن مقاييس الاستقرار APN و AD و ADM و FOM حيث تتطلب التقليل من هذه المقاييس في كل حالة ،يتطلب التحقق من الاستقرار(الثبات) مزيدًا من الوقت أكثر من التحقق الداخلي ،حيث يلزم إعادة تجميع المجموعات لكل مجموعة بيانات مع إزالة عمود واحد من مجموعة البيانات.

والتحقق من استقرار التجميع هو إصدار خاص من التحقق الداخلي يقوم بتقييم اتساق نتيجة التجميع من خلال مقارنتها بالمجموعات التي تم الحصول عليها بعد إزالة كل عمود على حدا.

جدول (8):مقاييس التحقق من الثبات(*stab*)للعنقود

Numberof clusters=3 عدد العناقيد	Validation Measures مقاييس التحقق	Clustering method طرق التعنقد
0.1464	APN	Hierarchical (الهرمي)
2.1898	AD	
0.7070	ADM	
0.7521	FOM	
0.2196	APN	K−Means (غير الهرمي)
2.1697	AD	
0.7979	ADM	
0.7103	FOM	
0.1645	APN	Pam (غير الهرمي)
2.1040	AD	
0.6122	ADM	
0.7102	FOM	

اتضح من جدول(8) أن التجميع الهرمي(Hierarchical) يعطي أفضل أداء مع APN ،أما خوارزمية Pam(غير الهرمي)أعطت أفضل أداء مع AD و ADM و FOM ،من خلال المقاييس الاستقرار(الثبات) والداخلية تبين أن طريقة التجميع الهرمي(Hierarchical) أفضل طريقة لقياس

صلاحية العنقود حيث تكررت الأفضلية من خلال مقاييس الداخلية والثبات لهذه الطريقة من خلال البيانات المتوفرة لدينا وكانت طريقة Pam (غير الهرمي) تميزت بقوة في مقاييس الثبات Stab.

النتائج والتوصيات:

أولا: النتائج

يمكن توضيح أهم النتائج :

1. باستخدام التحليل العنقودي الهرمي التراكمي لتصنيف المحافظات الفلسطينية إلى عناقيد وبإجراء مقارنة بين الطرق الاربع لحساب مصفوفة الاختلاف:

(المسافة الإقليدية ،مسافة Gower ،مسافة Kulkulas ،مسافة Brary)،باستخدام تقنية (معامل مؤشر Rank) من أجل تحقيق أفضل طريقة لحساب المسافة ،فإن مسافة Gower هي أفضل طريقة لقياس التشابه.

2. عند اجراء مقارنة بين طرق الربط لاختيار أفضل طريقة للربط من الطرق (الربط المفرد، الربط الكامل، الربط المتوسط، الربط الوسيط، الربط المعدل، الربط الهرميward)،وباستخدام تقنية (معامل التكتل) من أجل تحقيق أفضل طريقة للتجميع يظهر أن طريقة الربط ward أفضل طريقة لهذه البيانات.
3. باستخدام طريقة الربط الهرمية(ward) تم عنقدة المحافظات الفلسطينية إلى ثلاثة عناقيد وتتكون هذه العناقيد من محافظات على النحو التالي:
   - يتكون العنقود الأولى من خمس محافظات (شمال غزة ،مدينة غزة ،دير البلح ،خان يونس ،رفح)
   - يتكون العنقود الثاني من خمس محافظات (الخليل ،طولكرم ، نابلس، سلفيت ،طوباس)
   - العنقود الثالث يتكون من ست محافظات (بيت لحم ،جنين ،رام الله ،القدس ،اريحا ،قلقيلية)
4. باستخدام طريقة التجميع (K –Means) لتصنيف المحافظات الفلسطينية إلى ثلاثة عناقيد ، والعناقيد الثلاثة تحتوي على 5 ، 5 و 6 عناصر على التوالي ،المجموعة الأولى تتكون من (شمال غزة، مدينة غزة، دير البلح، خان يونس، رفح)، والمجموعة الثانية تتكون من(طولكرم، الخليل، نابلس، طوباس، سلفيت)، أما المجموعة الثالثة تتكون من(جنين، رام الله، بيت لحم، القدس، اريحا، الخليل).
5. باستخدام طريقة التجميع (K–mediod) لتصنيف المحافظات الفلسطينية إلى ثلاثة عناقيد، والعناقيد الثلاثة تحتوي على 10، 1 و 5 عناصر، العنقود الأول يتكون من (الخليل، جنين،

طولكرم، طوباس، القدس، رام الله، سلفيت، قلقيلية، نابلس، بيت لحم)، أما العنقود الثاني يتكون من (أريحا)،والعنقود الثالث يتكون من (شمال غزة ،مدينة غزة ،دير البلح ،خان يونس ،رفح) .

6. يتضح أن أفضل طريقة تمثل البيانات هي التجميع الهرمي (Hierarchical) مع ثلاث عناقيد يؤدي أفضل أداء مع (Silhouette width، Dunn، Connectivity) في مقاييس التحقق الداخلية.
7. باستخدام مقاييس التحقق من الثبات (Stability Validation) يتضح أن أفضل طريقة تمثل البيانات كانت التجميع الهرمي مع ثلاثة عناقيد وتؤدي أفضل أداء مع (APN)، والتجميعPam مع ثلاث عناقيد يؤدي أفضل أداء مع (ADM، FOM، AD) .

ثانياً: التوصيات

وفقًا للنتائج المذكورة أعلاه، قد نوصي بما يلي:

1. استخدام التحليل العنقودي لتصنيف العديد من الظواهر الاقتصادية والاجتماعية والصحية.
2. استخدام طريقة التحليل العنقودي لتصنيف المحافظات والدول بشكل عام وفقاً لمؤشر التنمية البشرية.
3. تحديث هذا البحث في حالة تنفيذ مسح حديث لأفراد القوى العاملة في فلسطين.
4. التأكيد على ضرورة استخدام الأساليب الإحصائية المتقدمة في مثل هذه الدراسات لما لها من أهمية في الوصول إلى نتائج دقيقة تحقق أهدافاً إنسانية لبناء مجتمع أفضل.
5. دعوة الجهات التنفيذية المتخصصة (وزارة الاقتصاد والتخطيط) إلى تبني نتائج البحث التي تم التوصل إليها جاعلين منها أحد المصادر الأساسية في رسم سلم أولويات تصنيف القوى العاملة في فلسطين.

المراجع

أولاً: المراجع العربية

البابا، طه المختار(2014):تعيين المراكز الابتدائية بشكل مدروس في خوارزمية K-Medoids ،مجلة جامعة البعث-المجلد36، العدد 2، 57-80 .

الجاعوني، فريد، الغانم، عدنان. "التحليل الاحصائي متعدد المتغيرات (التحليل التجميعي) في دراسة تحديد مستويات الهيكل الاقتصادي الاجتماعي لأسر المجتمع" مجلة جامعة دمشق، المجلد السابع عشر، العدد الثاني،2001، 209-222.

الجهاز المركزي للإحصاء الفلسطيني(2018): مسح القوى العاملة الفلسطينية ،التقرير السنوي، ،رام الله ،فلسطين.

الشكرجي، ذنون يونس(2008): " استخدام التحليل العنقودي الهرمي في تصنيف المشاهدات إلى مجاميع متجانسة مع تطبيق على دوري كرة السلة ، مجلة أبحاث كلية التربية الأساسية ، السنة 7، العدد 2، 335-347.

جودة، محفوظ، 2008"، التحليل الإحصائي المتقدم باستخدام "SPSS دار وائل للنشر، الطبعة الاولى، الأردن-عمان.

جونسون، ريتشارد، وشرن ،دين (1998):التحليل الاحصائي للمتغيرات المتعددة من الوجهة التطبيقية ،تعريب .عبد المرضي حامد عزام ،دار المريخ للنشر، المملكة العربية السعودية ،1418هـ.

رشيد، أسيل ومهدي ،نبأ (2011) : تحليل واقع التربية والتعليم في العراق باستخدام طرائق التحليل العنقودي (دراسة مقارنة) ،مجلة القادسية للعلوم الإدارية والاقتصادية ،السنة 12، العدد 2، 194-217.

عزام، عبد المرضي حامد(1998) "التحليل الاحصائي للمتغيرات المتعددة من الوجهة التطبيقية " كتاب مترجم ،دار المريخ للنشر ،الرياض ،المملكة العربية السعودية.

عبد الله، فرح محمد ،(2010) :(تصنيف الولايات السودانية ذات الخصائص الديموغرافية المتشابهة باستخدام التحليل العنقودي للعام 2002 م) ،جامعة السودان للعلوم والتكنولوجيا - رسالة ماجستير.

مارتن وليز، وآخرون(2007):تجميع البيانات ،سلسلة إصدارات الجمعية الامريكية (فرجينيا، جمعية الصناعات والتطبيقات الرياضية)،الولايات المتحدة الأمريكية.

نامق، فيصل(2010) أسلوب التحليل العنقودي لتصنيف الإنفاق على السلع والخدمات الأساسية وفقا للمستوى البيئي حضر وريف للسنوات من1971_ 2007 ، مجلة كلية بغداد للعلوم الاقتصادية الجامعة، العدد25، 331-351.

ثانياً: المراجع الاجنبية

Bentler, P.M. (2004). EQS structural equations program manual. Encino, CA: Multivariate Software.

Dunham, M. H. 2003 -Data Mining :Introductory and Advanced Topics Prentice Hal Bazsalica, 328p.

Han J. ,Kamber M.(2001),Data Mining: Concepts and Techniques, Morgan Kaufmann Publishers.

Jess , shen , (2007)Using Cluster Analysis, Cluster Validation, and Consensus Clustering to Identify Subtypes of Pervasive Developmental Disorder, Canada.

Kaufman ,L. Rousseeuw,(2010)- Finding Groups in Data: an Introduction to Cluster Analysis. John,170p.

MacQueen, J. (1967). "Some methods for classification and analysis of multivariate observations", Proceeding of the 5th Symposium on Mathematical Statistics and Probability. 1: 281–297, Univ. of Calif. Press

R. Feinstein, (1996) ((Multivariable analysis: an introduction)) Yale University press- London.

S. Aldenderfer, K. Blashfield, (1984) ((Cluster analysis)) Sage Publicatios. California. U.S.A.

Takane, Y. (1995). Constrained principal component analysis. Tokyo: Asakurashoten, (in Japanese).

Zhang K. (2007), "Visual Cluster Analysis in Data Mining" Ph.D thesis in Philosophy, Department of Computing Division of Information and Communication Sciences, Macquarie University, NSW 2109, Australia