المجلة العراقية للعلوم الإحصائية (21) 2012
ص ص [141-165]

توظيف التحليل العنقودي وطريقة الجار الأقرب في التعرف على الأنماط مع تطبيق على نوعية المياه الجوفية في محافظة نينوى

د. مثنى صبحي سليمان [1] رياض محمود صالح[2]

المستخلص

إن الهدف الرئيس في التعرف على الأنماط هو تطوير قدرة الطرائق في محاكاة الأشياء المميزة الأكثر اختلافاً في الشكل، بهدف تصنيف هذه الأشياء إلى عدد من الأصنافِ أو الأنواع ، وهذه الأصناف يمكن أن تكون صور أو بيانات أو أي نوع من المقاييس التي تحتاج إلى التصنيف. تم في هذا البحث توظيف التحليل العنقودي وطريقة الجار الأقرب لعملية اتخاذ القرارات في التعرف على الأنماط اللامعلمية وذلك من خلال استخدام نظرية بيز كقاعدة قرار في التعرف على الأنماط، وتطبيقها في مجال تحديد نوعية المياه الجوفية في محافظة نينوى. وقد أعطت الطريقة المقترحة نتائج دقيقة في عملية تصنيف البيانات إلى مجموعتين: الأولى مياه الآبار الصالحة للشرب والأخرى غير الصالحة للشرب، مما أمكن دراسة خصائص وصفات كل مجموعة وتحديد صفاتها الأساسية بحيث أصبح من السهل التعرف على العينات المأخوذة من أي بئر في المحافظة لمعرفة نوعية المياه الجوفية الخاصة بالبئر قيد الدراسة.

[1] مدرس/ كلية علوم الحاسوب والرياضيات/جامعة الموصل

[2] مدرس مساعد/ مركز بحوث البيئة والسيطرة على التلوث/ جامعة الموصل

الكلمات الدالة: التعرف على الأنماط، طريقة الجار الأقرب، التحليل العنقودي، المياه الجوفية.

Employment of Cluster Analysis and Nearest Neighbor Method in Pattern Recognition With the Application on Groundwater Quality in the Province of Nineveh

ABSTRACT

The main objective of pattern recognition methods is to develop capacity in the simulation of the most special objects in a different form, in order to classify these objects to a number of classes or types, these classes can be images or data or any kind of standards that need to be rated. In this research, recruitment of and the nearest neighbor method to the decision-making process in non-parametric pattern recognition through the use of theory in the Bayse theorem as a decision rule of discrimination, and their application in the field of groundwater quality in the province of Nineveh. Has given way suggest accurate results in the process of classifying data into two groups: the first water wells for drinking and other non-potable water, which could study the characteristics and qualities of each group and to identify qualities essential so it is easy to distinguish samples from any well in the province to see the quality of groundwater for The well under study.

Keyword: pattern recognition, nearest neighbor method, cluster analysis, groundwater.

(1) مقدمة :

يعد التعرف على الأنماط *Patterns Recognition* من الأمور الحياتية اليومية المهمة التي تتميز بها الكائنات الحية لتتعايش مع البيئة المحيطة بها، وهو أحد فروع علم تعلم الآلة *Machine Learning* وبشكل عام الذكاء الاصطناعي *Artificial*

Intelligence . وتهدف البحوث والتقنيات الخاصة بهذا العلم إلى بناء أنظمة ذكية آلية *Intelligent Automated Systems* تدعى بالمصنفات *Classifiers* التي سوف تؤشر تلقائياً قياسات تلك الأصناف وهدفها تقسيم فضاء النمط *Pattern Space* إلى عدة فئات.

إن جميع القياسات الممكنة (والتي تمثل البيانات) تسمى فضاء النمط (والتي تمثل المدخلات)، والنمط هو عبارة عن وصف الشيء، أما التعرف فهو ميزة أساسية في الكائن البشري فضلاً عن الكائنات الحية الأخرى. ويعود موضوع التعرف على الأنماط إلى ما قبل الستّينات التي كانت في الغالب ناتج البحثِ النظري الإحصائي.

إن الهدف الرئيس للتعرف على الأنماط هو تطوير قدرة الطرائق في محاكاة الأشياء المميزة الأكثر اختلافاً في الشكل، بهدف تصنيف هذه الأشياء إلى عدد من الأصنافِ أو الأنواع، هذه الأصناف يمكن أن تكون صور أو بيانات أو أي نوع من المقاييس التي تحتاج إلى التصنيف. كما أن التطور العلمي والتقني في تصميم قابلية أداء الحاسبات أعطى لمهمة التعرف على الأنماط قابلية التطبيق العملي، التي تباعاً وضعت طلبات جديدة للتطورات النظرية بشكل أكبر.

(2) هدف البحث:

يلاحظ في تطبيقات التعرف على الأنماط أنها في الغالب تكون في الأنظمة الأمنية، إذ أن التعرف على الأنماط يعد عنصراً مكملاً في كثير من ألأنظمة الأمنية في عملية اتخاذ القرارات، كما في التعرف على بصمة إصبع اليد أو صورة شخص معين وتحليلها لإنتاج أوصاف شخص مطلوب، أو التعرف على بصمة العين. فضلاً عن التطبيق في القطاع الصناعي، وذلك من خلال مراقبة الإنتاج باستخدام آلة التصوير للتحقق من عيوب هذه المنتجات لرفض القطع المعيبة من خلال تحليل هذه الصور للتعرف على المنتج وتصنيفه إلى مجموعة من الفئات (مثلاً المعيبة وغير المعيبة).

إن هدف البحث هو توظيف التحليل العنقودي وطريقة الجار الأقرب لعملية اتخاذ القرارات في التعرف على الأنماط اللامعلمية وذلك من خلال استخدام نظرية بيز *Bayes Theory* كقاعدة قرار في التعرف على الأنماط, وتطبيقها في مجال تحديد نوعية المياه الجوفية في محافظة نينوى.

(3) التحليل العنقودي *Cluster Analysis*:

كانت طرائق التصنيف سابقاً تعتمد على الجهد الشخصي الكبير والخبرة الشخصية العلمية في المجال المعين (خاصة عند وفرة النماذج والمعلومات) لكي تؤدي إلى نتائج دقيقة وموثوق فيها، إلا أن اجتياح الحاسوب لمجالات العلم الواسعة جعل عملية السيطرة على المعلومات الكبيرة والواسعة أمراً سهلاً وبجهد قليل وكلفة ضئيلة .

إن التحليل العنقودي يقوم باستخدام بعض العمليات الإحصائية لتحديد كيفية إنشاء مجموعات تجمعـهم في مجـاميع يطلق عليها العنـاقيد *Clusters*، إذ يتم الجمع على أساس أدنى المسافة (أعلى التشابه) في العنقود. وسمي التحليل العنقودي بهذا الاسم لظهور نتائجه على شكل عناقيد بحيث تكون عناصر كل عنقود متشابهة نسبياً فيما بينها، أي تشابه داخل العنقود الواحد، إلا أنها مختلفة عن باقي العناقيد، أي اقل تشابهاً بين ألعناصر للعناقيد الأخرى.

على هذا الأساس يمكن تعريف التحليل العنقودي بأنه عبارة عن عملية تحليل إحصائي متعدد المتغيرات مبرمج يعتمد على حساب متغيرات متنوعة وعديدة (كأن تكون أجسام، صفات، خصائص، أحداث، ....الخ) لنماذج مختلفة وكثيرة بالاعتماد على موضوع الدراسة، ثم مقارنة تلك النماذج ببعضها اعتمادا على ما تحويه من المتغيرات، وترتيب ارتباطاتها على شكل عناقيد بحيث يعمـل علـى تصـغير التبـاين داخـل العنقـود الواحد وتعظيم تباين مـا بين العناقيـد المختلفة [بهنام، 2006] .

ويستخدم التحليل العنقودي في تصنيف فئات الأنماط غير المرئية، ولهذا نجد العنقدة تقع تحت اسم عملية التصنيف غير مشرف عليها *Unsupervised Classification* في التعرف على الأنماط. وعليه أصبح من المهم الآن هو كيفية تنظيم الأنماط إلى العناقيد المعقولة (مجموعات)، والتي ستسمح لنا لاكتشاف التشابهات والاختلافات بين الأنماط ولاشتقاق الاستنتاجات المفيدة عنهم. هذه الفكرة استخدمت في العديد من الحقول، مثل علوم الحياة والعلوم الطبية (مثل طب ألأمراض العقلية)، وعلم الاجتماع، وعلم التربة (مثل الجغرافية، وعلم طبقات أرض) [*Theodoridis and Koutroumbas*, 2006].

(1.3)الخطوات الأساسية لعملية العنقدة *Basic Steps of Clustering Process*:

في عملية العنقدة يوصف كل نمط بواسطة مجموعة من السمات *Set of Features* **والتي من الممكن أن نعبر عنها بقيم حقيقية. وفي مرحلة التعليم يقدم كل نمط كمتجه يسمى بمتجه السمات الأساسية** *Basic Features Vector*، **ويمكن إدراج الخطوات الأساسية لعملية العنقدة كما يلي** [*Theodoridis and Koutroumbas*, 2006]:

أ- اختيار السمات الأساسية *Basic Features Selection*:

يجب أن يتم اختيار السمات الأساسية بشكل صحيح بحيث تمثل قدر كبير من المعلومات المتعلقة بالنمط قيد الدراسة، أي إعداد السمات الأساسية قبل استخدامها في المراحل اللاحقة.

ب- معيار التعنقد *Clustering Criterion*:

يعتمد معيار العنقدة على الخبرة المكتسبة للخبير مستند على نوع العناقيد التي يتوقع أنها تقع تحت مجموعة من البيانات، فقد يتضمن احد العناقيد

مجموعة من متجهات ذات السمات الملائمة لمعيار معين، في حين أن عنقود آخر يعتمد على معيار يختلف عن المعيار الأول في تعنقده.

ج- خوارزميات العنقدة *Clustering Algorithms*:

بعد تبني معيار العنقدة الملائم للبيانات فإن الخطوة التالية هي اختيار مخطط الخوارزمية المناسب الذي يكتشف تركيب تجمع مجموعة المعلومات أو البيانات.

د- صحة النتائج *Validity of Results*:

إن صحة نتائج خوارزمية العنقدة تتحقق باستخدام التقنيات والمعايير المناسبة، ولأن خوارزميات العنقدة تعرف العناقيد الغير معروفة مسبقاً ، وبغض النظر عن طرائق العنقدة، لذا فان التجزئة النهائية للبيانات تتطلب بعض أنواع التقييم في بعض تطبيقات العنقدة.

ه- تفسير النتائج *Interpretation of the Results*:

في معظم الحالات، فإن الخبرة في مجال التطبيق توجب تكامل نتائج عملية العنقدة مع بقية التجارب المبرهنة والمحللة من أجل استخلاص الاستنتاجات الصحيحة.

(4) مصنف الجار الأقرب *K-Nearest Neighbour Classification* :

تعد طريقة الجار الأقرب في التصنيف من الطرائق اللامعلمية في تصنيف الأشياء أو الصفات أو البيانات باستخدام الحالات الأقرب لها (*closest*)، إذ أنها تُستخدم مع الأنماط ذات الطبيعة المستمرة *continuous*. وبالرغم من انه يمكن التعديل على خوارزمية الجار الأقرب للتعامل مع الأنماط ذات الطبيعة المتقطعة *categorical*، فإن هنالك طرائق أخرى لامعلمية لتصنيف الصفات أو الأشياء ذات الطبيعة المتقطعة منها خوارزمية

Naïve Bayes Algorithm والتي تستخدم نظرية الاحتمال *Probability Theory* لإيجاد التصنيفات المحتملة للصفات أو الأشياء ذات الطبيعة المتقطعة [*Bramer*,2007].

إن الفكرة الأساسية لطريق الجار الأقرب تكمن في تصنيف الحالات غير المرئية أو (غير المصنفة) إلى الحالات الأقرب لها ضمن حجم معين، وعليه نحتاج إلى تحديد قيمة k لتعيين الفئة التي تتلقى اكبر تصويت من بين k_i الأقرب من الجيران.

ولسهولة النهج التي تتعامل به طريقة الجار الأقرب فإنها لاقت على مر السنين اهتماما كبيراً من قبل معظم الباحثين، إذ إن خوارزمية الجار الأقرب تتطلب مجموعة من القوالب المصنفة والتي يتم استخدامها في التعرف على مجموعة من أنماط مختارة، وذلك من خلال حساب المسافة بين النمط المراد تصنيفه مع بقية أنماط مجموعة التدريب، واختيار الأنماط التي عددها k الأقرب لاتخاذ قرار. ومن المرجح أن تكون مكلفة حسابيا وذلك للحجم الكبير لمجموعة البيانات، ولكن العديد من التطبيقات قد أثبتت قبول الطريقة وذلك بسبب التطور العلمي والتقني في تصميم قابلية أداء الحاسبات [*Bramer*,2007].

(1.4) خوارزمية الجار الأقرب للتصنيف:[*Webb*,2002]

لتقدير دالة الكثافة الاحتمالية للقيمة x نحتاج إلى تحديد منطقة صغيرة حولها حجمها v، وعليه فان احتمالية وقوع x داخل المنطقة v هو:

$$\theta = \int_{v(x)} p(x)dx \qquad \ldots(1)$$

إذ أن التكامل يكون على الحجم v ، ولحجم صغير فان:

$$\theta \sim p(x)v \qquad \ldots(2)$$

إن الاحتمالية θ يمكن أن تقترب من نسبة العينات التي تقع داخل المنطقة المحددة v حول x ، فإذا كانت k تمثل عدد العينات (الحالات) المأخوذة من n من العينات داخل المنطقة v ، فان:

$$\theta \sim \frac{k}{n} \quad \ldots(3)$$

ومن الصيغتين في (2) و(3) نحصل على تقدير دالة الكثافة الاحتمالية لـ x كما يلي:

$$\hat{p}(x) = \frac{k}{nv} \quad \ldots(4)$$

إذ أن:

v: تمثل حجم المنطقة حول الحالة المدروسة x

k: تمثل عدد الحالات داخل المنطقة v

n: تمثل العدد الكلي للحالات.

(2.4) قاعدة قرار بيز *Bayesian Decision Rule* :

إن قاعد القرار المعتمدة على نظرية بيز في التعرف على الأنماط مبنية على أن الاحتمالات اللاحقة *Posterior Probability* تعتمد على الاحتمالات السابقة *Prior Probability*، على سبيل المثال، الحركة القادمة في لعبة الشطرنج مستندة على النمطِ الحالي على لوحة الشطرنج. أي أن التعرف على نمط (الحالة x) يعتمد على أعظم قيمة للاحتمالات اللاحقة $\hat{p}(\omega|x)$ للنمط ، وعلية يمكن تأشير النمط x إلى الفئة (*class*) أو الصنف ω_m إذا تحقق الشرط التالي:

$$\hat{p}(\omega_m|x) \geq \hat{p}(\omega_i|x) \quad \forall\ i \in c \quad \ldots(5)$$

إذ أن c تمثل عدد الفئات أو الأصناف، وباستخدام نظرية بيز:

$$\hat{p}(\omega_j|x) = \frac{p(x|\omega_j)\cdot p(\omega_j)}{p(x)} \quad \ldots(6)$$

نحصل على قاعدة القرار الخاصة بالتعرف على الأنماط :

$$p(x|\omega_m) \cdot p(\omega_m) > p(x|\omega_i) \cdot p(\omega_i) \quad \ldots(7)$$

إذ أن الدليل $p(x)$ حذف من طرفي المتباينة.

مما تقدم يمكن استخدام دالة الكثافة الاحتمالية في الصيغة (4) لتحديد قاعدة للقرار للتصنيف باستخدام الجار الأقرب، فلو فرضنا أن أول أقرب مجموعة من الأنماط عددها k هنالك k_m تقع في الصنف أو الفئة ω_m بحيث أن:

$$\sum_{m=1}^{c} k_m = k \quad \ldots(8)$$

ولو فرضنا أن العدد الكلي للعينات في الفئة ω_m هو n_m بحيث أن:

$$\sum_{m=1}^{c} n_m = n \quad \ldots(9)$$

وعليه يمكن تقدير دالة الكثافة الشرطية للفئة *Class-Conditional Density* كما يلي:

$$\hat{p}(x|\omega_m) = \frac{k_m}{n_m \cdot v} \quad \ldots(10)$$

أما الاحتمال الأولي *Prior Probability* والذي يمثل احتمال الفئة، فهو:

$$\hat{p}(\omega_m) = \frac{n_m}{n} \quad \ldots(11)$$

وبتعويض كل من المعادلتين (10) و(11) في قاعدة قرار بيز (7) ، نحصل على الآتي:

$$\frac{k_m}{n_m \cdot v} \frac{n_m}{n} > \frac{k_i}{n_i \cdot v} \frac{n_i}{n} \quad \forall\ i \in c \quad \ldots(12)$$

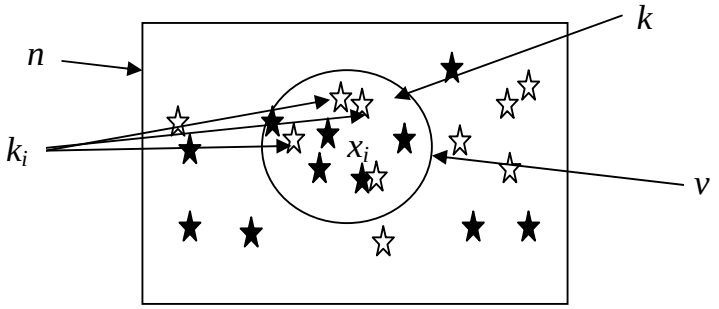
وعلية سيتم تحديد النمط x إلى الفئة (*class*) أو الصنف ω_m إذا تحقق الشرط التالي:

$$k_m > k_i \quad \forall\ i \in c \quad \ldots(13)$$

هذا يعني انه يتم التعرف على نمط حالة معينة بالاعتماد على عدد الجيران الأقرب ضمن فئة معينه، أي أن الدالة المميزة *Discriminate Function* لخوارزمية الجار الأقرب

تمثل نسبة عدد الحالات الخاصة بفئة معينه إلى عدد الحالات الكلية داخل المنطقة v ، أي:

$$h(x) = \frac{k_i}{k} \quad \forall\ i \in c \qquad \ldots(14)$$



الشكل رقم (1) : يوضح خوارزمية الجار الأقرب

(5) مقاييس التشابه والمسافة *Similarity and Distance Measures*:

[*Bramer*,2007]

تستخدم مقاييس التشابه والمسافة لتحديد الجار الأقرب للحالة المدروسة غير المرئية، فضلاً عن استخدامها في قياس التقارب والتماثل بين العناصر التي تمتلك أعلى قيمة تشابه داخل العنقود الواحد، ففي حالة تكوين العناقيد فان أزواج القيم تكون متشابه (تمتلك أقل مسافة) ضمن العنقود الواحد ومختلفة (لها مسافات كبيرة) مع قيم في عناقيد أخرى.

إن استخدام فكرة مقاييس التشابه لا يكون سهلاً في عملية العنقدة، لذا يستخدم في أغلب الأحيان بدلاً من مقاييس التشابه ، مقياس عدم التشابه *Dissimilarity* ، أو المسافة *Distance*، ويرمز للمسافة بين x_i و x_j بالرمز $dist(x_i, x_j)$.

وهنالك عدة طرائق لقياس المسافة بين عنصرين أو حالتين في n من الخواص أو الصفات، أهمها وأكثرها شيوعاً المسافة الاقليدية *Euclidean Distance* والتي نسبت للعالم الرياضي الاسكندينافي اليوناني أقليديس، وكما في الصيغة التالية:

$$dist(x_i, x_j) = \sqrt{\sum_{t=1}^{n}(x_{it} - x_{jt})^2} \qquad \ldots(15)$$

أما شروط المسافة فهي:

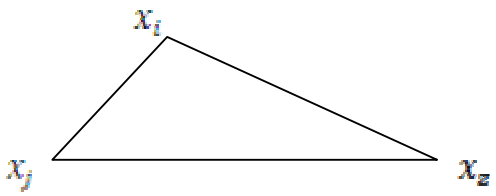
1. المسافة بين النقطة ونفسها يساوي صفر، أي أن $dist(x_i, x_i) = 0$.
2. المسافة بين النقطتين x_i و x_j هي نفسها المسافة بين النقطتين x_j و x_i أي أن:

$$dist(x_i, x_j) = dist(x_j, x_i)$$

3. المسافة الأقصر بين أي نقطتين تُمثل بخط مستقيم، أي أن:

$$dist(x_i, x_j) \leq dist(x_i, x_z) + dist(x_z, x_j)$$

ويدعى هذا الشرط بمتباينة المثلث، كما في الشكل:



الشكل رقم (2) : يوضح متباينة المثلث

(6) الجانب التطبيقي:

منذ أقدم العصور أُستخدمت المياه الجوفية كمصدر رئيسي للاستهلاك البشري, وخصوصاً في المناطق التي لا توجد فيها أو بالقرب منها مياه سطحية (سواء انهار أو بحيرات)، وأحياناً كثيرة استخدمت هذه المياه حتى في حال توفر مياه سطحية خصوصاً إذا كانت مياه الأنهار هذه ملوثة أو بنوعية أردأ من نوعية المياه الجوفية. وعرفت المياه الجوفية على أنها تلك المياه التي تخرج من باطن الأرض سواء على شكل ينابيع أو عيون أو آبار محفورة، وتعتمد نوعية المياه الجوفية بالدرجة الرئيسية على طبيعة الطبقات الجيولوجية التي تمر خلالها المياه من مناطق التغذية (*Recharge Zones*) وصولاً إلى مناطق الاستخراج.

لقد لوحظ تفاوت كبير جداً في نوعية المياه الجوفية ففي الوقت الذي تكون فيه المياه الجوفية في مناطق معينة بنوعية تضاهي نوعية المياه المعالجة بدرجة متقدمة نجد

أن هناك مناطق أخرى تكون المياه الجوفية فيها أشبه بمياه المطروحات الثقيلة (مع أنها خام *Fresh*)، ذلك بسبب وجود شوائب وعناصر محددة تجعل المياه ممنوعة الاستخدام لكل الأغراض. وقد تم ملاحظة أن هنالك تفاوت في نسب تلك الشوائب، ففي قسم من المياه الجوفية يكون تركيز النترات عالياً جداً وهو سبب رداءتها، ومياه أخرى يكون الحديد والمنغنيز هما سبب الرداءة، وأحيانا تكون الكبريتات هي السبب الرئيسي للرداءة، كما قد تكون العسرة هي سبب رداءة نوعية المياه الجوفية وأحياناً تجتمع كل تلك العوامل في موقع أخر لتجعل المياه الجوفية المستخرجة محرمة الاستخدام.

وفي محافظة نينوى ظهرت مياه جوفية متفاوتة في نوعيتها باحتوائها على شوائب مختلفة وبتراكيز متباينة وقد دفع هذا التفاوت إلى أهمية متابعة نوعية هذه المياه، ولكن ظهرت مشكلة رداءة النوعية للمياه الجوفية بشكل مقارب لما ذكر أعلاه وذلك يجعل نوعية المياه محرمة الاستخدام مما أدى إلى حرمان بعض الأراضي الزراعية من الإرواء بسبب تلك المشكلة.

وهناك العديد من الدراسات حول موضوع المياه الجوفية في محافظة نينوى منها [المشهداني وآخرون، 1989] الذي درس صلاحية استخدام المياه الجوفية للاستخدامات المختلفة للمنطقة الممتدة من ناحية بعشيقة إلى مدينة الموصل (المنطقة الشرقية للمحافظة)، وتبين لهم أن هناك تجاوزاً للخصائص النوعية لهذه المياه عن ما هو مسموح للأغراض المختلفة. كما بحث [*Mahmood*, 1994] الخصائص الفيزياوية والكيمياوية لمناطق متفرقة من محافظة نينوى مثل (عقرة، القوش، ربيعة، سنجار) ووجد تفاوتاً كبيراً في خصائص هذه المياه بسبب اختلاف العوامل الجيولوجية والهيدرولوجية، وظهر ارتفاع لكثير من المؤشرات عن ما هو مقبول للاستخدام البشري. وقد قام الباحث [الكنَة، 2001] بدراسة العيون الكبريتية في محافظة نينوى وأقضيتها فضلاً عن قسم من الآبار في 87 موقعاً مختلفاً. وتوصل إلى إن معظم هذه المياه لا تصلح للشرب بسبب محتواها العالي

من الكبريتات. ودرس [الحيالي، 2004] خصائص المياه الكبريتية في حمام العليل وطرق معالجتها.

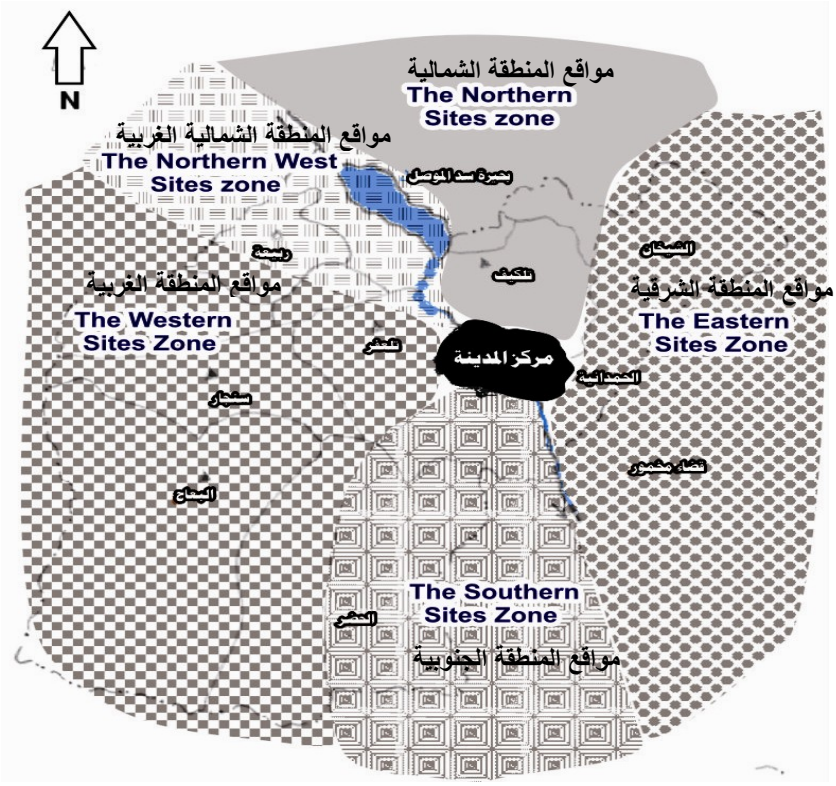
(1.6) جمع البيانات :

تم اخذ نماذج عشوائية من مناطق مختلفة من محافظة نينوى لآبار محفورة حديثاً وآبار قديمة وبأعماق مختلفة وقد جرى تحليل هذه النماذج في مختبرات جامعة الموصل وفقاً للطرائق القياسية للفحوصات لمنظمة الصحة العالمية عام 1998 [*APHA,AWWA,WEF*,1998]، كما جرت متابعة عدة مؤشرات نوعية للنماذج المأخوذة.

لقد قسمت المحافظة جغرافياً إلى عدة مناطق لتحديد أماكن النوعية الرديئة للمياه الجوفية وقد تفاوتت أعداد النماذج المأخوذة من كل منطقة كما تم استثناء تأثير العمق، وهذا التقسيم هو لتوضيح الصورة ولو بشكل سريع لأولئك الذين يرغبون بحفر آبار في أماكن معينة من المحافظة ولإعطاء نصيحة مسبقة لهم قبل أن يكون حفرهم للآبار مشكلة بحد ذاته (وهذا هو الهدف الرئيس للبحث) وهذه المناطق هي:

1- المناطق الشمالية (والشمال غربية).
2- المناطق الغربية.
3- المناطق الجنوبية.
4- المناطق الشرقية.

وهذا التقسيم أعتمد على كون مدينة الموصل مركزاً للمحافظة على أنها جغرافياً تقع في النصف الشرقي من المحافظة كما يتبين من الخارطة في الشكل (3):



الشكل رقم (3) : خارطة محافظة نينوى مبينا عليها التقسيم الجغرافي لمنطقة الدراسة

والجدول التالي يمثل البيانات الخاصة بالبحث والتي تمثل أثنى عشر مؤشراً (تمثل السمات الأساسية لنوعية المياه الجوفية) تم أخذها من آبار في إحدى وعشرون منطقة مختلفة في محافظة نينوى، والتي تم تحليلها في الجانب التطبيقي من البحث.

الجدول رقم (1):الخصائص العامة لنماذج المياه الجوفية في مناطق مختلفة من محافظة نينوى.

code	نوع التركيز / المنطقة	العسرة الكلية	المواد الصلبة الكلية	المواد الصلبة العالقة	المواد الصلبة الذائبة	الدالة الحامضية	التوصيلة الكهربائية	عسرة الكالسيوم	الكلوريدات	القاعدية	الفوسفات	النترات	الكبريتات
		T.H	*T.S*	*T.D.S*	*S.S*	*PH*	*EC*	*Ca.H*	*Cl^-*	*Alk.*	*PO_4*	*NO_3*	*SO_4*
1	شرق ربيعة	184	1129	699	430	8.45	1450	32	35	12	0.03	5.4	190
2	أم الكهف	1766	10164	4331	5833	7.82	7613	633	1311	7.3	0.04	7.8	476.6
3	قرية خرمل	600	1834	1560	274	7.72	2340	205	151.7	8	0.04	8.2	500
4	قرية عوينات	376	1555	1292	263	7.99	1780	160	167.7	6	0	14	560
5	العياضية	1350	4469	4251	218	7.97	7770	420	938	12	0.25	22	2150
6	بعاج خويتلة	2350	4524	4252	272	6.98	5500	1450	275.5	10	0.29	27	2520
7	ربيعة قرية السعدة	390	682	650	32	8.6	1850	130	147.6	10	0	8.6	210
8	وادي عكاب موصل	2000	3274	3000	274	6.5	3560	1500	179.4	80	0.38	0.5	2138
9	زمَار	500	873	843	30	7.7	1100	230	127.5	328	0.86	13.7	212
10	الخازر	481.5	754	742	12	7.3	1079	152	571	14	0.08	20.4	141
11	الشاقولي	561.5	1293	1257.5	35.5	7.37	2850	405.5	159	240	0.046	0	446
12	الشلالات	1612	2024	1952	72	7.13	2330	1224	13	12	0.16	0	750
13	بعويزة	2840	5708	5196	512	6.85	7300	710	441	30	0	14.4	3300
14	برطلة	1460	3131	3120	11	6.8	3470	908	370	110	0.05	14.5	1700
15	حقل دواجن برطلة	2385	4583	4384	200	6.98	6560	1085	642	275	0.02	14.3	2037
16	مخمور	480	830	754	76	6.88	1025	288	31	152	0.03	0	375
17	كوكجلي	1150	1863	1709	154	6.89	3690	570	238.5	468	0.04	6.2	692
18	النمرود	2240	10996	5350	5646	7.3	7450	925	511.7	287.5	0.15	13.89	2589
19	مفرق حمام العليل	2440	3082	3021	61	7.4	3250	1730	69	255	2.85	0.16	1850
20	القوش	286	579	500	79	7.66	538	162	4.7	300	0.05	2.25	36
21	فايدة	270	396	390	6	7.98	712	116	17	284	0.2	6.75	17

(2.6) تصنيف البيانات باستخدام التحليل العنقودي:

لقد تم تصنيف البيانات باستخدام التحليل العنقـودي وعلـى الحزمـة البرمجيـة *Minitab*(14)، إذ أن مصفوفة البيانات تمثل أعمدتها المؤشرات المأخوذة من عينـات المياه والتي تمثل المتغيرات (*variable*) ومصفوفها تمثل المناطق قيد الدراسة والتي تمثل المشاهدات (*observation*).

تم تحويل المشاهدات في مصفوفة البيانات إلى الصيغة القياسية وذلك لأن وحدات القياس الخاصة بالمؤشرات مختلفة، وباستخدام طريقـة وارد (*Ward's Method*) فـي أسلوب التكتل (*Agglomerative Method*) تم تـصنيف البيانـات باسـتخدام عنقـدة المشاهدات (*cluster observation*) إلى فئتين (*two classes*)، وكما يلي:

Cluster Analysis of Observations:

T.H; T.S ; T.D.S ; S.S ; PH ; EC ; Ca.H ; Cl^- ; Alk. ; PO_4 ; NO_3 ; SO_4

Standardized Variables, Euclidean Distance, Ward Linkage

Amalgamation Steps

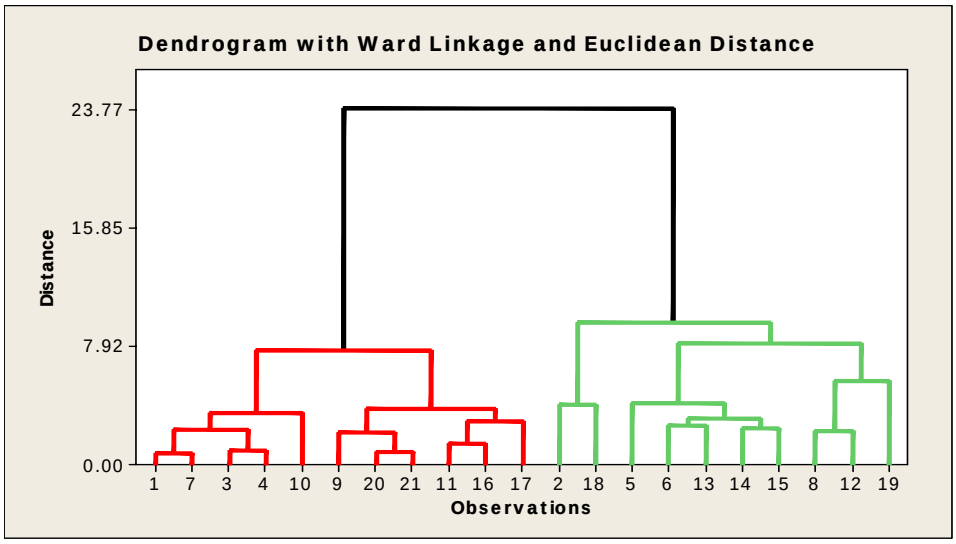
Step	Number of clusters	Similarity level	Distance level	Clusters joined		New cluster	Number of obs. in new cluster
1	20	90.932	0.7379	1	7	1	2
2	19	89.426	0.8604	20	21	20	2
3	18	88.119	0.9667	3	4	3	2
4	17	82.754	1.4033	11	16	11	2
5	16	73.532	2.1537	9	20	9	3
6	15	72.638	2.2264	8	12	8	2
7	14	71.172	2.3456	1	3	1	4
8	13	70.670	2.3865	14	15	14	2

9	12	67.526	2.6424	6	13	6	2
10	11	64.437	2.8937	11	17	11	3
11	10	61.970	3.0944	6	14	6	4
12	9	57.813	3.4326	1	10	1	5
13	8	54.503	3.7020	9	11	9	6
14	7	50.602	4.0194	2	18	2	2
15	6	50.236	4.0492	5	6	5	5
16	5	30.993	5.6149	8	19	8	3
17	4	6.303	7.6239	1	9	1	11
18	3	0.017	8.1354	5	8	5	8
19	2	-16.883	9.5105	2	5	2	10
20	1	-192.168	23.7730	1	2	1	21

Final Partition
Number of clusters: 2

	Number of observations	Within cluster sum of squares	Average distance from centroid	Maximum distance from centroid
Cluster1	11	37.761	1.78676	2.91990
Cluster2	10	106.017	3.08735	4.88369

والشكل التالي يمثل المخطط الشجري (*Dendrogram*) أو الهرمي لعملية العنقدة:



الشكل رقم (4) : الشكل الهرمي (*Dendrogram*) **لبيانات الدراسة**

من الشكل السابق والذي يوضح عملية العنقدة، نلاحظ أن مجموعة المشاهدات المتمثلة بالآبار في إحدى وعشرين منطقة مختلفة أنها توزعت على عنقودين (نمطين) مختلفين وكما يلي:

1. العنقود الأول (النط الأول ω_1) والذي يمثل تجمع لأحدى عشر منطقة متجانسة مع بعضها بنسبة تشابه (7.6239) .
2. العنقود الثاني (النمط الثاني ω_2) والذي يمثل تجمع لعشرة مناطق متجانسة مع بعضها بنسبة تشابه (9.5105) .

والجدول التالي يوضح توزيع المناطق قيد الدراسة على عنقودين:

الجدول رقم (2): توزيع المناطق قيد الدراسة على عنقودين.

	العنقود الأول ω_1	العنقود الثاني ω_2
.1	شرق ربيعة	أم الكهف
.2	قرية خرمل	العياضية
.3	قرية عوينات	بعاج خويتلة
.4	ربيعة قرية السعدة	وادي عكاب موصل
.5	زمَار	الشلالات
.6	الخازر	بعويزة
.7	الشاقولي	برطلة
.8	مخمور	حقل دواجن برطلة
.9	كوكجلي	النمرود
.10	القوش	مفرق حمام العليل
.11	فايدة	

وتجدر الإشارة إلى أن أهم مظاهر تلوث المياه هي زيادة نسبة الكبريتات (SO_4)، إذ أن للكبريتات تأثيراتٍ جمة سلبية على استخدامات المياه ويعتمد هذا التأثير على تركيزها في الماء، وقد صنّفت منظمة الصحة العالمية الكبريتات بتركيز(250 ملغم/لتر) كحد أقصى مقبول لمياه الشرب، وذلك لان ارتفاعها عن هذا الحد يتسبب بمشاكل صحية

للإنسان والحيوان [*WHO*,1985]. فصحياً قد تُسبب الكبريتات الإسهال لأولئك الذين يشربون مياه كبريتية بصورة مفاجئة وخصوصاً الأطفال الذين يكونون أكثر حساسية للمياه الكبريتية من الكبار، ويجب أن لا يكون تركيز الكبريتات أكثر من (500 ملغم/لتر) كحد أقصى تقريباً في مياه الشرب.

قد لا تكون مشكلة الكبريتات لوحدها سبب رداءة نوعية المياه الجوفية ولكن باقترانها مع شوائب أخرى تصبح للمشكلة أبعاد أكبر، إذ أن الكبريتات تكون متناسبة طردياً غالباً مع العسرة الكلية (*T.H*) (عسرة الكالسيوم أو المغنيسيوم أو كلاهما)، ومعروف أن العسرة من الشوائب المهمة والتي تجعل الماء غير ملائم للاستهلاك لأسباب صحية واقتصادية معروفة، وقد صنّفت منظمة الصحة العالمية العسرة الكلية بتركيز(500 ملغم/ لتر) كحد أقصى مقبول لمياه الشرب، وذلك لان ارتفاعها عن هذا الحد يتسبب بمشاكل صحية للإنسان والحيوان [*WHO*,1985].

مما تقدم وبعد دراسة خصائص الفئتين الناتجتين من عملية العنقدة تم ملاحظة ما يلي:

1. إن الآبار في العنقود الأول (ω_1) تعد <u>صالحة للشرب</u> وذلك لأن نسبة الكبريتات ونسبة العسرة الكلية فيها تقترب بشكل كبير من النسبة المعتمدة في منظمة الصحة العالمية، كما أن مجموعة البيانات في هذه الفئة ظهرت بشكل متجانس فيما بينها مما يدل على أن عملية العنقدة كانت ملائمة بشكل كبير.
2. كذلك بالنسبة للآبار في العنقود الثاني (ω_2) فإنه تعد <u>غير صالحة للشرب</u> وذلك لأن نسبة الكبريتات ونسبة العسرة الكلية فيها قد تجاوزت النسبة المعتمدة في منظمة الصحة العالمية، كما أن مجموعة البيانات في هذه الفئة ظهرت أيضاً بشكل متجانس فيما بينها مما يؤكد على أن عملية العنقدة إلى عنقودين كانت دقيقة.

وعليه أصبح لدينا فئتين (نمطين) من عملية العنقدة لبيانات الدراسة، الفئة الأول هي مجموعة الآبار(البيانات) التي تمثل خصائص (السمات الأساسية) للمياه الصالحة للشرب، والفئة الثانية هي مجموعة الآبار (البيانات) التي تمثل خصائص (السمات الأساسية) للمياه غير الصالحة للشرب.

(3.6) تطبيق خوارزمية الجار الأقرب في التعرف على الأنماط:

لإجراء عملية التعرف على للعينات المأخوذة من مناطق مختلفة في محافظة نينوى ودراسة خصائصها ومعرفة نوعية المياه الجوفية في تلك المنطقة، تم اخذ عينة من مياه أحد الآبار في المحافظة، وبعد قياس المؤشرات الخاصة بنوعية المياه، تم الحصول على متجه السمات الأساسية كما يلي:

$\underline{x}_0$ = [2050 3590 3426 164 7.29 4980 1500 252 30 0.06 14.2 2000]

إذ أن متجه السمات الأساسية يمثل المؤشرات (المتغيرات) حسب التسلسل التالي:

$\underline{x}_0$ = [*T.H T.S T.D.S S.S pH EC Ca.H* Cl^- *Alk.* PO_4 NO^3 SO_4]

تم كتابة برنامج بلغة (*Matlab*(*R*2008*b*، كما موضح في الملحق، لقياس المسافة بين متجه السمات $\underline{x}_0$ الخاص بعينة أحد الآبار في المحافظة وبين جميع الآبار في المحافظة ضمن الفئتين قيد الدراسة ، وتم الحصول على النتائج التالية بعد ترتيب الجدول بالاعتماد على التسلسل التصاعدي للمسافات:

الجدول رقم (3) : المسافات الاقليدية بين متجه السمات $\underline{x}_0$ وبين جميع الآبار قيد الدراسة

المنطقة	المسافة Distance	رمز المنطقة Code	التصنيف للمناطق Classification
بعاج خويتلة	1483.5	1	ω_2
وادي عكاب موصل	1529.5	2	ω_2
برطلة	1848.4	3	ω_2
مفرق حمام العليل	1933.2	4	ω_2
حقل دواجن برطلة	2213.5	5	ω_2
كوكجلي	3342.6	6	ω_1
العياضية	3374.9	7	ω_2
الشلالات	3680.2	8	ω_2
بعويزة	4012	9	ω_2
قرية خرمل	4425.8	10	ω_1
الشاقولي	4518.1	11	ω_1
قرية عوينات	5062	12	ω_1
ربيعة قرية السعدة	5815.9	13	ω_1
شرق ربيعة	5914.8	14	ω_1
زمار	6036.8	15	ω_1
مخمور	6086.5	16	ω_1
الخازر	6190.5	17	ω_1
القوش	6801.4	18	ω_1
فايدة	6840.7	19	ω_1
أم الكهف	9347.8	20	ω_2
النمرود	9775	21	ω_2

وباستخدام مصنف الجار الأقرب *9-Nearest Neighbour Classification* أي تم فرض أن (k=9) من الجيران الأقرب لعينة الدراسة، تم ملاحظة أن هنالك (8) من

الجيران من النمط الثاني ω_2 (مياه غير صالحة للشرب)، وجار واحد فقط من النمط الأول ω_1 (مياه صالحة للشرب)، أي أن الدالة المميزة ستكون بالشكل التالي:

$$h_{\omega_1}(\underline{x}_0) = \frac{1}{9} = 11.11\% \quad , \quad h_{\omega_2}(\underline{x}_0) = \frac{8}{9} = 88.89\%$$

مما تقدم نلاحظ أن الدالة المميزة للنمط الثاني تفسر 88.89% من السمات الأساسية للحالة $\underline{x}_0$ في حين أن الدالة المميزة للنمط الأول تفسر11.11% من السمات الأساسية للحالة، وعلية فان العينة المأخوذة من أحد الآبار تقع ضمن النمط الثاني أي أن العينة غير صالحة للشرب.

(7) منافشة النتائج والاستنتاجات:

1. إن التحليل العنقودي باستخدام طريقة وارد (*Ward's Method*) في أسلوب التكتل (*Agglomerative Method*) في تصنيف المشاهدات (*cluster observation*) أعطى نتائج مرضية وجيدة وذلك لأنة صنّف المشاهدات الكلية إلى مجموعتين متجانستين ولكن مختلفتين عن بعضهما البعض ، كما يوضحه الشكل (4) الهرمي، إذ أن:
   - نسبة التشابه بين العنقودين ω_1 و ω_2 هي (192.168-) وهي نسبة قليلة جداً مما يدل على الاختلاف الكبير بين العنقودين، كما أن نسبة الاختلاف (المسافة) بين العنقودين هي (23.773) مما يؤكد هذا الاختلاف .
   - أن اندماج العنقود الأول ω_1 مع بعضة كان بنسبة اختلاف (7.6239)، كما أن اندماج العنقود الثاني ω_2 مع بعضة كان بنسبة اختلاف (9.5105)، وهذه النسب قليلة مما يدل على التجانس الكبير داخل العنقود الواحد، وبنفس الوقت يكون العنقود الأول ω_1 أكثر تجانساً داخله من العنقود الثاني ω_2 .
2. إن التعرف على الأنماط بطريقة الجار الأقرب تُعد من الطرائق السريعة والبسيطة في التعرف على الأنماط في حالة توظيفها مع التحليل العنقودي وذلك لأنها لا تحتاج إلى

ملائمة البيانات لتوزيعات احتمالية شائعة، وإنما تعتمد على المسافات بين الحالة غير المرئية والحالات المصنفة، إذ أنها تعد من الطرائق اللامعلمية.

3. إن تطبيق الطريقة المقترحة على نوعية المياه الجوفية في محافظة نينوى أعطى نتائج مقبولة في عملية تصنيف البيانات إلى مجموعتين: الأولى مياه الآبار الصالحة للشرب والأخرى غير الصالحة للشرب، مما أمكن دراسة خصائص وصفات كل مجموعة وتحديد صفاتها الأساسية بحيث أصبح من السهل التعرف على العينات المأخوذة من أي بئر في المحافظة لمعرفة نوعية المياه الجوفية الخاصة بالبئر قيد الدراسة.

مما تقدم نوصي بالآتي:

1. نوصي باستخدام التحليل العنقودي مع طريقة الجار الأقرب في التعرف على البيانات أو الحالات التي ليس لها توزيعات شائعة خاصة بها.
2. نوصي باعتماد الطريقة المقترحة في مراكز البيئة المختلفة للتعرف على نوعية المياه الجوفية لآبار مجاورة للمنطقة المراد حفر بئر فيها لمعرفة نوعية المياه الجوفية قبل الحفر.

المصادر

1. بهنام، بيـداء سليمـان (2006)،" استخدام خوارزميات العـنقدة المحسنة في تنقـيب ألـبيانـات"، رسالة ماجستير غير منشورة، كلية علوم الحاسوب والرياضيات، جامعة الموصل.
2. الحيالي ،عبد الله اسماعيل (2004)، " دراسة المياه الكبريتية في مدينة الحمام العليل، خصائصها وتأثيراتها وإمكانية معالجته"، رسالة ماجستير غير منشورة ، كلية الهندسة، جامعة الموصل .
3. الكنّة، عبدالمنعم محمد علي (2001)، " نوعية المياه الكبريتية في محافظة نينوى"، رسالة ماجستير غير منشورة ، كلية العلوم، جامعة الموصل .

4. المشهداني، يحيى داود، وطليع، عبد العزيز، والحفوظي، سعد الدين ماجد (1989)، " المياه الجوفية الممتدة بين مدينة الموصل وناحية بعشيقة ومدى صلاحيتها للاستخدامات المدنية والزراعية"، مجلة التربية والعلم ، العدد 9 ،11-20 .

5. APHA,AWWA,WPCF (1985), "*Standard Methods for Water Examination and Tests*" New-York .
6. Bramer, M. (2007), "*Principles of Data Mining*", Springer-Verlag London Limited.
7. Mahmood, F.Y. (1994), "*Physio-chemical evaluation of groundwater of some wells in Nineveh district used for drinking and domestic purpose*" . Al-Rafidain engineering journal. Vol. 2, No.2 .
8. Theodoridis S. and Koutroumbas K. (2006), "*Pattern Recognition*", Third edition, Elsevier, USA.
9. Webb, Andrew R.(2002), " *Statistical Pattern Recognition*", Second Edition, John Wiley & Sons, Ltd.
10. WHO (1985), "*Guidelines for drinking water quality* " Vol. 3,Geneva**.**

الملحق

برنامج لقياس المسافة بين متجه السمات $\underline{x}_0$ الخاص بعينة أحد الآبار في المحافظة وبين جميع الآبار في المحافظة ضمن الفئتين قيد الدراسة

```
clear
clc
x1=[184.00  1766.0  600.00  376.00  1350.00 2350.00 390.0
2000.00 500.00  481.50  561.50  1612.00 2840.00 1460.00 2385.00
480.00  1150.00 2240.0  2440.00 286.00  270.00];
x2=[1129.00 10164.0 1834.00 1555.00 4469.00 4524.00 682.0
3274.00 873.00  754.00  1293.00 2024.00 5708.00 3131.00 4583.00
830.00  1863.00 10996.0 3082.00 579.00  396.00];
x3=[699.00  4331.0  1560.00 1292.00 4251.00 4252.00 650.0
3000.00 843.00  742.00  1257.50 1952.00 5196.00 3120.00 4384.00
754.00  1709.00 5350.0  3021.00 500.00  390.00];
x4=[430.00  5833.0  274.00  263.00  218.00  272.00  32.0    274.00
30.00   12.00   35.50   72.00   512.00  11.00   200.00  76.00
154.00  5646.0  61.00   79.00   6.00];
x5=[8.45    7.8 7.72    7.99    7.97    6.98    8.6 6.50    7.70
7.30    7.37    7.13    6.85    6.80    6.98    6.88    6.89
7.3 7.40    7.66    7.98];
x6=[1450.00 7613.0  2340.00 1780.00 7770.00 5500.00 1850.0
3560.00 1100.00 1079.00 2850.00 2330.00 7300.00 3470.00 6560.00
1025.00 3690.00 7450.0  3250.00 538.00  712.00];
x7=[32.00   633.0   205.00  160.00  420.00  1450.00 130.0
1500.00 230.00  152.00  405.50  1224.00 710.00  908.00  1085.00
288.00  570.00  925.0   1730.00 162.00  116.00];
x8=[35.00   1311.0  151.70  167.70  938.00  275.50  147.6   179.40
127.50  571.00  159.00  13.00   441.00  370.00  642.00  31.00
238.50  511.7   69.00   4.70    17.00];
x9=[12.00   7.3 8.00    6.00    12.00   10.00   10.0    80.00
328.00  14.00   240.00  12.00   30.00   110.00  275.00  152.00
468.00  287.5   255.00  300.00  284.00];
x10=[0.03   0.0 0.04    0.00    0.25    0.29    0.0 0.38    0.86
0.08    0.05    0.16    0.00    0.05    0.02    0.03    0.04
0.2 2.85    0.05    0.20];
x11=[5.40   7.8 8.20    14.00   22.00   27.00   8.6 0.50    13.70
20.40   0.00    0.00    14.40   14.50   14.30   0.00    6.20
13.9    0.16    2.25    6.75];
x12=[190.00 476.6   500.00  560.00  2150.00 2520.00 210.0
2138.00 212.00  141.00  446.00  750.00  3300.00 1700.00 2037.00
375.00  692.00  2589.0  1850.00 36.00   17.00];

f=[2050 3590 3426 164 7.29 4980 1500 252 30 0.06 14.2 2000];

for i=1:21
    w=(f(1)-x1(i))^2+(f(2)-x2(i))^2+(f(3)-x3(i))^2+(f(4)-
x4(i))^2+(f(5)-x5(i))^2+(f(6)-x6(i))^2+(f(7)-x7(i))^2+(f(8)-
x8(i))^2+(f(9)-x9(i))^2+(f(10)-x10(i))^2+(f(11)-x11(i))^2+(f(12)-
x12(i))^2;
    d=sqrt(w)
end
```