المجلة العراقية للعلوم الإحصائية

www.stats.mosuljournals.com

الجمع بين التحليل العنقودي والتحليل الانحدار الخطي المتعدد لإنشاء نموذج التنبؤ الأكثر دَقَّة للتبخر في إقليم كردستان العراق

بخشان احمد حمد

قسم الرياضيات، كلية التربية، جامعة صلاح الدين ، اربيل، العراق

معلومات النشر

تاريخ المقالة:
تم استلامه في 18 ايلول 2023
تم القبول في 12 تشرين الثاني 2023
متاح على الإنترنت في 1 كانون الاول 2023

الكلمات الدالة:
التحليل العنقودي
تحليل الانحدار الخطي المتعدد
التبخر

المراسلة:
بخشان احمد حمد
Paxshan.ahmad1@su.edu.krd

الخلاصة

تهدف هذه الدراسة الى بناء نموذج تنبؤ بالمتغيرات المؤثرة للتبخر في إقليم كردستان –العراق باستخدام مفهوم تحليل الانحدار والعنقودي. إن الأساليب المشتركة توجه العمل إلى إبراز نقاط القوة في كل تقنية، و إمكانية استخدام التحليل العنقودي الهرمي (الجار الاقرب، الجار الابعد والمتوسط) لتحسين الدقة ألتنبؤيه لنماذج الانحدار. تم تصنيف المتغيرات المؤثرة في معدل التبخر باستخدام بيانات الطقس لمحطة الأرصاد الجوية في إقليم كردستان، العراق للفترة من كانون الثاني 2020 حتى كانون الأول 2022، واستخدام قيم R^2 المعدلة، MSE و RMSE كمؤشر لكفاءة أداء النموذج.

توصلت الدراسة الى أن التعنقد قبل تحليل الانحدار يؤدى إلى تحسين دقة التنبؤ من خلال تصنيف وتحديد متغيرات مستقلة متجانسة ضمن العنقود الواحد مختلفة عن باقي العناقيد.

DOI: https://doi.org/10.33899/iqjoss.2023.0181226

1. المقدمة :

إن اتجاه القياس الكميه والمنهاج العام في البحوث هو استخدام الطرائق الإحصائية وذلك لان التصنيف العلمي للظواهر الطبيعية والمناخية وتحليل العلاقات المتبادلة بين الظواهر على أساس موضوعي والتنبؤ بنموذج يفسر العلاقة بين الظواهر أصبحت ضرورة .

الانحدار من أكثر طرائق استخراج البيانات الإحصائية المهمة استخدما في مختلف مجالات العلوم بما في ذلك الذكاء الاصطناعي وعلم الأعصاب و بيانات الطقس. فقد تم استخدام نموذج الانحدار على نطاق واسع من قبل العديد من الباحثين لتقدير التبخر على أساس معلمات الأرصاد الجوية مثل دراسة Da Silva (2016) و المولى (2017) المستخدمة لطرائق الانحدار لإنشاء نماذج شهرية إقليمية ومتوسط التبخر كدالة العوامل المتاحة بما في ذلك درجة الحرارة وخط الطول والارتفاع. (Mohammd, *et al.*, 2022).

تقنية الانحدار الخطي تقوم بتحليل العلاقة بين المتغيرات و الفرق الرئيسي بين تقنيات التعنقد والانحدار هو عملية التعلم. فالانحدار يحتوي على متغير استجابة (Y) مرتبط بالمتغيرات المستقلة (X) (التعلم الخاضع للإشراف) بينما يؤدي التحليل العنقودي إلى التعلم غير الخاضع للإشراف (Mohammed & Hannon, 2019).ويعد التحليل العنقودي من التحليلات الإحصائية المهمة والشائعة الاستخدام في مختلف مجالات العلوم التطبيقية والصرفة من خلال مجموعة من الخطوات تهدف إلى تصنيف مجموعة حالات (Cases) أو متغيرات

(Variables) بطرائق معينة وترتيبها داخل عناقيد (Cluster) بحيث تكون الحالات المصنفة داخل عنقود الواحد متجانسة فيما يتعلق بخصائص محددة وتختلف عن حالات أخرى موجودة في عنقود آخر.

الانحدار العنقودي تم اقتراحه كطريقة لتحديد نموذج لكل قسم من البيانات. والنظر إليها على أنها خليط معين أو نموذج فئة كامنة، من بيانات المنظور التحليلي كمزيج من التحليل العنقودي والانحدار. كلا النهجين الانحدار الخطي والعنقودي هي تقنية مفيدة عندما يكون عدم التجانس موجودا في البيانات. التحليل العنقودي هو أحد أساليب التحليل الإحصائي متعدد المتغيرات حيث يتم الحصول على العنقود بطريقة خاضعة للإشراف من أجل أن يكون نموذج الانحدار "الأفضل" لكل مجموعة(De Carvalho, *et al.*, 2010) .

التبخر هو عنصر رئيسي في الدورة الهيدرولوجية وتعتبر عاملا رئيسيا في إدارة الموارد المائية للمناطق الزراعية وشبه الزراعية. تقدير فقد المياه من خلال التبخر ضروري لنمذجة ومسح وإدارة العديد من مشروعات النظم الهيدرولوجية والمائية. التبخر هو متغير يجمع أو يتضمن تأثير العديد من عناصر الغلاف الجوي، مثل درجة الحرارة، ودرجة الحرارة الرطبة وسرعة الرياح واتجاهها والضغط الجوي على البحر والرطوبة النسبية وشدة سطوع الشمس. يزداد التبخر مع سرعة الرياح العالية، كحد أقصى لدرجات الحرارة والرطوبة المنخفضة (Mohammd, *et al.*, 2022) .

النماذج المطورة من بيانات الأرصاد الجوية تنطوي على علاقات تجريبية إلى حد ما، وتعطي هذه النماذج نتائج موثوقة عند تطبيقها على الظروف المناخية (Al–Mukhtar, 2021). وبالتالي فإن استخدام الصيغ أو النماذج الرياضية التي تتنبأ بالتبخر من البيانات المناخية المتاحة أمر وارد يوفر نتائج أكثر دقة. وتمت دراسة التنبؤ بالتبخر من قبل العديد من الباحثين منها:

دراسة (Almedeij, 2016) تطوير نموذج التبخر لمنطقة القاحلة في دولة الكويت بناء على الأساليب الإحصائية الكلاسيكية، منها الانحدار الخطي المتعدد وتحليل السلاسل الزمنية. أظهرت الدراسة أن قيم التبخر هو دالة لدرجة الحرارة والرطوبة النسبية وسرعة الرياح.

دراسة (Adnan, *et al.*, 2019) لقدرة ثلاث طرائق تكيفية للضبابية العصبية في تقدير التبخر الشهري باستخدام المدخلات المناخية لدرجات حرارة الهواء الدنيا والقصوى، وسرعة الرياح، وساعات سطوع الشمس، والرطوبة النسبية. تم تقييم الطرائق من خلال استخدام معيار جذر متوسط مربع الخطأ (RMSE) ومتوسط الخطأ المطلق (MAE) ومعامل التحديد (R^2).وأشارت النتائج إلى الدقة الفائقة لطريقة التجميع الضبابي (FCM) لنفس متغيرات الإدخال وتقديرات أفضل مقارنة بطريقتي قسم الشبكة المضمنة (GP) والتجميع المطروح (SC).

دراسة (Alsumaiei, 2020) لنمذجة معدلات التبخر اليومية في المناخات شديدة الجفاف باستخدام الشبكات العصبية الاصطناعية (ANNS). تم تطبيق الشبكات العصبية لنمذجة مثل هذه الظروف المناخية في دولة الكويت، لتعزيز أداء ANN وتم تحسين هيكل شبكة ANN من خلال اختبار مجموعات مدخلات الأرصاد الجوية المختلفة لنمذجة التبخر في تلك الظروف المناخية شديدة الجفاف.

دراسة (Al–Mukhtar, 2021) بحثت في إمكانية تطبيق استخدام غابة الانحدار الكمي. تم تشكيل النموذج باستخدام بيانات من ثلاث محطات أرصاد جوية مختلفة تقع في مناخات مختلفة قاحلة وشبه قاحلة في العراق. وكانت هذه المحطات في مدن بغداد، والبصرة والموصل. تمت مقارنة أداء غابات الانحدار الكمي مع ثلاثة من طرائق الذكاء الاصطناعي وهي الغابات العشوائية وآلة ناقلات الدعم والشبكة العصبية الاصطناعية بالإضافة إلى نماذج الانحدار الخطي المتعدد ،تم تقييم النتائج باستخدام معايير للأداء: معامل التحديد (R^2) ، جذر متوسط مربع الخطأ (RMSE) ، أظهرت النتائج أن نموذج غابات الانحدار الكمي حقق الأداء الأمثل بين الطرائق التي تم تقييمها.

دراسة (Mohammd, *et al.*, 2022) لنمذجة القياسات الشهرية على مدى 18عاما بين يناير 2000 وديسمبر 2017. تم استخدام تقنيات الانحدار الخطي المتعدد ، استخدمت درجة الحرارة وسرعة الرياح والرطوبة النسبية وساعات سطوع الشمس كمتغيرات مستقلة لإنشاء أفضل تنبؤا لنموذج التبخر في المناطق القاحلة في وادي حوران. وثقت الدراسة السابقة قيام هانسون بالتحقيق للتبخر اليومي في ثلاثة مواقع جنوب غرب ولاية أيداهو الهند.

أشارالباحثون (Essa, *et al.*, 2023) الذين استخدموا التحليل العنقودي كسلوب للبحث عن مجموعة من المتغيرات واقتراحها كخطوة أولية لتصنيف البيانات وجهوزيتها للتنبؤ(Taha, 2022).

2. مفاهيم أساسية:

تضمن هذه الدراسة جانبين، الاول مفهوم التحليل العنقودي الهرمي و طرائق التجميع (الجار الاقرب، الجار الابعد و المتوسط) مع مقياس المسافة الاقليدية. اما الجانب الثاني فيدور حول تحليل الانحدار الخطي المتعدد ومعايير المقارنة و جودة نموذج الانحدار.

2.1 التحليل العنقودي:

التحليل العنقودي من التحاليل الإحصائية التي تستخدم في البحث العلمي لغرض وصف الأساليب التي تبحث عن تعنقد البيانات المتعددة لتكون مجموعات متجانسة فيما بينها. كذلك لتوصيف ومقارنة مجتمعات (عناقيد) في البيئات غير المتجانسة، تعتمد على نقاط التشابه والاختلاف بين البيانات حيث يندرج هذا الأسلوب ضمن أساليب التنقيب اللامعملي للبيانات (Shahab & Rashed, 2021) والذي يعتبر من المجالات المهمة والحديثة في علم الإحصاء وأهم استخدامات التحليل العنقودي هي (استكشاف البيانات، التصنيف، التحديد، توليد فرضيات خاصة لكل دراسة والتنبؤ).

2.1.1 قياس المسافة الإقليدية (Euclidean Distance):

هي المسافة بين العناصر المختلفة لتحديد درجة التقارب بين متغيرين $\boldsymbol{x}$ و $\boldsymbol{y}$ ، حيث تعتبر المسافة الأكثر استخداما ، وهي مسافة الخط المستقيم بين نقطتين في فضاء متعدد الأبعاد، (Khattreer & N, 2020) ويتم حسابها وفقا للصيغة التالية:

$$D_{(x,y)} = \sqrt{\sum_{d=1}^{p} (x_{id} - y_{jd})^2} \quad ,p = \text{عدد البيانات} \qquad (1)$$

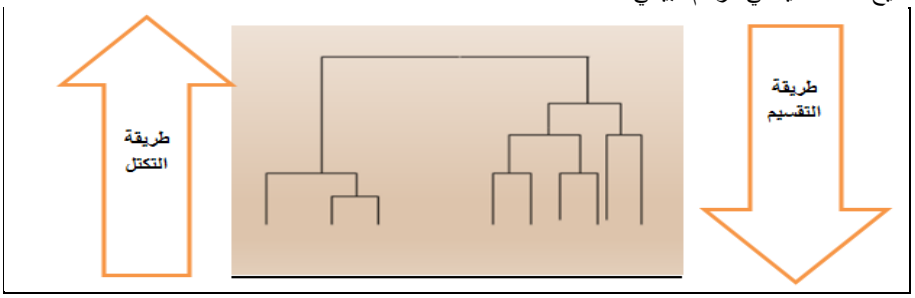
x_{id} : عبارة قيمة المتغير x_d للعنصر i، y_{jd} : عبارة قيمة المتغير y_d للعنصر j

$D_{(x,y)}$: عبارة قيمة المسافة بين النقطتين x_{id} و y_{id}

2.1.2 التحليل العنقودي الهرمي (Hierarchal Cluster Analysis):

تعد هذه الطريقة من أكثر الطرائق استخداما حيث لا يتم تقسيم بيانات الدراسة إلى عدد من العناقيد في خطوة واحدة بل هي عبارة عن تسلسل هرمي للعناقيد المترابطة مما يوضح عملية ربط العناقيد بعضها مع بعض من خلال سلسلة متداخلة لإعطاء شكل هرمي يكون Dendrogram، لا تتطلب هذه الطريقة معرفة مسبقة بعدد العناقيد التي يقوم هذا التحليل الهرمي للحالات بموجبها بتناسب العينات التي عددها قليل وصغير، وينقسم هذا التحليل إلى قسمين وهما التحليل الهرمي للحالات (Cases) والتحليل الهرمي للمتغيرات (Variables) وهناك طريقتان للتجمع الهرمي، بما في ذلك دمج المجموعات (العناقيد) الصغيرة في مجموعات أكبر ،وهي تقنية التجميع (The Agglomerative Technique) ، أوعن طريق فصل العناقيد الكبيرة إلى مجموعات صغيرة، وهي التقنية المثيرة للانقسام (The divisive technique) (Essa *et al*., 2023; Mohammed & AL-Rawi, 2019).

ويمكن توضيح هذه العملية في الرسم البياني أدناه:



شكل (1):المخطط الشجري الهرمي التجميعي و ألتقسيمي للعناصرضمن مجموعة من العناقيد

2.1.3 طرائق التجميع الهرمي:

هناك العديد من طرائق التحليل العنقودي ،من أهمها والتي استخدمت في هذه الدراسة منها:

- **طريقة الربط المفرد أقرب جار (Single Linkage Method):**

هذه الطريقة هي أبسط وأقدم طريقة، وتعتبر أيضا الأكثر انتشارا، حيث تجمع الأقرب في المسافة إلى العناصر لتكوين نواة العناقيد، ثم تضاف باقي العناصر التي هي أكثر تشابها وقربية في المسافة، مما يؤدي إلى سلسلة طويلة من الترابط. لتحديد المسافة بين المجموعات (Mohammed & AL-Rawi, 2019)، يتم حسابها وفقا للصيغة التالية:

$$D(A,B) = Min_{x_i \in A, y_i \in B}(d(x_i, y_i)) \quad (2)$$

حيث يمثلx_i عنصر في A و y_i عنصر في B، $D_{(I,J)}$ المسافة الإقليدية للمتغيرات في المجموعات i,j.

- **طريقة الربط الكامل أبعد جار (Complete Linkage Method):**

يعرف بطريقة الارتباط التام أو الكامل أو الجار الأبعد، في هذه الطريقة يتم تشكيل الكتلة بطريقة تعكس الطريقة الأولى، ذلك لأنها تبدأ بتجميع العناقيد (العناصر) المنفردة لتشكل عنقودا واحدا فقط عندما ترتبط جميع العناصر بصورة تامة (أي تشكل زمرة)، يتم تحديد التماثل بين العناقيد المختلفة عن طريق إيجاد المسافة الأبعد ما بين أي عنصرين أي أنها تعتمد على الأقل تشابها بين المتغيرات أو الحالات، وفقا للصيغة التالية:

$$D(A,B) = Max_{x_i \in A, y_j \in B}\left(d(x_i, y_j)\right) \quad (3)$$

- **طريقة المتوسط (Average Linkage) :**

تستخدم هذه الطريقة بالاعتماد على متوسط المسافة بين نقطة من العنقود الاول (A) n_A ونقطة من العنقود الثاني (B) n_B وفق الصيغة التالية:

$$D(A,B) = \sum_{i=1}^{nA}\sum_{j=1}^{nB} d(x_i, y_j) \Big/ n_A n_B \quad (4)$$

2.2 الانحدار الخطي المتعدد:

يعرف الطريقة الإحصائية المستخدمة لفهم العلاقات السببية بين المتغيرات الكمية المختلفة في الحياة اليومية بتحليل الانحدار الخطي المتعدد. ولإيجاد العلاقة بين المتغير التابع والمتغيرات المستقلة المختلفة بأسلوب رياضي.

يتم استخدام قيم المعلمات المقدرة لإنشاء معادلة تحليل الانحدار. يتم استخدام اختبارات مختلفة لتحديد ما إذا كان النموذج مقبولا أم لا. إذا تم اعتبار النموذج مقبولا، فيمكن استخدام معادلة الانحدار المقدرة للتنبؤ بقيمة المتغير التابع بالنظر إلى قيم المتغير المستقل (Ngo & Puente, 2012).

الأهداف الرئيسية لتحليل الانحدار هي: الوصف، التقدير، التنبؤ والتحكم. يصف الانحدار العلاقة بين المتغيرات التابعة والمستقلة (Ali & Younas, 2021). والقدرة على تقدير قيمة المتغير التابع بناء على القيم المرصودة للمتغيرات المستقلة، لإسقاط النتائج والتغيرات يعتمد على التفاعلات بين المتغيرات التابعة والمستقلة مع تقليل تأثير متغير مستقل (Esmaeel & Rashed, 2022).

تضع معظم الاختبارات الإحصائية افتراضات حول المتغيرات في التحليل التي يجب الوفاء به. ويستخدمون الاختبارات الإحصائية لتقييم مدى ملاءمة البيانات ودقة النموذج، وأخطاء النموذج المحتملة، والصعوبات في فهم النتائج (Ngo & Puente, 2012). عندما يتم انتهاك هذه الافتراضات، قد لا تكون النتائج جديرة بالثقة، مما يؤدي إلى خطأ من النوع الأول أو النوع الثاني. يتمثل هذه الفروض (Ali & Younas, 2021) بخطية العلاقة بين المتغيرات التابعة والمستقلة، عدم وجود ارتباط ذاتي بين المتغيرات المستقلة، توزيع البواقي بشكل طبيعي، تجانس تباين البواقي وعدم وجود قيم متطرفة.

2.2.1 انموذج الانحدار:

إن تحليل الانحدار أحد اساليب التحليل الإحصائي متعدد المتغيرات التي تحتوي على صيغتين، فتحليل الانحدار يقوم على أنموذجين أساسيين هما (خطي، غير خطي) ويعتبر الانموذج الخطي من الانموذجات الأكثر شيوعا وهذا الأنموذج ينطوي على نوعين هما (بسيط، متعدد) البسيط يقوم على العلاقة بين متغير معتمد أو تابع (Y) ومتغير مستقل واحد (X) وأما المتعدد فيقوم على العلاقة بين متغير تابع (Y) مع أكثر من متغير مستقل (X_i)، وفي كلتا الحالتين أن كان بسيطا أو متعددا فالغرض هو إيجاد التنبؤات (Best & Wof, 20l5; Esmaeel & Rashed, 2022) لا بد من توفر معادلة تقديرية يعتمد عليها أو تكون الأساس في الاختبارات والتعويضات ومثل هذه المعادلة يتم إيجادها بعد تقدير معالم الأنموذج الخاص بالانحدار وهوالأنموذج الذي يأخذ الشكل الاتي:

$$y = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \cdots + \hat{\beta}_p x_{ip} + \varepsilon_i \quad (5)$$

y يمثل المتغير التابع، $x_{ip} \ldots, x_{i2}, x_{i1}$ عبارة عن المتغيرات التفسيرية ومعلمات النموذج هي $\hat{\beta}_p \ldots, \hat{\beta}_1, \hat{\beta}_0$. ε_i يشير الى مصطلح "البواقي" وهو الفرق بين القيم المرصودة والمتوقعة ل y في معادلة الانحدار المقدرة. لتقليل مجموع البقايا التربيعية، يتم تحديد تقديرات المعلمات باستخدام نهج المربعات الصغرى (Kor & Altun, 2020).

في الانحدار الخطي ، نسعى إلى تقدير المعلمات $\hat{\beta}_p \ldots, \hat{\beta}_1, \hat{\beta}_0$ التي تقلل من مجموع الأخطاء التربيعية (Seber & Lee, 2012; Mohammed & Hannon, 2019).

$$\text{Min} \rightarrow \sum_{i=1}^{N} e_i{}^2 \quad , e_i = y_i - \hat{y}_i$$

والجدول (1) يشير الى مصادر التباين للانحدار المتعدد .

الجدول (1): جدول تحليل التباين للانحدار الخطي المتعدد

Source of Variation	DF	Type of Sum Squares	Mean Square	F-test
Regression	p	$\text{SSR} = \sum_{i=1}^{N} (\hat{y}_i - \bar{y})^2$	$MSR = SSR/p$	$\frac{\text{MSR}}{MSE}$
Residual (error)	$n-p-1$	$\text{SSE} = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$	$MSE = SSE/n-p-1$	
Total	$n-1$	$\text{SST} = \sum_{i=1}^{N} (y_i - \bar{y})^2$		

2.2.2 معايير المقارنة :

دقة التنبؤ وتقليل الأخطاء المتوقعة الى ادنى حد هو الجانب الحيوي المهم من عملية التنبؤ، و عملية اتخاذ القرار بجودة الانموذج يتطلب تقدير الأخطاء بين القيم المرصودة والمتوقعة، فيما يخص تحليل الانحدار فإن هناك ثلاثة مؤشرات مهمة للجودة منها:

- **الخطأ المعياري للانحدار (MSE) :**

الخطأ المعياري للانحدار يعرف أيضا باسم الخطأ القياسي المتبقي. يمكن التعبير عن جودة الانحدار على النحو التالي:

$$\text{MSE} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2/n \quad (6)$$

- **معامل التحديد (R-squared):**

أن معامل التحديد R–squared مقياس إحصائي يمثل نسبة التباين لمتغير تابع يتم تفسيره بوساطة المتغيرات المستقلة في الانموذج الانحدار. هو مقياس جودة الملاءمة الأكثر استخداما.

$$R^2 = 1 - \frac{\text{Unexplained Variation}}{\text{Total Variation}}$$

$$R^2 = 1 - \frac{\text{SSR}}{\text{SST}} \quad (7)$$

حسب التعريف R^2 تقع بين 0 و 1 وتمثل نسبة تباين العينة في y موضحة بوساطة xs. يزيد R^2 عند اضافة كل متغير مستقل جديد الى الانموذج الانحدار، لذلك فهو ليس معيارا مفيدا لاختيار الانحدار (Frost, 2023). $\text{Adj.}\, R^2$ هو نسخة معدلة لا تزداد دائما مع المزيد من متغيرات مستقلة فى انموذج الانحدار.

$$Adj\ R^2 = 1 - \frac{SSR/N-k-1}{SST/N-1}$$

$$Adj.R^2 = 1 - \frac{N-1}{N-k-1}\frac{SSR}{SST} \quad (8)$$

- **جذر متوسط مربع الخطأ (RMSE) :**

أحدى مقاييس الأداء الرئيسيين لانموذج الانحدار هو جذر متوسط الخطأ التربيعي (RMSE). يحسب متوسط الفرق بين القيم المتوقعة والفعلية للانموذج. يعطي تقديرا لمدى توقع الانموذج للقيمة المستهدفة (الدقة).

$$RMSE = \sqrt{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 / n} \quad (9)$$

$$\text{RMSE} = \sqrt{\text{MSE}} \quad (10)$$

يتميز جذر متوسط الخطأ التربيعي بميزة عرض الخطأ في نفس وحدة العمود المتوقع، مما يجعل من السهل تفسيره. كلما انخفضت قيمة جذر متوسط الخطأ التربيعي، كان الانموذج أفضل.

أثناء تضمين متغيرات مؤثرة إضافية في مجموعة بيانات، فإننا نقوم بتقليل جذر متوسط الخطأ التربيعي. ينخفض MSE مع اقتراب نقاط البيانات من خط الانحدار مع عدد أقل من الأخطاء(Frost, 2023).

3. الجانب العملي:

في هذه الدراسة تم أخذ البيانات من وزارة النقل والاتصالات لإقليم كردستان– العراق المديرية العامة للأنواء الجوية والرصد الزلزالي بموجب كتابهم المرقم 77 في 2023/1/9. بيانات التبخر الشهرية لمحطات محافظات (أربيل، السليمانية ودهوك) خلال فترة (36) شهرا للفترة من كانون الثاني 2020 لغاية كانون الأول 2022. تم استخدم البرنامج الإحصائي (SPSS Version 24) لاستخراج النتائج للتحليل العنقودي أولا. وثانيا تطبيق تحليل الانحدار لكل عنقود تم تشكيله باعتبار التبخر متغيرا معتمدا مع باقي المتغيرات المستقلة (درجة الحرارة، درجة الحرارة الرطبة، شدة سطوع الشمس، الضغط الجوي على مستوى البحر، اتجاه وسرعة الرياح والرطوبة النسبية).

جدول (2): يوضح المتغير التابع و المتغيرات التفسيرية المستخدمة في الدراسة

	Variables	
1	E-Evaporation	التبخر
2	T-Temperature	درجة الحرارة (درجة مئوية)
3	WT-Wet Temperature	درجة الحرارة الرطبة (درجة مئوية)
4	SS-Sunshine	شدة سطوع الشمس (واط/ م^2)
5	SP-Pressure at Sea level	الضغط الجوى على مستوى البحر (ملم زئبق)
6	Wind Direction	اتجاه الرياح
7	RT-Relative Humidity	الرطوبة النسبية (٪)
8	WS-Wind Speed	سرعة الرياح (م/ ثانية)

الجدول (2) يبين المتغيرات قيد الدراسة، تم قياس البيانات بمعدلال شهرية للفترة 36 شهرا، حيث تضمنت المتوسط، الخطأ المعياري للمتوسط، الوسيط، المنوال، الانحراف المعياري، التباين، الالتواء و التفلطح للمتغيرات. والجدول(3) يبين الاحصاء الوصفي للمتغيرات .

الجدول(3): الاحصائيات الوصفية للمتغيرات قيد الدراسة للفترة من كانون الاول 2020 ولغاية كانون الثاني 2022

Variables	E	T	WT	SS	SP	W D	RH	W S
N	36.000	36.000	36.000	36.000	36.000	36.000	36.000	36.000
Mean	8.608	23.115	15.925	8.936	13.528	247.478	51.925	1.514
Std. Err. of Mean	0.909	1.578	0.925	0.498	0.539	10.681	3.080	**0.043**
Median	8.000	24.650	17.000	8.550	14.000	270.000	48.100	1.500
Mode	2.30	11.40	7.00	5.100	8.90	270.000	29.500	1.400
Std. Deviation	5.454	9.467	5.552	2.990	3.235	64.087	18.480	**0.255**
Variance	29.741	89.630	30.829	8.942	10.464	4107.134	341.500	0.065
Skewness	0.217	-0.073	**-0.192**	-0.087	0.112	0.387	0.229	0.087
Kurtosis	-1.427	-1.560	-1.467	-1.468	-1.147	-0.900	**-1.610**	-1.251

بلغ متوسط درجة الحرارة (23.115) درجة مئوية مقارنة بوسيط درجات الحرارة لفترة الدراسة كانت (24.650) درجة مئوية، في حين متوسط درجة شدة سطوع الشمس بلغت (8.936) مقارنة بوسيط (8.550)، متوسط الضغط الجوي على مستوى سطح البحر بلغت (13.528) مقارنة بوسيط (14.000)، متوسط الرطوبة النسبية بلغت (51.925) مقارنة بوسيط (48.100). متوسط درجة حرارة الرطبة (15.925) درجة مئوية مقارنة بوسيط درجة الحرارة الرطبة (17.000)، متوسط سرعة الرياح كانت (1.514، م/ ث) مقارنة بالوسيط (1.500، م/ ث).

أقل خطأ معياري للمتوسط وأقل انحراف معياري للمتغير سرعة الرياح (0.043)، أقل قيمة للالتواء كانت للمتغير درجة الحرارة الرطبة (−0.192) في حين أقل قيمة التفلطح كانت لمتغير الرطوبة النسبية (−1.610).

3.1 نتائج التحليل العنقودي:

باستخدام طريقة الجار الاقرب، الجار الابعد والمتوسط (معدل الربط بين المجموعات) مستخدما مربع المسافة الاقليدية لتصنيف قابلية المتغيرات الداخلة في انموذج الانحدار المتعدد ومن ثم تقييم قوة ودقة الانموذج. يتم الجمع بين مصفوفة المسافة والعلاقة بين العناقيد، وتقسيم عدد البيانات إلى عنقودين منفصلين باستخدام نمط التكتل. تكون درجة التجانس قوية داخل المجموعات المختلفة،

الجدول (4) يبين مصفوفة مربع المسافة الاقليدية بين المتغيرات قيد الدراسة لكل عنصر عبارة عن معاملات المسافة بين المتغيرات، مع زيادة المسافة بين أي متغيرين يزداد الفرق.

الجدول(4):مصفوفة مربع المسافة الاقليدية بين المتغيرات

Variables	T	WT	SS	SP	WD	RH	WS
Temperature (T)	.000						
Wet Temperature (WT)	**0.609**	.000					
Sunshine (SS)	**3.244**	**3.822**	.000				
Pressure at Sea level (SP)	**5.258**	**2.906**	**8.979**	.000			
Wind Direction (WD)	38.714	39.975	41.065	38.110	.000		
Relative Humidity (RH)	**137.758**	**135.801**	**135.744**	**127.010**	95.933	.000	
Wind Speed (WS)	91.698	89.226	93.956	82.936	101.171	43.979	.000

مربع المسافة الاقليدية على التوالي من الأقرب إلى الأبعد كان أقل قيمة (0.609) للمسافة بين متغير درجة الحرارة ودرجة الحرارة الرطبة، تليها قيمة (2.906) بين متغير درجة الحرارة الرطبة ومتغير الضغط الجوي على مستوى سطح البحر.

المسافة بين متغير درجة الحرارة وشدة سطوع الشمس كانت قيمتها (3.244)، ومن ثم قيمة المسافة بين شدة سطوع الشمس مع درجة حرارة الرطبة (3.822)، المسافة بين الضغط الجوى على مستوى سطح البحر مع درجة الحرارة كانت قيمتها (5.258) من جهة أخرى المسافة بين الضغط الجوى على مستوى سطح البحر وبين شدة سطوع الشمس بلغت (8.979).

ابعد مسافة كانت قيمتها (137.758) بين متغيري درجة الحرارة والرطوبة النسبية، تليها المسافة بين درجة الحرارة الرطبة والرطوبة النسبية كانت قيمتها (135.801)، المسافة بين متغيري شدة سطوع الشمس والرطوبة النسبية قيمتها (135.744) ومع المتغير اتجاه الرياح بلغ قيمته (95.933) نستنتج بأن متغير الرطوبة النسبية لها ابعد مسافة مع المتغيرات الاخرى قيد الدراسة.

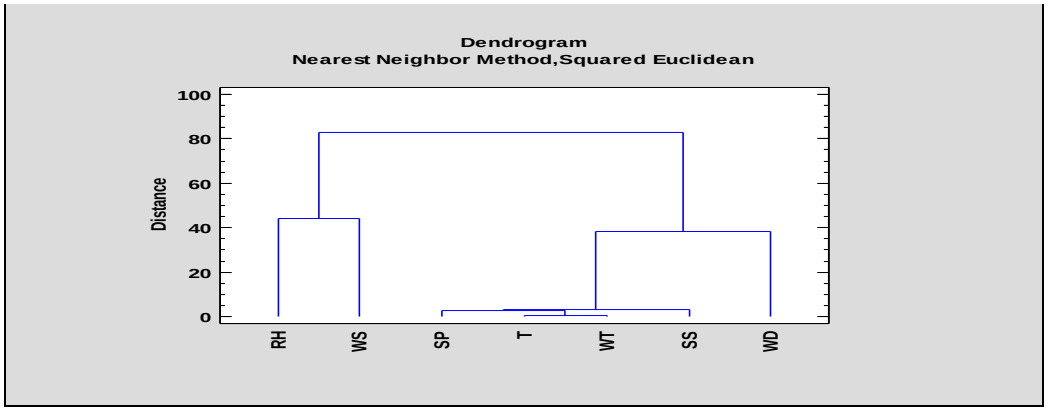
الجدول (5) يوضح قيمة المسافة أو المعاملات (Coefficients) وفق مقياس المسافة وطريقة الربط المستخدمة في التحليل، وتحديد المفردات أو المجموعات التي يتم ربطها في كل خطوة من خطوات التحليل.

الجدول(5):جدول التكتل بطرائق الجار الاقرب،الجارالابعد و معدل الربط العنقودي بمقياس مربع المسافة الاقليدية

Agglomeration Schedule Clustering Method: Average Linkage (Between Groups) Distance Metric: Squared Euclidean				Agglomeration Schedule Clustering Method: Furthest Neighbor (Complete Linkage) Distance Metric: Squared Euclidean			Agglomeration Schedule Clustering Method: Nearest Neighbor (Single Linkage) Distance Metric: Squared Euclidean		
Stage	**Combined Cluster 1**	**Combined Cluster 2**	**Distance**	**Combined Cluster 1**	**Combined Cluster 2**	**Distance**	**Combined Cluster 1**	**Combined Cluster 2**	**Distance**
1	**4**	**7**	**0.609**	**4**	**7**	**0.609**	**4**	**7**	**0.609**
2	3	4	3.533	3	4	3.822	2	4	2.906
3	2	3	5.714	2	3	8.979	2	3	3.244
4	2	5	39.466	2	5	41.065	2	5	38.110
5	1	6	43.973	1	6	43.979	1	6	43.979
6	1	2	109.123	1	2	137.758	1	2	8 2.936

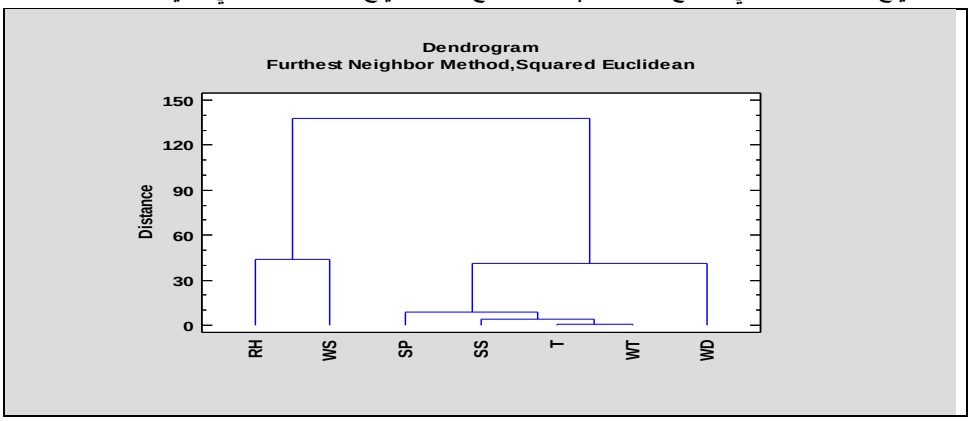
عملية الربط بين المتغيرات تتم على أساس مربع المسافة الأقليدية فيما بينها أي أن المسافة الأقصرو قيمتها (0.609) ستمثل المرحلة الأولى من العنقدة التي هي عبارة عن تجمع متغير درجة الحرارة الرطبة (4) ودرجة الحرارة (7) كخطوة أولى للتعنقد في طرائق الثلاث (الجار الا قرب، الجار الأبعد والمعدل) من ثم الخطوة الثانية الحرارة الرطبة (4) مع متغيرالضغط الجوى على مستوى سطح البحر (2) في طريقة الجار الأقرب بينما في طريقتي الجارالابعد والمتوسط يتم الربط بين متغير الحرارة الرطبة (4) مع متغير شدة سطوع الشمس (3) وفي الخطوة الثالثة الربط بين متغير شدة سطوع الشمس (3) مع متغير الضغط الجوى على مستوى سطح البحر (2) في الطرائق الثلاث. الخطوة الرابعة هي الربط بين متغير الضغط الجوى على مستوى سطح البحر (2) مع متغير اتجاه الرياح (5) في طرائق الثلاث، تتساوى الطرائق الثلاث في الخطوة الخامسة من الربط بين الرطوبة النسبية (1) وسرعة الرياح (6) وكذلك في الخطوة السادسة يتم الربط بين متغيرالرطوبة النسبية (1) متغيرالضغط الجوى على مستوى سطح البحر (2).

الشكل (2) يوضح المخطط الشجري لعدد ومراحل تشكل العناقيد بطريقة الجار الأقرب ومربع المسافة الاقليدية بالإضافة إلى ملاحظة المفردات أو المجموعات التي تم ربطها معا في كل خطوة من خطوات التحليل، وتعتمد مقياس المسافة في تقسيمها دون الرجوع إلى طريقة الربط بعكس نمط التكتل كما موضح في الجدول (4). إن التعنقد بين المتغيرات يعتمد على المسافة الأقصر بينهما وذلك لكونها أكثر تجانسا من الأزواج الأخرى من المتغيرات كما هو واضح في الشكل (2) درجة الحرارة ودرجة الحرارة الرطبة تشكل العنقود ألاول من ثم مع متغير الضغط الجوى على مستوى سطح البحر ومع شدة سطوع الشمس وأخيرا مع اتجاه الرياح. العنقود الثاني يجمع متغير الرطوبة النسبية مع سرعة الرياح.



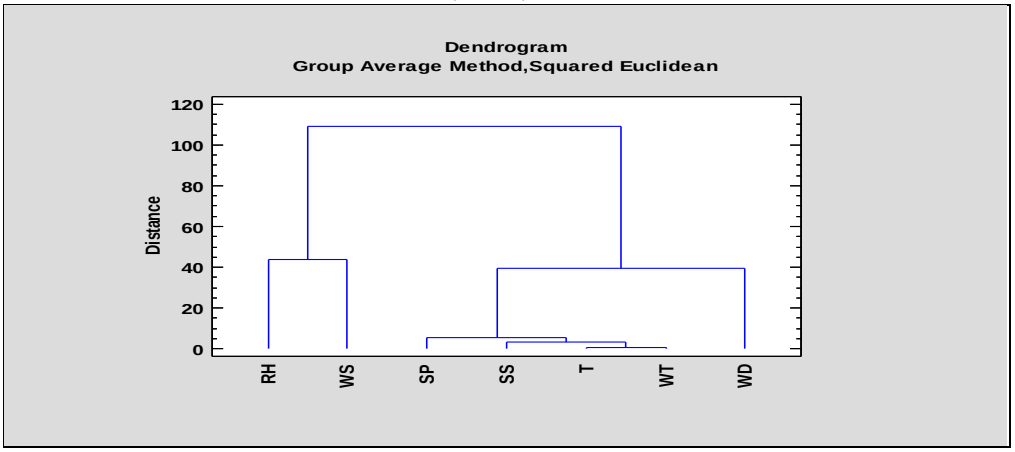
الشكل (2):المخطط الشجري حسب طريقة الجار الاقرب

المخطط الشجري في الشكل (3) يوضح عدد ومراحل تشكل العناقيد بطريقة الجار الأبعد ومربع المسافة الاقليدية، موضحا مراحل التعنقد يبدأ بعنقود درجة الحرارة مع درجة الحرارة الرطبة من ثم تعنقد مع شدة سطوع الشمس وتليها الضغط الجوى على مستوى سطح البحر وأخيرا مع اتجاه الرياح.أما العنقود الثاني فيجمع بين الرطوبة النسبية مع سرعة الرياح كما ورد أيضا في طريقة الجار الأقرب.



الشكل (3):المخطط الشجري حسب طريقة الجار الابعد (الربط الكامل)

المخطط الشجري في الشكل (4) يوضح عدد ومراحل تشكيل العناقيد بطريقة المتوسط ومربع المسافة الاقليدية، موضحا مراحل التعنقد التي تبدأ بعنقود درجة الحرارة مع درجة الحرارة الرطبة مع شدة سطوع الشمس ومن ثم ضغط البخار وأخيرا مع اتجاه الرياح. أما العنقود الثاني فيجمع بين الرطوبة النسبية مع سرعة الرياح كما ورد أيضا في طريقتي الجار الأقرب والجارالأبعد.



الشكل (4):المخطط الشجري حسب طريقة المتوسط (معدل الربط)

3.2 نتائج تحليل الانحدار الخطي المتعدد:

نتائج تحليل الانحدار الخطي المتعدد لمتغير التبخر معتمدا على نتائج التحليل العنقودي الهرمي ومقياس المسافة الاقليدية من خلال طرائق (الجار الاقرب، الجار الابعد و معدل الربط) في تحديد المتغيرات المستقلة تبين بأن المعادلة التنبؤية الأولى تضمنت درجة الحرارة الرطبة ودرجات الحرارة من ثم شدة سطوع الشمس، الضغط الجوي على مستوى البحر واتجاه الرياح والمعادلة الثانية شملت متغير الرطوبة النسبية وسرعة الرياح.

الجدول (6) يبين معايير المقارنة معادلتى الانحدار المتعدد من حيث نسبة التغير المفسرة من قبل المتغيرات المستقلة لكل معادلة، متوسط الخطأ التربيعي و جذر متوسط الخطأ التربيعي.

الجدول(6): نتائج تحليل الانحدار الخطي المتعدد للتنبؤ بالتبخر

Model	Input parameter	R^2	Adju. R^2	RMSE	MSE	Sig.
1	WT,T,SS,SP,WD	0.970	0.966	1.012	1.026	0.000
2	RH,WS	876.0	0.868	1.979	3.919	0.000

اعلى نسبة للتغير المفسرة من قبل متغيرات درجة الحرارة الرطبة ودرجة الحرارة، شدة سطوع الشمس، الضغط الجوي على مستوى البحر واتجاه الرياح بلغت قيمتها (% 96.6)، اقل قيمة لمربع متوسط الخطأ كانت (1.026) وجذر متوسط الخطأ التربيعي (1.013) للمعادلة التنبؤية الاولى :

$$\hat{E}1 = -6.685 - 1.474 * WT + 1.031 * T + 1.017 * SP + 0.277 * SS - 0.005 * WD \quad (11)$$

اماالمعادلة التنبؤية الثانية فكانت نسبة التغير المفسرة (% 86.8) من قبل متغيرات الرطوبة النسبية وسرعة الرياح ومربع متوسط الخطأ كانت قيمته (3.919) أما جذر متوسط الخطأ التربيعي فقيمته (1.979) للمعادلة الثانية. فيما يخص معنوية المعادلتين فقد أثبتا معنوياتهما فكانت قيمة الاحتمالية لرفض فرضية العدم (0.000).

$$\hat{E}2 == 19.649 - 0.287 * \mathrm{RH} + 2.564 * \mathrm{WS} \quad (12)$$

قيمة معدلات التبخر الحقيقية والقيم المقدرة لها في المعادلة التنبؤية الأولى و الثانية موضح في جدول (7).

الجدول(7):القيم الحقيقية و المقدرة لمعدلات التبخر في المعادلة التنبؤية الأولى و الثانية

#	E	$\hat{E}_1$	$\hat{E}_2$	#	E	$\hat{E}_1$	$\hat{E}_2$
1	1.600	1.713	2.857	19	16.300	15.701	14.504
2	1.800	1.457	3.944	20	14.600	15.066	12.895
3	2.200	2.565	3.139	21	12.800	13.459	12.986
4	4.300	4.619	3.774	22	7.800	8.743	11.290
5	10.300	10.250	11.588	23	3.600	4.081	5.156
6	14.700	14.335	14.838	24	2.300	2.544	2.293
7	14.000	14.761	13.587	25	2.300	1.785	0.595
8	18.500	16.459	14.584	26	15.600	12.977	13.731
9	11.600	12.050	13.784	27	17.200	17.126	13.755
10	9.700	9.683	10.755	28	15.600	15.198	15.530
11	4.800	5.102	8.350	29	11.300	11.020	12.236
12	2.200	2.946	1.369	30	8.000	8.549	9.672
13	1.800	1.146	0.845	31	4.100	5.335	6.944
14	4.300	1.991	1.846	32	15.600	15.198	15.530
15	4.700	3.866	3.113	33	11.300	11.020	12.236
16	6.400	6.476	4.583	34	8.000	8.549	9.672
17	9.900	10.793	7.773	35	4.100	5.335	6.944
18	14.300	16.218	12.605	36	2.300	1.785	0.595

يلاحظ بان القيم المقدرة لمعدلات التبخر في المعادلة التنبؤية الأولى اقرب الى القيم الحقيقية مقارنة بالقيم المقدرة للمعادلة التنبؤية الثانية خلال فترة الدراسة.

4. الاستنتاجات:

أظهرت الدراسة ان الأساليب المشتركة توجه العمل إلى إبراز نقاط القوة في كل تقنية، وأن التعنقد قبل تحليل الانحدار يؤدى إلى تحسين دقة التنبؤ من خلال تصنيف وتحديد متغيرات مستقلة متجانسة ضمن العنقود الواحد مختلفة عن باقي العناقيد.

من ثم تليها الخطوة الثانية بناء انموذج الانحدار لكل عنقود تم تشكيله، أي أن كل عنقود لديها انموذج انحدار محدد بما يضمن الحصول على أعلى قيمة R^2 المعدلة و اقل قيمة للاخطاء نتيجة استخدام التحليل العنقودي. لذلك فإن تقسيم المتغيرات المستقلة السبعة قيد

الدراسة على العنقود الأولى (الأنموذج الأول) للانحدار الخطي المتعدد، علاقة التبخر مع متغيرات مستقلة يشمل المتغيرات درجة الحرارة، درجة الحرارة الرطبة، شدة سطوع الشمس، الضغط الجوي على سطح البحر مع اتجاه الرياح. أظهرت علاقة التبخر للأنموذج الثاني مع متغيرين مستقلين هما الرطوبة النسبية وسرعة الرياح في العنقود الثاني.

أظهرت النتائج أن الأنموذج الأول المطور أثبت كفاءته وقدرته على التنبؤ وتفوقه على أهمية الأنموذج الثاني لتقدير التبخر في المناطق الجبلية. من جهة أخرى أثبت التحليل العنقودي أنه تقنية مناسبة لتطوير نماذج للتنبؤ بالانحدار الخطي المتعدد أكثر دقة لجميع التطبيقات الهيدرولوجية. بناء على نتائج هذه الدراسة، سيكون الجمع بين استخدام التحليل العنقودي والانحدارالمتعدد مفيدا في الدراسات التنبؤية.

5. Reference

1. Adnan, R. . M., Malik, A. & Kumar, A., 2019. Pan Evaporation Modeling by Three Different Neuro-Fuzzy Intelligent Systems Using Climatic Inputs. *Arabian Journal of Geosciences,* 12(606).

2. Ali, P. A. & Younas, . A. A., 2021. Understanding and Interpreting Regression Analysis. *Evid Based Nurs,* 24(4), pp. 116-118.

3. Almawla , A., 2017. Predicting the daily evaporation in Ramadi city by using artificial neural netwok.. *Anbar Journal of Engineering Science,* 7(2), pp. 134-139.

4. Almedeij, J., 2016. Modeling Pan Evaporation for Kuwait using Multiple Linear Regression and Tme-Series Techniques.. *American Journal of Applied Sciences,* 13(6), pp. 739-747.

5. Al-Mukhtar, M., 2021. Modeling the Monthly Pan Evaporation Rates Using Artificial Intelligence Methods: a Case Study in Iraq. *Environmental Earth Sciences,* 80(39).

6. Alsumaiei, A. A., 2020. Utility of Artificial Neural Networks in Modeling Pan Evaporation in Hyper-Arid Climates. *Water,* 12 (1508).

7. Best, H. & Wolf, C., 2015. Regression Analysis and Causal Inference. In: *Regression Analysis and Causal Inference.* London:Los Angeles : s.n.

8. Da Silva, H. d. S. M. J. J. S., 2016. Modeling of reference evapotranspiration by multiple linear regression.. *Journal of Hyperspectral Remote Sensing,* 6(1), pp. 44-58.

9. De Carvalho, F. d. A., S. G. & Queiroz, D. N., 2010. A Clusterwise Center and Range Regression Model for Interval-Valued Data.

10. Esmaeel , S. M. & Rashed, S. N., 2022. Detection of outliers in the linear regression model with application to well water pollution data on the. *Iraqi Journal of Statistical Sciences,* 19(1), pp. 76-84.

11. Essa, A. K., Fadhil, L. & Shihab, D. H., 2023. A comparison between the hierarchical clustering methods for postgraduate students in Iraqi universities for the year 2019-2020 using the cophenetic and delta correlation coefficients. *Periodicals of Engineering and Natural Sciences ISSN 2303-4521,Original Research,* 11(1), pp. 174-185.

12. Frost, J., 2023. *Statistics By Jim Making statistics intuitive*. [Online] Available at: https://statisticsbyjim.com/regression/mean-squared-error-mse/

13. Khattreer, R. & N, D. N., 2020. *Multivariate Data Reduction and Discrimination with SAS Software.* Cary, NC, USA: SAS Press and John WileyISBN.

14. Kor, . K. & Altun, G., 2020. Is Support Vector Regression method suitable for predicting rate. *Journal of Petroleum Science and Engineering,194.*

15. Mohammd, A. S., Said, M. A. M. & Kaml, A. H., 2022. Develop Evaporation Model Using Multiple Linear Regression in the Western Desert of Iraq. *International Journal of Design & Nature and Ecodyna,* 17(1), pp. 137-143.

16. Mohammed , M. . A. & AL-Rawi, D. A. . G., 2019. Using Some of Hierarchical Approach of Cluster Analysis for Classification of Agricultural Lands by Area and the Amount of Production for some Agricultural Crops in the Iraqi Governorates for the Years (2005) and (2010). *Journal of Al-Rafidain University College for Sciences,* 44(1), pp. 52-74.

17. Mohammed, F. A. R. & Hannon, O. B., 2019. Using the Hybrid MLR-GA Approach for Air Pollution Forecasting. *Iraqi Journal of Statistical Sciences Special Issue,* 16(2), pp. 25-36.

18. Ngo, T. H. & Puente, L., 2012. The Steps to Follow in a Multiple Regression Analysis. *SAS Global Forum, Statistics and Data Analysis,* p. 333.

19. Seber, . G. . A. & Lee, . A. J., 2012. *Linear Regression Analysis.*. s.l.:John Wiley, Sons.

20. Shahab, Z. . A. & Rashed, S. N., 2021. Using the linear and non-linear discriminant function with cluster analysis to study the level of education for the completed stages (governmental – private) In Nineveh Governorate. *Iraqi Journal of Statistical Sciences,* 18(1), pp. 89-104.

21. Taha, A. . H., 2022. Use of cluster analysis to study the reality of E-learning due to the Corona-19 pandemic on Nineveh Technical Institute students. *Entrepreneurship Journal for Finance and Business,* 3(4), pp. 40-51.

Combining Cluster Analysis with Multiple Linear Regression Analysis to Create the Most Accurate Prediction Model for Evaporation in the Kurdistan Region of Iraq

Bakhshan Ahmed Hamad
Department of Mathematics, College of Education, Salahaddin University, Erbil, Iraq

Abstract:

This study aims to build a prediction model for the influential variables of evaporation in the Kurdistan region - Iraq, using the concept of regression and cluster analysis. The methods common guide the work to highlight the strengths of each technique, and the possibility of using hierarchical cluster analysis (nearest neighbor, furthest neighbor, and median) to improve the predictive accuracy of regression models. The variables affecting the evaporation rate were classified using weather data from meteorological stations in the Kurdistan Region, Iraq for the period from January 2020 to December 2022, and The adjusted R^2, MSE, and RMSE values were used as indicators of the efficiency of the model's performance.

The study found that clustering before regression analysis leads to improve prediction accuracy by classifying and identifying homogeneous independent variables within one cluster that are different from the rest of the clusters.

Keywords: Cluster Analysis (CA), Multiple Liner Regression Analysis (MLR), Evaporation (E).