وزارة التعليم العالي والبحث العلمي
جامعة الموصل
كلية علوم الحاسوب والرياضيات
قسم الإحصاء والمعلوماتية

العنقدة الحصينة لبيانات المناخ ذات الأبعاد العالية

رسالة مقدمة

إلى مجلس كلية علوم الحاسوب والرياضيات في جامعة الموصل
وهي جزء من متطلبات نيل شهادة ماجستير
علوم في الإحصاء

من قبل
مـروان ميسر محمـود عبود

بإشراف
أ.م. د بشار عبد العزيز مجيد علي

1446هـ **2025م**

الملخص

تحظى التحليلات الإحصائية المتعلقة بدراسة التغيرات المناخية بأهمية متزايدة نظرًا لتأثيرها المباشر على اتخاذ القرارات الزراعية، خاصة في تحديد مواعيد الزراعة واختيار المتغيرات الأكثر تأثيرًا على تقلبات الطقس. كما تسهم هذه التحليلات في تشخيص الظواهر المناخية المتطرفة، مثل ارتفاع درجات الحرارة وزيادة سرعة الرياح، والتي قد تتسبب في حرائق أو أضرار للكائنات الحية. تعتمد الرسالة على طرق إحصائية حصينة للكشف عن عناقيد الضوضاء. إذ تم التركيز على مناخ مدينة الموصل للفترة من 2013 إلى 2022، وقد استخدمت عنقدة المتغيرات حول المكونات الكامنة Clustering of Variables Around Latent Components(CLV). يتميز هذا الأسلوب بمعالجة البيانات المناخية اليومية عبر ترتيب عرضي للبيانات، مما يجعل عدد المتغيرات أكبر من عدد المشاهدات ($P > N$)، وهو ما يُعرف ببيانات عالية الأبعاد. واستخدمت تحليل المكونات الرئيسة Principal Components Analysis(PCA) لتقليل الأبعاد، وتقنية البوتستراب (Bootstrap) لضمان استقرار العناقيد. اعتمدت الرسالة على مجموعة من المؤشرات لتقييم كفاءة العنقدة، منها مؤشر راند Rand Index (RI) ومؤشر راند المعدل Adjusted Rand Index (ARI) لصلاحية العنقدة الخارجية، بالإضافة إلى مؤشرات التماسك الداخلي للعناقيد (Cohesion Partition) والانفصال بين العناقيد (Isolation Partition). تم تحليل البيانات من خلال قياس التغيرات اليومية لمتغيرات مناخية كدرجات الحرارة والرطوبة لمدة 365 يومًا سنويًا على مدار عشر سنوات. ركزت الدراسة على فصلي الصيف والشتاء كونهم الأكثر تطرفًا عن باقي الفصول، وتم تحديد الأنماط المناخية الرئيسية لكل متغير مناخي بصورة منفصلة وكذلك الكشف عن المتغيرات الضوضاء(الأيام المناخية الشاذة) باستخدام طريقة عنقدة حصينة بديلة تعرف بـ "K+1". أظهرت النتائج ارتباطات قوية بين المتغيرات داخل العناقيد، فقد تجاوزت معاملات الارتباط 0.90، مما يدل على وجود أنماط سلوكية مشتركة بين هذه المتغيرات. كما تم تحديد المتغيرات الضوضائية باستخدام عتبات ارتباط أقل من 0.40. تعزز هذه النتائج قدرة المجتمعات المحلية على التكيف مع التغيرات المناخية من خلال تحسين دقة التنبؤ بالظواهر المناخية المتطرف.

Ministry of Higher Education and
Scientific Research
University of Mosul
College of Computer Science and Mathematics
Department of Statistics and Informatics

Robust Clustering of High Dimensional Climate Data

A Thesis Submitted to the Council of the College of
Computer Science and Mathematics
University of Mosul
as a Partial Fulfillment of Requirements
for the Degree of Master of Science
in Statistics

By

Marwan Myasar Mahmood Aboud

Supervised by

Assist. Prof. Dr. Bashar Abdul-Aziz Majeed Ali

1446 A.H. 2025 A.D.

Abstract

Statistical analyses related to studying climate changes are gaining increasing importance due to their direct impact on agricultural decision-making, particularly in determining planting schedules and identifying the most influential variables affecting weather variability. These analyses also contribute to diagnosing extreme climatic phenomena, such as rising temperatures and increased wind speeds, which may lead to fires or harm living organisms.

This thesis employs robust statistical methods to detect noise clusters, focusing on the climate of Mosul City for the period from 2013 to 2022. The method of Clustering of Variables Around Latent Components (CLV) was used, which is distinguished by processing daily climate data in a cross-sectional arrangement where the number of variables exceeds the number of observations (P > N), a characteristic of high-dimensional data. Principal Components Analysis (PCA) was applied for dimensionality reduction, and the bootstrap technique was used to ensure cluster stability.

The study relied on several indicators to assess clustering efficiency, including the Rand Index (RI) and the Adjusted Rand Index (ARI) for external clustering validity, as well as indicators for internal cluster cohesion (Cohesion Partition) and separation between clusters (Isolation Partition). The data analysis involved measuring daily changes in climatic variables such as temperature and humidity over 365 days annually for ten years. The study focused on the summer and winter seasons, as they exhibit the most extreme variations compared to other seasons.

The primary climate patterns for each climatic variable were identified separately, and noise variables (anomalous climatic days) were detected using a robust alternative clustering method known as “K+1.” The results revealed strong correlations among the variables within the clusters, with correlation coefficients exceeding 0.90, indicating shared behavioral patterns among these variables. Noise variables were identified using correlation thresholds below 0.40.

These findings enhance local communities' ability to adapt to climate changes by improving the accuracy of predicting extreme climatic events.

.