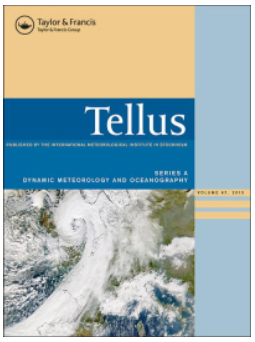

Tellus A: Dynamic Meteorology and Oceanography

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/zela20

Identification of homogeneous rainfall regions in New South Wales, Australia

Shahid Khan, Ijaz Hussain & Ataur Rahman

To cite this article: Shahid Khan, Ijaz Hussain & Ataur Rahman (2021) Identification of homogeneous rainfall regions in New South Wales, Australia, Tellus A: Dynamic Meteorology and Oceanography, 73:1, 1-11, DOI: 10.1080/16000870.2021.1907979

To link to this article: https://doi.org/10.1080/16000870.2021.1907979

Tellus
SERIES A
DYNAMIC
METEOROLOGY
AND OCEANOGRAPHY
PUBLISHED BY THE INTERNATIONAL METEOROLOGICAL INSTITUTE IN STOCKHOLM

Identification of homogeneous rainfall regions in New South Wales, Australia

By SHAHID KHAN[1], IJAZ HUSSAIN[1]*, and ATAUR RAHMAN[2], [1]*Department of Statistics, Quaid-i-Azam University, Islamabad, Pakistan;* [2]*Centre for Infrastructure Engineering, Western Sydney University, Australia*

(Manuscript Received 26 February 2021; in final form 22 March 2021)

ABSTRACT

Identifying homogeneous regions based on spatial variables is vital for providing a certain and fixed region's spatial and temporal behavior. However, a significant problem of non-separation rises when the geographic coordinates are utilized for clustering, just because the Euclidean distance is not suitable for clustering when considering the geographic coordinates. Therefore, this study focuses on employing such methods where the non-separation is minimum for identifying homogenous regions. The average annual rainfall data of 226 meteorological monitoring stations for 1911–2018 of New South Wales (NSW), Australia, was considered for the current study. The data is standardized with zero mean and unit variance to remove the effect of different measurement scales. The geographical coordinates are then converted to rectangular coordinates by the Lambert projection method. Using the Partition Around Medoid (PAM) algorithm, also known as the k-medoid algorithm (which minimizes the sum of dissimilarities instead of the sum of squares of Euclidean distances) on rectangular Lambert projected coordinates, 10 well-separated clusters are obtained. The Mean Squared Prediction Error (MSPE) is comparatively smaller if the prediction of unobserved locations in cluster 3 is made. However, this error increases if the prediction is made for a complete monitoring network. The identified 10 homogeneous regions or clusters provide a good separation when the lambert coordinates are used instead of geographical coordinates.

Keywords: New South Wales, precipitation, partition around medoid clustering algorithm, Lambert projection method, geographical coordinates, Lambert coordinates

1. Introduction

Hydrological variables carry a vital and key role in the management of water resources. Among different hydrological variables, precipitation is one of the prime and crucial variables. For example, the scarcity of precipitation affects irrigation and public drinking water supply, while the surplus of the precipitation in a specific region may cause floods and soil erosion. Lack of monitoring and spatial and temporal information on precipitation significantly affects the planning and water resources management, particularly water supply reservoirs, irrigation projects, flood control systems, drought monitoring systems, and urban drainage design. Thus, it is essential to know the spatial and temporal variability of precipitation for efficient planning and management of water resources.

This study focuses on the precipitation of New South Wales (NSW) state in Australia. NSW has a diverse spatial and seasonal climate. The NSW's rainfall is highly affected by ocean winds and mountain ranges called the Great Dividing Range (GDR). Considering such factors, the climate regions of NSW are further divided into four categories: the coastal belt, the ranges and tablelands of the GDR, the western slopes and plains, and the arid plains (Meteorology, 2020).

The coastal belt receives the most rainfall in NSW (800–2000 mm) per year. Peaks along with the GDR are usually cooler than the rest of NSW. The GDR belt receives a moderate to a high volume of rainfall after the coastal belt, ranging from 600 mm to 1,500mm per year. In winter, snowfall and frosts are not uncommon in this belt. However, in the summer months, the belt is warmer, but not as hot as the rest of the state. The western slopes and plains are the central band of NSW, experience moderate rainfall. On average, the area receives a 300 mm to 1000 mm rainfall per annum. The arid plains on the very west side of NSW have a particularly harsh and hot

*Corresponding author. e-mail: ijaz@qau.edu.pk

Citation: Tellus A: 2021, **73**, 1907979, https://doi.org/10.1080/16000870.2021.1907979

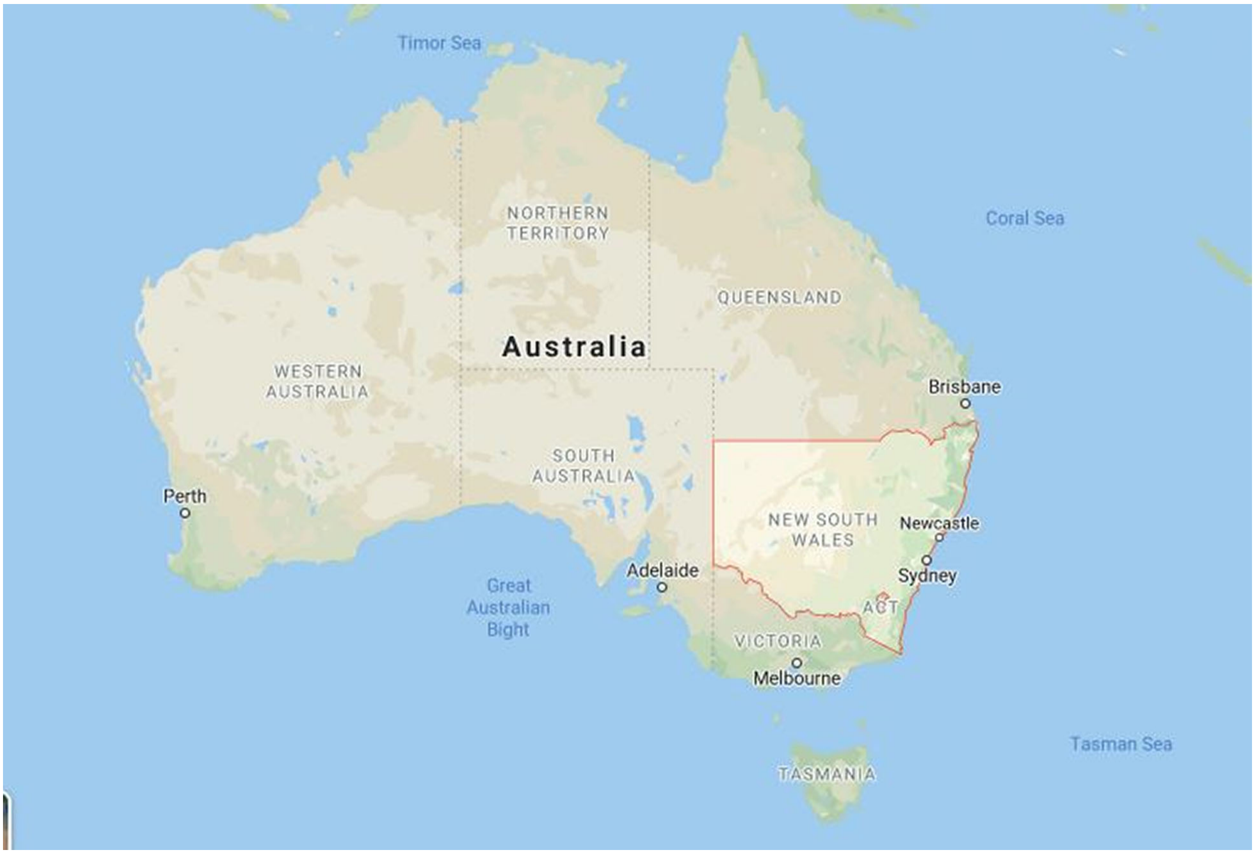

Fig. 1. Map of New South Wales (NSW), Australia, surrounded by Queensland, Victoria, South Australia, Coral, and Tasman seas.

climate. Rainfall in this part of the state averages from 150 mm to 500 mm in a year. Hot temperatures in summer and freezing nights in winters with frequent droughts and water shortage during dry months are not uncommon in this belt (Bushmans, 2020; Meteorology, 2020).

Identifying the homogeneous regions based on precipitation is a key tool for providing the spatial and temporal behavior of precipitation. Additionally, the homogeneous rainfall regions' determination is one of the significant and essential steps towards obtaining regional rainfall patterns. Such homogeneous sub-regions can help estimate the total rainfall, predict the rainfall in the sub-regions and optimize the number of monitoring sites. The term *homogeneous regions* refer to regions with some hydrological similarity (Patil and Stieglitz, 2011; Wazneh et al., 2013; Swain et al., 2016).

Several studies have been carried out to recognize homogenous regional rainfall, like the study carried by Hussain et al. (2011), which identified homogeneous climate regions in Pakistan using the Partition Around Medoid (PAM) algorithm. Similarly, the study developed by Dikbas et al. (2012), compared the Fuzzy c-means and k-means clustering method and noted that the Fuzzy c-means method gives promising results for homogeneous regions formation. Both the Fuzzy c-means and k-means clustering methods were also compared by Goyal and Gupta (2014) for precipitation in Northeast India, and they concluded that the Fuzzy c-means provide better results in the formation of homogenous regions. Sadri and Burn (2011) consider the L-moments and Fuzzy c-means method for identifying homogenous regions of rainfall in the Canadian province of Alberta, Saskatchewan, and Manitoba. Using Fuzzy c-means Satyanarayana and Srinivas (2011) regionalized and recognized twenty-four homogeneous precipitation regions throughout the Chinese territory. For medoids-based clustering, Estivill-Castro and Murray (1997) developed a genetic heuristic algorithm based on genetic recombination upon random assorting recombination. Using hierarchical and divisive cluster analysis Soltani and Modarres (2006) were able to categorize twenty-eight rainfall monitoring stations into eight homogeneous regions in Iran. They considered only average rainfall at the sites and did not consider spatial coordinates. However, while clustering the spatial data, ignoring the spatial coordinates and time replications may lead to false and misleading results. To overcome such a problem, Kerby et al. (2007) developed a spatial clustering method

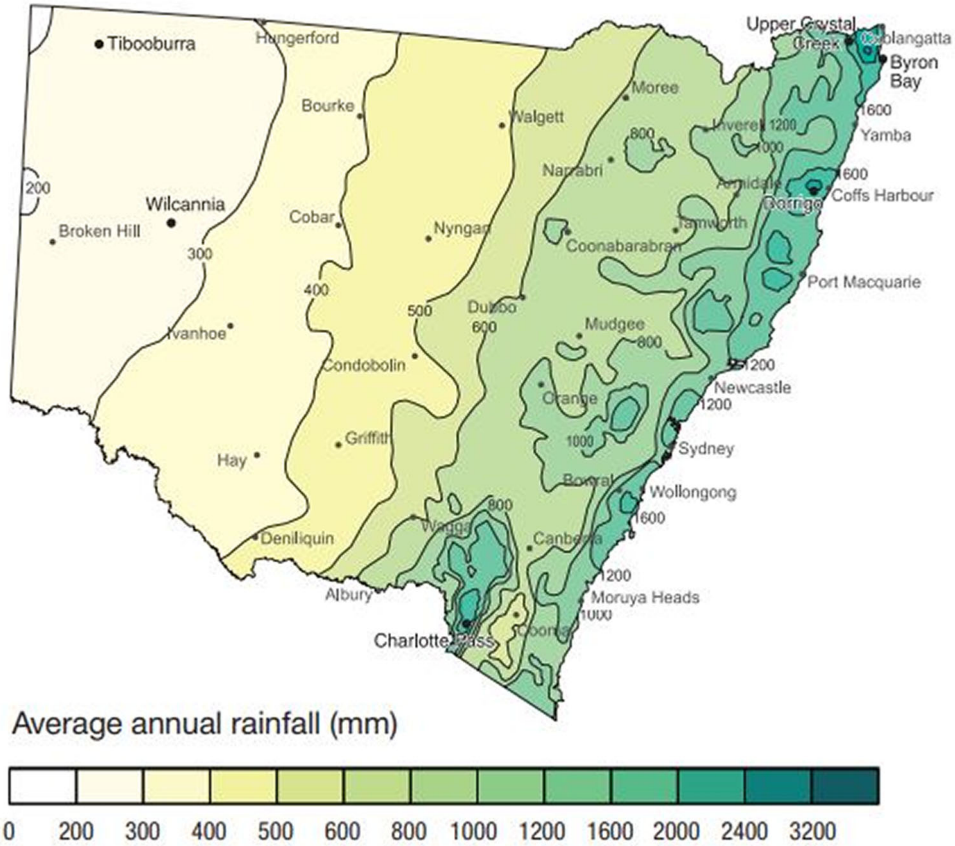

Fig. 2. Four distinct climate zones: the coastal belt, the ranges and tablelands of the Great Dividing Range, the Western slopes and plains, and the arid plains of New South Wales, Australia.

considering the likelihoods. This novel method considers spatial coordinates by means of assessing the variance-covariance matrix between observations. However, such a method fails and does not perform well for data with time replication and data with many sites. Identifying homogeneous regions in data having outliers, noisy data, and having auto-correlation in spatial data, a robust weighted kernel k-means algorithm was developed by Sap and Awan (2005). This novel algorithm can handle noisy data, outliers, and auto-correlation in data, more efficiently and effectively than any other algorithm.

In the current study, using the simple clustering approach, a separable and homogeneous region of precipitations was identified within NSW. Mainly most of the clustering methods consider Euclidian distance between samples. Since it is obvious that the geographic coordinates are spherical, Euclidean distance with spherical coordinates is inappropriate to be used. So, a Lambert projection method was adopted to transform the geographical to rectangular coordinates. After transforming the coordinates, PAM was applied to the transformed data for identifying homogenous regions. Additionally, OK was implemented on a single cluster (cluster 3) to validate clustering performance measures.

2. Materials and methods

2.1. Study area

The considered region (NSW state in Australia) for the study lies at latitude 32° and longitude 147°. The region is surrounded by Queensland, Victoria, South Australia, Coral, and Tasman seas, as shown in Fig. 1. The state has a total area of 810,000 km^2, making it one of the country's smallest states. However, by population, NSW is the only state in Australia having 8 million populations. The average rainfall in NSW is 554.5 mm per annum (Bushmans, 2020). However, with much diversity in the state varying from seashores and peak mountains to dry and arid plains, the state climate is diverse. Due to this, the state is further divided into four sub-categories, as shown in Fig. 2.

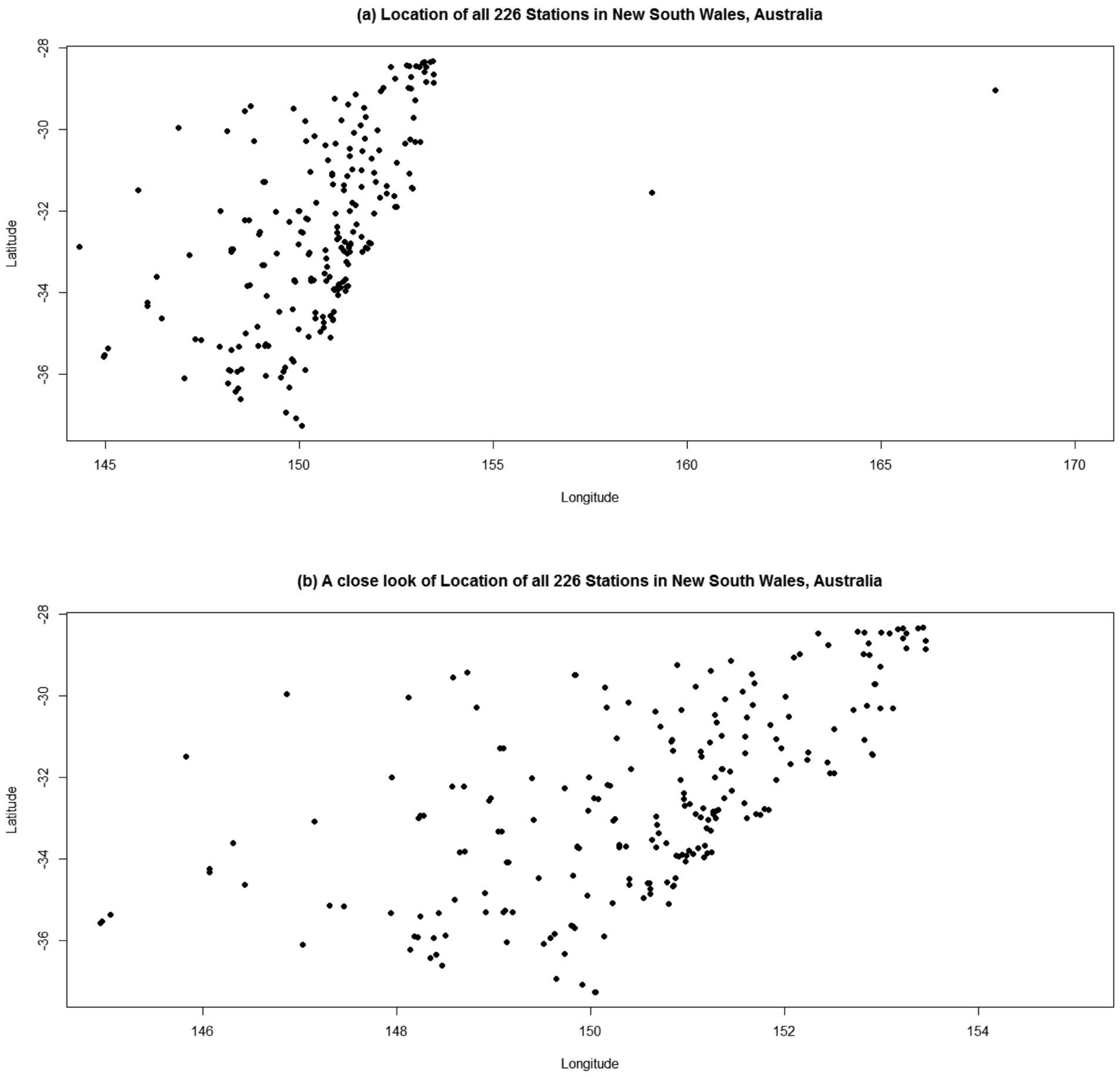
(a) Location of all 226 Stations in New South Wales, Australia
Latitude
-28
-30
-32
-34
-36
Longitude
145
150
155
160
165
170
(b) A close look of Location of all 226 Stations in New South Wales, Australia
Latitude
-28
-30
-32
-34
-36
Longitude
146
148
150
152
154

Fig. 3. (a) Location of 226 rainfall stations in New South Wales, Australia. (b) A close look at the location of 223 stations, excluding two stations from Lord Howe Island and one from Norfolk Island, Australia.

2.2. Data description

Rainfall data of 226 rainfall gauging stations were obtained from the Bureau of Meteorology, Australia (BOM, 2020). All the 226 locations of the gauging stations are shown in Fig. 3. Average annual rainfall data were considered for all the stations. For spatial clustering, along with the average annual rainfall variable, both the geographic coordinates (Latitude, Longitude) and transformed Lambert coordinates (transformed Lambert Latitude, transformed Lambert Longitude) of the monitoring stations are considered. To remove the effect of different measurement scales, the average annual rainfall data along with the coordinates are standardized to zero mean and unit variance. Lastly, as the geographic coordinates play a crucial role in spatial data, therefore a 50% weight is assigned to the geographic coordinates, and the remaining 50% weight is assigned to the average annual rainfall variable.

2.3. The partition around medoids clustering (PAM)

Spatial clustering algorithms can be classified into four basic categories: partition-based, hierarchal-based, density-based, and grid-based (Mandal et al., 2007; Hamad-

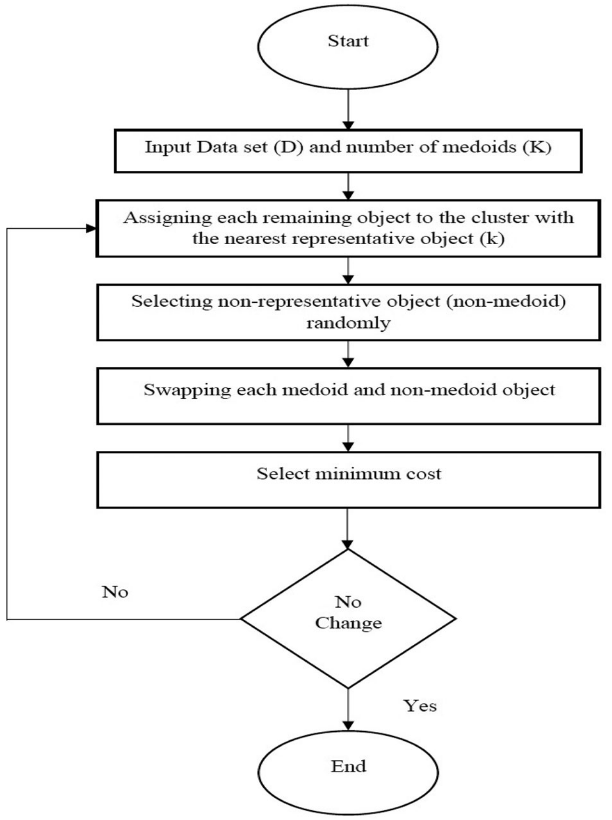
Start
Input Data set (D) and number of medoids (K)
Assigning each remaining object to the cluster with the nearest representative object (k)
Selecting non-representative object (non-medoid) randomly
Swapping each medoid and non-medoid object
Select minimum cost
No Change
No
Yes
End

Fig. 4. A flow chart of iterative steps of the Partition Around Medoid (PAM) algorithm.

Ameen, 2008). Our main objective is to identify and discover the homogenous locations hidden in the data, so among all the clustering partitioning-based algorithm was found to be the most suitable method for our study. The partitioning-based clustering algorithms are further categorized into two major categories, the k-mean, and k-medoid methods. The k-mean and k-medoid methods are based on randomly partitioning the database into k subsets and rectifying the cluster centers to reduce the cost function. The spatial domain's considered cost function is just the sum of each data point's squared distances to its assigned centers. All the remaining data points are assigned to the nearest centers. The k-means algorithms being one of the initial clustering algorithms are known for their quick termination. It is easy to implement and understand. The cluster centers in the k-means method are considered the gravity center of all the data points in the same cluster.

Additionally, in regular planar space, the gravity center of the cluster assures the minimum sum of the distance between the cluster members and itself. However, in obstacle planner space, the gravity center does not behave the same (Nanopoulos et al., 2001). The k-medoid algorithm overcomes such a problem. An actual object in the cluster is chosen as a cluster representative (medoid), instead of some gravity centers in such algorithms. Use of the actual nominee in k-medoids results to decrease the sensitivity of outliers. Furthermore, this technique also assures the accessibility of the center by all data objects within the cluster.

The PAM algorithm, also known as the k-medoid algorithm, is one of the most accurate and used algorithms in partitioning-based clustering. The PAM was first developed by Kaufman and Rousseeuw in 1990 (Kaufman and Rousseeuw, 2009). As from PAM's name, it is obvious that this algorithm represents a cluster by a medoid (Dunham, 2003). Unlike the k-means algorithm, the k-medoid minimizes the sum of dissimilarities instead of the sum of squares of Euclidean distances and is more robust in nature.

The very first step in performing PAM clustering is to select the medoids. Once the medoids have been selected, each non-selected object (O_i) is grouped with the most similar and selected medoids (O_j). More broadly, O_i belongs to a cluster represented by O_j if and only if $d(O_j,\ O_i)\ =\ min\ O_e\ d(O_i,\ O_e)$, where $min\ O_e$ is minimum overall medoid O_e and $d(O_1,\ O_2)$ is the dissimilarity or distance between O_1 and O_2. All the dissimilar values are used as inputs for PAM. Finding k-medoids, the PAM initializes with an arbitrary and random k object. Then at every step, a swap between the selected object O_j and a non-selected object O_i is made till the swap result in an improvement of the quality of the clustering.

2.3.1. Algorithm of partitioning around medoid. The PAM being the most common realization of the k-medoids clustering works as follows.

1. Initialize by randomly choosing k representative of n data points as medoids.
2. Join up each non-selected object (non-medoid data point) O_i to the closest and selected object (selected medoid) O_j.
3. For all pairs of O_i and O_j, compute the total cost change TC_{ji}.
4. Select the pair O_j, O_i which corresponds to $\min TC_{ji}(O_j, O_i)$. However, if the minimum TC_{ji} is negative, replace O_j with O_i, and go back to step 3.
5. Otherwise, for each O_i, find the most similar representative object.
6. Repeat steps 2 to 3 until there is no change in the medoid. Halt.

The experimental results prove that the PAM only performs better for small data sets (e.g. 200 datasets in 5 clusters). The algorithm fails to show promising results in larger data sets. While applying PAM for clustering in step 3 and 4, a total of $k(n-k)$ pairs of $O_j,\ O_i$ is formed.

So, for each pair, computing TC_{ji} requires the inspection of $(n-k)$ non-selected objects. Combining Steps 2 and 3 a $O(k(n-k)^2)$ complexity is required for only one iteration. So, for a larger value of n and k, it is obvious that the PAM becomes too costly. Thus, an alternative method for such larger data sets, a Clustering LARge Applications (CLARA), was developed. The iterative steps of the PAM algorithm can also be shown in Fig. 4.

2.4. The lambert projection method

As the earth has a spherical shape, so the geographical coordinates are spherical, too. Using the Lambert conformal conic projection, the spherical coordinates can be transformed into a rectangular form (Snyder, 1987). The geographical coordinates can be transformed into rectangular coordinates by the following steps.

Step 1: Conversion of geographical coordinates to radians

$$\lambda_0 = \frac{Origin\ of\ latitude \times \pi}{180},$$
$$\lambda_1 = \frac{Latitude \times \pi}{180},$$
$$\lambda_2 = \frac{Longitude \times \pi}{180},$$

where,

$$Origin\ of\ latitude = \frac{maximum\ latitude - minimum\ laltitude}{2},$$
$$\pi = 3.1416,$$
$$\Phi_1 = \frac{(Origin\ of\ latitude - 0.3 \times Range\ of\ latitude) \times \pi}{180},$$
$$\Phi_2 = \frac{(Origin\ of\ latitude + 0.3 \times Range\ of\ latitude) \times \pi}{180},$$
$$\theta = n(\lambda_1 - \lambda_2).$$

Step 2: Since the earth's surface has an elliptical shape, then determining the meridian distance, which is the distance from the equator to a point at a latitude on the ellipsoid, will be

$$D = \frac{(1-e^2)}{\left(1-e^2\ \sin\ \Phi^2\right)^{\frac{3}{2}},}$$

where, $e = \sqrt{\frac{(a^2-b^2)}{a^2}},$ is the eccentricity, $a = 6378137,$ is called the length of the significant radius or semi-major axis, and $b = 6356752,$ is the length of the minor radius or semi-minor axis. a will always be greater than b that is $a > b$.
Step 3: The coordinates can be converted to D-plane coordinates by

$$x = \rho \sin(\theta),$$
$$y = \rho_0 - \rho \cos(\theta),$$

where x is the D-plane longitude and y is the D-plane latitude and

$$\rho = D \times F\ \cot^n\left(\frac{\pi}{4} + \frac{\lambda_1}{2}\right),$$
$$\rho_0 = D \times F\ \cot^n\left(\frac{\pi}{4} + \frac{\lambda_0}{2}\right),$$

where,

$$F = \frac{\cos\ \Phi_1\ \tan^n\left(\frac{\pi}{4} + \frac{\Phi_1}{2}\right)}{n},$$
$$n = \frac{ln(\cos\ \Phi_1 \sec\Phi_2)\cos\ \Phi_1\ \tan^n\left(\frac{\pi}{4} + \frac{\Phi_1}{2}\right)}{ln\left[\tan\left(\frac{\pi}{4} + \frac{\Phi_2}{2}\right) cot\left(\frac{\pi}{4} + \frac{\Phi_1}{2}\right)\right]},$$

Step 4: For appropriate scaling, the D-plane coordinate equation can be multiplied with some constant. However, such multiplication will not affect the results, it will only visualize the D-plane coordinate a better way. Though it is an optional step, one can multiply the D-plane results by 0.001 for better visualization

$$x = 0.001 \times\ \rho \sin(\theta),$$
$$y = 0.001 \times \rho_0 - \rho \cos(\theta).$$

2.5. Ordinary kriging

Kriging is the estimation or prediction of unknown values at of a random variable, Z, at one or less unsampled points from less or more sparse sample data, say $Z(s_1),\ Z(s_2),\ Z(s_3), \ldots,\ Z(s_n),$ at point $s_1,\ s_2,\ s_3, \ldots,\ s_n$. More precisely, kriging is used to interpolate random fields Z at unobserved locations (Matheron, 1963). In 1951, a South African engineer, D.G. Krige, laid the foundations for kriging and was named kriging after his name. However, in 1960, nine years after the foundation of kriging, kriging's main developments came from G. Matheron. So far, several spatial interpolation methods or types of kriging are discovered like Ordinary Kriging (OK), Universal Kriging (UK), Simple Kriging (SK), Lognormal Kriging (LK), Regression Kriging (RK), Indicator Kriging (IK), Disjunctive Kriging (DK), Co-Kriging (CK), and Multiple Indicator Kriging (MIK). All these kriging types have their advantages and disadvantages and can be used in different conditions related to the problem's sort and nature. OK, one of the most common types of kriging is based on the assumption of an unknown constant mean. The kriging prediction of Z at a point s_0 by $\hat{Z}(s_0)$ is given by

$$\hat{Z}(s_0) = \sum_{i=1}^{n} \lambda_i Z(s_i),$$

where, $\hat{Z}(s_0)$ is the estimated value at a point $s_0, Z(s_i)$ are the observed values at points $s_i,$ n is the sample size, and λ_i are weights chosen for s_i to satisfy the following two statistical conditions.

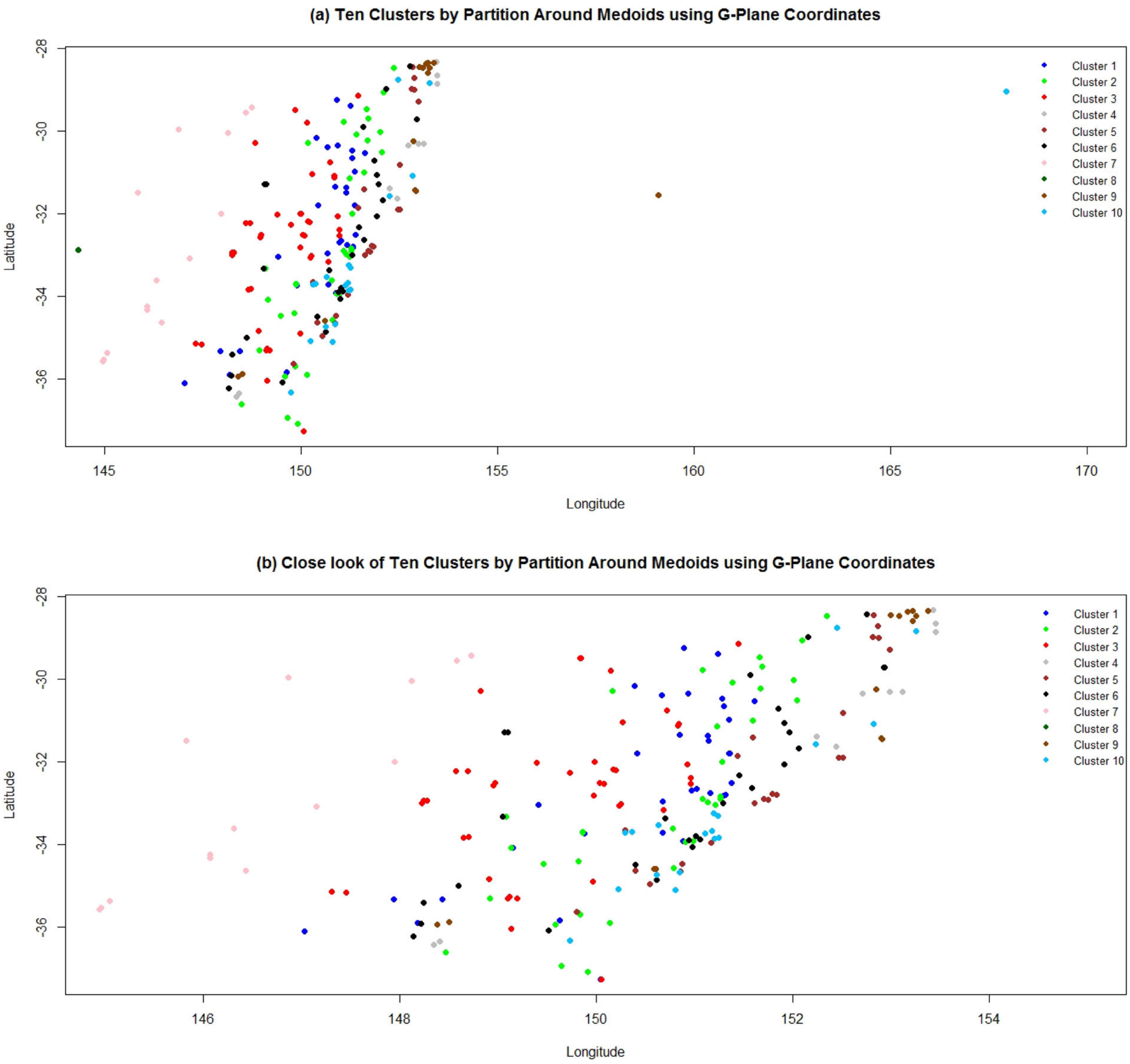

Fig. 5. (a) Map of allocations of 226 rainfall stations to 10 clusters by Partition Around Medoid (PAM) using G-plane coordinates, the different colors represent the memberships of stations to a specific cluster. (b) A close look into the allocations of 223 rainfall stations to 10 clusters by Partition Around Medoid (PAM) using G-plane coordinates.

2.5.1. Unbiasedness. Ensuring the unbiasedness of the estimate, the sum of weights is made equal to 1

$$\sum_{i=1}^{n} \lambda_i = 1,$$

and the expected error is

$$E\left[\hat{Z}(s_0) - Z(s_0)\right] = 0.$$

2.5.2. Minimum variance.

$$var\left[\hat{Z}(s_0) - Z(s_0)\right] = minimum.$$

The optimum weights λ_i can be obtained by solving the following equations simultaneously

$$\begin{bmatrix} C_{11} & C_{21} & \ldots & C_{N1} & 1 \\ C_{12} & C_{22} & \ldots & C_{N2} & 1 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ C_{1N} & C_{2N} & \ldots & C_{NN} & 1 \\ 1 & 1 & \ldots & 1 & 0 \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \ldots \\ \lambda_N \\ \mu \end{bmatrix} = \begin{bmatrix} C_{1s_0} \\ C_{2s_0} \\ \ldots \\ C_{Ns_0} \\ 1 \end{bmatrix},$$

which assures that the OK predicator is a minimum variance unbiased predictor.

where, C_{ij}, i, $j = 1, 2, 3, \ldots, N$ are the covariates between the data, C_{is_0} $i = 1, 2, 3, \ldots, N$ are the covariates

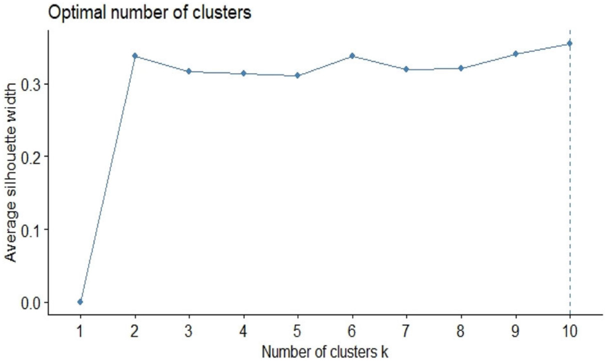

Fig. 6. Average silhouette width comparison for selecting the optimal number of clusters.

between the datum to be predicated and the observed data, and μ is the Lagrange multiplier accounting for unbiasedness. Lastly, for predication of unobserved locations, the variance can be determined by

$$\sigma^2(s_0) = \sigma^2_{(s_0)} - \lambda^t C_{s_0} - \mu_0,$$

where, $\sigma^2_{(s_0)}$ is the variance of $Z(s_0)$, $\lambda = (\lambda_1, \lambda_2, \lambda_3, ..., \lambda_N)^t$, and $C_{s_0} = (C_{1s_0}, C_{2s_0}, C_{3s_0}, ..., C_{Ns_0})^t$.

3. Results and discussion

3.1. Clustering by PAM using geographic coordinates

For identifying homogeneous clusters in NSW, Australia, several clustering methods like the Fuzzy c-means clustering method, the k-means technique, PAM, and CALRA were applied to the selected data. However, PAM giving the most promising results was considered for both geographic and rectangular coordinates. The results and comparison of PAM using geographic and rectangular coordinates are discussed here.

Considering the geographic coordinates, the PAM clustering method was applied to the average annual rainfall data. The different sites show membership to more than one cluster. Using the geographic coordinates, the allocation of rainfall stations to concern clusters is indicated by numerical numbers in Fig. 5. All 226 rainfall sites were classified into ten clusters as suggested by silhouette comparison. All the variables affecting climate, being space-time fields (measured only concerning time and space), should be spatially dependent. Hence all the sites of different clusters should be spatially separable. However, as we are using G-plane coordinates, the sites are mixed up with each other, showing no separation. Figure 5 shows that all the rainfall sites being classified into ten clusters have minimum separation.

3.2. Clustering by PAM using rectangular coordinates

After a failed attempt to separate clusters from geographic coordinates, all the geographic coordinates were transformed to rectangular coordinates using Lambert's conformable conic projection method. Besides the coordinates' transformation, the transformed rectangular coordinates and average annual rainfall were standardized to zero mean and unit variance. Once the data was standardized, the PAM clustering technique was applied for identifying the homogeneous sites.

The optimal number of clusters K is selected based on silhouette criterion, presented in Fig. 6. Using rectangular coordinates, the average silhouette for K = 10 is comparatively higher than K = 9, 8, 7, …, 1, and so a total of ten clusters were selected to be optimal for our study.

While using PAM with rectangular coordinates, all 226 monitoring sites were allocated to ten clusters. It can be clearly shown in Fig. 7. Using rectangular coordinates shows more separable homogeneous clusters as compare to geographic coordinates.

The center location or the medoid of the clusters can be shown in Table 1. Cluster 1 contains 34 rainfall stations located in the southernmost part of NSW, bordering Victoria. Cluster 1 receives moderate average rainfall of 797 mm annually. Cluster 2, which lies towards the east in GDR, consists of 36 rainfall stations and receives a bit higher average rainfall (839 mm) annually.

Cluster 8 is the collection of rainfall stations of the westernmost part of NSW. All the sites of cluster 8 lie in the arid plains, so this cluster receives the lowest average rainfall (381 mm) annually. Twenty-nine rainfall stations adjacent to cluster 8 on the eastside constitute cluster 3. As cluster 3 is far away from sea and GDR, this cluster receives the second-lowest average rainfall in NSW (590 mm) annually.

Cluster 9 consists of seven rainfall stations situating on the lower side of GDR and receives the maximum average rainfall (1628 mm). Seventeen stations on the northernmost side of NSW and thirty-one stations from the same part of NSW adjacent to the coastal belt constitute cluster 4 and cluster 7, respectively. These clusters receive an average rainfall of 1686 mm and 1192 mm, receptively. Cluster 6 consists of sixteen stations, lies on the northernmost side of the GDR of NSW. Because of GDR and approximately high elevation, this cluster also receives a little higher average annual rainfall (1060 mm). Two stations form the Lord Howe Island, and one station of Norfolk Island constitute cluster 10. All the three sites of cluster 10 lie in small islands, situated on the East side of NSW, between the Coral and Tasman seas. Due to the effect of both seas, this cluster receives the third most

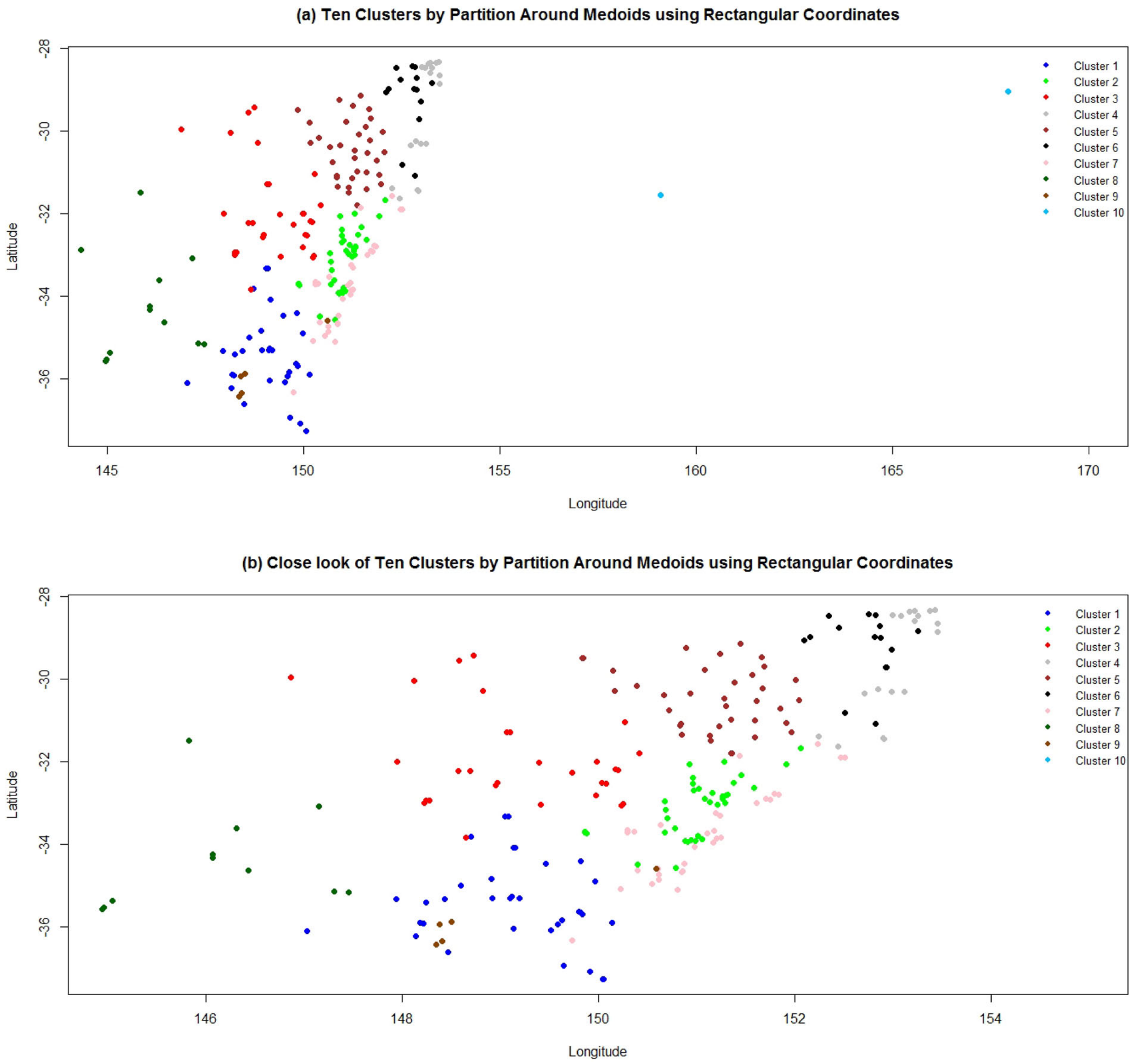

Fig. 7. (a) Map of allocations of 226 rainfall stations to 10 clusters by Partition Around Medoid (PAM) using rectangular coordinates; the different colors represent the memberships of sites to a specific cluster. (b) A close look into allocations of 223 rainfall stations to 10 clusters by Partition Around Medoid (PAM) using rectangular coordinates.

Table 1. Summary information of all ten cluster's medoids using Partition Around Medoid.

Cluster no.	Medoid cluster	Station name	G-Lat	G-Lon	D-Lat	D-Lon	Average annual rainfall
Cluster 1	201	Uriarra Forest	−35.2994	148.9222	523.0008	−292.438	814.2
Cluster 2	216	Wollombi (Blair)	−32.9667	151.1333	331.8256	−25.2381	825.2
Cluster 3	65	Dunedoo Post office	−32.0165	149.3956	498.4693	73.11653	612.4
Cluster 4	85	Green Pigeon (Morning View)	−28.4738	153.0861	158.0061	477.3265	1632.2
Cluster 5	200	Uralla (Lana)	−30.6417	151.3002	324.6122	232.7578	773.3
Cluster 6	196	Upper Mongogarie (Kimberley)	−28.9667	152.8167	183.3333	422.1592	1075.9
Cluster 7	176	Sydney (Observatory Hill)	−33.8607	151.205	321.889	124.0043	1215.7
Cluster 8	87	Griffith Airport Aws	−34.2487	146.0695	790.0565	−193.941	397.6
Cluster 9	35	Cabramurra Smhea	−35.9383	148.3842	567.3771	−365.959	1681.5
Cluster 10	116	Lord Howe Island Aero	−31.5421	159.0786	−412.978	129.6496	1478.6

Table 2. Summary statistics of all ten clusters by Partition Around Medoid using rectangular coordinates.

Cluster no.	Sample size	Average	Standard deviation	CV (%)	Minimum	Maximum
Cluster 1	34	797.76	140.14	17.57	566.50	1058.60
Cluster 2	36	839.81	123.91	14.75	585.80	1035.50
Cluster 3	29	590.95	79.49	13.45	411.80	785.30
Cluster 4	17	1686.14	165.20	9.80	1421.30	2015.30
Cluster 5	37	785.77	128.79	16.39	538.80	1080.70
Cluster 6	16	1060.93	122.11	11.51	856.10	1262.30
Cluster 7	31	1192.03	106.95	8.97	1002.20	1424.70
Cluster 8	16	381.38	107.87	28.28	225.80	577.50
Cluster 9	7	1628.63	173.74	10.67	1288.70	1808.90
Cluster 10	3	1468.90	175.05	11.92	1289.20	1638.90
Complete Data	226	921.20	362.34	39.33	225.80	2015.30

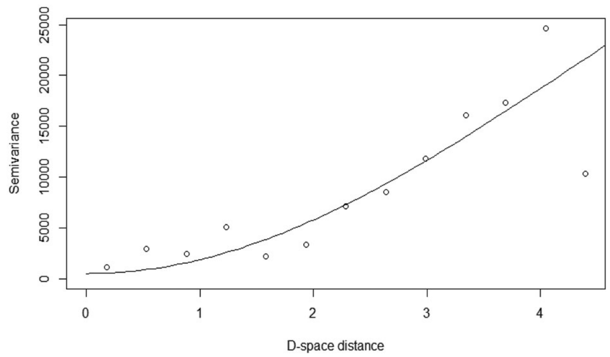

Fig. 8. Fitted Gaussian variogram for Cluster 3 using rectangular coordinates.

average annual rainfall (1468 mm). The detailed exploratory analysis of all the clusters can be seen in Table 2. The coefficient of variation of all the clusters is smaller than the coefficient of variation for the complete data set, suggesting that the heterogeneity in the data is reduced by clustering. The coefficient of variation for the complete data set is 39.33%, while the highest coefficient of variation in the cluster is that of cluster 8, which is 28.28%.

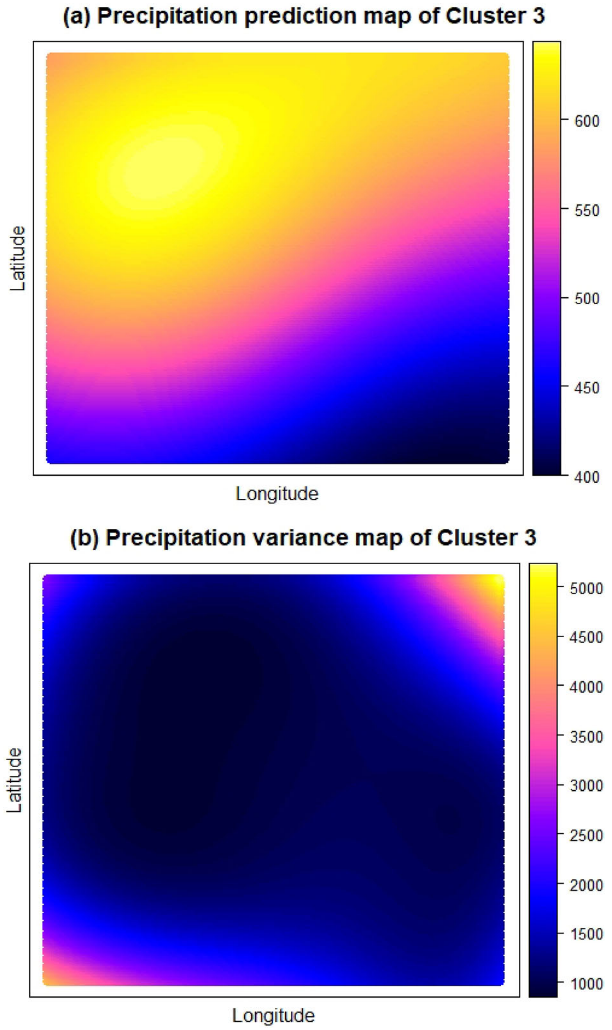
(a) Precipitation prediction map of Cluster 3
Latitude
Longitude
600
550
500
450
400
(b) Precipitation variance map of Cluster 3
Latitude
Longitude
5000
4500
4000
3500
3000
2500
2000
1500
1000

Fig. 9. (a) Prediction (b) Variance map of precipitation for Cluster 3 using ordinary kriging.

3.3. Ordinary kriging

Once the 226 rainfall stations of NSW, Australia, are split into 10 homogeneous clusters, the prediction maps of the average annual rainfall of cluster 3 and the overall monitoring stations are estimated by OK. Different variogram models like Spherical, Gaussian variogram, Exponential, and Linear bounded are applied. However, the Gaussian variogram fit best for cluster 3 stations. Figure 8 shows the fitted variogram from the D-plane for all the stations of cluster 3. A total of 841 grid points within cluster 3 are created, and then OK is applied to predict the average annual rainfall, using the variogram model of cluster 3. The prediction map and the prediction variance of cluster 3 are shown in Fig. 9. The results of

OK reveal that most of the parts in cluster 3 are predicted to receive precipitation between 400 and 650 mm.

A cross-validation method is applied to compare prediction accuracy between cluster 3 and the overall monitoring stations. The results show that the Mean Squared Prediction Error (MSPE) for the overall domain is 3859, whereas, for cluster 3, this error is 1733, which indicates that the MSPE of the overall monitoring network is approximately reduced by 45% when compared with cluster 3. This reduction in MSPE indicates that the desirable homogenization of precipitation regions is achieved.

4. Conclusion

Hydrological variables carry a key role in the management of water resources. These variables are measured concerning time and space, and therefore, they should be considered spatially dependent and space-time random fields. Among different hydrological variables, precipitation is one of the prime and crucial variables which affects the climate. Identifying homogeneous regions based on precipitation is a key tool for providing precipitation's spatial and temporal behavior. Therefore, the homogeneous regions should be identified so that the spatially closed stations should belong to similar clusters.

The non-separation issue happened when the geographic coordinates are utilized for clustering, just because the Euclidean distance is not suitable for clustering when geographic coordinates are considered. Hence, the coordinates are converted to rectangular coordinates by the Lambert projection method, and rectangular Lambert measures distance projected coordinates for obtaining more separated clusters. Using the PAM algorithm, also known as the k-medoid algorithm (which minimizes the sum of dissimilarities instead of the sum of squares of Euclidean distances) on rectangular Lambert projected coordinates, well-separated clusters are obtained. The MSPE is comparatively smaller if the prediction of unobserved locations in cluster 3 is made. However, this error increases if the prediction is made for a complete monitoring network.

References

BOM. 2020. Climate data online. Online at: http://www.bom.gov.au/

Bushmans. 2020. Rainfall by region: NSW. Online at: https://www.bushmantanks.com.au/blog/rainfall-by-region-nsw/

Dikbas, F., Firat, M., Koc, A. C. and Gungor, M. 2012. Classification of precipitation series using fuzzy cluster method. *Int. J. Climatol.* **32**, 1596–1603. doi:10.1002/joc.2350

Dunham, M. 2003. *Data Mining: Introductory and Advanced Topics*. Pearson Education (Singapore) Pte, Ltd.

Estivill-Castro, V. and Murray, A. T. 1997. *Spatial Clustering for Data Mining with Genetic Algorithms*. Queensland University of Technology Australia.

Goyal, M. K. and Gupta, V. 2014. Identification of homogeneous rainfall regimes in Northeast Region of India using fuzzy cluster analysis. *Water Resour. Manage.* **28**, 4491–4511. doi:10.1007/s11269-014-0699-7

Hamad-Ameen, J. J. 2008. Cell planning in GSM mobile. *WSEAS Trans. Commun.* **7**, 393–398.

Hussain, I., Pilz, J. and Spoeck, G. 2011. Homogeneous climate regions in Pakistan. *IJGW.* **3**, 55–66. doi:10.1504/IJGW.2011.038369

Kaufman, L. and Rousseeuw, P. J. 2009. *Finding Groups in Data: An Introduction to Cluster Analysis*. North America: John Wiley & Sons.

Kerby, A., Marx, D., Samal, A., and Adamchuck, V. 2007. Spatial clustering using the likelihood function. Seventh IEEE International Conference on Data Mining Workshops (ICDMW 2007), Omaha, NE: IEEE.

Mandal, S., Saha, D., Mahanti, A. and Pendharkar, P. C. 2007. Cell-to-switch level planning in mobile wireless networks for efficient management of radio resources. *Omega* **35**, 697–705. doi:10.1016/j.omega.2005.09.008

Matheron, G. 1963. Principles of geostatistics. *Econ. Geol.* **58**, 1246–1266. doi:10.2113/gsecongeo.58.8.1246

Meteorology, B. O. 2020. Stormy weather. Online at: http://www.bom.gov.au/nsw/sevwx/facts/stormy-weather.pdf

Nanopoulos, A., Theodoridis, Y., and Manolopoulos, Y. 2001. *C2P: Clustering Based on Closest Pairs*. VLDB, Roma, Italy.

Patil, S. and Stieglitz, M. 2011. Hydrologic similarity among catchments under variable flow conditions. *Hydrol. Earth Syst. Sci.* **15**, 989–997. doi:10.5194/hess-15-989-2011

Sadri, S. and Burn, D. 2011. A fuzzy C-means approach for regionalization using a bivariate homogeneity and discordancy approach. *J. Hydrol.* **401**, 231–239. doi:10.1016/j.jhydrol.2011.02.027

Sap, M. N. M. and Awan, A. M. 2005. Finding spatio-temporal patterns in climate data using clustering. International Conference on Cyberworlds (CW'05), Singapore, IEEE.

Satyanarayana, P. and Srinivas, V. 2011. Regionalization of precipitation in data sparse areas using large scale atmospheric variables–A fuzzy clustering approach. *J. Hydrol.* **405**, 462–473. doi:10.1016/j.jhydrol.2011.05.044

Snyder, J. P. 1987. *Map Projections–A Working Manual*. Washington, D.C: US Government Printing Office.

Soltani, S. and Modarres, R. 2006. Classification of spatio-temporal pattern of rainfall in Iran using a hierarchical and divisive cluster analysis. *J. Spat. Hydrol.* **6**, 1–12.

Swain, J. B., Sahoo, M. M. and Patra, K. C. 2016. Homogeneous region determination using linear and nonlinear techniques. *Phys. Geogr.* **37**, 361–384. doi:10.1080/02723646.2016.1211460

Wazneh, H., Chebana, F. and Ouarda, T. B. M. J. 2013. Depth-based regional index-flood model. *Water Resour. Res.* **49**, 7957–7972. doi:10.1002/2013WR013523