Scientia Iranica A (2013) 20 (2), 259–269

Sharif University of Technology
Scientia Iranica
Transactions A: Civil Engineering
www.sciencedirect.com

Rain gauge network design using coupled geostatistical and multivariate techniques

M.R. Shaghaghian, M.J. Abedini *

Department of Civil and Environmental Engineering, Shiraz University, P.O. Box 71348-51156, Shiraz, Iran

Received 30 June 2012; revised 6 October 2012; accepted 12 November 2012

KEYWORDS
Point kriging;
Block kriging;
Factor analysis;
Cluster analysis;
Variance-based approach;
Network design.

Abstract A methodology for the design of a rain gauge network is developed in this study. To the best of the authors' knowledge, this is the first time a combination of geostatistical tools and factor analysis, along with a clustering technique, has been used to prioritize rain gauge stations in terms of information content over the study area. The whole study area is divided into homogeneous subregions and a conventional variance-based approach is implemented in each subregion to rank rain gauge stations. For this purpose, factor analysis coupled with ordinary block kriging is used to identify the number of homogeneous subregions, and then, ordinary point kriging is used to assign rain gauge stations to each subregion. The developed scheme is quite time-efficient as it is not sensitive to initial guesses on cluster centers, there is no need to specify the number of clusters in advance and, above all, it is highly relevant to the overall objective stipulated in rain gauge network design. The proposed methodology is implemented on real data set in the south west of Iran. The results show that the proposed approach compares well with existing paradigms in rain gauge network design and only six rain gauge stations are required to provide the necessary information. In particular, the measure of network accuracy lies somewhere in between the so called time consuming and more simplified approaches used in rain gauge network design.

1. Introduction

Hydrologists and water resources managers are frequently confronted with problems that require estimating the spatial variation of a rainfall field from sparse information distributed in space and/or time. These estimates are often used at various spatial and temporal scales for a variety of applications, including water budget studies, reservoir operation, and flood forecasting and control. The background data required for such estimations are often collected via either ground-based measurements (i.e., rain gauge stations) or air-based instruments (i.e., radar and satellite imagery). The accuracy of both point and regional-wise rainfall estimation is highly coined with the number and spatial distribution of rain gauges or radar stations. The rain gauge densities and distributions have to be sufficient to allow valid information reflecting spatial and temporal variations of the rainfall field in a river basin. As a result, precipitation monitoring and subsequent network design are considered an inevitable part of any study aimed at providing the background data required for planning and management of water resources projects. There are numerous inter-related factors affecting a typical rain gauge network design. These factors include, but are not limited to, the overall objective of designing a network (e.g., water balance studies, reservoir operation and flood forecasting), the process considered (e.g., evaporation, rainfall, radiation), the attribute under consideration (e.g., rainfall depth, rainfall duration, rainfall hyetograph), the temporal scale or sampling interval in time (e.g., hourly, daily, monthly, annually. etc.), the spatial scale (e.g., catchment, regional, countrywide), the topographic setting (flat, rolling and mountainous), types of precipitation (e.g., orographic, convective and cyclonic), the nature of the objective function used for optimization (e.g., variance-based, entropy-based, fractal-based and distance-based techniques) and the optimization algorithm used for minimization or maximization purposes (e.g., exhaustive search algorithm, genetic algorithm, simulated annealing, Tabu search, etc.). While the study area itself (of course, if the data are not

* Corresponding author. Tel.: +98 711 6474604; fax: +98 711 6473161.
E-mail address: abedini@shirazu.ac.ir (M.J. Abedini).

Peer review under responsibility of Sharif University of Technology.

Production and hosting by Elsevier

1026-3098 © 2013 Sharif University of Technology. Production and hosting by Elsevier B.V. Open access under CC BY-NC-ND license.
http://dx.doi.org/10.1016/j.scient.2012.11.014

synthetic) would dictate factors such as spatial scale, topographic setting, and types of precipitation, a number of other important decisions have to be made in order to narrow the discussion on rain gauge network design to a manageable size.

Practically speaking, there is a close connection between such factors as the overall objective of designing a network, processes and attributes under consideration, and the temporal scale or sampling interval in time. As an example, while for flood forecasting and warning, accurate representation of the depth of the point rainfall and its time-wise variation is considered an important task [1,2], in reservoir operation, monthly rainfall is the correct time interval for inclusion in reservoir design and operation [3]. Spatial variability and sampling interval are highly related to each other. As the sampling interval increases, the attribute under consideration becomes less variable and subsequently lower rain gauge density is required to monitor the rainfall. In our case, as the sole objective of designing a network is to synthesize and assess the long-term water balance studies, accurate representation of mean areal rainfall is considered a major task. For this purpose, evaluating the spatial variation of annual rainfall depth and selecting the number and spatial locations of rain gauge stations is the sole objective of network design.

Having established all those pertinent factors, two key terms remain, i.e., an objective function and a typical algorithm for its optimization. Rain gauge network design assumes a variety of approaches as pertains to the selection of these two key terms. These approaches concerning the objective function are generally known as variance-based methods [2,4–11], entropy-based techniques [3,12–15], fractal-based methods [16,17], and distance-based approaches [18,19]. After casting the required objective function, an optimization algorithm has to be employed to either minimize or maximize the corresponding objective function. Early studies were mostly based on random searches and enumeration [20]. However, for the past three decades or so, researchers have considered some other more systematic approaches, including the simplex method [6], the gradient method [8], simulated annealing [19,21], Tabu search [22,23] genetic algorithm [24] and Ant Colony [25] as common optimization techniques in multiple fields of network design.

Interaction between the nature of the objective function, the associated optimization algorithm and the subsequent simplification involved in rain gauge network design will be the focus of the current paper. It might help to restate the problem in conventional form and see where the issue to be addressed in this paper, is. As a rule, when the mean areal rainfall over a river basin is to be estimated, there is generally a network of rain gauges in place. In light of the existing network, three different types of problem could be delineated in rain gauge network design.

1. Prioritizing the existing rain gauge network in terms of its contribution to estimation accuracy;
2. Choosing the location of some additional potential rain gauges to improve the estimation accuracy as much as possible;
3. Selecting an optimal subset, n, of an existing dense rain gauge network containing N stations.

In either case, the estimation accuracy can be expressed in terms of information content or variance of residuals. These features are, in turn, a function of the number n and the spatial location of rain gauges. When the number N is small, all possibilities [i.e., $\binom{N}{n}$], with reference to computation of the variance of residuals or quantification of information content, can be examined thoroughly [7,14]. However, when N is large, for intermediate values of n, this exhaustive analysis of each combination is not generally possible, resulting in the problem of dimensionality issue to be addressed in this paper. Different investigators consider different simplifying assumptions to reduce the computational cost involved. Conventional paradigms in rain gauge network design surmount this issue by resorting to Bellman's principle of optimality [26] aimed at thoroughly searching the feature space for each combination with due attention to results obtained in earlier combinations [7,10,14]. This simplification would not necessarily lead to the same result considering all scenarios for $\binom{N}{n}$. Then, a logical query would be to quantify the discrepancy involved between these two scenarios.

Ironically, almost all entropy-based network design tried to avoid this curse of the dimensionality problem by making some further implicit or explicit simplifying assumptions [3,12–15]. While Krstanovic and Singh [13] and Yoo et al. [15] considered geographical/topographical boundaries to partition the rain gauge stations, assuming each subregion represents the area of similar climatological characteristics, A1-Zahrani and Husain [14] addressed the combinational problem by assuming the number of stations in each imaginary zone to be less than 10. The plot of the measure of accuracy versus the number of stations pursued a totally different route in variance-based methods [7,10]. While Bastin et al. [7] used block kriging and forward addition to sequentially add and delineate various combinations of optimum rain gauge stations, Kassim and Kottegoda [10] considered point kriging and backward elimination to sequentially eliminate and delineate appropriate rain gauge combinations. Both studies suffer from the simplification cited above.

The main contribution of the current paper is in using a combination of factor and cluster analysis to objectively subdivide the study area into zones of similar characteristics and then using the variance-based approach to prioritize stations in each cluster. Total number of rain gauge stations within the study area are obtained by adding the optimum number of stations in each cluster. The next section of the paper is devoted to a discussion of the geostatistical framework regarding both ordinary point and block krigings which have been chosen for clustering and application of variance-based techniques. Then, the following section describes a rationale for using factor analysis to cluster stations into subregions. Section 3 is devoted to materials and methods summarizing the study area, nature of data used, and the proposed methodology. In Section 4, the results of various paradigms on network design are compared and contrasted with the proposed scheme. The last section includes the conclusions which can be drawn from this study.

2. Theoretical background

The network design pursued in this paper intends to use a variance-based approach coupled with multivariate techniques (e.g., factor and cluster analyses) to prioritize rain gauge stations, and then compare the results with conventional paradigms in network design. As a result, a brief account of geostatistical analysis and multivariate techniques are in order.

2.1. An overview on ordinary point and block kriging

At this stage, it might help to have an overview of various types of kriging, in particular, ordinary point and block kriging. A regionalized variable, such as annual rainfall depth, observed at known spatial locations, $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$, can be considered a realization of a random function given by:

$$\mathbf{P} = [P(\mathbf{x}_1), P(\mathbf{x}_2), \ldots, P(\mathbf{x}_n)]^T. \tag{1}$$

This function can be decomposed into a large-scale variation, $m(\mathbf{x})$, and a small-scale variation with zero expectation, $W(\mathbf{x})$, to be modeled as deterministic and stochastic processes, respectively. The parent random function can be expressed as:

$$P(\mathbf{x}) = m(\mathbf{x}) + W(\mathbf{x}). \tag{2}$$

The upper case letter is used to denote the random function or its corresponding random variable and the lower case letter for its particular realization.

In both simple and ordinary kriging, the deterministic component is considered to be independent of spatial location. On the contrary, in universal kriging, $m(\mathbf{x})$ will be a function of spatial location, $\mathbf{x}$. In simple kriging, m is constant and known. In contrast, in ordinary kriging, m is constant and unknown. In ordinary point kriging, the support size is a point as large as the size of the rain gauge catch, while in ordinary block kriging, the support size could be a block as large as the study area itself. At every point in space, one has to differentiate between three types of random variable:

$P_\circ(\mathbf{x_i})$: The observed value of P at spatial location $\mathbf{x}_i$,

$P(\mathbf{x}_0)$: The true value of P at spatial location $\mathbf{x}_0$,

$\hat{P}(\mathbf{x}_0)$: The estimated value of P at spatial location $\mathbf{x}_0$,

where $P(\mathbf{x}_0)$ would not be accessible and $\hat{P}(\mathbf{x}_0)$ is given by the following formula in the case of ordinary point kriging:

$$\hat{P}(\mathbf{x}_0) = \sum_{i=1}^{N} \lambda_i(\mathbf{x}_0) P_\circ(\mathbf{x}_i), \tag{3}$$

where λ_i is the weight associated with observed rainfall depth at $\mathbf{x}_i$, i.e. $P_\circ(\mathbf{x}_i)$ and N is the total number of rain gauge stations. In this paper, a methodology will be proposed to effectively utilize these weighting coefficients, i.e. λ_i to cluster the whole study area into clusters of homogeneous characteristics. In the same way, the mean areal rainfall over block V, centered at $\mathbf{x}_0$, i.e. $\hat{P_\mathrm{V}}(\mathbf{x}_0)$, can be obtained through the following relationship:

$$\hat{P_\mathrm{V}}(\mathbf{x}_0) = \sum_{i=1}^{N} \lambda_i^{\mathrm{BK}}(\mathbf{x}_0) P_\circ(\mathbf{x}_i), \tag{4}$$

where λ_i^{BK} is the weighting coefficient associated with observed rainfall depth at $\mathbf{x}_i$, i.e. $P_\circ(\mathbf{x}_i)$. Once again, in this paper, these weighting coefficients will be effectively utilized to identify the optimum number of clusters corresponding to accurate representation of mean areal rainfall over the study area. It might be helpful to differentiate between two types of residual (i.e., estimation error), namely, point residual (i.e., $R(\mathbf{x}_0)$) and block residual (i.e., $R_\mathrm{V}(\mathbf{x}_0)$), to be defined as:

$$R(\mathbf{x}_0) = \hat{P}(\mathbf{x}_0) - P(\mathbf{x}_0),$$
$$R_\mathrm{V}(\mathbf{x}_0) = \hat{P_\mathrm{V}}(\mathbf{x}_0) - P_\mathrm{V}(\mathbf{x}_0). \tag{5}$$

Using Eq. (3), along with two criteria for point residual (i.e., zero mean and minimum variance), leads to an ordinary point kriging system in terms of the variogram:

$$\sum_{j=1}^{N} \lambda_j(\mathbf{x}_0)\gamma(\mathbf{x}_i, \mathbf{x}_j) - \mu(\mathbf{x}_0) = \gamma(\mathbf{x}_i, \mathbf{x}_0) \quad \forall\, i = 1, 2, \ldots, N,$$
$$\sum_{i=1}^{N} \lambda_i(\mathbf{x}_0) = 1, \tag{6}$$

where $\mu(\mathbf{x}_0)$ is the Lagrange multiplier. After obtaining λ's, Eq. (3) can be used to find the estimated value at the prescribed location. Subsequently, the variance of residual at spatial location $\mathbf{x}_0$ is given by:

$$\mathrm{VAR}[R(\mathbf{x}_0)] = \sigma_\mathrm{R}^2(\mathbf{x}_0) = \sum_{i=1}^{N} \lambda_i(\mathbf{x}_0)\gamma(\mathbf{x}_i, \mathbf{x}_0) - \mu(\mathbf{x}_0). \tag{7}$$

Using Eq. (4), along with two criteria for block residual (i.e., zero mean and minimum variance), leads to an ordinary block kriging system in terms of the variogram:

$$\sum_{j=1}^{N} \lambda_j^{\mathrm{BK}}(\mathbf{x}_0)\gamma(\mathbf{x}_i, \mathbf{x}_j) - \mu(\mathbf{x}_0) = \frac{1}{M}\sum_{k=1}^{M} \gamma(\mathbf{x}_i, \mathbf{x}'_k) \quad \forall\, i = 1, 2, \ldots, N,$$
$$\sum_{i=1}^{N} \lambda_i^{\mathrm{BK}}(\mathbf{x}_0) = 1, \tag{8}$$

where M is the number of discretized points inside a typical block. After obtaining λ^{BK}'s, Eq. (4) can be used to find the mean areal rainfall over the study area. Subsequently, the variance of block residual over block V centered at $\mathbf{x}_0$ is given by:

$$\mathrm{VAR}[R_\mathrm{V}(\mathbf{x}_0)] = -\mu(\mathbf{x}_0) + \frac{1}{M}\sum_{i=1}^{N}\sum_{k=1}^{M} \lambda_i^{\mathrm{BK}}(\mathbf{x}_0)\gamma(\mathbf{x}_i, \mathbf{x}'_k) - \frac{1}{M^2}\sum_{k=1}^{M}\sum_{j=1}^{M} \gamma(\mathbf{x}'_j, \mathbf{x}'_k), \tag{9}$$

where "′" corresponds to the discretized points inside a typical block. The backbone of almost all variance-based techniques considers the point variance of the residual (i.e., $R(\mathbf{x}_0)$) and the block variance of the residual (i.e., $R_\mathrm{V}(\mathbf{x}_0)$) to be a function of number and spatial location of rain gauge stations in place.

2.2. An overview on factor analysis

Factor analysis is considered an extremely useful multivariate statistical technique to rearrange and organize original variables (i.e., mean annual rainfall at each station) into fewer underlying factors, $F_1, F_2, \ldots, F_m$ (also called common factors), to retain as much information contained in the original variables as possible [27]. Unlike the original variables, which might have strong spatial correlation, factors with no spatial coordinate associated whit them are completely uncorrelated with each other. As a result, substituting these factors for the original variables can effectively reduce the overall complexity of a large data set. Assuming the original data are standardized (each data is subtracted from spatial mean and then divided by spatial standard deviation), the eigenvalue quantifies the contribution of a factor to total variance for the attribute under consideration. Factors are produced according to an eigenvalue analysis of the correlation matrix, and factor loadings and factor scores are the main calculations of Factor Analysis (FA). The key idea is to assume each variable as a linear combination of a set of unobserved, underlying, and latent variables plus an error component. The first step of FA is to standardize the raw data and compute a correlation matrix of the variables from the standardized variables [28, p. 413]. Due to the availability of only one realization at each rain gauge station, data standardization has to be performed in a slightly different context. For this purpose, spatial data series are assumed to be stationary and ergodic. As a result, data standardization implies subtraction from the spatial mean and division by the global standard deviation.

Expressing each standardized observation as a linear combination of common factors and N additional sources of variation, $\epsilon_1, \epsilon_2, \ldots, \epsilon_N$, we have:

$$\begin{cases} \overline{P_\circ}(\mathbf{x}_1) = L_{11}F_1 + L_{12}F_2 + \cdots + L_{1m}F_m + \epsilon_1 \\ \overline{P_\circ}(\mathbf{x}_2) = L_{21}F_1 + L_{22}F_2 + \cdots + L_{2m}F_m + \epsilon_2 \\ \vdots \\ \overline{P_\circ}(\mathbf{x}_N) = L_{N1}F_1 + L_{N2}F_2 + \cdots + L_{Nm}F_m + \epsilon_N \end{cases} \tag{10}$$

where L_{ij} is a factor loading which relates the ith observation to the jth factor. The second step is to estimate the factor loadings that express the degree of closeness between the factor and variables. For this purpose, some assumptions have to be considered in order to make the above formulation applicable. At first, common factors and error terms are assumed to be independent and normally distributed random variables whose means are equal to zero and whose variances are equal to 1 and Ψ_i, respectively. Hence, regarding Eq. (10), variance of an observation and correlation between two observations can be determined as follows:

$$\begin{aligned} \mathrm{COV}\left[\overline{P}(\mathbf{x}_i), \overline{P}(\mathbf{x}_j)\right] &= \mathrm{COR}\left[\overline{P}(\mathbf{x}_i), \overline{P}(\mathbf{x}_j)\right] \\ &= R(\mathbf{x}_i, \mathbf{x}_j) \\ &= \begin{cases} \sum_{k=1}^{m} L_{ik}L_{jk} + \Psi_i & \text{if } i \neq j, \\ \sum_{k=1}^{m} L_{ik}^2 + \Psi_i & \text{if } i = j. \end{cases} \end{aligned} \tag{11}$$

There are four common methods for estimating the non-unique matrix of factor loading, $[L_{ij}]$ [28,29]. One of the most popular is the principle component method. In this method, firstly, Ψ_i's are neglected, then, the correlation matrix is factored into LL^T with the aid of the spectral decomposition method:

$$R_{N\times N} = C_{N\times m} D_{m\times m} C_{m\times N}^T, \tag{12}$$

where C is an orthogonal matrix, so called a truncated modal matrix, constructed with m major normalized eigenvectors ($\sum_{i=1}^{N} C_{ij}^2 = 1$ and $\sum_{i=1}^{N} C_{ij}C_{ik} = 0$) of R, and D is a diagonal matrix, so called spectral matrix, constructed with the m largesteigenvalues, $\theta_1, \theta_2, \ldots, \theta_m$, of the correlation matrix. Since the eigenvalues, θ_i, of a positive semi-definite matrix, R, are all positive or zero, we can write D as $D^{1/2}D^{1/2}$. So, substituting this in Eq. (12) and comparing it with relations in Eq. (11), after incorporating the cited assumptions, components of the loading factor matrix can be calculated as follows:

$$L_{ij} = C_{ij}\sqrt{\theta_j}. \tag{13}$$

Each component of the loading factor matrix (L_{ij}) represents the correlation between the ith variable and the jth factor. Factor loadings range from -1 to $+1$, with a larger absolute value indicating a stronger relationship between the respective factor and variable. Considering Eq. (11), the variance of each observation may be calculated by summation of the square of components in the corresponding row of the loading factor matrix. Consequently, the proportion of the total sample variance due to the jth factor is:

$$\frac{\sum_{i=1}^{N} L_{ij}^2}{\sum_{i=1}^{N} R(\mathbf{x}_i, \mathbf{x}_i)} = \frac{\sum_{i=1}^{N} (C_{ij}\sqrt{\theta_j})^2}{N} = \frac{\theta_j \sum_{i=1}^{N} C_{ij^2}}{N} = \frac{\theta_j}{N}. \tag{14}$$

In every application, a decision must be made on how many dominant factors (m) should be retained in order to effectively summarize, cluster or interpret data. The scree graph, whose plots describe sorted eigenvalues versus eigenvalue number, is a good tool in selecting the number of dominant factors. Since the eigenvalues serve as variances of the factors, if the graph drops sharply, followed by a straight line with a much smaller slope, m may be chosen equal to the number of eigenvalues before the straight line begins. This corresponds to retaining eigenvalues greater than one. Therefore, with regard to Eq. (14), the high proportion of variations can be explained by retaining the first m dominant eigenvalues.

The last step linearly transforms factors associated with the initial set of loadings by factor rotation to maximize variable variances and to obtain a better interpretable loading pattern. The factor loading matrix was rotated to obtain uncorrelated factors by varimax rotation [30]. In subsequent paragraphs, the entries of this new matrix will be denoted by $[L'_{ij}]$. This study utilizes FA coupled with ordinary block kriging to identify the number of clusters with similar characteristics, and then uses a combination of FA and ordinary point kriging to assign objects to various clusters.

2.3. Data clustering via coupled factor analysis and kriging

Cluster analysis is the organization of a collection of patterns (usually represented as a vector of measurements, or a point in a multidimensional space) into clusters based on similarity. The clustering technique determines optimum partitions based on a certain similarity and/or dissimilarity function that measures the global error extent between data points and cluster centers in a feature space. Here, FA may be employed for data clustering [31]. The approach benefits from many advantages including no requirement for any initial guess, self-detection of number of clusters, and above all, objective-oriented clustering. Therefore, coupling FA, as a characterization tool, and kriging, as a spatial estimator, seems to be an efficient way for data clustering in a spatial context. Let us see how a combination of FA and ordinary block kriging can be utilized to identify the number of clusters.

In the Appendix it is shown that the standardized mean areal rainfall can be written as a linear combination of standardized observed point rainfall. However, standardized observed point rainfall itself can be written as a linear combination of factors. As a result, standardized mean areal rainfall can be expressed as a linear combination of factors via

$$\hat{\overline{P}}_{\mathrm{V}}(\mathbf{x}_0) = \sum_{j=1}^{m} \beta_j F_j,$$

where:

$$\beta_j = \sum_{i=1}^{N} \lambda'_i L'_{ij}, \tag{15}$$

where β_j can be interpreted as the sensitivity of factor j to the standardized mean areal rainfall. As $\mathrm{VAR}[\hat{\overline{P}}_{\mathrm{V}}(\mathbf{x}_0)] = \sum_{j=1}^{m} \beta_j^2 \approx 1$, β_j^2 can also be interpreted as the contribution of factor j to total variance of the standardized mean areal rainfall. Hence, the number of factors can be further reduced by retaining dominant contributing factors. This new reduced number can be taken as the number of clusters associated with mean areal rainfall.

Taking factors as representative of each cluster, we can now switch to partitioning the whole study area into these

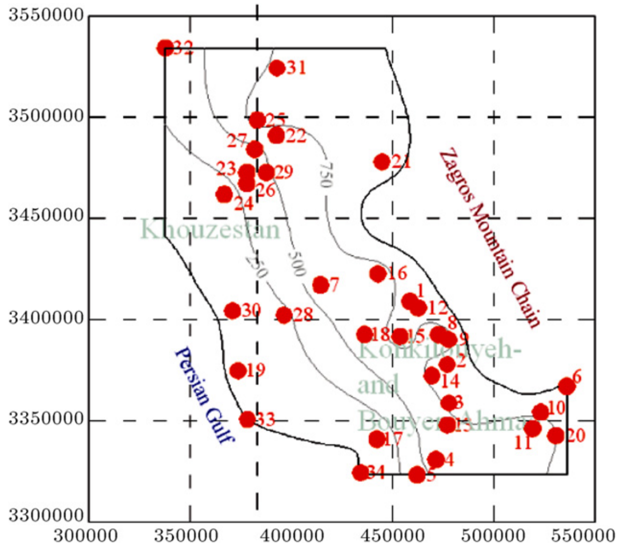

Figure 1: Location of study area and rain gauge stations.

identified clusters. For this purpose, the whole study area has to be discretized into a mesh, and ordinary point kriging has to be used to compute weighting coefficients for each nodal point. Based on similar reasoning, the standardized point estimation of rainfall at $\mathbf{x}_0$ can be written as:

$$\hat{\hat{P}}(\mathbf{x}_0) = \sum_{j=1}^{m} \alpha_j F_j,$$

where:

$$\alpha_j = \sum_{i=1}^{N} \lambda_i'' L'_{ij}, \tag{16}$$

where α_j can be interpreted as the sensitivity of factor j to the standardized point rainfall estimate. As $\mathrm{VAR}[\hat{\hat{P}}(\mathbf{x}_0)] = \sum_{j=1}^{m} \alpha_j^2 \approx 1$, α_j^2 can also be interpreted as the contribution of factor j to total variance of the standardized point rainfall estimate. A threshold can be defined for α_j and assignment can be made based on the numerical value of the selected threshold. Needless to say, if $\mathbf{x}_0$ coincides with a rain gauge station, then, due to the exact interpolant property of kriging [32], α_j becomes proportional to L'_{ij} and the assignment process becomes much simpler. In this case, one can resort to entries of rotated loading factors to allocate rain gauge stations to various clusters. Very low values of threshold imply that every node belongs to all clusters. As the threshold increases, the assignment becomes more distinct.

3. Materials and methods

3.1. Description of the study area

The study area selected for this study is the plain region of Kohkiloyeh-Bouyerahmad and Khouzestan provinces in South west of Iran with a total area of about 25,000 km^2, as illustrated in Figure 1. It is located between longitude 49° 17′ and 51° 22′ east, and between latitude 30° 2′ and 31° 56′ north. The physiography within the study area is the near-horizontal depositional surfaces of the Gachsaran and Dehdasht regions. Elevations in the study area range from 1000 m on the slopes of the Zagros Mountain Chain to 0 m on the coasts of the Persian Gulf.

The overall rain pattern in the region is strongly affected by Mediterranean low pressure systems which enter from the west throughout the year. The precipitation mostly occurs in the form of rain, which usually results from frontal storm systems traveling eastward. However, summer precipitation, which has usually no major contribution to total annual precipitation, results from localized convective-type storms.

A total of 34 rainfall gauge stations providing monthly observations, for at least a 10-years period, are used for analysis in this study. Figure 1 shows the location of each rain gauge in the study area, and Table 1 summarizes the UTM coordinates, elevation and average annual rainfall depths (mm) of each rain gauge station. The mean annual rainfall ranges from 249.9 mm (Rain-gauge No. 19) to 901.4 mm (Rain-gauge No. 6). This shows, roughly, the great spatial rainfall variability over the region. Based on monthly rainfall records, summer precipitation includes, at most, thirteen percent of the total annual precipitation. The observed values also indicate that average annual precipitation becomes more significant with an increase in elevation.

3.2. Variogram modeling

Unlike the majority of conventional deterministic methods, stochastic rainfall estimation, i.e., geostatistical approaches, would take the spatial structure of rainfall observations into account and use semi-variance (invariably called variograms) as a measure of spatial variability. In brief, various types of geostatistical method consist of the following three steps:

- Exploratory spatial data analysis;
- Variogram modeling;
- Estimation.

In exploratory spatial data analysis, a few tasks, including delineation and removal of outliers, check for normality, and the need for any possible data transformation is pursued. Variogram modeling amounts at computation of experimental variograms and finding an admissible theoretical variogram which will best fit the experimental variogram. The experimental variogram, $\hat{\gamma}(\mathbf{x_i}, \mathbf{x_j})$, is computed as half the average squared difference between the components of data pairs:

$$\hat{\gamma}(\mathbf{x_i}, \mathbf{x_j}) = \frac{1}{2N(\mathbf{h_{ij}})} \sum [P(\mathbf{x_i}) - P(\mathbf{x_j})]^2, \tag{17}$$

where $N(\mathbf{h_{ij}})$ is the number of data pairs, a separation vector, $\mathbf{h_{ij}} = \mathbf{x_i} - \mathbf{x_j}$, apart. After trying a number of permissible theoretical variograms, the analysis confirmed that an exponential structure provided the best goodness of fit among other competing models. Once the basic model was chosen, modeling the sample experimental omnidirectional variogram became an exercise in nonlinear curve fitting. The selected model consisted of an exponential structure with $\sigma^2 = 37\,511$ mm^2 and a range of 206 991 m (206.991 km). Figure 2 shows the results of variogram modeling. In this study, the variogram is used for two purposes. First, stochastic estimation of point and mean areal rainfall calls for variograms to be used in the kriging system. Second, due to the particular nature of rainfall spatial series, a correlation function is derived from the variogram function, assuming second-order stationarity is applicable. Then, the correlation function is digitized to obtain the correlation matrix, taking the rain gauge station topology into account.

Table 1: Precipitation data used for characterizing rainfall spatial variability.

Station number[a]	Station name	UTM coordinates		Station elevation (m)	Average annual precipitation AAP[b] (mm)
		Easting (m)	Easting (m)		
1	Dehdasht	458 392	3 409 273	829	540.9
2	Dail	476 871	3 378 002	870	785.9
3	Dogonbadan	477 738	3 358 946	776	385.5
4	Benpir	471 312	3 331 042	670	682.4
5	Bibihakimeh	461 972	3 323 468	380	402.2
6	Golbabakan	536 009	3 367 171	920	901.4
7	Likak	414 428	3 417 231	650	458.4
8	Seyedabad	472 592	3 392 696	650	508.3
9	Nazmkan	477 643	3 390 376	650	497.5
10	Tangebirim	522 992	3 354 484	800	732.9
11	Hajghlandar	519 081	3 346 351	640	594.1
12	Samghan	462 473	3 406 087	800	514.2
13	Bibijanabad	477 021	3 348 051	717	385.1
14	Abchirak	469 288	3 372 480	793	539.1
15	Bouyeri	453 508	3 391 838	820	574.5
16	Eidanak	442 680	3 422 368	600	666.3
17	Khaibad	442 247	3 341 099	38	342.2
18	Behbahan	436 135	3 392 851	650	362.2
19	Dehmola	373 646	3 374 928	32	232.1
20	Batoun	530 475	3 342 837	735	645.1
21	Barez	444 614	3 478 011	815	682.4
22	Baghmalek	392 428	3 491 115	675	605.2
23	Mashin	377 973	3 472 798	380	403.7
24	Ramhormoz	366 738	3 461 849	155	289.6
25	Delibakhtiyari	383 029	3 498 608	850	617.4
26	Jokonak	377 909	3 467 256	330	368.3
27	Dehsadat	381 925	3 484 151	429	441.9
28	Chamnezam	396 307	3 402 398	190	344.7
29	Meydavoud	387 483	3 472 691	480	395.5
30	Omidiyeh	370 802	3 404 522	34.9	265.9
31	Izeh	392 773	3 524 369	767	694.1
32	M. Soleyman	337 724	3 534 332	321	467.9
33	Hendijan	378 177	3 350 861	3	249.9
34	B. Deylam	434 126	3 324 522	4	326.4

[a] Station numbers are the same as Figure 1.
[b] Average annual precipitation.

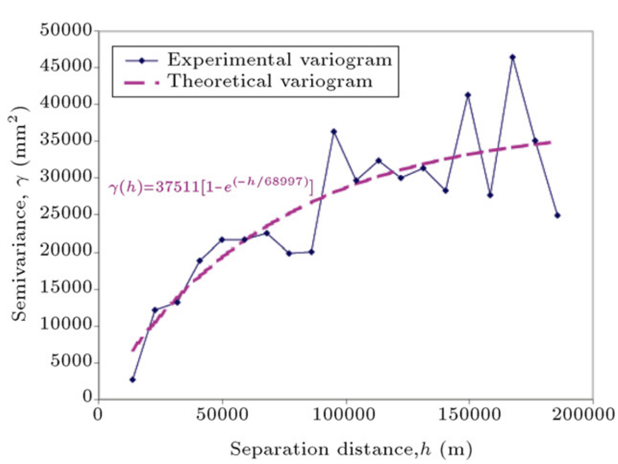

Figure 2: Experimental variogram along with best fit theoretical exponential model.

3.3. Methodology

In this research, the key issue is to objectively challenge a few variance-based rain gauge network design paradigms in practical use and then propose a methodology whereby prioritization of rain gauge stations can be made by a combination of geostatistical tools and multivariate techniques. For this purpose, a more detailed analysis of two conventional paradigms in rain gauge network design is offered first. Then, a summary of the step by step procedure to prioritize rain gauge stations, based on the proposed methodology, will be provided. In what follows, the two conventional paradigms in rain gauge network design are invariably referred to as the "time consuming approach" and "Bastin's simplified approach". These approaches are compared and contrasted with the proposed approach. In either case, the network design problem consists of prioritizing the rain gauge stations, which results in the best estimate of the mean areal rainfall, i.e. $\hat{P}_V(\mathbf{x}_0)$, in a relative sense. Furthermore, in all three approaches, the study area has to be discretized into a mesh to be able to compute the variance of the residual over the whole region (i.e., Eq. (9)) or some portion of it [7,9,21].

3.3.1. *Time consuming approach*

This approach aims at monitoring the numerical value of the variance of the residual over the whole study area for various combinations of rain gauge station, i.e. $\binom{N}{n}$, starting with a single rain gauge station. For either small or large values of n, these various combinations can be thoroughly searched. However, for intermediate values of n and large values of N, the comprehensive analysis of each combination would be impossible, resulting in the curse of dimensionality issue. Pardo-Igúzquiza [21] proposed simulated annealing to surmount this issue. However, he worked exclusively with synthetic data.

3.3.2. Bastin's simplified approach

Bastin et al. [7] tried to address the curse of dimensionality issue by making simplifying assumptions. Their approach aims at monitoring numerical values of regional variances of the residual, just like the time consuming approach. However, they assumed that Bellman's principle of optimality would be applicable while moving from one combination to another. For a single rain gauge station, both approaches lead to identical results. For the two rain gauge combinations, while a time consuming approach needs to search among $\frac{N(N-1)}{2}$ combinations for a minimum variance of residual, Bastin's approach requires only $N-1$ combinations. The degree of difference between these two approaches for rain gauge network design will be highlighted in the results and discussion section.

3.3.3. Proposed approach

In the proposed methodology, at first, the whole study area is divided into subregions with reference to the objective of rain gauge network design. Then, the so called "time consuming approach" is implemented in each subregion to prioritize the rain gauge stations. A step by step procedure can be summarized as follows:

Step 1. Identify the variogram function for the attribute under consideration and then, assuming a second-order stationarity, find the correlation function. Keep in mind that the correlation function for original data is the same as the correlation function for standardized data;

Step 2. In reference to network topology, digitize the correlation function to obtain the correlation matrix;

Step 3. Compute eigenvalues of the correlation matrix and plot the scree diagram. Using the scree diagram, identify dominant eigenvalues greater than one.

Step 4. Identify the truncated modal matrix ($C_{N\times m}$) and the truncated diagonal matrix ($D_{m\times m}$);

Step 5. Compute the loading factor matrix via: $\mathbf{L} = [L_{ij}] = [C_{ij}\sqrt{\theta_j}]$;

Step 6. Use the varimax method to rotate the loading factor matrix and obtain a better interpretable loading pattern, i.e. $\mathbf{L}'$;

Step 7. Discretize the study area into a generic mesh (5×5 km). Compute the scaled block weighting coefficient via appropriate equations and then compute β_j from $\beta_j = \sum_{i=1}^{N} \lambda_i' L_{ij}'$;

Step 8. In reference to numerical values of β_j, decide on dominant factors corresponding to the number of clusters;

Step 9. Calculate scaled point weighting coefficients for each generic point and then compute α_j from $\alpha_j = \sum_{i=1}^{N} \lambda_i'' L_{ij}'$ and assign each generic node, as well as rain gauge stations, to each cluster;

Step 10. Perform conventional network analysis in each cluster and draw the variance of residual versus the number of rain gauges for each cluster;

Step 11. Count the effective number of rain gauges in each cluster;

Step 12. Combine the results.

4. Results and discussion

Precision of a rain gauge network depends on the number and spatial location of the rain gauges within the study area. In other words, due to climatological and topographical features of the basin, each rain gauge has its own unique contribution to the precision of the network. Thus, removing, adding or reshuffling these rain gauges wil affect the precision of rain gauge networks. An ideal rain gauge network would neither be over-saturated with redundant rain gauges, nor suffer from lack of rain gauges. Therefore, a typical procedure of rain gauge network design has to look for a combination among the existing rain gauges which maximizes the information content or minimizes the variance of residual. In addition, the procedure should be capable of offering spatial locations for further addition of rain gauges to obtain more significant information. Consequently, in an optimal rain gauge network design, the purpose is to have a rain gauge configuration in order to achieve maximum precision with a minimum number of rain gauges.

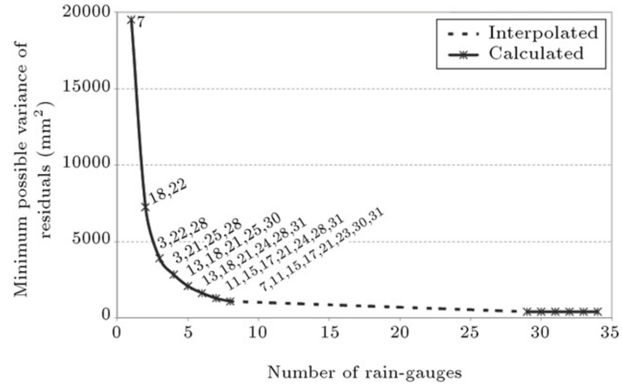

Figure 3: Variance of residual (accuracy) versus the number of points (N)-Time consuming approach.

The main objective of rain gauge networks in this study is to monitor the average annual precipitation of the aforementioned region. As a result, implementation of various rain gauge network paradigms requires discretization of the study area. For this purpose, the area is superimposed by a 5 km $\times$ 5 km square grid. Figure 3 shows the results for selecting the best combination of a certain number of rain gauges using a time consuming viewpoint. As the graph clearly demonstrates, at early stage of rain gauge addition, the degree of variance of residual reduction is remarkable. However, as the number of rain gauges increases, the rate of variance reduction diminishes considerably. Furthermore, for rather small or large values of n, various rain gauge combinations can be thoroughly searched. However, due to high computational effort, the procedure cannot be implemented for intermediate values of n. The figure depicts that after the best combination of a certain number of rain gauges is achieved, the variance of residuals cannot be further reduced significantly. Therefore, if the figure is simulated with a bilinear curve, the number of rain gauges corresponding to break point in the slope can be considered the best number of rain gauges within the study area. In this study, implementation of a time consuming approach shows that only six rain gauge stations (i.e., 13, 18, 21, 24, 28, and 31) are required to obtain a variance of residual as low as 1622 mm^2. Figure 4 intends to compare and contrast Bastin's simplified approach [7] with that of a more time consuming one. The two approaches share only two points, one at the beginning and another toward the end. Our experience with various implementations shows that the one rain gauge scenario corresponds to a rain gauge located near the center of the study area. The six rain gauge scenario associated with Bastin's simplified approach (i.e., 7, 25, 13, 19, 21, and 24) provides a variance of residual as low as 3412 mm^2 and

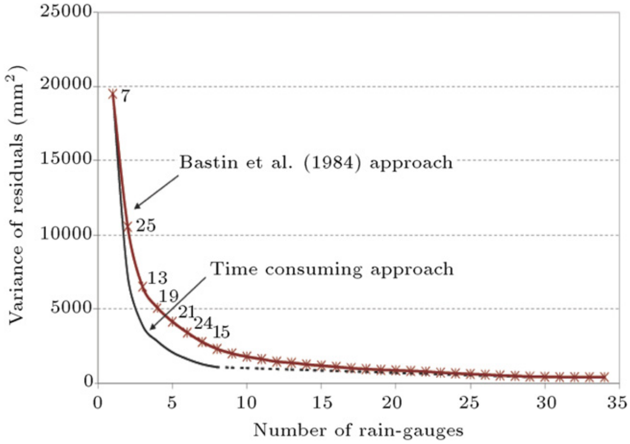

Figure 4: Variance of residual (accuracy) versus the number of points (n)-Bastin's simplified approach.

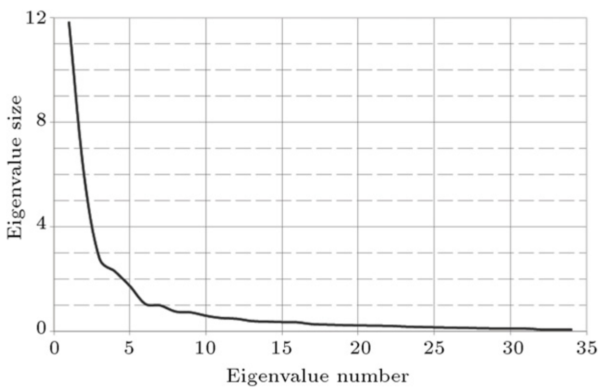

Figure 5: Variation of eigenvalue size versus eigenvalue number–Scree diagram.

shares only three stations with that of the time consuming one. The degree of departure between the two approaches seems to be quite distinct.

As for the proposed methodology, following the step by step procedure mentioned above, the correlation function for standardized rainfall data was found to be $R(h) = \rho(h) = \exp\left(\frac{-h}{68\,997}\right)$. In reference to network topology, the correlation function can be digitized to obtain the correlation matrix. Figure 5 demonstrates a scree diagram for the correlation matrix. Using this graph, it seems the first six factors explained more than 75% of the variance of observations, and can be considered dominant factors (factors with eigenvalues equal to or more than one). Tables 2 and 3 summarize the entries of original and rotated factor loading matrices associated with the first six dominant factors. After obtaining scaled ordinary block kriging weighting coefficients (i.e., λ_i''s), Eq. (15) can be used to find β_j. Table 4 summarizes the numerical values of β_j in decreasing order. With regard to these weighting coefficients, it is observed that while the six factors explain 98.9% of the variance, the first three explain 97.9% variance of the standardized mean areal annual precipitation over the study area. Therefore, just the first three factors will be considered as the number of clusters for the remainder of this study.

As for clustering the study area, while the content of Table 3 can be directly utilized to assign rain gauge stations to each cluster, for other generic points, scaled ordinary point kriging weighting coefficients have to be combined with rotated factor

Table 2: Factor loading or correlation between rain gauge observations and dominant factors before rotation.

Rain gauge #	Factors					
	1	2	3	4	5	6
1	0.75	0.10	−0.33	−0.37	−0.01	−0.07
2	0.77	0.34	0.00	−0.21	0.14	0.12
3	0.73	0.39	0.17	0.04	0.19	0.11
4	0.61	0.37	0.24	0.35	0.29	0.01
5	0.57	0.33	0.19	0.43	0.33	−0.03
6	0.47	0.32	0.47	−0.12	−0.42	−0.05
7	0.64	−0.22	−0.36	0.04	−0.23	−0.03
8	0.77	0.26	−0.16	−0.35	0.09	0.06
9	0.76	0.29	−0.10	−0.34	0.08	0.08
10	0.53	0.37	0.54	−0.07	−0.41	−0.04
11	0.53	0.38	0.54	−0.02	−0.37	−0.03
12	0.75	0.14	−0.30	−0.38	0.02	−0.04
13	0.69	0.40	0.23	0.17	0.23	0.08
14	0.77	0.34	−0.02	−0.12	0.19	0.11
15	0.78	0.18	−0.30	−0.24	0.06	−0.01
16	0.68	−0.08	−0.34	−0.25	−0.10	−0.11
17	0.63	0.27	−0.01	0.43	0.30	−0.06
18	0.74	0.06	−0.37	−0.06	−0.04	−0.06
19	0.47	−0.13	−0.34	0.50	−0.33	−0.09
20	0.47	0.35	0.55	−0.03	−0.43	−0.06
21	0.45	−0.33	−0.02	−0.20	0.00	−0.26
22	0.45	−0.71	0.24	−0.07	0.12	−0.12
23	0.49	−0.74	0.19	0.02	0.03	0.28
24	0.46	−0.65	0.11	0.08	−0.05	0.32
25	0.42	−0.70	0.26	−0.06	0.15	−0.21
26	0.50	−0.72	0.15	0.04	−0.01	0.33
27	0.47	−0.75	0.24	−0.03	0.10	0.07
28	0.59	−0.21	−0.38	0.28	−0.33	−0.01
29	0.50	−0.72	0.17	−0.01	0.03	0.21
30	0.49	−0.28	−0.30	0.38	−0.36	0.05
31	0.32	−0.54	0.23	−0.09	0.16	−0.49
32	0.21	−0.40	0.18	−0.02	0.12	−0.47
33	0.44	−0.03	−0.28	0.52	−0.20	−0.14
34	0.53	0.23	0.00	0.51	0.28	−0.10

loadings to obtain α_j, via Eq. (16). In reference to numerical values of α_j, a threshold can be defined, and assignments can be made based on the selected threshold. Generally speaking, there is no distinct criterion for minimum correlation (loading) to determine if a rain gauge (variable) is related to a certain factor or not. However, a value of 0.5 is considered in many studies [33,34]. Therefore, the values equal to, or more than, 0.5 are illustrated in bold in Table 3. Figure 6 shows the clustered region, with regard to the elaboration made above. Careful synthesis of Figure 6 shows that a few spots were assigned to more than one cluster, while some others were not allocated to any cluster at all. The rain gauge stations, along with generic points in each cluster, can be effectively utilized to prioritize rain gauge stations in each cluster. Figure 7 shows the variance of residuals versus the number of rain gauges for the three significant factors, with the number of rain gauges illustrated in bold in Table 3. This information may be effectively utilized to obtain the maximum number of rain gauges representing each cluster. If a variation of the variance of residuals with the number of rain gauges for each factor is simplified by bilinear curves, the point at which the slope changes and variation of the variance of residuals with the number of rain-gauges decreases considerably can be considered as the optimal number of rain gauges. Hence, for each subregion (cluster), two rain gauges are enough. Subsequently, the six rain gauge scenario associated with the proposed approach (i.e., 13, 18, 26, 28, 31 and 33) provides a variance of residual as low as 1916 mm^2 and shares only four stations with that of the time consuming one. Based on the proposed scheme, the selected rain gauge stations are fairly uniformly distributed over the study area, while the ones

Table 3: Factor loading or correlation between rain gauge observations and dominant factors after rotation.

Rain gauge #	Factors					
	1	2	3	4	5	6
1	**0.70**	−0.16	−0.17	−0.11	−0.45	0.24
2	**0.84**	−0.07	0.10	0.08	−0.12	0.22
3	**0.83**	−0.06	0.14	0.10	0.18	0.12
4	**0.72**	−0.03	0.06	0.01	**0.51**	0.06
5	**0.67**	−0.02	−0.02	−0.03	**0.57**	0.08
6	**0.57**	−0.12	0.38	0.02	−0.10	−0.49
7	**0.46**	−0.33	**−0.51**	−0.02	−0.26	−0.05
8	**0.80**	−0.09	0.01	0.02	−0.32	0.27
9	**0.80**	−0.09	0.06	0.04	−0.28	0.24
10	**0.65**	−0.13	0.42	0.04	−0.03	**−0.53**
11	**0.65**	−0.12	0.41	0.04	0.03	**−0.50**
12	**0.73**	−0.14	−0.13	−0.09	−0.44	0.26
13	**0.80**	−0.05	0.13	0.07	0.33	0.08
14	**0.84**	−0.06	0.04	0.08	−0.02	0.23
15	**0.77**	−0.11	−0.19	−0.05	−0.30	0.25
16	**0.56**	−0.26	−0.29	−0.13	−0.41	0.13
17	**0.68**	−0.03	−0.20	−0.07	0.46	0.12
18	**0.67**	−0.15	−0.37	−0.08	−0.24	0.13
19	0.35	−0.16	**−0.68**	−0.04	0.08	−0.32
20	**0.59**	−0.12	0.41	0.03	0.00	**−0.56**
21	0.23	−0.47	−0.09	−0.27	−0.21	0.09
22	0.06	**−0.87**	−0.04	−0.14	0.01	0.12
23	0.08	**−0.89**	−0.15	0.27	0.01	0.09
24	0.10	**−0.77**	−0.23	0.32	0.00	0.02
25	0.03	**−0.86**	−0.01	−0.23	0.04	0.11
26	0.10	**−0.86**	−0.19	0.32	−0.01	0.07
27	0.06	**−0.91**	−0.08	0.05	0.02	0.12
28	0.42	−0.28	**−0.65**	0.02	−0.12	−0.21
29	0.10	**−0.88**	−0.15	0.20	−0.02	0.10
30	0.29	−0.32	**−0.64**	0.09	−0.03	−0.28
31	0.03	**−0.67**	0.04	**−0.51**	0.03	0.09
32	0.00	−0.49	0.01	−0.48	0.07	0.03
33	0.37	−0.09	**−0.59**	−0.10	0.19	−0.26
34	**0.58**	−0.02	−0.23	−0.10	**0.53**	0.07

Table 4: Sensitivity coefficients of factors with respect to standardized mean areal rainfall.

	Factors					
	1	2	3	4	5	6
β_i	0.6753	−0.6518	−0.3133	−0.0965	−0.0241	−0.0086

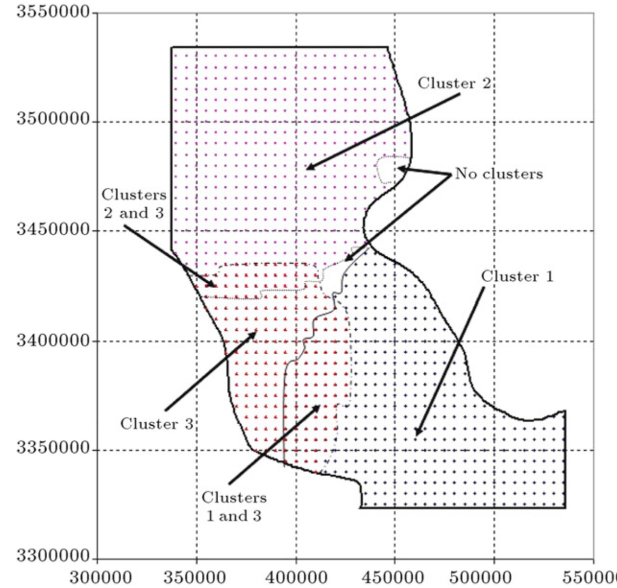

Figure 6: Clustering of study area using coupled geostatistical and multivariate techniques.

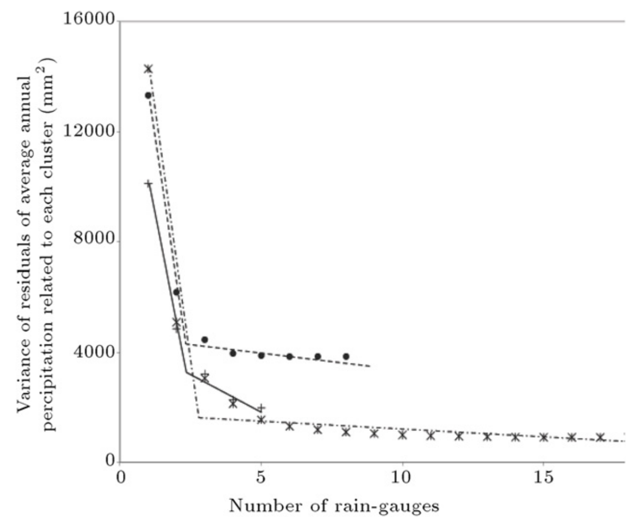

Figure 7: Variance of residual (accuracy) versus the number of points (N)-proposed approach.

proposed by Chen et al. [3] are clustered toward the basin outlet. In flat regions of temperate, Mediterranean and tropical zones, the World Meteorological Organization (WMO) recommends 1 station for 900–3000 km^2. In our case, cluster 3 meets this criterion, but clusters, one and two do not, with 1 station for 5000 km^2 [35].

5. Summary and conclusions

Spatio-temporal mapping of precipitation at various scales in time and space is considered a prerequisite in almost all studies concerning the operation and management of water resources systems. The accuracy of precipitation maps is directly linked to the number and spatial distribution of rain gauge stations. Quantitative assessment of rainfall accuracy calls for a systematic and objective procedure in rain gauge network design. Critical and concise review of existing literature on rain gauge network design highlights the need for clarifying the interaction and subsequent simplification involved between development of the objective function and the optimization algorithm used for its optimization.

In this paper, after identifying the gap in relevant literature, a methodology has been developed and implemented to design and analyze a rain gauge data collection network with the purpose of obtaining the best estimate of long-term mean annual areal rainfall over the study area. According to the proposed approach, for the first time, multivariate statistical techniques have been coupled with geostatistical tools to identify the number of homogeneous zones, to assign objects (i.e., rain gauges or generic nodal points) to various subregions and, finally, to prioritize rain gauges in terms of information content in each subregion. After implementing the proposed methodology on an existing rain gauge network in south west of Iran, the results show that the proposed approach compares well with existing paradigms in rain gauge network design. In particular, the measure of network accuracy lies somewhere in between the, so called, time consuming and more simplified approaches. The following more specific conclusions appear to be in order from the current study:

1. Coupled factor analysis and ordinary block kriging seems to be an appropriate and objective approach to identify the number of homogeneous subregions. There is no need to cluster data for various numbers of clusters and then to identify the optimum number of clusters, as is routine in k-mean clustering.
2. Coupled factor analysis and ordinary point kriging seems to be an appropriate and objective approach to assign rain gauge stations and generic nodal points to various clusters. The assignment is not sensitive to initial guesses on cluster centers, as is the case in k-mean clustering.
3. In the proposed approach, both the number of clusters and the assignment process is objective dependent, while in k-mean clustering, the overall objective has no role to play in clustering. This objective dependency in the proposed scheme is quite important, as, in network design, the network density will be affected by numerous factors mentioned in the introduction.
4. Contrary to popular belief, the simplified paradigm in rain gauge network design does not necessarily lead to the same results as that of more time consuming paradigms.
5. The usefulness of the proposed scheme becomes more distinct if N, the number of rain gauge stations, becomes enormously large.
6. The methodology can be submitted to rigorous physical interpretation with regard to the factors involved, e.g. degree of variability.

An interesting exercise would be to cluster the study area, based on some other clustering procedures, such as k-mean or fuzzy k-mean clustering, and see if it would give rise to the same results as those proposed in this study. Furthermore, it is suggested to implement simulated annealing on the real data of this study, and see how the proposed methodology results are compared with those of simulated annealing.

Appendix

This appendix intended to derive scaled weighting coefficients in terms of ordinary block kriging weighting coefficients. To start with, estimated value of rainfall over block V indexed at $\mathbf{x}_0$ is given by:

$$\hat{P}_{\mathrm{V}}(\mathbf{x}_0) = \sum_{i=1}^{N} \lambda_i^{\mathrm{BK}}(\mathbf{x}_0) P(\mathbf{x}_i),$$

subject to:

$$\sum_{i=1}^{N} \lambda_i^{\mathrm{BK}}(\mathbf{x}_0) = 1. \tag{A.1}$$

Furthermore, variance of estimated value of rainfall over block V indexed at $\mathbf{x}_0$ is given by:

$$\begin{aligned}\mathrm{VAR}[\hat{P}_{\mathrm{V}}(\mathbf{x}_0)] &= \sum_{i=1}^{N}\sum_{j=1}^{N} \lambda_i^{\mathrm{BK}}\lambda_j^{\mathrm{BK}} \mathrm{COV}[P(\mathbf{x}_i), P(\mathbf{x}_j)] \\ &= \sum_{i=1}^{N}\sum_{j=1}^{N} \lambda_i^{\mathrm{BK}}\lambda_j^{\mathrm{BK}} R(\mathbf{x}_i, \mathbf{x}_j).\end{aligned} \tag{A.2}$$

On the other hand, standardized value of $\hat{P}_{\mathrm{V}}(\mathbf{x}_0)$ is given by:

$$\hat{\overline{P}}_{\mathrm{V}}(\mathbf{x}_0) = \frac{\hat{P}_{\mathrm{V}}(\mathbf{x}_0) - m}{\mathrm{VAR}[\hat{P}_{\mathrm{V}}(\mathbf{x}_0)]^{1/2}}, \tag{A.3}$$

where:

$$m = E[P(\mathbf{x})].$$

However:

$$\begin{aligned}\hat{\overline{P}}_{\mathrm{V}}(\mathbf{x}_0) &= \frac{\hat{P}_{\mathrm{V}}(\mathbf{x}_0) - m\sum_{i=1}^{N}\lambda_i^{\mathrm{BK}}(\mathbf{x}_0)}{\mathrm{VAR}[\hat{P}_{\mathrm{V}}(\mathbf{x}_0)]^{1/2}} \\ &= \frac{\sum_{i=1}^{N}\lambda_i^{\mathrm{BK}}[P(\mathbf{x}_i) - m]}{\mathrm{VAR}[\hat{P}_{\mathrm{V}}(\mathbf{x}_0)]^{1/2}} \\ &= \frac{\sum_{i=1}^{N}\lambda_i^{\mathrm{BK}}\mathrm{VAR}[P(\mathbf{x}_i)]^{1/2}\overline{P}(\mathbf{x}_i)}{\mathrm{VAR}[\hat{P}_{\mathrm{V}}(\mathbf{x}_0)]^{1/2}}.\end{aligned} \tag{A.4}$$

As a result:

$$\hat{\overline{P}}_{\mathrm{V}}(\mathbf{x}_0) = \sum_{i=1}^{N} \lambda_i' \overline{P}(\mathbf{x}_i),$$

where:

$$\begin{aligned}\lambda_i' &= \frac{\mathrm{VAR}[P(\mathbf{x}_i)]^{1/2}}{\mathrm{VAR}[\hat{P}_{\mathrm{V}}(\mathbf{x}_0)]^{1/2}}\lambda_i^{\mathrm{BK}} \\ &= \left\{\frac{\mathrm{VAR}[P(\mathbf{x}_i)]}{\mathrm{VAR}[\hat{P}_{\mathrm{V}}(\mathbf{x}_0)]}\right\}^{1/2}\lambda_i^{\mathrm{BK}}.\end{aligned} \tag{A.5}$$

However:

$$\begin{aligned}\rho(\mathbf{x}_i, \mathbf{x}_j) &= \frac{R(\mathbf{x}_i, \mathbf{x}_j)}{\{\mathrm{VAR}[P(\mathbf{x}_i)].\mathrm{VAR}[P(\mathbf{x}_j)]\}^{1/2}} \\ &= \frac{R(\mathbf{x}_i, \mathbf{x}_j)}{\sigma^2},\end{aligned} \tag{A.6}$$

where:

$$\sigma^2 = \mathrm{VAR}[P(\mathbf{x})].$$

As a result:

$$\lambda_i' = \left\{\frac{1}{\sum_{i=1}^{N}\sum_{j=1}^{N}\lambda_i^{\mathrm{BK}}\lambda_j^{\mathrm{BK}}\rho(\mathbf{x}_i, \mathbf{x}_j)}\right\}^{1/2}\lambda_i^{\mathrm{BK}}. \tag{A.7}$$

References

[1] Ghahraman, B. and Sepaskhah, A.R. "Autographic rain gauge network design for Iran by kriging", *Iran. J. Sci. Technol.*, 25(B3), pp. 653–660 (2001).
[2] Cheng, K.S., Lin, Y.C. and Liou, J.J. "Rain gauge network evaluation and augmentation using geostatistics", *Hydrol. Process.*, 22, pp. 2554–2564 (2008).
[3] Chen, Y.C., Wei, C. and Yeh, H.C. "Rainfall network design using kriging and entropy", *Hydrol. Process.*, 22, pp. 340–346 (2008).
[4] Rodriguez-Iturbe, I. and Mejia, J.M. "The design of rainfall networks in time and space", *Water Resour. Res.*, 10(4), pp. 713–728 (1974).
[5] Bras, R.L. and Rodriguez-Iturbe, I. "Network design for the estimation of areal mean of rainfall events", *Water Resour. Res.*, 12(6), pp. 1185–1195 (1976).
[6] Hughes, J.P. and Lettenmaier, D.P. "Data requirements for kriging: estimation and network design", *Water Resour. Res.*, 17(6), pp. 1641–1650 (1981).
[7] Bastin, G., Lorent, B., Duque, C. and Gevers, M. "Optimal estimation of the average rainfall and optimal selection of rain gauge locations", *Water Resour. Res.*, 20(4), pp. 463–470 (1984).
[8] Bogárdi, L. and Bárdossy, A. "Multicriterion network design using geostatistics", *Water Resour. Res.*, 21(2), pp. 199–208 (1985).
[9] Rouhani, S. "Variance reduction analysis", *Water Resour. Res.*, 21(6), pp. 837–846 (1985).

[10] Kassim, A.H.M. and Kottegoda, N.T. "Rainfall network design through comparative kriging methods", *Hydrol. Sci. J.*, 36(3), pp. 223–240 (1991).
[11] van Groenigen, J.W. and Stein, A. "Constrained optimization of spatial sampling using continous simulated annealing", *J. Environ. Qual.*, 27, pp. 1078–1086 (1998).
[12] Krstanovic, P.F. and Singh, V.P. "Evaluation of rainfall networks using entropy: I. Theoretical development", *Water Resour. Manage.*, 6, pp. 279–293 (1992).
[13] Krstanovic, P.F. and Singh, V.P. "Evaluation of rainfall networks using entropy: II. Application", *Water Resour. Manage.*, 6, pp. 295–314 (1992).
[14] Al-Zahrani, M. and Husain, T. "An algorithm for designing a precipitation network in the southwestern region of Saudi Arabia", *J. Hydrol.*, 205, pp. 205–216 (1998).
[15] Yoo, C., Jung, K. and Lee, J. "Evaluation of rain gauge network using entropy theory: comparison of mixed and continuous distribution function applications", *J. Hydrol. Eng. ASCE*, 13(4), pp. 226–235 (2008).
[16] Korvin, G., Boyd, D.M. and O'Dowd, R. "Fractal characterization of the South Australian gravity station network", *Geophys. J. Int.*, 100, pp. 535–539 (1990).
[17] Mazzarella, A. and Tranfaglia, G. "Fractal characterisation of geophysical measuring networks and its implication for an optimal location of additional stations: an application to a rain gauge network", *Theor. Appl. Climatol.*, 65, pp. 157–163 (2000).
[18] van Groenigen, J.W., Pieters, G. and Stein, A. "Optimizing spatial sampling for multivariate contamination in urban areas", *Environmetrics*, 11, pp. 227–244 (2000).
[19] Barca, E., Passarella, G. and Uricchio, V. "Optimal extension of the rain gauge monitoring network of the Apulian regional consortium for crop protection", *Environ. Monit. Assess.*, 145, pp. 375–386 (2008).
[20] Lenton, R.L. and Rodriguez-Iturbe, I. "Rainfall network systems analysis: the optimal estimation of total areal storm depth", *Water Resour. Res.*, 13(5), pp. 825–836 (1977).
[21] Pardo-Igúzquiza, E. "Optimal selection of number and location of rainfall gauges for areal rainfall estimation using geostatistics and simulated annealing", *J. Hydrol.*, 210, pp. 206–220 (1998).
[22] Lee, Y.M. and Ellis, J.H. "Comparison of algorithms for nonlinear integer optimization: application to monitoring network design", *J. Environ. Eng.-ASCE*, 122(6), pp. 524–531 (1996).
[23] Li, M.H., Tung, C.P., Sui, C.H. and Yang, F.H. "Estimating seasonal basin rainfall using tabu search", *Terr. Atmos. Ocean. Sci.*, 17(1), pp. 295–316 (2006).
[24] Wu, J., Zheng, C., Chien, C.C. and Zheng, L. "A comparative study of Monte Carlo simple genetic algorithm and noisy genetic algorithm for cost-effective sampling network design under uncertainty", *Adv. Water Resour.*, 29, pp. 899–911 (2006).
[25] Li, Y. and Chan Hilton, A.B. "Reducing spatial sampling in long-term groundwater monitoring networks using ant colony optimization", *Int. J. Comput. Intell. Res.*, 1(1), pp. 19–28 (2005).
[26] Bellman, R.E., *Dynamic Programming*, Princeton University Press, Princeton, NJ (1957).
[27] Hair, J.F., Black, W.C., Babin, B.J., Anderson, R.E. and Tatham, R.L., *Multivariate Data Analysis*, 6th Edn., p. 730. Prentice-Hall, Inc., Upper Saddle River, NJ (2005).
[28] Rencher, A.C., *Methods of Multivariate Analysis*, 2nd Edn., p. 708. John Wiley and Sons, Inc., New York (2002).
[29] Johnson, R.A. and Wichern, D.E., *Applied Multivariate Statistical Analysis*, 5th Edn., p. 767. Prentice-Hall, Inc., Upper Saddle River, NJ (2002).
[30] Kaiser, H.F. "The varimax criterion for analytic rotation in factor analysis", *Psychometrika*, 23, pp. 187–200 (1958).
[31] Dinpashoh, Y., Fakheri-Fard, A., Moghaddam, M., Jahanbakhsh, S. and Mirnia, M. "Selection of variables for the purpose of regionalization of Iran's precipitation climate using multivariate methods", *J. Hydrol.*, 297, pp. 109–123 (2004).
[32] Chilés, J.-P. and Delfiner, P., *Geostatistics: Modeling Spatial Uncertainty*, p. 695. John Wiley and Sons, Inc., New York (1999).
[33] Kumru, M.N. and Bakac, M. "R-mode factor analysis applied to the distribution of elements in soils from the Aydin basin, Turkey", *J. Geochem. Explor.*, 77, pp. 81–91 (2003).
[34] Chen, K., Jiao, J.J., Jianmin, H. and Huang, R. "Multivariate statistical evaluation of trace elements in groundwater in a coastal area in Shenzhen, China", *Environ. Pollut.*, 147, pp. 771–786 (2007).
[35] WMO *Guide to Hydrological Practices*, In *WMO No. 168*, 1, WMO, Geneva, Switzerland (1981).

Mahmood Reza Shaghaghian was born in 1978, in Shiraz, Iran. He obtained his B.S. degree in Civil Engineering from Shiraz University, Iran, in 2000 and his M.S. degree in Geotechnical Engineering from Ferdowsi University of Mashhad, Iran, in 2003. Currently, he is working on rain gauge network design for his Ph.D. degree program.

Mohammad Javad Abedini was born in 1958, in Shiraz, Iran. He obtained his B.S. and M.S. degrees in Civil Engineering and Hydraulic Structures from Shiraz University, Iran, in 1985 and 1991, respectively, and his Ph.D. degree in Water Resources Engineering from Guelph University, Canada, in 1998. He is currently Associate Professor in the Department of Civil amd Environmental Engeering at Shiraz University, Iran. His research interests include computational hydrology and hydraulics, analysis of temporal and spatial series and data assimilation.