# Multivariate delineation of rainfall homogeneous regions for estimating quantiles of maximum daily rainfall: A case study of northwestern Mexico

FABIOLA ARELLANO-LARA and CARLOS A. ESCALANTE-SANDOVAL
*Facultad de Ingeniería, Universidad Nacional Autónoma de México, Ciudad Universitaria, 04360 México, D.F.*
Corresponding author: F. Arellano-Lara; e-mail: fabi_arelara@yahoo.com.mx

Received October 2, 2012; accepted October 21, 2013

RESUMEN

La escasez de información en el análisis de frecuencias de lluvias máximas diarias puede generar estimadores ineficaces para propósitos de diseño. Una forma de reducir estos errores es la aplicación de técnicas regionales, las cuales requieren que las estaciones involucradas pertenezcan a la misma región homogénea. En este trabajo se realiza una delimitación de regiones homogéneas de precipitación empleando un método multivariado basado en las técnicas de análisis de componentes principales y de agrupamiento jerárquico ascendente. La metodología propuesta se aplicó a una región del noroeste de México. Se concluyó que sólo se requieren los coeficientes de variación de los momentos-L y de la latitud, longitud y altitud de cada estación climatológica para definir las regiones homogéneas de precipitación, y que la inclusión o exclusión de información en las técnicas regionales tiene un impacto directo en la estimación de eventos asociados a diferentes periodos de retorno.

ABSTRACT

Lack of data in maximum daily rainfall frequency analysis can generate inefficient estimates for design purposes. An approach to diminish these errors is to apply regional estimation techniques, which require that all stations be located at the same homogeneous region. In this paper, a delineation of homogeneous precipitation regions was made based on the multivariate methods of principal component analysis and hierarchical ascending clustering. A region in northwestern Mexico was selected to apply this methodology. It was concluded that only the coefficients of variation of the L-moments, along with latitude, longitude and altitude at each climatological station are sufficient to define the homogeneous rainfall regions, and that either the inclusion or exclusion of information in the regional techniques has a direct impact on the estimation of events associated to different return periods.

**Keywords:** Homogeneous rainfall regions, principal component analysis, hierarchical ascending clustering, regional frequency analysis.

## 1. Introduction

The North American Monsoon System (NAMS) is defined as a pronounced increase in rainfall from an extremely dry June to a rainy July over large areas of the southwestern United States and northwestern Mexico (Adams and Comrie, 1997). The occurrence of NAMS is associated to atmospheric dynamics conditions and topographic characteristics, which interact with each other to cause a convective environment. This phenomenon can generate a high potential danger of flooding to residents in the country. In order to protect their lives and goods, it is very important to have a mathematical tool that may reduce the uncertainties in estimating design events for different return periods, which are needed in many hydraulic studies and projects such as flood plain delineation or drainage works in cities.

In maximum daily rainfall frequency analysis, when information exists but not with the length of record required to provide accurate parameter estimates, the error of the estimated value for some return periods can be very large and inefficient for design purposes. A way of reducing this error is by applying a joint estimation model where information from nearby sites in the same region may be combined with the record of inadequate length. This approach will increase the amount of information and will provide a regional at-site estimate. An example of these regional models is the station-year technique, which is used to obtain a regional at-site estimate of the maximum daily rainfall for different return periods (Cunnane, 1988). These events are necessary to shape the intensity-duration-frequency curves (IDF) whose intensities $i$ (mm/h) associated to certain duration $d$ (h) and return period $T$ (years) are used for designing hydraulic works.

The regional analysis correlates hydrological variables with the physiographical and climatological characteristics. Through these regional relations it is also possible to obtain flow estimates in rivers, as it can be seen in Wiltshire (1985), Stedinger (1983), Gingras and Adamowsky (1993), Burn (1988), Robinson (1997), Gutiérrez-López (1996), Escalante and Reyes (1998, 2000), Pandey and Nguyen (1999), Ouarda *et al.* (2001), Gómez (2003), Skaugen and Vaeringstad (2005), and Ouarda *et al*. (2008).

The regional techniques require that the involved stations belong to the same homogeneous region. Since the inclusion or exclusion of information has a direct impact on the estimation of events associated to different return periods, adequately establishing that such homogeneity is achieved is an essential step to reduce the associated uncertainties.

A homogeneous region can be delineated by using geographical characteristics or statistical tests. Some works also have proposed indexes to evaluate the uncertainty and applicability of these methods: Nouh (1987), Cunnane (1988), Rosbjerg and Madsen (1995), GREHYS (1996a, b), Campos (1999), and Lin and Chen (2003).

In this work, the delineation of homogeneous regions is based on multivariate methods: principal component analysis (PCA) and hierarchical ascending clustering (HAC).

## 2. Materials and methods

### *2.1 Principal component analysis*

PCA is a multivariate statistical technique highly descriptive, which is used to identify patterns on data in such a way as to highlight their similarities and differences. PCA can reduce the dimensionality of the data, transforming the set of $r$ original variables or attributes in another set of $s$ uncorrelated variables called principal components. The $r$ variables are measured on each of the $m$ sites. The order of the initial matrix of data is $mr$ and it is restricted to $m > r$. After applying the PCA technique, the order of the resulting matrix is $ms$. This reduction of dimensionality is achieved with a little loss of information, which is considered non-significant to preserve the principal components.

PCA allows using either the correlation matrix or the covariance matrix. The first option gives the same importance to all and each of the variables. This can be convenient when the researcher considers that all the variables are equally relevant. The second option can be used when all the variables have the same units of measure.

The $s$ new variables (principal components) are obtained as linear combinations of the $r$ original variables. Components are arranged according to the percentage of variance that can be explained. In this sense, the first component will be the most important since it explains the largest percentage of the variance of data. Each researcher will decide how many components will be elected in the study.

PCA is performed in the space of the $r$ variables and, in dual form, in the space of $m$ sites. Variables and sites can be graphically represented by considering the first and second component as coordinate axes. A point-variable is represented by the coordinate corresponding to that variable in each of these components. The cloud of points-variables is located in a circular area of radius 1. The proximity between the point-variables indicates the degree of correlation between them. When the correlation is equal to one, the points coincide.

When the $r$ variables are uncorrelated, $r$ equally important components will be obtained. In contrast, when all variables have a perfect correlation, a simple component is generated. This component is a linear combination of the $r$ equally weighted variables and explains 100% of the total variation.

The cloud of points-sites is not enclosed in a circle of radius 1. A point-site located at the extreme of one axis means that such station is closely related to the respective component. The opposite case indicates that the site has no relation with the two