Master's Thesis of Science in Agriculture

Identification of Homogeneous Precipitation Regions with Time Series Gauge and Satellite Data Using Machine Learning Methods

기계 학습을 이용한 관측 및 위성 강수 시계열 데이터 기반 동질 강수 지역 분석 연구

August 2023

Munyensanga Shimwa Desire

Department of International Agricultural Technology
Graduate School of International Agricultural Technology
Seoul National University

Identification of Homogeneous Precipitation Regions with Time Series Gauge and Satellite Data Using Machine Learning Methods

Supervised by
Prof. Hakkwan Kim

Submitting a master's thesis of
Science in Agriculture

Major of Green Ecosystem Engineering
Department of International Agricultural Technology
Graduate School of International Agricultural Technology
Seoul National University

Confirming the master's thesis written by
Munyensanga Shimwa Desire

August 2023

Chair	**Kyo Suh, Ph.D.**	_____
Vice Chair	**Hakkwan Kim, Ph.D.**	_____
Examiner	**Joon Weon Choi, Ph.D.**	_____

Abstract

Identification of Homogeneous Precipitation Regions with Time Series Gauge and Satellite Data Using Machine Learning Methods

Munyensanga Shimwa Desire
Department of International Agricultural Technology
Graduate School of International Agricultural Technology
Seoul National University

Homogeneous regions are often needed for region frequency analysis and precipitation estimation, but the formation of those regions is often associated with a lot of uncertainties due to temporal and spatial variability of precipitation.

This study tackles two challenges related to the formation of homogeneous precipitation regions from ground gauge data. The first challenge is the temporal variability of precipitation which is not often considered in the formation of homogeneous regions. It is well known that precipitation varies a lot in time and space. However, many past studies on the formation of homogeneous precipitation regions did not capture the important aspect of temporal variability of precipitation because they usually use other variables such as averages, and location features instead of time series data. To overcome the temporal variability challenge, this study used timeseries precipitation data to form homogeneous precipitation regions.

The second challenge is the variation of precipitation in space. Rain gauge had been traditionally the main source of precipitation data as they were considered more accurate than other source of precipitation data; however, rain gauge measures precipitation at a point in space. It is challenging to accurately interpolate point data over an area, given that the density of gauges is often scarce in many regions of the

world and interpolation technics may introduce errors.

To overcome the spatial variability challenge, this study used satellite data to form homogeneous precipitation regions. Satellite derived data is a relatively recent source of precipitation data where precipitation is indirectly estimated from infrared and passive microwaves information received from several satellite sensors. The estimates products were surface data because they are released in the form of surface grids.

A machine learning approach was provided in this study to form homogeneous precipitation regions using gauge and satellite daily time series data. The ground precipitation data used in this study were provided by Korea Meteorological Agency (KMA). Data from the Automated Synoptic Observing System (ASOS) and Automatic Weather Station (AWS) were used respectively. Satellite data used in this study was the Integrated Multi-satellitE Retrievals for GPM (IMERG) from National Aeronautics and Space Administration (NASA).

Precipitation regions were formed using two clustering methods, K-Means and Self Organizing Maps (SOM). Both clustering algorithms were able to define homogeneous precipitation regions from time series gauge and satellite data. Spatial maps of the regions were provided in the results and discussion section of the present study. Heterogeneity results were compared by using Hosking and Wallis homogeneity test. Based on the clusters formed by SOM and K-Means in ASOS dataset, it was observed that the performance of SOM in defining homogeneous regions is greatly affected by the size of the map. SOM was able to identify a bigger number of homogeneous regions when the number of nodes was increased. It was able to identify 6 homogeneous regions when the number of nodes was increased to 16 while K-Means identified 5 homogeneous regions for the same number of clusters. K-Means was able to identify a greater number of homogeneous regions when

cluster number was small. For example, when the number of clusters was 10, K-Means identified 3 homogeneous regions while SOM identified 2 homogenous regions.

However, the number of homogeneous and possibly heterogeneous regions identified by SOM gradually increased as the number of nodes increased from 10 to 16.

Based on the number of homogeneous regions identified by SOM and K-Means in AWS datasets, both clustering methods identified similar number of regions in AWS dataset. The number of homogeneous regions identified by both clustering methods did not improve when the number of clusters were increased to 12, 14 or 16

Based on the number of homogeneous regions identified by SOM and K-Means in satellite dataset, both were able to identify almost the same number of homogeneous regions, although there were differences between SOM and K-Means according to the number of clusters. K-Means identified 2 homogeneous regions among 9 clusters while SOM identified 4 homogeneous regions and 2 possibly homogeneous regions in the same number of clusters Both clustering methods were able to identify 10 homogeneous regions when the number of nodes was increased to 16 however K-Means also identified 2 possibly heterogeneous regions.

Overall, it was observed that SOM was slightly more efficient in identifying a greater number of homogeneous regions in ASOS and satellite datasets.

Keywords: Precipitation homogeneous regions, Satellite data, Machine learning, Gauge data, Time series data

Student Number: 2021-22782

Table of Contents

List of Tables

List of Figures

Chapter 1. Introduction

1.1. Background

The ability to accurately estimate precipitation has increasingly become a necessity in different disciplines (Claps et al., 2022). The need for an accurate global estimation of rainfall has been amplified by recent awareness of climate change and its effects. Thus, being able to accurately estimate the amount of precipitation in a given region is a necessity in various engineering fields such as civil engineering, for the construction of water management structures, and agriculture for the scheduling of irrigation systems or hydrology to monitor climate change (Claps et al., 2022). Furthermore, precipitation estimates at the watershed level are very important input for hydrological modelling to estimate and forecast stream flow (Claps et al., 2022). It is therefore very important to focus on improving the input precipitation to expect a good performance of hydrological models.

The distribution of precipitation in time and space can be more accurately estimated if information from other related site are used instead of using information from one sample or site. This principle of regional frequency analysis can be applied to whenever there are several samples of the same kind of data. The process of grouping the sites which exhibit similar behavior into clusters is called regionalization and has many advantages in environmental sciences where the same kind of data are observed at different sites (Hosking & Wallis, 1997).

Regionalization of precipitation is frequently used in different domains such as planning of land use, floods mitigation, drought analysis, prediction of precipitation, and downscaling of precipitation (Hosking & Wallis, 1997). Regionalization is also used for frequency analysis to get design values of infrastructures related to water resources engineering such as dam and sewer systems when at site observation are not available. It is similarly used to improve the reliability of observations at site .

Regionalization is likely to give improved accuracy in estimates because more information is used in the case of regional frequency analysis as compared to single site analysis. This is very advantageous especially in remote or mountainous areas often characterized by complex precipitation patterns and scarcity of precipitation gauge (Hosking & Wallis, 1997).

Defining homogeneous precipitation regions using statistics computed from point precipitation gauge is relatively more precise because the accuracy of gauge data proved to be superior to other data commonly used to estimate precipitation such as satellite derived data. Point rain gauge has traditionally been the main source of rainfall data but the networks of those gauges are limited in terms of spatial distribution. In addition, those time series are often incomplete in terms of time distribution. Hydrologists use regionalization to be able to estimate records of an ungauged or insufficiently gauged area or to fill missing data of incomplete sites, by using records of other sites which are believed to behave in a similar manner. When sites are grouped into cluster of similar behavior, the records of those sites can be used with confidence to predict variable of ungauged sites within the same group (Nathan & McMahon, 1990). The steps involved in identifying homogeneous regions can be grouped into following categories, I) To select variables, II) To select which grouping method to use, III) To define homogeneous regions, IV) Evaluation of homogeneous groups.

Various methods have been used to group precipitation stations into clusters that exhibit required statistical homogeneity. Methods based on statistical analysis of rain gauge observations were traditionally used in defining homogeneous precipitation regions. The statistical methods can be classified into following categories, (i) correlation analysis, (Arthur & Vassilvitskii)principal component analysis, (iii) factor analysis. (iv) hierarchical approach, (v) region of influence and (v) cluster analysis (Srinivas, 2013).

Many researchers recently used cluster analysis because it can easily identify patterns in very complicated dataset.

The choice of grouping methods significantly affects the results of homogeneous regions because different methods may produce different results (Kalkstein et al., 1987). Earlier researches used several interpolation technics to estimate precipitation in ungauged sites, these technics includes Thiessen polygons, Lagrange approach, Inverse Distance Technics, multiquadric interpolation, optimal interpolation, kriging techniques, and others. Various studies have compared the results of different interpolation and grouping technics(Haddad et al., 2015; Jackson & Weinand, 1995; Kalkstein et al., 1987). Tabios and Salas (1985) compared different technics used in earlier interpolation research and concluded that geotechnical techniques (Kriging and optimal interpolation) produce better results than other technics including multiquadric, inverse distance interpolation, Thiessen polygon and polynomial interpolation. Alam and Paul (2020) concluded that Fuzzy C means performed better than Agglomerative hierarchical clustering and K-means clustering algorithm. However, those studies were conducted using other variables such as latitudes, longitudes, and annual averages of rainfall; but timeseries data have not been extensively used to form homogeneous precipitation regions. Regionalization of precipitation using time series data have not been extensively conducted mainly because clustering of time series data present unique challenge caused by the high dimensionality of timeseries dataset (Roushangar & Alizadeh, 2018).

1.2. Purpose

The purpose of this study is to provide an approach to overcome challenges related to temporal and spatial variability of precipitation in the formation of precipitation regions. This study has three objectives. The first objective is to define homogeneous regions using time series data to consider temporal variability of precipitation in the formation of homogeneous regions. The second objective is to use satellite data to overcome the longstanding challenge of defining homogeneous regions in ungauged areas. The third objective is to compare two clustering algorithms and recommend best method for defining homogeneous regions using time series data. This study provides an approach to define homogeneous region using time series precipitation from gauge observation and satellite estimates. Satellite data are gridded dataset, meaning that they are surface data while gauge data are point data. There is therefore a great advantage in delineating homogeneous regions using satellite data because regions can be defined even in ungauged areas and boundary are clearly delimited by grid cell boundaries. The comparison of different technics will provide a better understanding of the advantage and disadvantage of different grouping methods and thus help engineers and scientists to make better choice of proper grouping methods to use in regionalization of precipitation.

Chapter 2. Review of literature

2.1. Regionalization of precipitation

The importance of regionalization of precipitation and related applications has been the subject of many researches in the past. Different approaches have been used in literature for regionalization of precipitation. Annual, monthly, daily data or other form of precipitation data have been used as input for regionalization of precipitation. Mallants and Feyen (1990) conducted a study to define homogenous rainfall regions in the Ijzer watershed in norther France and western Belgium. They used principal component analysis and daily precipitation data collected from 11 stations for 3 years, a dry year: 1973, a wet year: 1977, and an average year: 1978. Four regions were delineated by using principal components in the watershed of Ijzer.

Rasheed et al. (2019) used event-based characteristics such rainfall intensity, antecedent dry days, total rainfall, and rainfall duration to define homogeneous regions. They used cluster analysis and data from 17 stations for a period from 2011 and 2015 to group precipitation stations into homogeneous regions. The homogeneity of the regions was tested by using Hosking- Wallis heterogeneous tests. They found out that the entire region of Southeast Queensland was homogenous based on conventional delineation of homogenous regions, but it was also found that the region could be divided into 2 homogeneous regions when delineation of homogenous regions was based on event-based rainfall.

A number of researches related to regionalization of precipitation have been conducted in Korea. Nam et al. (2015) delineated climatic rainfall regions in Korea using multivariate and regional frequency analysis. Factor analysis, and fuzzy C-Means clustering were used to cluster annual maximum data from 67 stations across Korea. They compared the at site frequency analysis with regional frequency

analysis and concluded that the regional frequency analysis estimates were more accurate. Kim et al. (2012) identified 6 homogeneous regions by using Self Organizing Maps and considering 61 gauges stations with data from 1980-2010.

Most of previous studies uses annual, seasonal or monthly precipitation in the formation of homogeneous regions even though It has been proved that regionalization of precipitation based on different time scales produces different results (Saikranthi et al., 2013); and it has been recommended to choose a temporal resolution based on the final utilization of the precipitation regions because all homogeneous regions may not be suitable for every purpose (Irwin et al., 2017). There was therefore a need to study the formation of homogeneous regions based on time series with small scale temporal resolution (Irwin et al., 2017). The present study used time series daily precipitation data to consider small scale temporal variability in the regionalization of precipitation.

2.2. Homogeneous regions using other source of data.

Conventional methods of regionalization which use statistics to define homogenous regions has different limitations, as it cannot be used in regions with few or no precipitation gauges. The study by Satyanarayana and Srinivas (2011) used fuzzy clustering analysis to define homogenous precipitation regions in scarcely gauged areas using other source of data instead of precipitation data. Large scale atmospheric variable (LSAV) such as location parameters, latitude, longitude, altitude, and seasonality of precipitation were used as variables instead of precipitation to form homogeneous regions. To validate the regions, they used dataset from India Meteorological Department (IMD) which contains annual gridded rainfall data of 2140 stations with records between 1959 and 2004. The regions identified through this method can be validated using statistical analysis of data collected on site.

The methods commonly used in past studies on the regionalization of precipitation was to form homogeneous regions by using data from rain gauge stations (Claps et al., 2022). But those methods are known to have a major limitation in ungauged or scarcely gauged areas because they cannot be used in areas with no gauges and cannot produce meaningful results in areas with few gauges (Satyanarayana & Srinivas, 2011). In order to overcome the challenge of forming homogeneous regions in sparsely gauged areas, Satyanarayana and Srinivas (2011) used large-scale atmospheric variable (LSAV) as input data for the formation of homogeneous regions. However, the variables used as input were not precipitation variables but other variables which influence precipitation. There are a lot of uncertainties associated with using variables which influence precipitation instead of using precipitation estimates in regionalization of precipitation.

To overcome the spatial variability challenge, this study used satellite data to form homogeneous precipitation regions. Satellite data present a huge potential in the formation of precipitation regions because they are surface data and are widely available so they can be used to form homogeneous regions in ungauged areas.

2.3. Satellite Precipitation data

Many of the past studies on satellite precipitation data focused on improving the accuracy of released satellite precipitation products. Chen et al. (2020) estimated daily precipitation for the valley of Xijiang in the Southeast of China for a period of 8 years starting from 2010. Data from rainfall gauging stations were merged with precipitation data from 4 satellites products, the TRMM multi-satellite precipitation analysis (TMPA), Climate Prediction Center (CPC) morphing technic (CMORPH), Precipitation Estimation from Remote Sensed Information using Artificial Neural Network (PERSIANN), and Global Satellite Mapping of Precipitation (GSMaP), using a combination of geographically weighted regression (GWR) and ridge regression. The study concluded that the estimation of precipitation was greatly improved in accuracy by merging rain gauge data with SPP data. Furthermore, the study suggests that the use of multiple SPP give better results that the use of a single SPP.

Zhang et al. (2021) used a new method of double machine learning to combine precipitation estimate from different satellites data and gauged data from China. The research used semi daily data from 697 rain gauge station distributed in China. The satellite products used were IMERG, PERSIANN, GSMap and the satellite product derived from ASCAT soil moisture product. The resolution of all those products were uniformized to $0.1^{\circ}\times0.1^{\circ}$ The regression models of random forest (RF), artificial neural network (Rasheed et al.), support vector machine (SVM) and extreme learning machine (ELM) were used to develop SML algorithms. The classification model of RF in combination with the regression model of RF, ANN, SVM and ELM were used to developed the double machine learning (DML) algorithms. 70% of the gauges, were randomly sampled in each subregion and their data were us as training dataset while 30% of the gauges data were randomly sampled and used as test data group.

The continuous and categorical metrics including the Kling-Gupta efficiency (KGE), Probability of detection (POD), success ratio (SR), bias score (BS) and critical success index (CSI) were used to evaluate those products and the original SPPs. The study generated 12 products among them, 4 precipitation products were obtained using DML algorithms, 4 products were generated using SML algorithms, 3 products were obtained using linear merging methods and 1 product was obtained using gauge only interpolated product. A recent study by Wang and Yong (2020) evaluated two satellite-based products, IMERG and GSMaP with gauge observations on 6 continents. The daily time resolution was used while spatial resolution of 0.5 degree were used for the period start by 2015 to the end of 2018. They concluded that IMERG perfomers better than GSMaP.

2.4 Self organizing maps

Self-organizing map (SOM) is a neural network published by Kohonen (1990). As explained by Miljkovic (2017), SOM is one of the most common neural networks. SOM is commonly used to map high dimensional data to a two-dimensional grid for visualization purposes to overcome the limited ability of humans in visualizing such dimensional data (Miljkovic, 2017). SOM is used to identify and classify patterns in spatial-temporal space because of their ability to produce a map of features. SOM has been used in the classification and interpretation of satellite imagery(Giacco et al., 2010), including identification of land use classes from satellite imagery or identification of sources of dusts from satellite imagery(Lary et al., 2016). SOM has been used for environmental analysis to analyze spatial and temporal patterns of pollutants (Licen et al., 2023). SOM has been used in other fields including medical imaging to analyze diseases from medical images or in maritime applications for planning ship trajectories and in robotics for learning the motion map and solving traveling salesman problem (Miljkovic, 2017). SOM has been used in hydrology to analyze variation of ground water (Varouchakis et al., 2023).

SOM has also been used to define homogeneous precipitation regions. Annual precipitation data from 31 rain gauges and spanning a period from 1960 to 2010 were used in a clustering approaches proposed by Roushangar and Alizadeh (2018) to define homogeneous precipitation regions in Iran. Discreet wavelet transforms were used to get the features of the time-frequency of the time series. Homogeneous regions were defined using k-means and Self Organizing Maps (SOM) clustering techniques. The annual precipitation time series was pre-processed to get time related coefficients and the input layer was determined using the same coefficients instead of the time series data. The efficiency of the model in clustering was verified using 3 indexes: the silhouette coefficient, the Dunn index and Davis Bouldin index.

Outcomes of the studies showed that K-means clustering performed better in comparison to SOM.

SOM was anticipated to be an appropriate choice to work with high dimensionality of time series data because of the proved ability of SOM in identifying patterns in high dimensional data, and ability in identifying and classifying patterns in spatial-temporal space. Therefore, SOM was chosen in this study to form homogeneous regions by using time series data.

Chapter 3. Material and Methods

3.1. Data

3.1.1. Gauge data

The data used in this study were from the dataset of Korea Meteorological Agency (https://data.kma.go.kr). Two sources of gauge data were considered in this study as shown in Figure 1 and Figure 2. 103 ASOS (Automated Synoptic Observing System) stations and 511 AWS (Automatic Weather Station). After a thoroughly checking of the dataset, 57 ASOS stations were retained for further analysis because they have continuous records of 30 years daily precipitation. Finally, 66 ASOS stations and 375 AWS were combined to get the total number of 441 stations with 20 years continuous precipitation records.

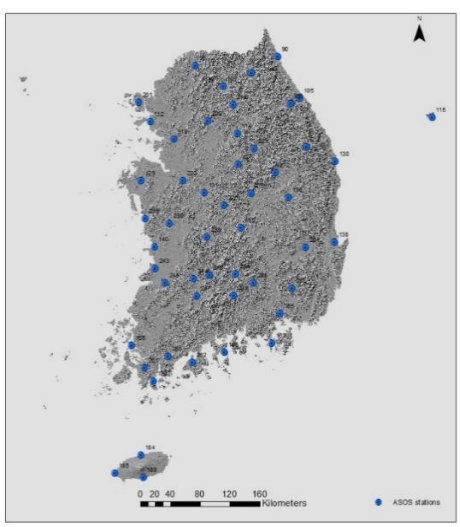


Figure 1. ASOS stations

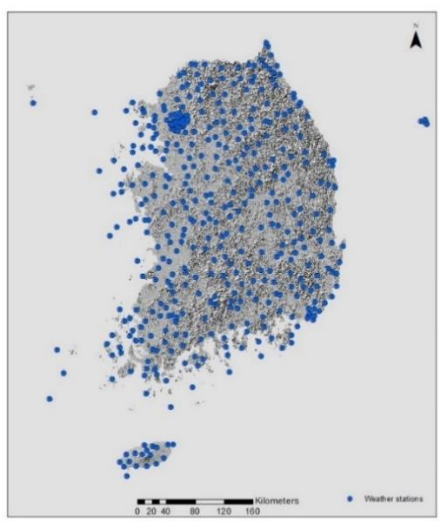


Figure 2. AWS stations

AWS and ASOS stations (441 stations) spread across Korea were analyzed to give a better insight into the formation of regions from time series data and permit to make conclusions about the abilities of grouping technics. The number of stations and corresponding records periods were stated in Table 1.

Table 1. Number of gauge stations

No	Name	No of stations (2001-2020)	No of stations (1991-2020)
1	ASOS	66	57
2	AWS	375	-
Total	AWS+ ASOS	441	57

All the ASOS series have 10, 958 days (30 years daily records). The series which have missing values have been excluded from the analysis. The variables considered were the daily precipitation data from 1991-2020 in the first analysis. In the second analysis, a dataset of 20 years daily precipitation was used as input for the clustering algorithms. 375 AWS stations and 66 ASOS stations with 20 years daily precipitation continuous records were grouped into homogeneous regions. The precipitation time series have equal length of 7,305 days (20 years). The 441 stations were grouped into homogeneous regions using two different technics, K-Means clustering and SOM and the resulting regions were compared for a better understanding of the grouping technics advantage and disadvantage. The data were processed using Python and R packages. The pre-processing of the data started by checking the length of the time series, series with missing data were removed from the analysis, then all the time series were normalized to have a scale between 0 and 1. The normalization of the data was done using python skit learn preprocessing functions.

3.1.2. Satellite data

The satellite data used in this research was the Integrated Multi-satellitE Retrievals for GPM (IMERG). A map accumulated dataset of daily precipitation of South Korea for 20 years from January 2001 to December 2020 were used in the present study. The data were downloaded from NASA's satellite precipitation data repository (https://giovanni.gsfc.nasa.gov/giovanni/). The data were extracted from satellite files to csv file using Python code. IMERG dataset combines precipitation data from a constellation of satellites from Global Precipitation Measurement (GPM), a mission of National Aeronautics and Space Administration (NASA), to produce a gridded precipitation estimate on a global level with a spatial resolution of 0.1 ° to 0.1 ° and time resolution of 30 minutes or monthly. The diagram of the GPM core observatory was shown in Figure 3.

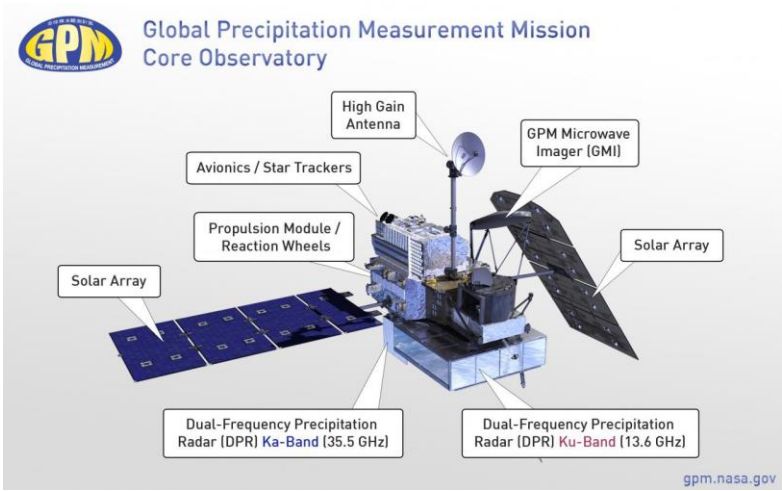


Figure 3. GPM core diagram (Source: https://gpm.nasa.gov/missions/GPM/core-observatory)

The data used in this study were from IMERG version 6, more specifically the IMERG derived product of daily estimates formed by the Goddard, Earth Science (GES), Data and Information Services Center (DISC) at NASA's Goddard Distributed Active Archive Center(DAAC) (Leptoukh et al., 2001).

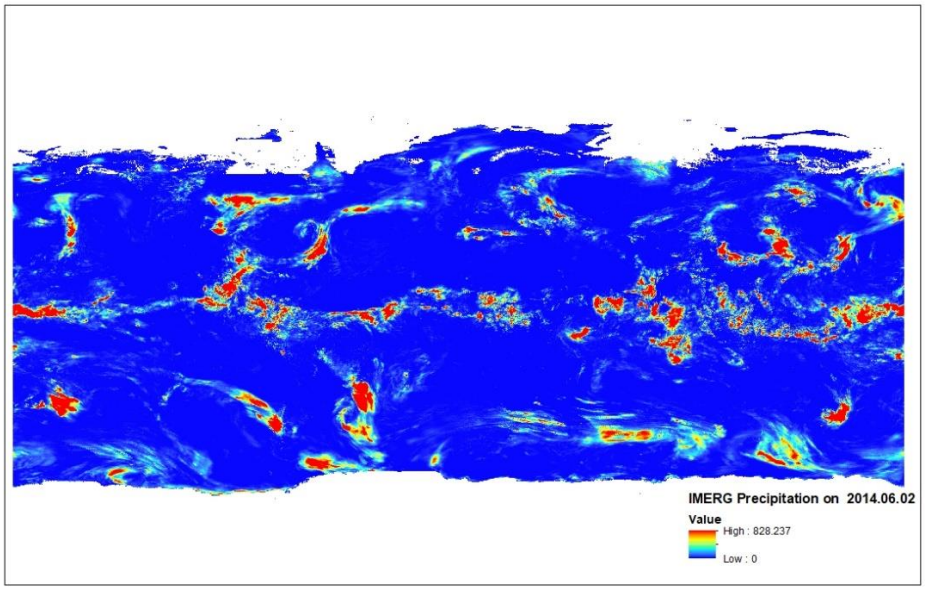


Figure 4. Example of IMERG world Precipitation on 2014.06.2

IMERG was released in 3 batches by the Precipitation Processing Center (PPS) of NASA Goddard. The first batch Early IMERG was released few hours after the satellite observations and is intended to be used for disaster monitoring, the late IMERG, and the final IMERG which released months after the satellite observation. The final IMERG was used in the present study because it was the most accurate and was recommended for research purposes.

Satellite derived precipitation was traditionally estimated from 2 main types of sensors: The passive microwaves sensors mounted on low earth-orbit (leo) satellite and Infrared (IR) sensors mounted on geosynchronous-Earth-orbit (Haddad et al., 2015). Satellite precipitation were traditionally estimated from passive microwaves

sensor passive microwave (PMW) sensors which have limited sampling capabilities. The IMERG on the other hand used many leo satellite and combine the outcome with geosynchronous-Earth-orbit (Haddad et al., 2015).

The raw data from IR on geo satellites were in form of brightness temperature (Tb). IR does not measure directly surface precipitation but rather measure the temperature or radiation reflection on the top of clouds (Figure 5). One way to improve the relationship estimated between the Tb measured from IR and surface precipitation was to use estimates from PMW sensors.

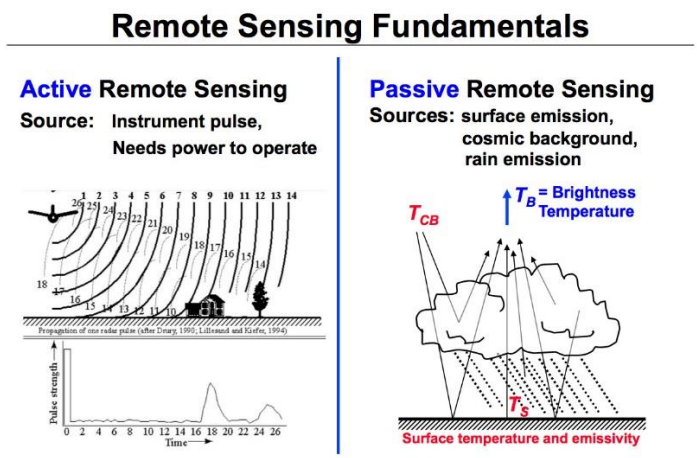


Figure 5. Remote sensing diagram (Source: https://gpm.nasa.gov/image-gallery/active-and-passive-remote-sensing-diagram Source NASA)

The satellite products were obtained as level 2 precipitation estimates. The PPS collect estimates from different PMW provider in the form of brightness temperature, intercalibrate them and compute the next level of estimate of brightness temperature. After intercalibration, precipitation estimates were computed using internal algorithm. The estimates were gridded and were used by different institution to produce their own estimates. An example of study area IMERG precipitation imagery was shown on Figure 6.

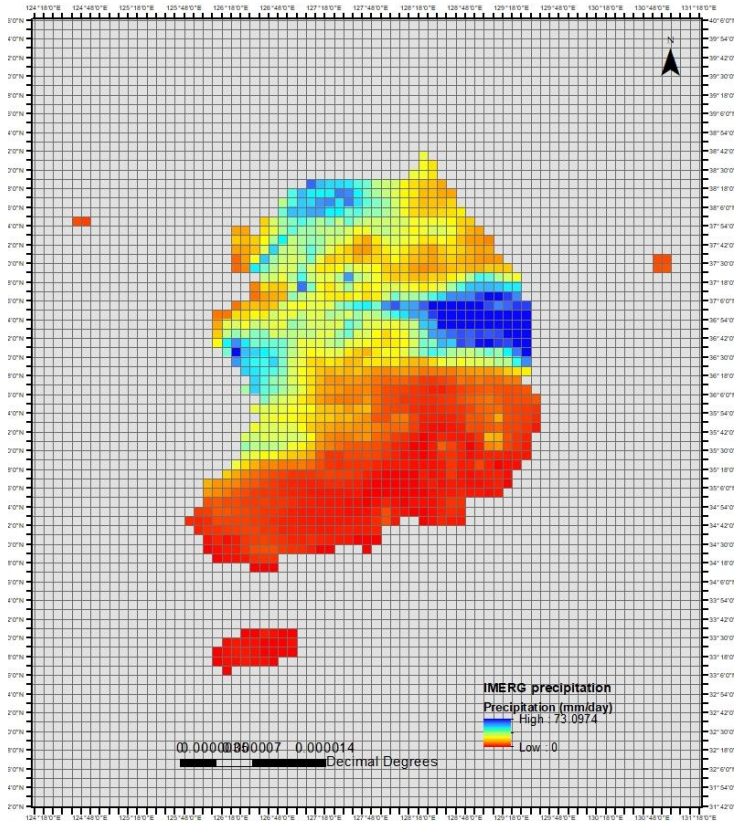


Figure 6. Example of study area satellite precipitation imagery

3.2. Clustering methods

Different methods have been used in past to group precipitation stations into clusters that display required statistical homogeneity. Methods based on statistical analysis of rain gauge observation have been extensively used in defining homogeneous precipitation regions in the past but new methods involving machine learning algorithms are increasingly used to define homogeneous precipitation regions(Carvalho et al., 2016). The present research used two clustering methods, K-Means clustering and Self-Organizing Maps to form homogeneous precipitation regions in Korea. The study scheme was summarized in Figure 7.

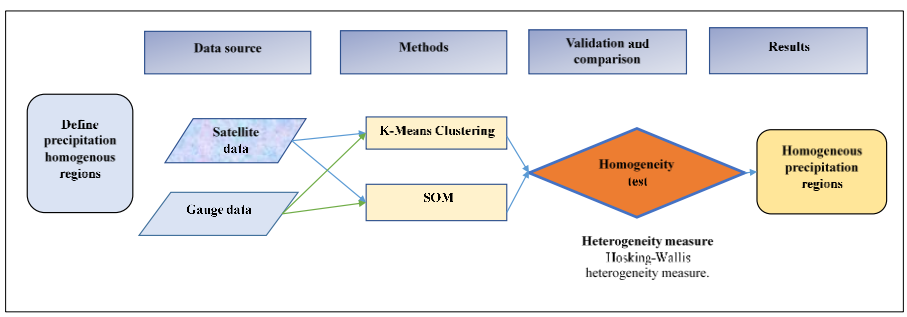


Figure 7. Research scheme

Two sources of data, gauge and satellite data were used as input for the two clustering algorithms. The resulting clusters were tested by using Hosking and Wallis homogeneity test. Clusters from both clustering methods were mapped and results were compared in the results and discussion section of this study.

The clustering methods used and compared in this study vary in their implementation principles, but they represent the main conventional methods commonly used in clustering precipitation data. K-Means clustering represents the simple and classic clustering methods commonly used in defining homogeneous

precipitation regions while Self-Organizing Maps represent the machine learning model increasingly used to define homogeneous precipitation regions. Their comparison will provide an insight into the performance of simple algorithms in clustering timeseries precipitation data compared to more complex neural network machine learning algorithms.

3.2.1. K-Means Clustering

K-Means clustering algorithm (Ralambondrainy, 1995) was used to cluster precipitation stations into homogeneous regions. K-means clustering is a centroid based machine learning algorithm widely used in data mining community to partition a dataset into corresponding clusters.

K-means clustering was implemented using python machine learning library called tslearn. The Euclidean distance was chosen as the methods to calculate distance between points. The tslearn package provides different tools for machine learning analysis of time series data. The TimeSeriesKMeans was used to partition the dataset into corresponding clusters. The metric used for cluster assignment is Euclidean distance. The initialization method which uses specific probabilities to choose random initial centers is utilized in this study for initialization of K-Means clustering. The method is commonly known as k-means++ and was proposed by Arthur and Vassilvitskii (2007). This initialization methods is more accurate and faster than default random initialization(Arthur & Vassilvitskii, 2007). The number of clusters for K-Means clustering algorithm was chosen to be the optimum number of clusters determined by Elbow method.

The centroid based cluster is represented by center point called centroid which is often the mean or median of all the points in the clusters. K-means algorithm initially assigns each data points to a centroid and iteratively updates the centroids until each

data point is assigned to the nearest centroid. The objective of K-means cluster analysis is to minimize the sum of squares with in the cluster.

If the target object is $\boldsymbol{x}$, and the average of cluster $\boldsymbol{C_i}$ is x_i, the criterion function is given by the following formula:

$$\boldsymbol{E} = \sum_{\boldsymbol{i=1}}^{\boldsymbol{k}} \sum_{\boldsymbol{x \in C_i}} |\boldsymbol{x} - \boldsymbol{x_i}|^{\boldsymbol{2}} \qquad \textbf{(1)}$$

$\boldsymbol{E}$ is the sum of the squared of all the points in the database. Euclidean distance is used to calculate the distance between each data point and the center of the cluster.

Euclidean distance d between point x_1 and y_1 from two vectors x = (x_1, x_2,…. x_n) and vector y = (y_1,y_2,…… y_1). The Euclidean distance between two points can be calculated as below:

$$\boldsymbol{d}\,(\boldsymbol{x_i}, \boldsymbol{y_i}) = \left[\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} (\boldsymbol{x_i} - \boldsymbol{y_i})^{\boldsymbol{2}}\right]^{\boldsymbol{1/2}} \qquad \textbf{(2)}$$

Assuming a data set of n statistical individuals for which we know the value of p variables. The strategy of classification consists of determining the distance between points and distance between a group of points. One of the algorithms, hierarchical ascending start by searching the closest elements and put them into a new object (group of objects) generated by the algorithm. The algorithm calculates distance between the object and remaining objects to be classified as in previous step, but with only n-1 object remaining. The algorithm searches again for the closest objects and groups them and calculates the new distances again, the process continues until one object remains.

3.2.2. Self-Organizing Maps (SOM)

Self-Organizing Maps (SOM) is an unsupervised machine learning algorithms made known by Kohonen (1990). They are used to visualize data and identify patterns from high dimensional dataset. The patterns are visualized by reducing the high dimensional dataset to a lower dimension space, mostly into 2-dimension space. Self-organizing maps has been used by many researchers to identify patterns form larger dataset in many fields including clustering, data mining and others fields.

SOM was implemented using one of the popular SOM package called Minisom (Vettigli, 2018). The package is based on python programming language and NumPy library and has the capability of representing high dimensional dataset into a low dimension map. Gaussian function was chosen as neighborhood function and Euclidian distance was used as the distance measure. The hyperparameter called sigma controls the initial spread, while the hyperparameter knows as learning rate controls the model learning rate. The 2-dimensional rectangular map was chosen for this study. The size of the map depends on the targeted number of clusters. Details on optimization of SOM hyperparameter were given in section 3.4. Basic principles of SOM algorithm were presented in the section below.

To represent input data spatially, a cell is found for the best match of the input and a neighborhood of grids is formed around that same cell. SOM starts by initializing every weight of each node, then it chose randomly a vector from the training data set. The weight of vector that resembles the input vector is calculated; the neighborhood of the winning node is calculated. The weight that resembles the input vector is rewarded and its neighbor are also awarded accordingly. The steps are iterated for several times.

The self-organizing maps is different from other neural networks because it uses competitive learning instead of error correction commonly used by backpropagation

gradient. The input node directly linked to output node. Principles of SOM have been extensively explained in a series of paper by Kohonen (1990, 1998, 2001).

3.3. Optimization of clustering methods for K-Means

One of the major challenges in K-Means clustering is to determine the optimal cluster number in cluster analysis. The potential optimal cluster number is an important input parameter and needs to be determined in advance for K-Means clustering algorithms. Elbow method, one of the oldest and commonly methods to determine the potential optimal number of clusters was used to determine the optimal number of clusters for the present study.

Elbow analysis was implemented using python package called Yellowbricks (Bilbro, 2019). Yellowbricks is the python library that depends on scikit-learn and matplotlib libraries to provide interesting tools for machine learning visualization, model selection and hyperparameter tuning.

Elbow method is a visual method where the potential number of clusters correspond to the elbow in line chart. A cost function J is set up

$$J = \sum_{i=1}^{k} \sum_{x \in C_I} |x - C_i|^2 \qquad \textbf{(3)}$$

where J is the cost function; C_i is the cluster; x is the element of cluster; k is the number of clusters.

The sample partition will be refined with the increase of clustering number. J will decrease when the degree of each cluster gradually increases. When the number of clusters k is less than the optimal clustering number, the increase of k will largely increase the degree of each cluster and thus J will be greatly increased. However, when k reaches the optimal number of clusters any further increase of k will not increase J substantially. The k value of this elbow is considered as the optimal number of clusters.

3.4. Optimization of clustering methods for SOM

A trial and error method were used to determine the optimum number of clusters. Different combination of hyperparameters including learning rate, sigma, and map size (X, and Y) were tried for every number of clusters. Two measure of internal errors, Quantization Error (TE), and Topographical Error (QE) were calculated for every trial and the combination of hypermeters that minimize TE were chosen as optimal hyperparameters for each trial.

3.4.1. Quantization error

SOM does not have a direct method to determine optimum number of nodes. The number of nodes is often determined by a trial and error methods(Kohonen, 1991). Increasing the number of nodes means increasing the capacity of the model to represent the underlying structure of the data. However, a very large number of nodes would lead to overfitting. It is therefore very important to find the optimal number of nodes. One of the practices to determine the optimal number of nodes is to minimize the quantization error. The quantification error measure the distance between the best-matching units on the SOM and the average of the corresponding input vectors. The quantification error decreases when the number of nodes increases. When the number of nodes is increased, the quantification error decreases. The relationship between the number of nodes and quantification error is used to determine the optimal number of clusters. In the beginning the quantification error decreases sharply as the number of nodes increases. However, the rate of improvement slows down and reach a point where there is no more substantial improvement. That point is known as elbow point and is considered as the optimal number of nodes. A plot of quantification error against the number of nodes is one

approach that give an indication of how quantification error changes when the number of nodes increases.

The quantization error is expressed in the formula below:

$$\boldsymbol{QE = 1/N \sum_{i=0}^{n} \|x_i - m_{ci}\|} \qquad \textbf{(4)}$$

where N is the number of input vector x; x_i is the original input vector; m_c is the best matching unit; $\|x_i - m_c\|$ is a measure of how close is the original input vector to the SOM matching unit.

3.1.2. Topographical error

Topographical error is a measure of how SOM preserves the spatial relationship between input data points. In calculating topographical error (TE), the number of inputs points whose closest units in the SOM are not neighbors are divided by the total number of inputs points. A high TE value means that there were a lot of cases where SOM has not succeeded in preserving the topographical relationship of input points. The optimal number of nodes were decided based on the minimum TE.

3.5. Homogeneity test

It is very important to test homogeneity of formed regions to test if they are truly homogeneous. Homogeneity of clusters in regional frequency analysis was tested using a widely known method developed by Hosking and Wallis in 1991.

Implementation of Hosking and Wallis heterogeneity test was implemented using lmomRFA package, version 3.5. The package depend on R programming language and was developed by the author of the same test (Hosking & Wallis, 1997). Heterogeneity measure H_1 is calculated using function regsimh from lmomRFA. The function uses Monte Carlo simulations to estimate the sampling variability of L-moments ratios of a homogeneous region which has same average L-moment rations and record length as the data. The present study uses 1000 number of simulations in the calculation of heterogeneity measure for every cluster. Note that 500 number of simulation were considered as sufficient to get accurate results(Hosking & Wallis, 1997).The algorithm calculate heterogeneity measure based on a measure of dispersion of L-moment ratios between sites. Heterogeneity measure H_1 and can be expressed as:

$$\boldsymbol{H_1 = \frac{V - \mu V_1}{\sigma_{V1}}} \qquad \textbf{(5)}$$

where μV_1 is the average computed from simulated data; σ_{V1} is the standard deviation computed from the same simulated data.

The sample variance (V) is computed as:

$$\boldsymbol{V = \frac{\sum_i^p n_i \left(L_{CV}^{\ i} - \overline{L}_{CV}^{\ i}\right)^2}{\sum_{i=1}^p n_i}} \qquad \textbf{(6)}$$

where n_i is the number of points in rain gauge station I; L_{CV}^{i} is the L-coefficient of variation (L_{cv}) calculated for rain gauge station *I;* $\overline{L}_{CV}^{i}$ is the average of L_{CV}^{i} considering all *p* rain gauge stations.

The cluster is considered, homogeneous if the calculated H_1 is inferior to 1, possibly homogeneous if H_1 is between 1 and 2, and heterogeneous if H_1 is superior to 2. Alternatives measures called H_2 or H_3 can be used in place of H_1; H_2 is calculate from L skewness instead of L_{CV} and H_3 is calculate using L kurtosis instead of L_{CV}. However, H_1 is usually enough.

Chapter 4. Results and Discussion

4.1. ASOS stations

SOM and K-Means clustering methods were used to group 57 ASOS stations into homogeneous precipitation clusters using daily precipitation data. Clusters formed by each method were tested by Hosking and Wallis heterogeneity measure H_1. The measure was used to confirm if the clusters were indeed homogeneous hydrologically. Homogeneity test results were used to assess and compare the ability of each clustering method in defining homogeneous precipitation clusters using time series data.

4.1.1. K-Means Clustering

Twelve clusters were initially determined by Elbow method as the optimal number of clusters for K-Means clusters and ASOS dataset as shown on the Figure 8.

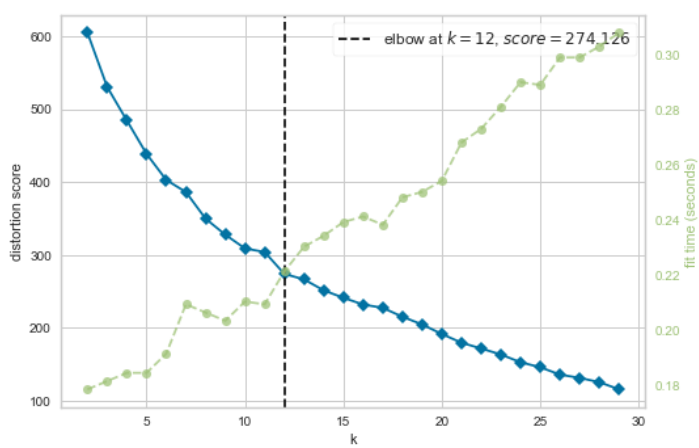


Figure 8. Elbow for K-Means- ASOS stations

It is well known that traditional methods commonly used to find optimal number of clusters in other datasets may not produce good results for time series datasets. The fact was confirmed by the results of homogeneity test for initial K-Means clustering as listed in Table 2. The number of clusters was subjectively changed to 10, 14, and then to 16 based on the results of homogeneity test.

Cluster analysis was conducted on ASOS stations using K-Means clustering with initial cluster number of 10. The results of homogeneity test showed that one cluster was identified as homogeneous out of 10 clusters. This may indicate that 10 clusters may not be the optimal number of clusters for ASOS dataset.

Table 2. Homogeneity test for K-Means- ASOS stations (N=10)

Clusters	No of sites	H_1	Homogeneity
Cluster 1	3	0.8	Homogeneous
Cluster 2	10	3.7	Heterogeneous
Cluster 3	9	5.1	Heterogeneous
Cluster 4	5	3.2	Heterogeneous
Cluster 5	6	15.5	Heterogeneous
Cluster 6	6	0.3	Homogeneous
Cluster 7	6	2.7	Heterogeneous
Cluster 8	8	4.6	Heterogeneous
Cluster 9	1	0.0	-
Cluster 10	3	0.4	Homogeneous

The number of clusters was increased to 12 clusters for K-Means clustering. The results of homogeneity test in Table 3 showed that 4 clusters out of 12 clusters were homogeneous and one cluster was possibly heterogeneous. The number of homogeneous clusters was improved compared to previous clustering when the number of clusters was 10.

But the fact that a greater number of clusters were heterogeneous may indicate a reduced ability of the clustering method to deal with time series data, or it may be an indication that the chosen cluster number was still too small for the dataset.

Table 3. Homogeneity test for K-Means- ASOS stations (N=12)

Clusters	No of sites	H_1	Homogeneity
Cluster 1	3	-0.8	Homogeneous
Cluster 2	6	2.8	Heterogeneous
Cluster 3	6	4	Heterogeneous
Cluster 4	10	5.7	Heterogeneous
Cluster 5	1	0	-
Cluster 6	6	0.3	Homogeneous
Cluster 7	5	23.8	Heterogeneous
Cluster 8	3	0.3	Homogeneous
Cluster 9	6	4.2	Heterogeneous
Cluster 10	3	0.8	Homogeneous
Cluster 11	4	1.3	Possibly Heterogeneous
Cluster 12	4	2.8	Heterogeneous

The number of clusters was increased to 14 and the clustering process was repeated. The number of homogeneous clusters identified by K-Means was increased to 5 as shown in Table 4.

Table 4. Homogeneity test for K-Means -ASOS stations (N=14)

Clusters	No of sites	H_1	Homogeneity
Cluster 1	4	2.8	Heterogeneous
Cluster 2	6	4.6	Heterogeneous
Cluster 3	7	2.7	Heterogeneous
Cluster 4	3	0.8	Homogeneous
Cluster 5	6	0.3	Homogeneous
Cluster 6	6	2.7	Heterogeneous
Cluster 7	3	-0.5	Homogeneous
Cluster 8	2	0.8	Homogeneous
Cluster 9	6	15.7	Heterogeneous
Cluster 10	1	0.0	-
Cluster 11	7	4.5	Heterogeneous
Cluster 12	1	0.0	-
Cluster 13	2	1.3	Possibly Heterogeneous
Cluster 14	3	0.4	Homogeneous

The number of clusters was further increased to 16 K-Means clustering and the algorithm was rerun to assess if homogeneity results will be improved. The number of homogeneous clusters remained the same even after the number of clusters was increased to 16 as shown in Table 5.

Table 5. Homogeneity test for K-Means-ASOS stations (N=16)

Clusters	No of sites	H_1	Homogeneity
Cluster 1	2	5.8	Heterogeneous
Cluster 2	7	3.9	Heterogeneous
Cluster 3	4	-0.1	Homogeneous
Cluster 4	5	1.2	Possibly Heterogeneous
Cluster 5	6	2.6	Heterogeneous
Cluster 6	3	4.6	Heterogeneous
Cluster 7	3	2.4	Heterogeneous
Cluster 8	3	0.9	Homogeneous
Cluster 9	1	0.0	-
Cluster 10	6	1.2	Possibly Heterogeneous
Cluster 11	5	0.6	Homogeneous
Cluster 12	5	11.2	Heterogeneous
Cluster 13	1	0.0	-
Cluster 14	1	0.0	-
Cluster 15	3	1.6	Possibly Heterogeneous
Cluster 16	2	-1.1	Homogeneous

4.1.2. SOM

The optimal parameters for SOM were investigated using two internal Minisom package metrics: the quantization error (QE), and topographical error (TE). The topographical error was used as a metrics to determine the optimal combination of hyperparameters such as sigma and learning rate (Lr). Different combinations of hyperparameters, learning rate, sigma, and the map size (x, y) were analyzed and TE and QE were calculated for each combination. The combination of parameters that minimize TE was determined for the three trials as shown in Table 6. The number of nodes was decided to be the same as K-Means clustering for comparison purposes.

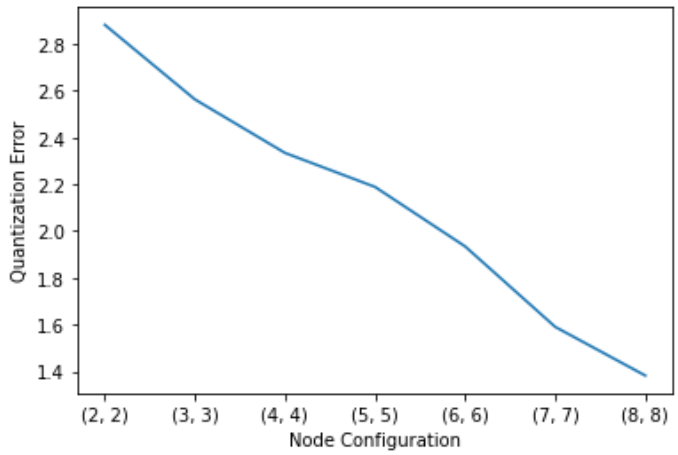


Figure 9. Quantization error SOM-ASOS

The quantization error always decreased when the number of nodes increased, therefore QE alone was not suitable to compare maps of different size as can be observed on the Figure 9.

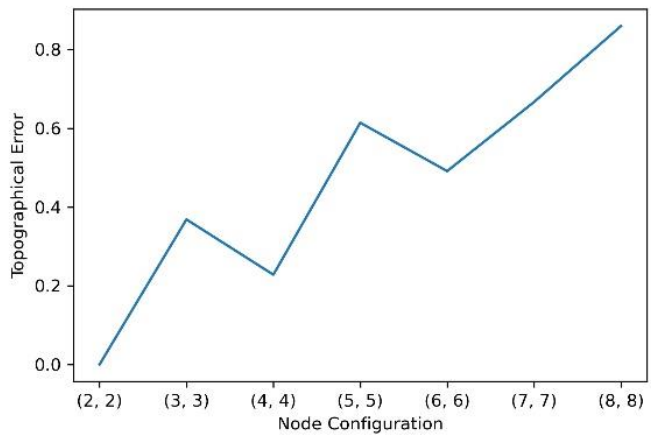


Figure 10. Topographical error SOM-ASOS

It was observed that different hyperparameters produce different homogeneity test results for SOM, even when the number of nodes was the same. The hyperparameters which minimize TE as shown in Table 6, were chosen as suitable for every test

because they tended to produce well distribute clusters while hyperparameters that minimize QE tended to produce many sole station clusters.

Table 6. Optimal hyperparameters for SOM-ASOS stations

No of clusters	QE	TE	Lr	Sigma	X	Y
10	2.599	0.018	0.5	1.5	2	5
12	2.524	0.018	0.5	1.5	2	6
16	2.562	0.018	0.3	2	6	4
14	2.597	0.035	0.5	2	2	8

The number of nodes for SOM was initially determined to be 10, then it was increase to 12,14 and 16. The results of homogeneity test for initial number of clusters showed that SOM identified 1 homogeneous cluster out of 10 clusters as listed in Table 7.

Table 7. Homogeneity test for SOM – ASOS stations (N=10)

Clusters	No of sites	H_1	Homogeneity
Cluster 1	6	6.0	Heterogeneous
Cluster 2	3	0.8	Homogeneous
Cluster 3	5	22.8	Heterogeneous
Cluster 4	5	2.4	Heterogeneous
Cluster 5	9	5.9	Heterogeneous
Cluster 6	5	-0.5	Homogeneous
Cluster 7	6	6.7	Heterogeneous
Cluster 8	8	6.9	Heterogeneous
Cluster 9	6	2.5	Heterogeneous
Cluster 10	4	3.4	Heterogeneous

When the number of nodes was increased to 12, SOM identified 2 homogeneous and 3 possibly heterogeneous clusters out of 12 clusters as shown in Table 8.

Table 8. Homogeneity test for SOM-ASOS stations (N=12)

Clusters	No of sites	H_1	Homogeneity
Cluster 1	8	1.2	Possibly Heterogeneous
Cluster 2	5	0.6	Homogeneous
Cluster 3	3	8.4	Heterogeneous
Cluster 4	4	1.3	Possibly Heterogeneous
Cluster 5	6	3.4	Heterogeneous
Cluster 6	3	4.1	Heterogeneous
Cluster 7	3	0.8	Homogeneous
Cluster 8	3	1.0	Possibly Heterogeneous
Cluster 9	6	5.2	Heterogeneous
Cluster 10	5	24.8	Heterogeneous
Cluster 11	3	7.9	Heterogeneous
Cluster 12	8	4.3	Heterogeneous

When the number of nodes was increased to 14, the number of homogeneous clusters increased to 4 as shown in Table 9. The increase in number of homogeneous clusters may indicate that the capacity of SOM to capture patterns in timeseries data was gradually improving as SOM map size was increased.

Table 9. Homogeneity test for SOM-ASOS stations (N=14)

Clusters	No of sites	H_1	Homogeneity
Cluster 1	3	0.84	Homogeneous
Cluster 2	7	1.02	Possibly Heterogeneous
Cluster 3	2	1.83	Possibly Heterogeneous
Cluster 4	4	-0.41	Homogeneous
Cluster 5	4	3.20	Heterogeneous
Cluster 6	5	10.20	Heterogeneous
Cluster 7	3	0.61	Homogeneous
Cluster 8	3	4.51	Heterogeneous
Cluster 9	3	3.53	Heterogeneous
Cluster 10	4	21.07	Heterogeneous
Cluster 11	5	3.35	Heterogeneous
Cluster 12	4	11.16	Heterogeneous
Cluster 13	5	-0.05	Homogeneous
Cluster 14	5	6.74	Heterogeneous

The number of homogeneous clusters identified by SOM was increased to 6 when the number of nodes was increased to 16 as listed in Table 10. The increase in the number of homogeneous clusters may indicate the capacity of SOM to better represent high dimensional data when the number of nodes was increased.

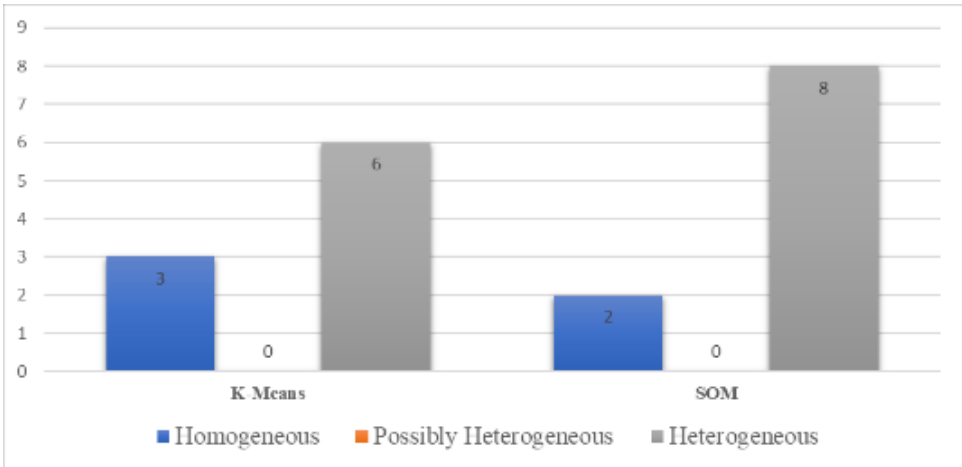
Table 10. Homogeneity test for SOM -ASOS stations (N=16)

Clusters	No of sites	H_1	Homogeneity
Cluster 1	5	1.1	Possibly Heterogeneous
Cluster 2	3	0.1	Homogeneous
Cluster 3	5	0.6	Homogeneous
Cluster 4	3	2	Heterogeneous
Cluster 5	3	0.1	Homogeneous
Cluster 6	3	21.3	Heterogeneous
Cluster 7	2	7.1	Heterogeneous
Cluster 8	4	7.3	Heterogeneous
Cluster 9	3	0.9	Homogeneous
Cluster 10	2	-0.6	Homogeneous
Cluster 11	3	1.2	Possibly Heterogeneous
Cluster 12	5	4.5	Heterogeneous
Cluster 13	4	3.4	Heterogeneous
Cluster 14	4	1.2	Possibly Heterogeneous
Cluster 15	3	4	Heterogeneous
Cluster 16	5	0.4	Homogeneous

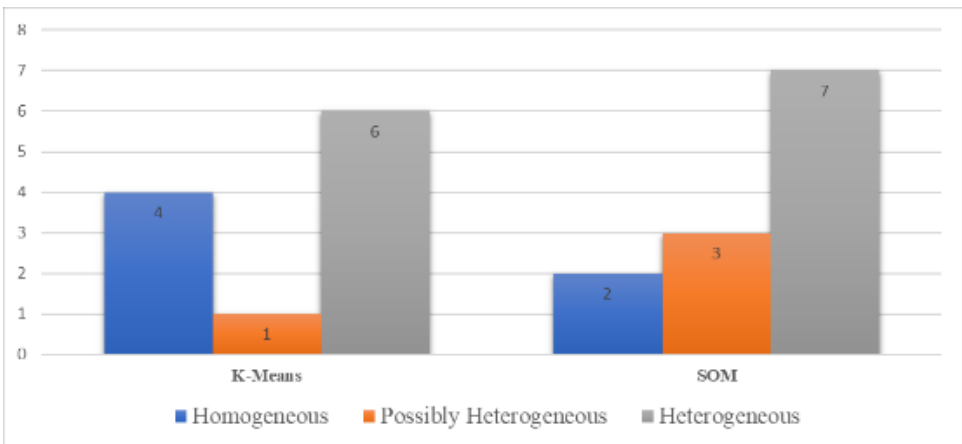
4.1.3. Comparison

Based on the clusters formed by SOM and K-Means, SOM was able to identify a bigger number of homogeneous clusters when the number of nodes is increased. It was able to identify 6 homogeneous clusters when the number of nodes was increased to 16 while K-Means identified 5 homogeneous regions for the same number of clusters as shown in Figure 11. K-Means was able to identify a greater number of homogeneous regions when cluster number was small. For example, when the number of clusters was 10, K-Means identified 3 homogeneous clusters while SOM identified 2 homogenous clusters as it can be observed on Figure 11 (a) and (b) however the number of homogeneous and possibly heterogeneous clusters identified

by SOM gradually increased as the number on nodes increased from 10 to 16 as show in Figure 11 (c) and (d). The performance of SOM in defining homogeneous regions was greatly affected by the size of the map.

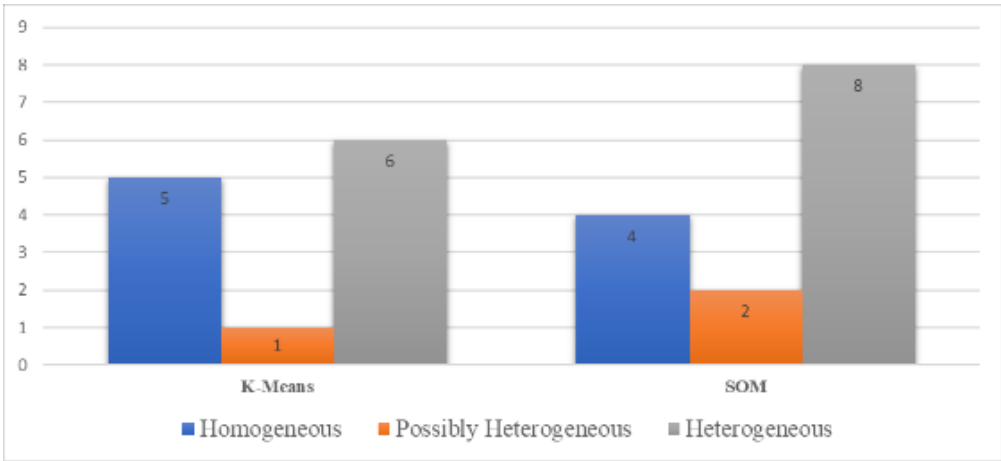


(a)

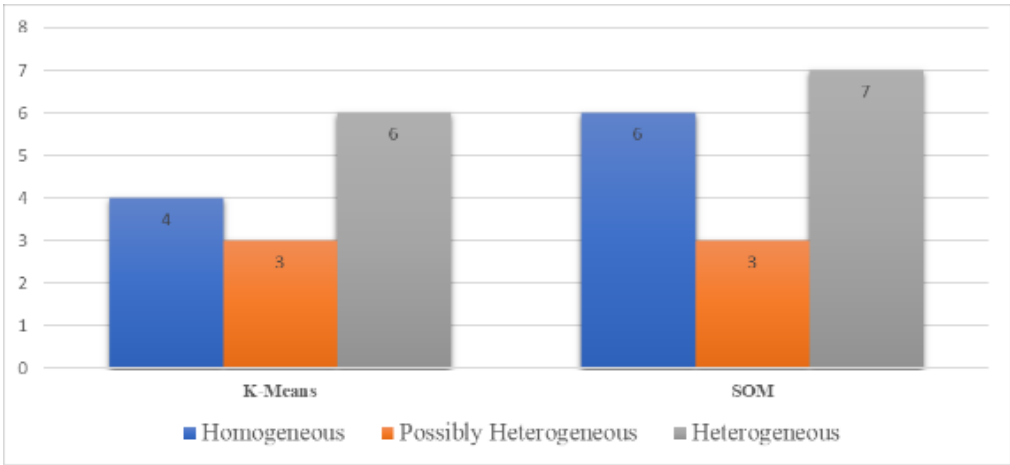


(b)

Figure 11. Homogeneity test for ASOS dataset for different number of clusters(N); (a) N equal 10, (b)



(c)



(d)

Figure 12. Homogeneity test for ASOS dataset for different number of clusters(N); (c) N equal 14 and (d) N equal 16.

4.2. AWS stations

SOM and K-Means clustering were used to group 441 stations into homogeneous clusters. The homogeneity of each cluster was tested using Hosking and Wallis homogeneity test to confirm if they were hydrologically homogeneous. The results of the homogeneity test were used to assess the ability of each clustering method in defining homogeneous precipitation clusters.

4.2.1. K-Means Clustering

Twelve clusters were initially determined by Elbow method as the optimal number of clusters for AWS dataset as shown on the Figure 13. However, the number of clusters was subjectively changed to improve homogeneity of clusters.

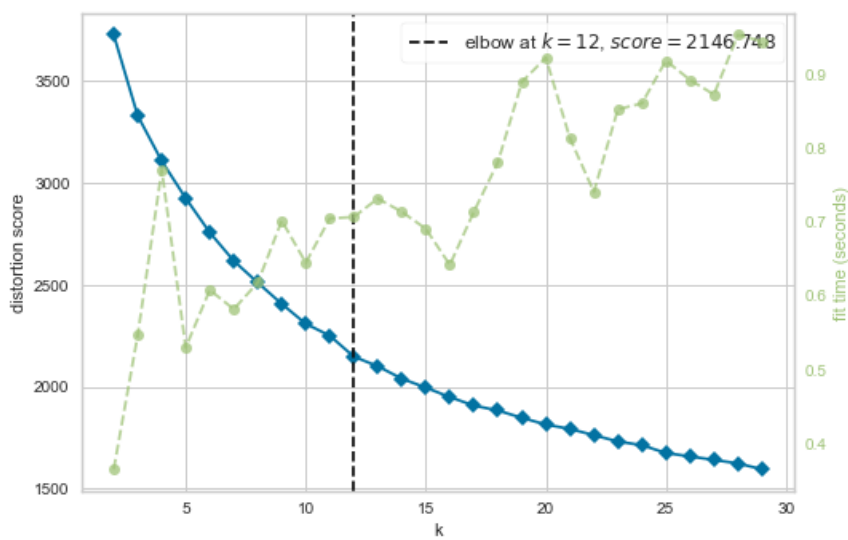


Figure 13. Elbow for K-Means clustering AWS stations

The number of clusters was therefore changed from 10 to 14, then to 16 to improve homogeneity based on initial results of homogeneity test. The results of homogeneity test in Table 11 showed that one clusters was homogeneous out of ten clusters when cluster number was set to 10 for K-Means clustering.

Table 11. Homogeneity test for K-Means- AWS stations (N=10)

Clusters	No of sites	H_1	Homogeneity
Cluster 1	41	34.4	Heterogeneous
Cluster 2	57	8.0	Heterogeneous
Cluster 3	45	8.2	Heterogeneous
Cluster 4	40	15.5	Heterogeneous
Cluster 5	33	-0.7	Homogeneous
Cluster 6	63	17.7	Heterogeneous
Cluster 7	31	19.1	Heterogeneous
Cluster 8	33	25.6	Heterogeneous
Cluster 9	61	19.5	Heterogeneous
Cluster 10	37	8.7	Heterogeneous

The low number of homogeneous clusters showed that 10 clusters were probably not a suitable number of clusters for AWS dataset. The number of clusters was increased to 12 clusters and the results of homogeneity test were presented in Table 12.

Table 12. Homogeneity test for K-Means- AWS stations (N=12)

Clusters	No of sites	H_1	Homogeneity
Cluster 1	29	1.2	Possibly Heterogeneous
Cluster 2	40	0.6	Homogeneous
Cluster 3	55	8.7	Heterogeneous
Cluster 4	28	1.3	Possibly Heterogeneous
Cluster 5	26	3.4	Heterogeneous
Cluster 6	45	4.1	Heterogeneous
Cluster 7	45	0.8	Homogeneous
Cluster 8	34	1.1	Possibly Heterogeneous
Cluster 9	28	5.6	Heterogeneous
Cluster 10	42	24.2	Heterogeneous
Cluster 11	20	7.9	Heterogeneous
Cluster 12	49	4.5	Heterogeneous

The results of homogeneity test for K-Means clustering showed that 2 out of 12 clusters were homogeneous and 2 clusters were possibly heterogeneous. The fact that a greater number of clusters were not homogeneous may indicate that the cluster number was still too small for the dataset. The number of clusters was increased to 14 and the clustering process was repeated. The number of homogeneous clusters identified by K-Means was 2 as shown in Table 13.

Table 13. Homogeneity test for K-Means -AWS stations (N=14)

Clusters	No of sites	H_1	Homogeneity
Cluster 1	39	12.2	Heterogeneous
Cluster 2	52	15.6	Heterogeneous
Cluster 3	29	11.2	Heterogeneous
Cluster 4	17	6.9	Heterogeneous
Cluster 5	33	5.3	Heterogeneous
Cluster 6	20	19.9	Heterogeneous
Cluster 7	27	5.0	Heterogeneous
Cluster 8	29	-0.4	Homogeneous
Cluster 9	37	22.3	Heterogeneous
Cluster 10	28	28.6	Heterogeneous
Cluster 11	43	-0.4	Homogeneous
Cluster 12	45	10.8	Heterogeneous
Cluster 13	31	7.8	Heterogeneous
Cluster 14	11	4.8	Heterogeneous

K-Means identified one homogeneous cluster for AWS station even after the number of clusters was increased to 16 as shown in Table 14.

Table 14. Homogeneity test for K-Means-AWS stations (N=16)

Clusters	No of sites	H_1	Homogeneity
Cluster 1	17	6.3	Heterogeneous
Cluster 2	28	-0.3	Homogeneous
Cluster 3	49	3.6	Heterogeneous
Cluster 4	20	20.2	Heterogeneous
Cluster 5	16	6.8	Heterogeneous
Cluster 6	27	6.2	Heterogeneous
Cluster 7	23	10.7	Heterogeneous
Cluster 8	29	7.2	Heterogeneous
Cluster 9	18	6.7	Heterogeneous
Cluster 10	26	28.4	Heterogeneous
Cluster 11	30	2.4	Heterogeneous
Cluster 12	27	12.6	Heterogeneous
Cluster 13	33	6.2	Heterogeneous
Cluster 14	33	13.5	Heterogeneous
Cluster 15	20	11.8	Heterogeneous
Cluster 16	45	10.8	Heterogeneous

Homogeneity of clusters did not improve even when the number of nodes was increased to 16. One homogeneous cluster was identified by K-Means as indicated by the results of homogeneity test in Table 14.

4.2.2. SOM

The optimal parameters for SOM in AWS dataset were investigated using two internal Minisom package metrics: the quantization error (QE), and topographical error (TE) as shown in Figure 14 and Figure 15. The topographical error was used as a metrics to determine the optimal combination of hyperparameters such as sigma and learning rate. A combination of different hyperparameters, including number of clusters, sigma, and learning rate was analyzed and QE and TE were calculated for every trial. The combination of parameters that minimize TE was determined for every clustering as shown in Table 15. The number of nodes was decided to be the same as K-Means clustering for comparison purposes.

Table 15. Optimal hyperparameters for SOM-AWS stations

No of clusters	QE	TE	Lr	Sigma	X	Y
14	2.688	0.003	0.4	2	2	7
16	2.727	0.003	0.3	2	4	4
12	2.589	0.005	0.5	1.5	3	4
10	2.636	0.003	0.4	1.5	5	2

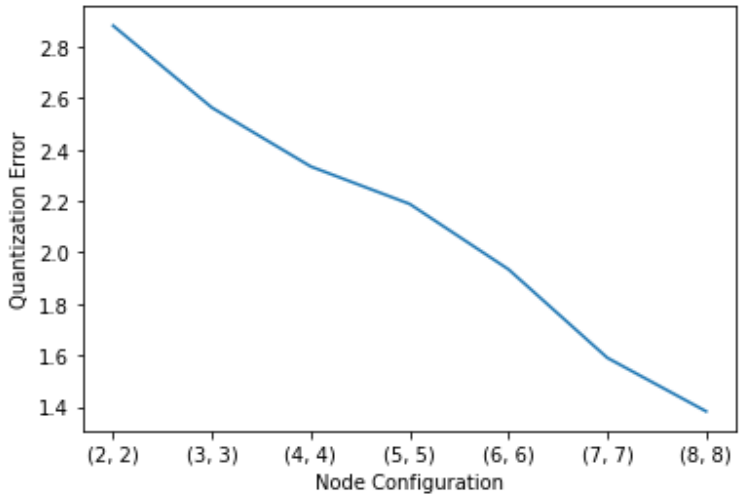


Figure 14. Quantization error SOM-AWS stations

The quantization error always decreases when the number of nodes increases, therefore QE alone was not suitable to compare maps of different size as can be observed in Figure 15.

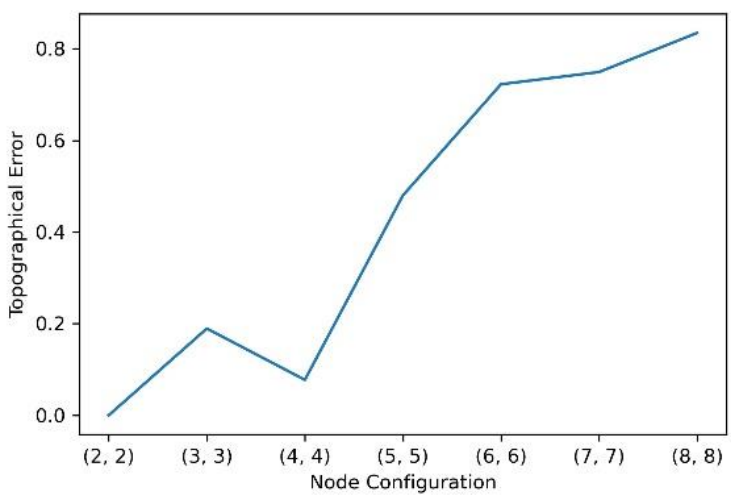


Figure 15. Topographical error SOM-AWS stations

The number of nodes for SOM was initially determined to be 10, then it was increase to 12, 14 finally to 16. The results of homogeneity test in Table 16 shows that SOM initially identified 1 homogeneous cluster out of 10 clusters. The small size of the map for a big dataset may be the cause for low number of homogeneous clusters.

Table 16. Homogeneity test for SOM - AWS stations (N=10)

Clusters	No of sites	H_1	Homogeneity
Cluster 1	46	7.9	Heterogeneous
Cluster 2	21	22.1	Heterogeneous
Cluster 3	63	30.2	Heterogeneous
Cluster 4	35	-0.8	Homogeneous
Cluster 5	52	13.3	Heterogeneous
Cluster 6	34	29.5	Heterogeneous
Cluster 7	41	3.2	Heterogeneous
Cluster 8	19	6.4	Heterogeneous
Cluster 9	28	9.4	Heterogeneous
Cluster 10	102	22.9	Heterogeneous

When the number of nodes was increased to 12. SOM identified 2 homogeneous clusters out of 12. (Table 17). The slight increase in number of homogeneous clusters may indicate that the capacity of SOM to capture patterns in timeseries data was enhanced when SOM map size was increased.

Table 17. Homogeneity test for SOM-AWS stations (N=12)

Clusters	No of sites	H_1	Homogeneity
Cluster 1	51	13.7	Heterogeneous
Cluster 2	40	4.9	Heterogeneous
Cluster 3	28	24.0	Heterogeneous
Cluster 4	36	3.6	Heterogeneous
Cluster 5	39	21.2	Heterogeneous
Cluster 6	50	11.3	Heterogeneous
Cluster 7	31	12.4	Heterogeneous
Cluster 8	66	15.0	Heterogeneous
Cluster 9	42	16.8	Heterogeneous
Cluster 10	25	-1.8	Homogeneous
Cluster 11	14	0.3	Homogeneous
Cluster 12	19	17.3	Heterogeneous

The number of homogeneous clusters was not improved when the number of nodes was increased to 14. SOM still identified one homogeneous cluster as shown in Table 18.

Table 18. Homogeneity test for SOM-AWS stations (N=14)

Clusters	No of sites	H_1	Homogeneity
Cluster 1	35	3.8	Heterogeneous
Cluster 2	18	8.4	Heterogeneous
Cluster 3	20	7.2	Heterogeneous
Cluster 4	25	24.8	Heterogeneous
Cluster 5	42	12.7	Heterogeneous
Cluster 6	28	19.0	Heterogeneous
Cluster 7	45	11.9	Heterogeneous
Cluster 8	20	3.4	Heterogeneous
Cluster 9	20	9.0	Heterogeneous
Cluster 10	72	13.2	Heterogeneous
Cluster 11	5	2.8	Heterogeneous
Cluster 12	58	-0.1	Homogeneous
Cluster 13	27	7.7	Heterogeneous
Cluster 14	26	27.9	Heterogeneous

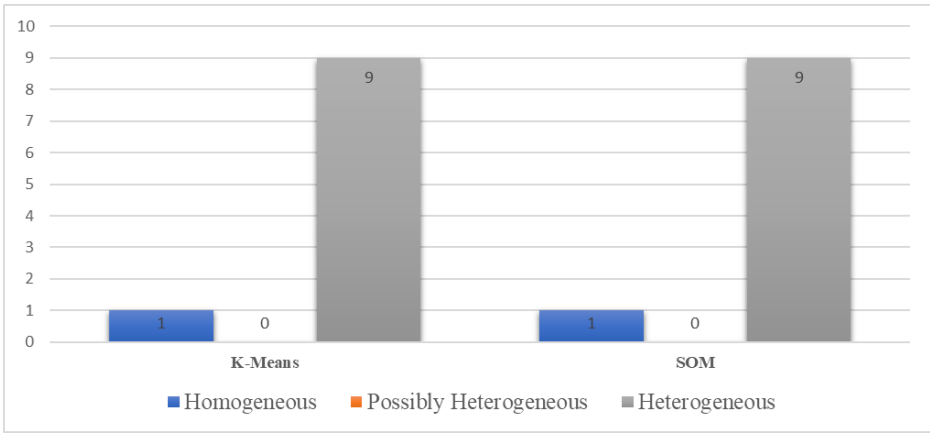
The number of nodes was increased to 16 for SOM, the number of homogeneous clusters identified by SOM did not change any further as shown in Table 19. The increase in map size does not produce any further change in identifying homogeneous clusters.

Table 19. Homogeneity test for SOM -AWS stations (N=16)

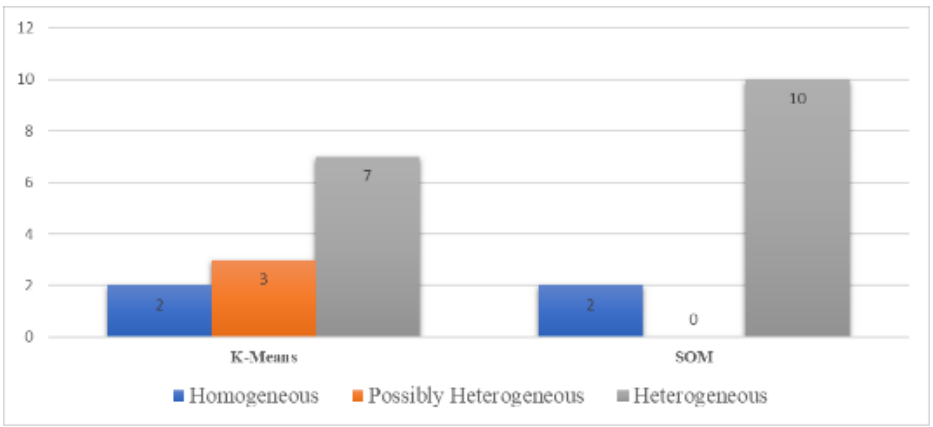
Clusters	No of sites	H_1	Homogeneity
Cluster 1	17	6.3	Heterogeneous
Cluster 2	28	-0.3	Homogeneous
Cluster 3	49	3.6	Heterogeneous
Cluster 4	20	20.2	Heterogeneous
Cluster 5	16	6.8	Heterogeneous
Cluster 6	27	6.2	Heterogeneous
Cluster 7	23	10.7	Heterogeneous
Cluster 8	29	7.2	Heterogeneous
Cluster 9	18	6.7	Heterogeneous
Cluster 10	26	28.4	Heterogeneous
Cluster 11	30	2.4	Heterogeneous
Cluster 12	27	12.6	Heterogeneous
Cluster 13	33	6.2	Heterogeneous
Cluster 14	33	13.5	Heterogeneous
Cluster 15	20	11.8	Heterogeneous
Cluster 16	45	10.8	Heterogeneous

4.2.3. Comparison

Based on the identification of homogeneous clusters by SOM and K-Means for AWS datasets, both clustering methods identified similar number of clusters in AWS dataset. The number of homogeneous clusters identified by both clustering methods did not improve when the number of clusters were increased to 12, 14 or 16 as it can be observed on Figure 16. This shows that the number of clusters or map size was probably not the cause of low yield in homogeneous clusters for AWS dataset.

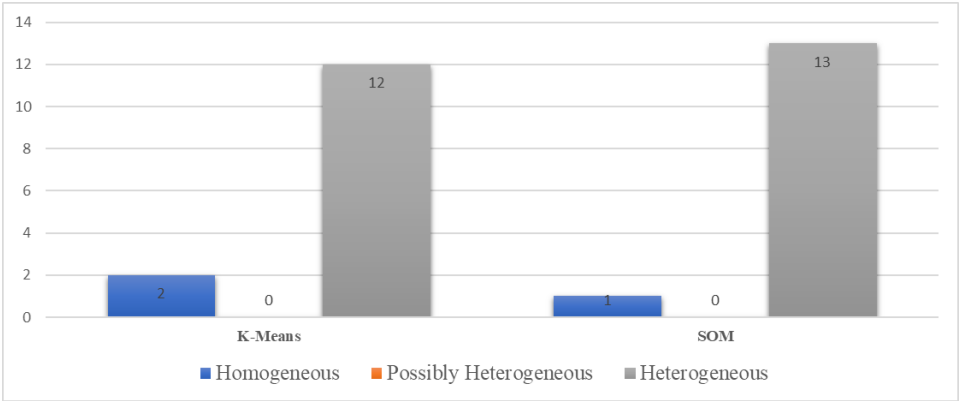


(a)

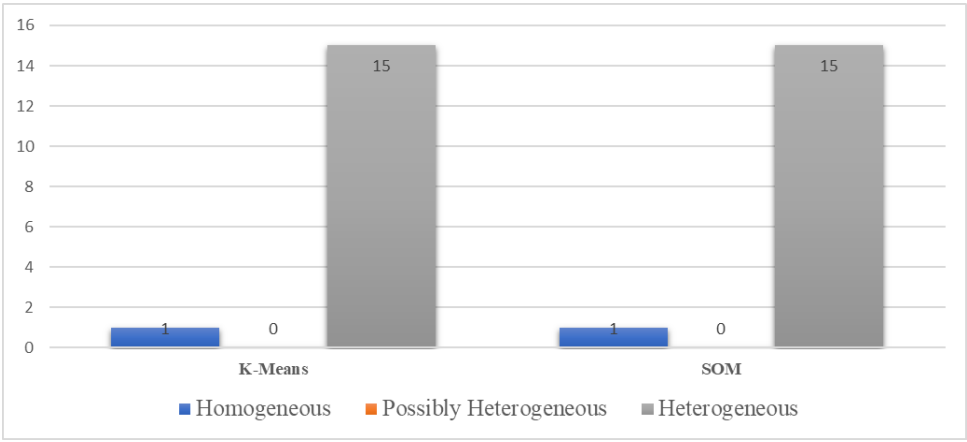


(b)

Figure 16. Homogeneity results of AWS dataset for different number of clusters (N); (a) N equal 10, (b) N equal 12.



(c)



(d)

Figure 17. Homogeneity results of AWS dataset for different number of clusters (N); (c) N equal 14 and (d) N equal 16.

4.3. Satellite precipitation dataset

SOM and K-Means clustering were used to group satellite dataset into homogeneous clusters. The clusters from each method were tested by means of Hosking and Wallis homogeneity test to confirm if they were hydrologically homogeneous. The results of the homogeneity test were used to assess the ability of each clustering method in defining homogeneous precipitation clusters using only timeseries data as input.

4.3.1. K-Means Clustering

Nine clusters were initially determined by Elbow method as the optimal number of clusters as shown on the Figure 18. However, the number of clusters was subjectively increased to 18 to improve homogeneity of clusters based on the results of initial homogeneity test.

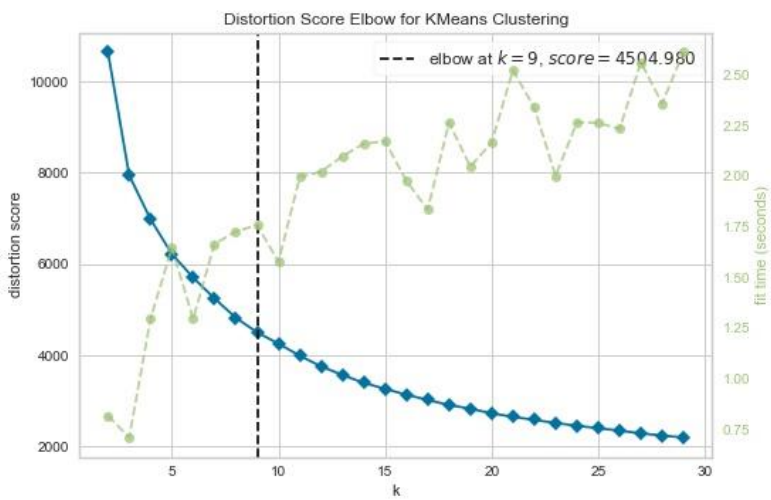


Figure 18. Elbow for K-Means- Satellite data

It is well known that traditional methods commonly used to find optimal number of clusters in other datasets may not produce good results for time series datasets. The fact was confirmed by the poor results of homogeneity test for initial K-Means clustering as listed in Table 20. Therefore, the number of clusters was increased to 16.

Table 20. Homogeneity test for SOM- Satellite data (N=9)

Clusters	No of grids	H_1	Homogeneity
Cluster 1	153	15.3	Heterogeneous
Cluster 2	157	-2.2	Homogeneous
Cluster 3	199	24.8	Heterogeneous
Cluster 4	152	0.2	Homogeneous
Cluster 5	89	10.5	Heterogeneous
Cluster 6	178	2.2	Heterogeneous
Cluster 7	133	2.5	Heterogeneous
Cluster 8	88	6.0	Heterogeneous
Cluster 9	31	7.7	Heterogeneous

The results of homogeneity test for K-Means and satellite dataset clustering showed that 10 out of 16 clusters were homogeneous and 2 clusters were possibly heterogeneous. This may indicate that for K-Means clustering, 16 clusters were more suitable for the dataset than 10 clusters.

Table 21. Homogeneity test for K-Means-Satellite data (N=16)

Clusters	No of sites	H_1	Homogeneity
Cluster 1	68	1.754574	Possibly Heterogeneous
Cluster 2	81	-5.68341	Homogeneous
Cluster 3	98	-13.8025	Homogeneous
Cluster 4	81	10.66735	Heterogeneous
Cluster 5	71	6.598001	Heterogeneous
Cluster 6	70	1.21077	Possibly Heterogeneous
Cluster 7	69	-6.85871	Homogeneous
Cluster 8	31	7.663366	Heterogeneous
Cluster 9	78	-2.99995	Homogeneous
Cluster 10	64	-0.4881	Homogeneous
Cluster 11	57	-3.84908	Homogeneous
Cluster 12	60	0.466897	Homogeneous
Cluster 13	93	0.466897	Homogeneous
Cluster 14	82	4.747098	Heterogeneous
Cluster 15	97	-7.08237	Homogeneous
Cluster 16	80	-7.20629	Homogeneous

4.3.2. SOM

The optimal parameters for SOM were investigated using two internal Minisom package metrics: the quantization error (QE), and topographical error (TE). The topographical error was used as a metrics to determine the optimal combination of hyperparameters such as sigma and learning rate.

Table 22. Optimal hyperparameters for SOM-Satellite data

No of clusters	QE	TE	Lr	Sigma	X	Y
9	3.8565000	0.0033898	0.5	1.5	3	3
16	3.774506963	0.054237288	0.5	1.5	4	4

The combination of parameters that minimize the TE was determined for every clustering analysis as show in Table 22. The number of nodes was decided to be the one which yield the same number of clusters as K-Means clustering for comparison purposes. The quantization error always decreases when the number of nodes increased, so QE alone was not suitable to compare maps of different size. TE was used to identify optimal parameters for different cluster numbers. The number of nodes for SOM was initially determined to be 9, then it was increase to 16.

Table 23. Homogeneity test for SOM-Satellite data (N=9)

Clusters	No of grids	H_1	Homogeneity
Cluster 1	92	-2.6	Homogeneous
Cluster 2	47	-0.3	Homogeneous
Cluster 3	75	-2.4	Homogeneous
Cluster 4	239	1.3	Possibly Heterogeneous
Cluster 5	5	-1.5	Homogeneous
Cluster 6	84	1.3	Possibly Heterogeneous
Cluster 7	297	29.6	Heterogeneous
Cluster 8	115	24.2	Heterogeneous
Cluster 9	226	49.1	Heterogeneous

When the number of nodes was set to 9, SOM identified 4 homogeneous clusters among the 9 clusters. The low number of homogeneous clusters may be attributed the size of the map being too small for the dataset. The number of nodes was increased to 16 for further analysis. When the number of nodes were increased to 16, SOM identified 10 homogeneous clusters among the 16 clusters as shown in Table 24.

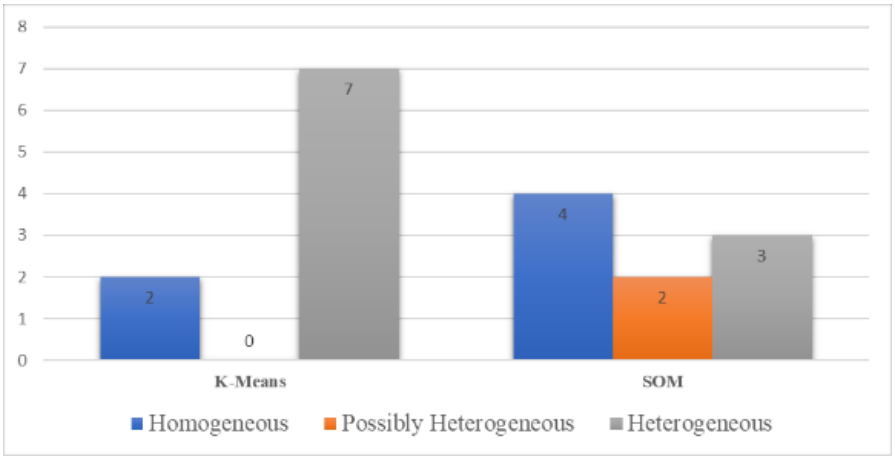
Table 24. Homogeneity test for SOM-Satellite data (N=16)

Clusters	No of grids	H_1	Homogeneity
Cluster 1	22	-2.5	Homogeneous
Cluster 2	34	0.0	Homogeneous
Cluster 3	47	-3.8	Homogeneous
Cluster 4	41	-3.9	Homogeneous
Cluster 5	21	0.1	Homogeneous
Cluster 6	70	-5.8	Homogeneous
Cluster 7	56	-1.8	Homogeneous
Cluster 8	45	-3.8	Homogeneous
Cluster 9	129	0.1	Homogeneous
Cluster 10	35	19.3	Heterogeneous
Cluster 11	261	6.1	Heterogeneous
Cluster 12	96	13.6	Heterogeneous
Cluster 13	59	1.0	Homogeneous
Cluster 14	201	28.3	Heterogeneous
Cluster 15	50	26.5	Heterogeneous
Cluster 16	13	9.3	Heterogeneous

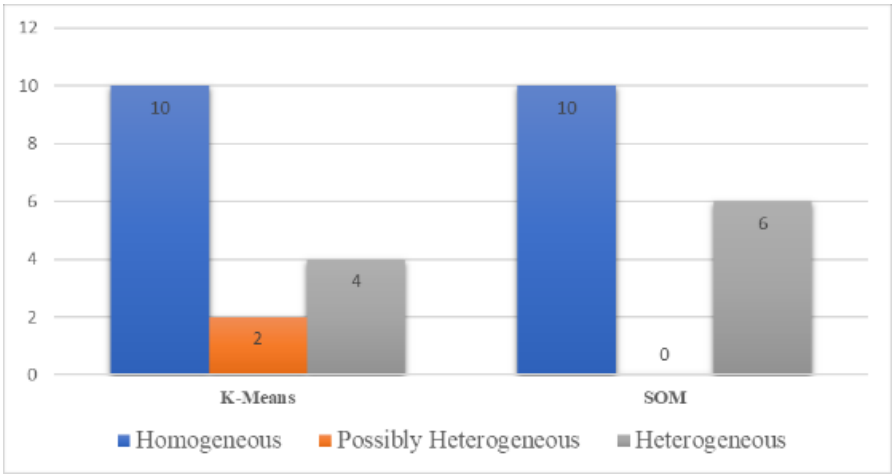
4.3.3. Comparison

Based on the number of homogeneous clusters identified by SOM and K-Means in satellite dataset, there were differences between SOM and K-Means according to the number of clusters. However, both were able to identify almost the same number of homogeneous clusters. K-Means identified 2 homogeneous clusters among 9 clusters while SOM identified 4 homogeneous clusters and 2 possibly homogeneous clusters in the same number of clusters as shown in Figure 19 (a). Both clustering methods were able to identify 10 homogeneous clusters when the number of nodes was increased to 16 however K-Means also identified 2 possibly heterogeneous

clusters.



(a)



(b)

Figure 19. Homogeneity results of satellite dataset for different number of clusters (N); (a) N equal 9, (b) N equal 16

4.4. Spatial mapping of clusters

K-Means and SOM defined homogeneous clusters with distinct geographical location. The stations located in the North were separated from the stations located in the East, West and South of the country as shown in Figure 20 and Figure 21 respectfully. It was observed that both clustering algorithms were able to cluster the stations in Jeju Island as a separate cluster. Previous studies have indeed classified the Island as a separate precipitation region (Nam et al., 2015).

4.4.1. Spatial mapping of ASOS clusters

Clusters identified by K-Means clustering and SOM have been grouped in 5 parts based on their geographical location : The North, East, South and inland clusters as it was shown in the spatial mapping of K-Means clusters (Figure 20) and the spatial mapping of SOM clusters (Figure 21).

In the North, K-Means identified 2 clusters, cluster 1 and 10. The clusters in the North contains stations located in the North of Gangwon-do and Gyeonggi-do. Cluster 1 was heterogeneous while cluster 10 was possibly heterogeneous. However, SOM identified 2 clusters in the same region, cluster 8 and cluster 16. The clusters identified by SOM had different patterns and both clusters were homogeneous. The patterns of clusters in the North were almost similar with the patterns identified by Nam et al. (2015).

In the East K-Means identified 1 cluster. The clusters covered east of Gangwon-do, Northeast and Southeast of Gyeongsangbuk-do. Although the same region was identified by (Nam et al., 2015), the clusters were found to be heterogeneous by the present study probably due to the large distance between stations in that cluster.

SOM identified 3 clusters in the East including cluster 6,7, and 10. The clusters

identified by SOM in the East were heterogeneous except cluster 10 which was possibly heterogeneous.

In the inland part, K-Means identified 6 clusters. Cluster 4,6,9,11,14, and cluster 15. Cluster 11 was homogeneous while cluster 4 and 15 were possibly heterogeneous, cluster 6 was heterogeneous and cluster 9 and 14 were single stations. SOM formed 4 clusters inland, cluster 3,4,5, and 14. Cluster 3 and 5 were homogeneous, cluster 4 was heterogeneous and cluster 14 was possibly heterogeneous.

In the South K-Means identified 2 clusters: cluster 3 and 5. The 4 stations in cluster 3 were homogeneous while 6 stations in cluster 5 were heterogeneous. It has been observed that the size of clusters greatly influences homogeneity results.

SOM formed 2 clusters in the South. Cluster contained 5 stations and was heterogeneous while cluster 2 which contained 3 stations was homogeneous. It has been observed that small clusters have more probabilities of being homogeneous.

In the West, K-Means formed 3 clusters, cluster 2, 7, and 16. Cluster 2 was heterogeneous probably because 2 stations were located inland while the other 4 stations were located on the West coast. Cluster 7 was also heterogeneous while cluster 16 was homogeneous. SOM formed 4 clusters in West: cluster 11, 12, 13, and 15. Cluster 11 was possibly heterogeneous while other clusters were heterogeneous. The clusters formed by SOM in the West tend to have stations located towards inland within a different climatic zone. The facts that some stations in the same cluster may be in a different climatic zone explains why the cluster was heterogeneous.

K-Means and SOM clustered Jeju island as a separate precipitation cluster. The clusters have indeed been considered as a separate climatic clusters based on weather characteristics of the island and finding and previous studies (Nam et al., 2015).

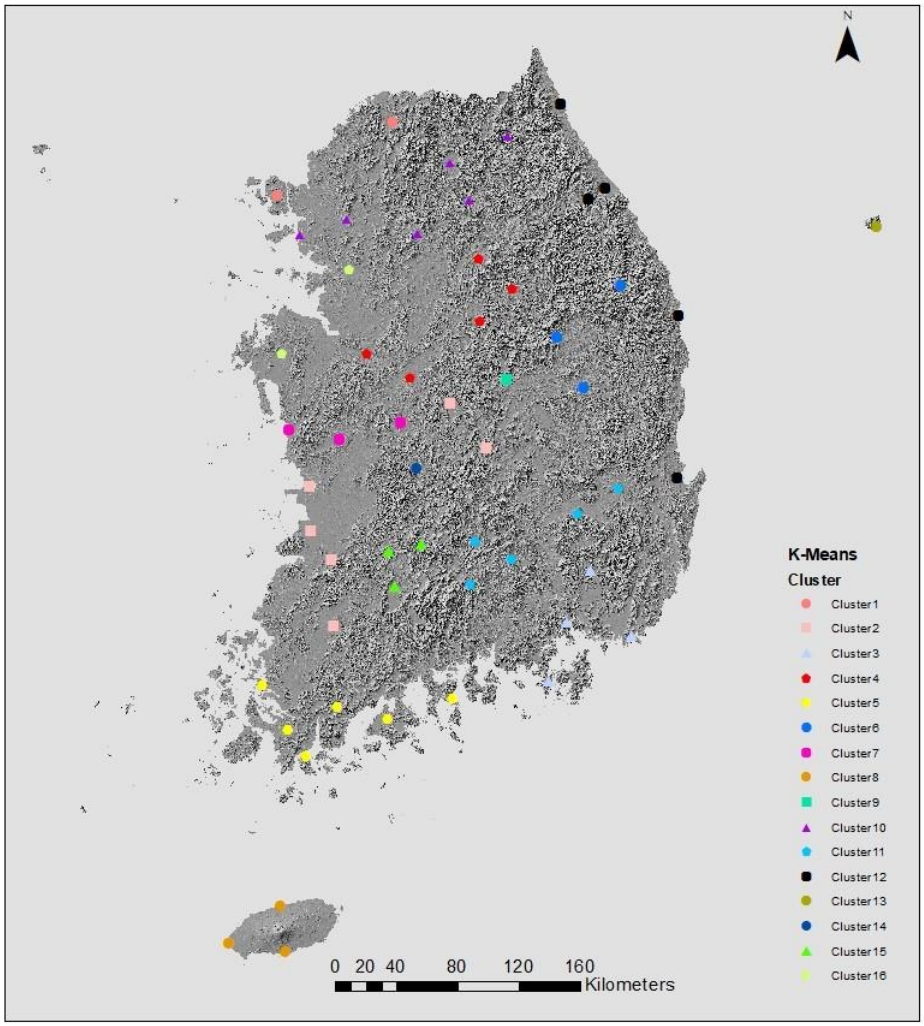


Figure 20. Map of ASOS clusters (K-Means)

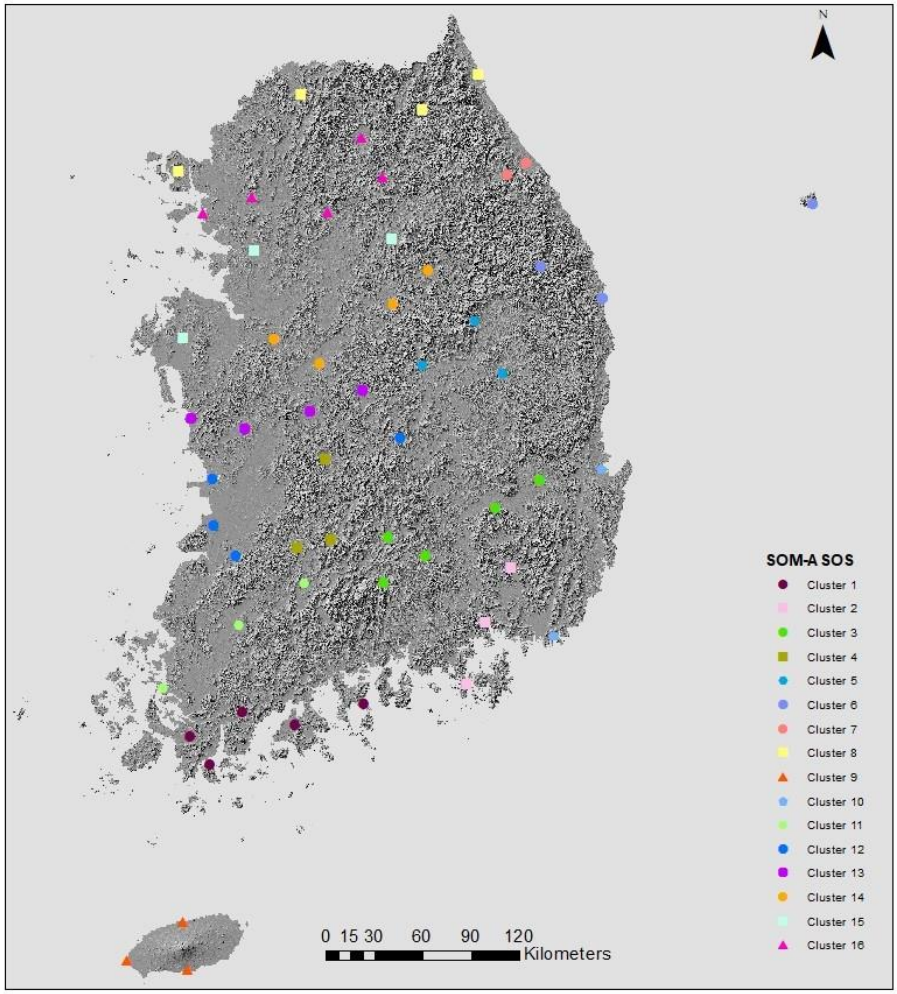


Figure 21. Map of ASOS clusters (SOM)

4.4.2. Spatial mapping of AWS clusters

Clusters identified by K-Means and SOM in AWS dataset have been grouped according to geographical location in 5 parts: The North, East, South and inland clusters as it was shown in the spatial mapping of K-Means clusters (Figure 22) and the spatial mapping of SOM clusters (Figure 23).

In the North, K-Means identified 2 clusters, cluster 10 cover the North-West and cluster 5 in the North-Est. The clusters in the North contained stations located in the North of Gangwon-do and Gyeonggi-do. Both clusters were heterogeneous. SOM identified 2 clusters in the same North, cluster 8 was heterogeneous and cluster 11 was homogeneous. As it can be observed in Figure 21, cluster 11 contained less stations and was homogeneous while the cluster 5 formed by K-Means in the same region, which contained more stations including stations in the Uleung-do was heterogeneous (Figure 20). The size of the clusters and geographical locations of the stations significantly affect homogeneity of the cluster.

In the East K-Means identified 3 clusters. The clusters cover East of Gangwon-do, North-East and South-East of Gyeongsangbuk-do and Gyeongsangnam-do. Clusters 5 was heterogeneous, while cluster 2 was homogeneous and cluster 4 was possibly heterogeneous. SOM formed 3 clusters in the East. Cluster 11 and 10 were homogeneous, while cluster 9 located in the South-Est of Gyeongsangnam-do was heterogeneous. SOM clustering showed that Uleung-do Island belonged to the cluster in the East coast of Gyeongsangbuk-do instead of belonging to Northeast coast of Gangwon-do as it was suggested by K-Means clustering in present study.

K-Means formed 3 clusters inland. Cluster 6,8, and cluster 2. Cluster 2 was in the East part of Gangwon-do and was found to be homogeneous while other inland clusters were heterogeneous. SOM identified 4 clusters. Cluster 2,3,6 and 7. The clusters were found to be heterogeneous.

In the South K-Means identified 2 clusters, cluster 12 and 7. Cluster 12 in South-East of Gyeongsangnam-do including Busan area was found to be heterogeneous while Cluster 7 in South East Jellanam-do, including Gwangju area, was homogeneous. SOM identified 2 heterogeneous clusters in the South, cluster 9 in East of Gyeongsangnam and cluster 5 in the West of Jellanam-do.

In the West, K-Means formed 3 clusters, cluster 1, 3, and 9. Cluster 1, in East of Chungcheongnam was possibly heterogeneous while other clusters in the East were found to be heterogeneous. SOM identified 3 clusters in the West. Cluster 2 was also extending inland from Jeollabuk-do to Gyeongsangbuk-do, cluster 4 was in the East of Chungcheongnam-do while cluster 12 was mainly located in West of Gyeonggi-do. The clusters in the West were found to be heterogeneous.

K-Means and SOM clustered Jeju island as a separate precipitation cluster but it was found to be heterogeneous probably because both clustering algorithms also includes some stations from the south of Jellanam-do in Jeju cluster.

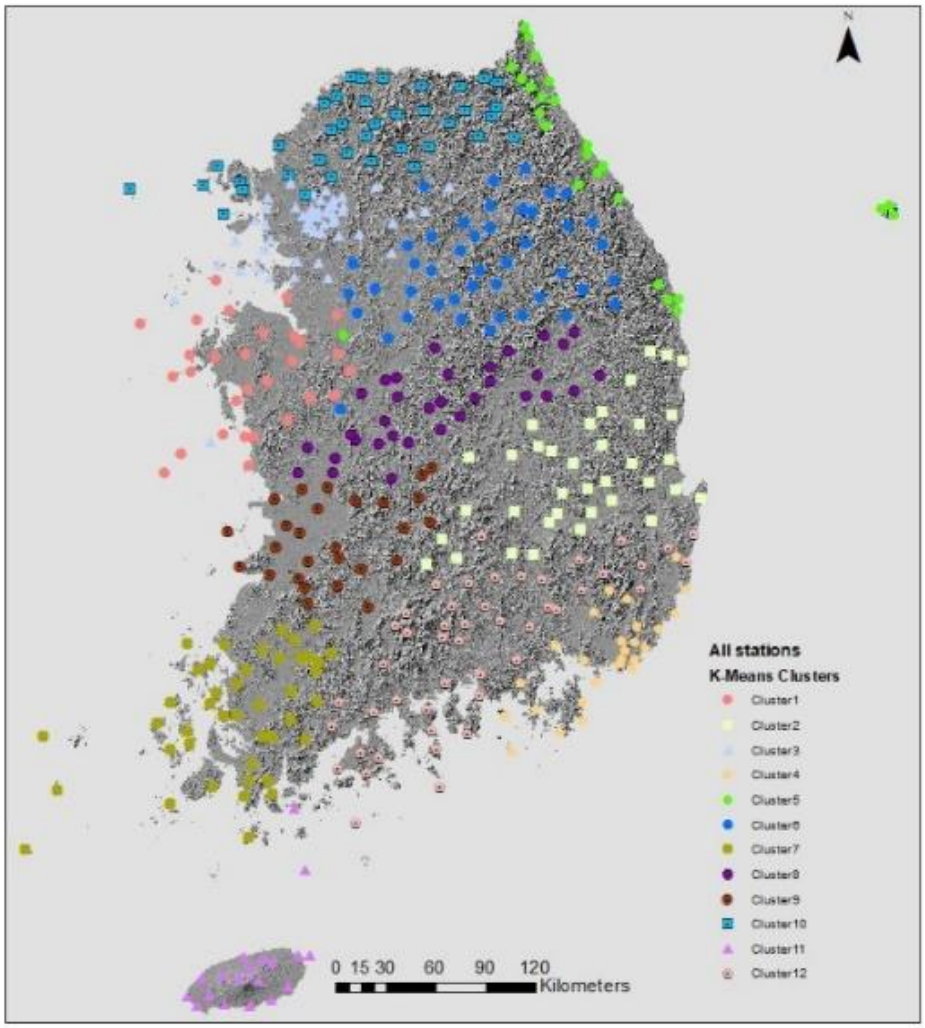


Figure 22. Map of AWS clusters (K-Means)

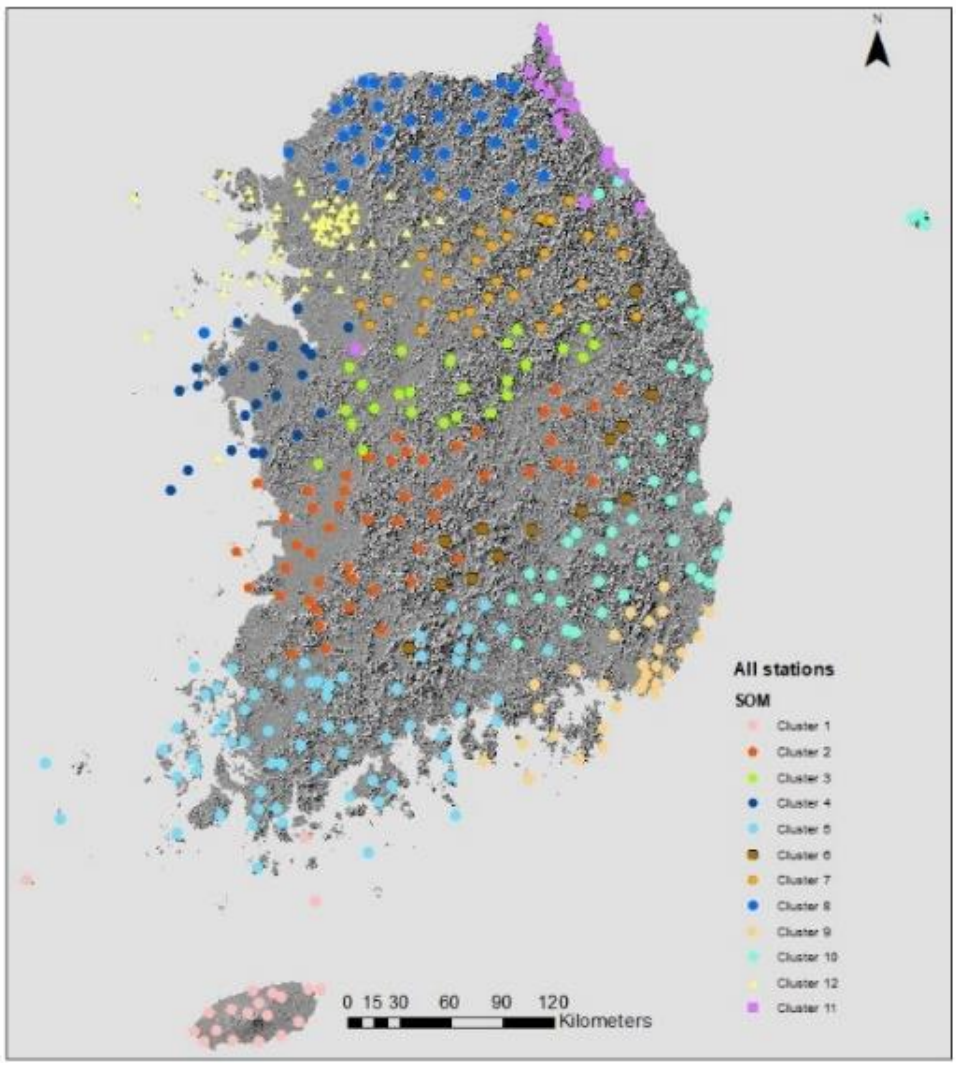


Figure 23. Map of AWS clusters (SOM)

4.4.3. Spatial mapping of Satellite dataset clusters

Clusters identified by K-Means clustering and SOM in satellite dataset have been grouped according to geographical location in 5 parts: The North, East, South and inland clusters as it was shown in the spatial mapping of K-Means clusters (Figure 24) and the spatial mapping of SOM clusters (Figure 25). Both clustering algorithms were able to form clusters that were geographically separated even though the clusters were formed based only on the long-term time series precipitation data from satellite dataset.

K-Means divided the North into 2 homogeneous clusters (cluster 3 and 15). Grids in those clusters were in the North of Gangwon-do and Gyeonggi-do. SOM grouped the same grids in the North in a larger cluster identified as cluster 14. The cluster was heterogeneous probably due to its large size.

K-Means identified 3 clusters in the East. The cluster covers the East of Gangwon-do, North-East and South-East of Gyeongsangbuk-do. Cluster 10 and 16 clusters in the East were homogeneous while cluster 6 was possibly heterogeneous. SOM identified 5 clusters in the East including cluster 15, 13, 5, and 1. Clusters identified by SOM in the East were homogeneous except cluster 15.

K-Means identified 3 inland homogeneous clusters, including cluster 2, 9 and cluster 12). SOM identified 4 inland clusters including cluster 8,4,3, and 6. All the inland clusters identified by K-Means and SOM were homogeneous.

K-Means identified 3 clusters in the South including cluster 1,7, and 15. Cluster 1, 7 and 15 were homogeneous. The clusters cover the south of Jellanam-do. SOM identified 2 clusters in the South, cluster 9 was homogeneous while cluster 14 was heterogeneous.

In the West part of Korea, K-means identified 5 clusters, cluster 4,5,13,11 and 14. Clusters 11 and 13 were homogeneous while others were heterogeneous. SOM

identified 2 clusters in South-West including cluster 7 and cluster 11. Cluster 7 was homogeneous and cluster 11 was heterogeneous. The North-West part of the country was covered by 2 big clusters identified by SOM, cluster 14 and 12 which were found to be heterogeneous.

K-Means and SOM clustered Jeju island as a separate cluster. The Island is indeed considered a separate regions based on the location of the Island and findings of previous study (Nam et al., 2015).

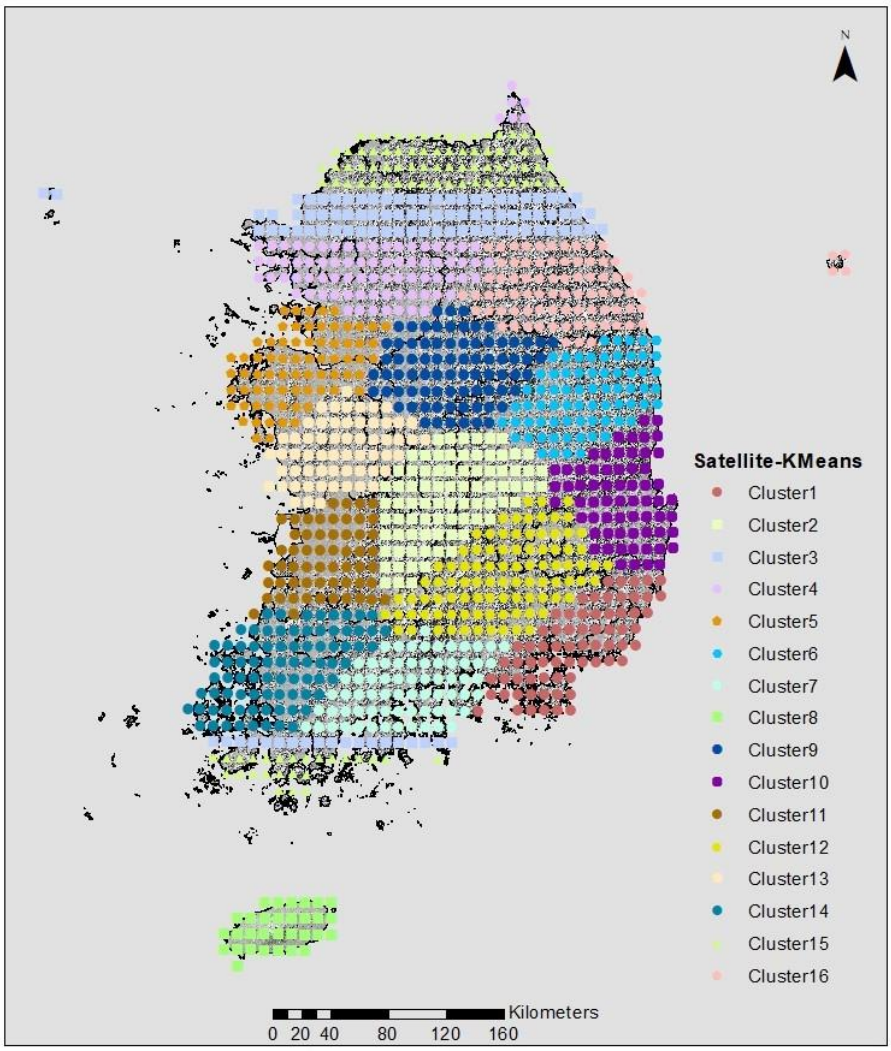


Figure 24. Map of clusters satellite data (K-Means)

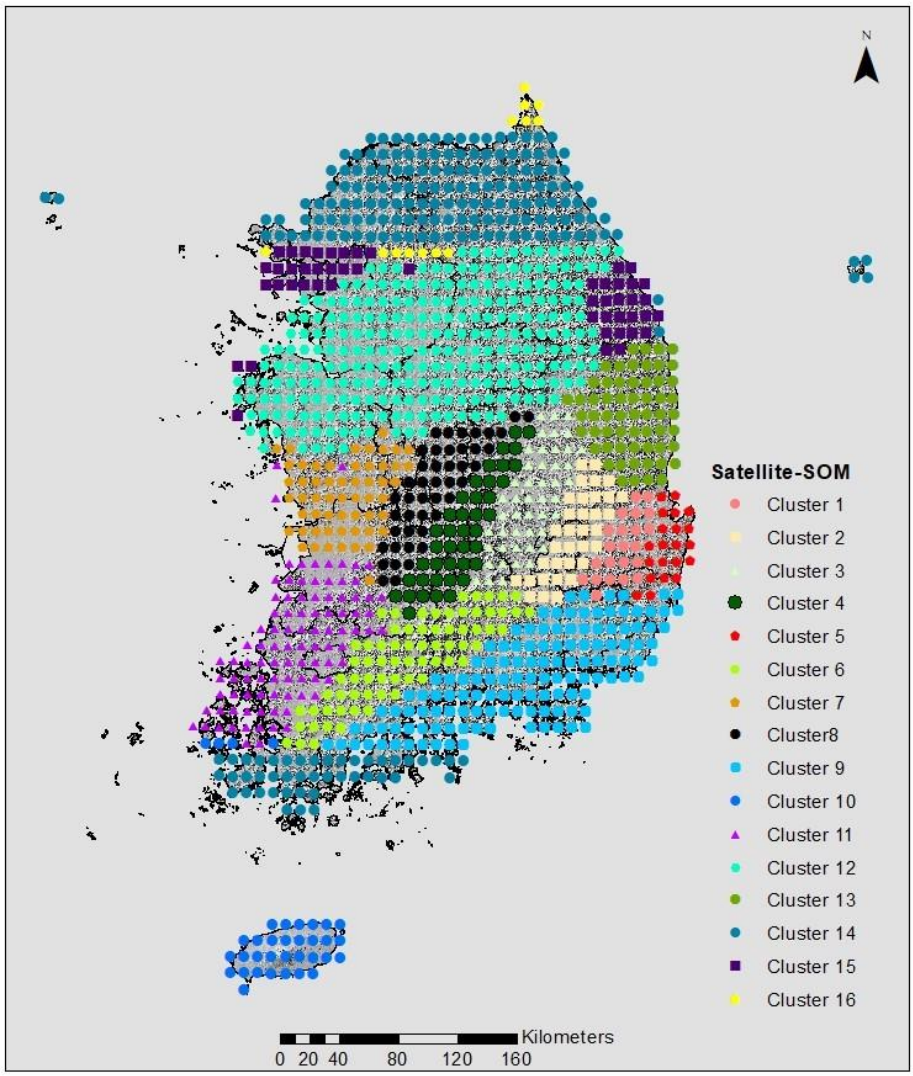


Figure 25. Map of clusters satellite data (SOM)

Chapter 4. Conclusion

An approach was developed in this study to define homogeneous precipitation regions using timeseries of gauge and satellite data. The clusters from each method were tested by means of Hosking and Wallis heterogeneity measure to confirm if they were hydrologically homogeneous. The regions formed by both clustering algorithms were spatially mapped and results of homogeneity test for each region were provided. The results of the homogeneity test have been used to evaluate and compare the ability of each clustering method in defining homogeneous precipitation regions using time series data.

K-means clustering and SOM yielded almost similar number of homogeneous regions. But SOM was able to identify greater number of homogeneous regions when map size was increased. Furthermore, SOM identified fewer sole stations compared to K-Means. Both clustering methods identified 10 homogeneous regions out of 16 regions in satellite precipitation dataset.

Homogeneous regions formed by both clustering methods can be used for precipitation estimation, hydrological modelling, and regional frequency analysis. It was observed that the number of clusters greatly affect the capacity of the clustering algorithm in defining homogeneous precipitation regions. Mainly because smaller clusters have more probability of being homogeneous. Since the number of clusters for K-Means or the number nodes for SOM is predefined by researcher before clustering, it was recommended to try and evaluate different clusters numbers before clustering timeseries data for precipitation regionalization purpose.

Bibliography

Alam, M. S., & Paul, S. (2020). A comparative analysis of clustering algorithms to identify the homogeneous rainfall gauge stations of bangladesh. *Journal of Applied Statistics*, *47*(8), 1460-1481. https://doi.org/10.1080/02664763.2019.1675606

Arthur, D., & Vassilvitskii, S. (2007). K-means plus plus : The advantages of careful seeding. *Proceedings of the Eighteenth Annual Acm-Siam Symposium on Discrete Algorithms*, 1027-1035.

Bilbro, B. B. a. R. (2019). *Yellowbrick*. https://www.scikit-yb.org/

Carvalho, M. J., Melo-Goncalves, P., Teixeira, J. C., & Rocha, A. (2016). Regionalization of europe based on a k-means cluster analysis of the climate change of temperatures and precipitation. *Physics and Chemistry of the Earth*, *94*, 22-28. https://doi.org/10.1016/j.pce.2016.05.001

Chen, S. L., Xiong, L. H., Ma, Q. M., Kim, J. S., Chen, J., & Xu, C. Y. (2020). Improving daily spatial precipitation estimates by merging gauge observation with multiple satellite-based precipitation products based on the geographically weighted ridge regression method. *Journal of Hydrology*, *589*. https://doi.org/ARTN 125156 10.1016/j.jhydrol.2020.125156

Claps, P., Ganora, D., & Mazzoglio, P. (2022). Chapter 11 - rainfall regionalization techniques. In R. Morbidelli (Ed.), *Rainfall* (pp. 327-350). Elsevier. https://doi.org/https://doi.org/10.1016/B978-0-12-822544-8.00013-5

Gaál, L., Kyselý, J., & Szolgay, J. (2008). Region-of-influence approach to a frequency analysis of heavy precipitation in slovakia. *Hydrology and Earth System Sciences*, *12*(3), 825-839.

Giacco, F., Thiel, C., Pugliese, L., Scarpetta, S., & Marinaro, M. (2010). Uncertainty analysis for the classification of multispectral satellite images using svms and soms. *Ieee Transactions on Geoscience and Remote Sensing*, *48*(10), 3769-3779. https://doi.org/10.1109/Tgrs.2010.2047863

Haddad, K., Johnson, F., Rahman, A., Green, J., & Kuczera, G. (2015). Comparing three methods to form regions for design rainfall statistics: Two case studies in australia. *Journal of Hydrology*, *527*, 62-76. https://doi.org/https://doi.org/10.1016/j.jhydrol.2015.04.043

Hosking, J. R. M., & Wallis, J. R. (1997). *Regional frequency analysis: An approach based on l-moments*. Cambridge University Press. https://doi.org/DOI: 10.1017/CBO9780511529443

Irwin, S., Simonovic, S. P., & Burn, D. H. (2017). Delineation of precipitation regions in two canadian study areas: The role of the temporal resolution of the precipitation data. *Hydrological Sciences Journal*, *62*(13), 2061-2071. https://doi.org/10.1080/02626667.2017.1353694

Jackson, I. J., & Weinand, H. (1995). Classification of tropical rainfall stations - a comparison of clustering-techniques. *International Journal of Climatology*, *15*(9), 985-994. https://doi.org/DOI 10.1002/joc.3370150905

Kalkstein, L. S., Tan, G., & Skindlov, J. A. (1987). An evaluation of three clustering procedures for use in synoptic climatological classification. *Journal of Applied Meteorology and Climatology*, *26*(6), 717-730. https://doi.org/10.1175/1520-0450(1987)026<0717:AEOTCP>2.0.CO;2

Kim, H. U., Sohn, C., & Han, S.-O. (2012). Identifying the optimal number of homogeneous regions for regional frequency analysis using self-organizing map. *Spatial Information Research*, *20*, 13-21.

Kohonen, T. (1990). The self-organizing map. *Proceedings of the IEEE*, *78*(9), 1464-1480.

https://doi.org/10.1109/5.58325
Kohonen, T. (1991). Self-organizing maps: Ophmization approaches. In T. Kohonen, K. MÄKisara, O. Simula, & J. Kangas (Eds.), *Artificial neural networks* (pp. 981-990). North-Holland. https://doi.org/https://doi.org/10.1016/B978-0-444-89178-5.50003-8
Kohonen, T. (1998). The self-organizing map. *Neurocomputing*, *21*(1-3), 1-6. https://doi.org/Doi 10.1016/S0925-2312(98)00030-7
Kohonen, T. (2001). Self-organizing maps of massive databases. *Engineering Intelligent Systems for Electrical Engineering and Communications*, *9*(4), 179-185.
Lary, D. J., Alavi, A. H., Gandomi, A. H., & Walker, A. L. (2016). Machine learning in geosciences and remote sensing. *Geoscience Frontiers*, *7*(1), 3-10. https://doi.org/10.1016/j.gsf.2015.07.003
Leptoukh, G., Ahmad, S., Eaton, P., Koziana, J., Ouzounov, D., Savtchenko, A., Serafino, G., Sharma, A., Sikder, M., & Zhou, B. (2001). Modis data ingest, processing, archiving and distribution at the goddard earth sciences daac. *Igarss 2001: Scanning the Present and Resolving the Future, Vols 1-7, Proceedings*, 2286-2288.
Licen, S., Astel, A., & Tsakovski, S. (2023). Self-organizing map algorithm for assessing spatial and temporal patterns of pollutants in environmental compartments: A review. *Science of the Total Environment*, *878*. https://doi.org/ARTN 163084 10.1016/j.scitotenv.2023.163084
Mallants, D., & Feyen, J. (1990). Defining homogeneous precipitation regions by means of principal components-analysis. *Journal of Applied Meteorology*, *29*(9), 892-901. https://doi.org/Doi 10.1175/1520-0450(1990)029<0892:Dhprbm>2.0.Co;2
Miljkovic, D. (2017). Brief review of self-organizing maps. *2017 40th International Convention on Information and Communication Technology, Electronics and Microelectronics (Mipro)*, 1061-1066.
Nam, W., Shin, H., Jung, Y., Joo, K., & Heo, J. H. (2015). Delineation of the climatic rainfall regions of south korea based on a multivariate analysis and regional rainfall frequency analyses. *International Journal of Climatology*, *35*(5), 777-793.
Nathan, R. J., & McMahon, T. A. (1990). Identification of homogeneous regions for the purposes of regionalisation. *Journal of Hydrology*, *121*(1), 217-238. https://doi.org/https://doi.org/10.1016/0022-1694(90)90233-N
Ralambondrainy, H. (1995). A conceptual version of the k-means algorithm. *Pattern Recognition Letters*, *16*(11), 1147-1157. https://doi.org/Doi 10.1016/0167-8655(95)00075-R
Rasheed, A., Egodawatta, P., Goonetilleke, A., & McGree, J. (2019). A novel approach for delineation of homogeneous rainfall regions for water sensitive urban design—a case study in southeast queensland. *Water*, *11*(3). https://mdpi-res.com/d_attachment/water/water-11-00570/article_deploy/water-11-00570.pdf?version=1552992285
Roushangar, K., & Alizadeh, F. (2018). A multiscale spatio-temporal framework to regionalize annual precipitation using k-means and self-organizing map technique. *Journal of Mountain Science*, *15*(7), 1481-1497. https://doi.org/10.1007/s11629-017-4684-5
Saikranthi, K., Rao, T. N., Rajeevan, M., & Rao, S. V. B. (2013). Identification and validation of homogeneous rainfall zones in india using correlation analysis. *Journal of Hydrometeorology*, *14*(1), 304-317. https://doi.org/10.1175/Jhm-D-12-071.1
Satyanarayana, P., & Srinivas, V. V. (2011). Regionalization of precipitation in data sparse areas using large scale atmospheric variables – a fuzzy clustering approach. *Journal of Hydrology*, *405*(3), 462-473.

https://doi.org/https://doi.org/10.1016/j.jhydrol.2011.05.044
Srinivas, V. (2013). Regionalization of precipitation in india–a review. *Journal of the Indian Institute of Science*, *93*(2), 153-162.
Tabios, G. Q., & Salas, J. D. (1985). A comparative-analysis of techniques for spatial interpolation of precipitation. *Water Resources Bulletin*, *21*(3), 365-380
Varouchakis, E. A., Solomatine, D., Perez, G. A. C., Jomaa, S., & Karatzas, G. P. (2023). Combination of geostatistics and self-organizing maps for the spatial analysis of groundwater level variations in complex hydrogeological systems. *Stochastic Environmental Research and Risk Assessment*. https://doi.org/10.1007/s00477-023-02436-x
Vettigli, G. (2018). *Minisom: Minimalistic and numpy-based implementation of the self organizing map*. https://github.com/JustGlowing/minisom/
Wang, H., & Yong, B. (2020). Quasi-global evaluation of imerg and gsmap precipitation products over land using gauge observations. *Water*, *12*(1). https://mdpi-res.com/d_attachment/water/water-12-00243/article_deploy/water-12-00243-v2.pdf?version=1579744699
Zhang, L., Li, X., Zheng, D., Zhang, K., Ma, Q., Zhao, Y., & Ge, Y. (2021). Merging multiple satellite-based precipitation products and gauge observations using a novel double machine learning approach. *Journal of Hydrology*, *594*, 125969. https://doi.org/https://doi.org/10.1016/j.jhydrol.2021.125969

Abstract(Korean)

기계 학습을 이용한 관측 및 위성 강수 시계열 데이터 기반 동질 강수 지역 분석 연구

문옌산가

국제농업기술학과

국제농업기술대학원

서울대학교

동질 강수 지역의 구분은 지역 빈도 분석과 강수량 추정에 필요하지만 이러한 지역의 구분은 강수량의 시간적, 공간적 가변성으로 인해 많은 불확실성을 내포하고 있다.

본 연구는 지상 관측 강수 데이터를 이용한 동질 강수 지역 분석과 관련된 문제점을 해결하고자 하였다. 첫 번째 문제는 동질 강수 지역 분석에서 자주 고려되지 않는 강수량의 시간적 변동성이다. 강수량은 시간과 공간에 따라 많은 차이가 나는 것으로 알려져 있다. 그러나 동질 강수 지역의 분석 및 구분을 위한 많은 선행 연구들은 일반적으로 시계열 자료 대신 강수량 평균, 관측소의 위치 특성과 같은 변수를 사용하기 때문에 강수의 시간적 변동성을 고려하지 못했다. 시간 변동성 문제를 극복하기 위해 본 연구에서는 시계열 강수 데이터를 사용하여 동질 강수 지역 구분 및 분석하였다.

두 번째 문제점은 공간적 강수량의 변화다. 강수량계를 이용한 강수량 측정은 다른 강수량 측정을 위한 여러가지 방법 중에 가장 정확한 것으로 알려져 있기 때문에 전통적으로 강수량 자료의 확보를 위해 가장

많이 활용되고 있다. 그러나 세계의 많은 지역에서 강수량 관측 밀도가 낮고 강수량 추정을 위한 보간 기술로 인해 오류가 발생할 수 있다는 점을 감안하면 한 지점에서 수집된 강수 자료를 정확하게 보간하여 강수량 자료를 추정하고 확보하는 것은 한계가 있다.

이러한 공간적 변동성 문제를 해결하기 위해 본 연구에서는 위성 강수 데이터를 사용하여 동질 강수 지역을 구분하고자 하였다. 위성 강수 데이터는 여러 위성 센서에서 수신한 적외선 및 수동 마이크로웨이브 정보를 통해 간접적으로 강수량을 추정하는 것으로 최근의 강수량 측정을 위한 방법으로 이용되고 있다. 위성 강수 데이터는 표면 그리드 형태로 제공된다.

본 연구에서는 지상 관측 및 위성 강수 시계열 일 자료를 이용하여 동질 강수 지역 분석을 위한 기계 학습 방법론을 제공하고자 하였다. 본 연구에 사용된 지상 관측 강수량 자료는 기상청에서 제공하고 있는 종관기상관측 (ASOS, Automated Synoptic Observing System) 및 방재기상관측 (AWS, Automated Weather Station) 자료가 각각 사용되었다. 본 연구에서 사용된 위성 데이터는 미국항공우주국 (NASA)의 IMERG(Integrated Multi-satellitE Retrievals for GPM)이다.

동질 강수 지역은 K-Means와 SOM (Self Organizing Maps)의 두 가지 클러스터링 방법을 이용하여 분석하였다. 동질 지역 구분에 따른 각 동질 지역의 이질성 분석은 Hosking과 Wallis homogeneity test를 이용하였다. 종관기상 (ASOS) 관측 자료를 이용하여 동질 강수 지역으로 구분된 지역의 이질성을 분석한 결과에 따르면 SOM의 동질 강수 지역 구분 성능이 맵의 크기에 따라 크게 영향을 받는 것으로 나타났다. SOM은 노드 수가 증가할수록 더 많은 수의 동질 강수 지역을 분류할 수 있었다.

노드 수가 16개로 증가했을 때 6개의 지역이 동질성을 가지는 것으로 나타났으나, 반면 K-Means는 5개의 지역이 동질 강수 지역인 것으로 나타났다. K-Means는 군집 수가 적을 때 더 많은 수의 동질 지역을 구분할 수 있었다. 예를 들어 클러스터 수가 10개일 때 K-Means는 3개의 지역이 동질한 것으로 나타났으나, 반면 SOM은 2개의 지역이 동질성이 있는 것으로 나타났다. 그러나 노드 수가 10개에서 16개로 증가함에 따라 SOM에 의해 분류된 동질 지역의 수는 점차 증가했다.

방재기상관측 (AWS) 자료를 이용한 결과에서는 SOM 및 K-Means 방법을 적용하여 구분된 지역의 동질성을 분석 결과가 유사한 것으로 나타났다. 두 방법으로 구분된 강수 지역의 동질성은 클러스터 수가 12, 14 또는 16으로 증가해도 개선되지 않았다.

위성 강수 자료를 이용한 동질 강수 지역 구분 및 동질성 분석 결과에서는 SOM과 K-Means는 군집의 수에 따라 동질 지역의 수가 차이가 있었지만 거의 동일한 수준이었다. K-Means는 9개의 동질 강수 지역 중에서 2개의 지역에서 동질성이 있는 것으로 나타났으며, SOM에 의한 동질 지역의 수는 4개의 지역에서 동질성이 있는 것으로 나타났다.

본 연구에서는 전반적으로 지상 관측 및 위성 강수 데이터를 이용하여 동질 강수 지역을 구분할 경우 SOM 방법이 K-Means 방법에 비해 더 많은 동질 강수 지역을 구분할 수 있는 것으로 나타났다.

주요어: **동질 강수 지역, 위성 강수 자료, 기계 학습, 지상 관측 자료, 시계열 자료**

학 번: 2021-22782