BASIN-BASED CLUSTERING OF HYDROELECTRIC POWER PLANTS IN TURKEY BY USE OF STREAM-FLOW AND HYDROELECTRIC ENERGY PRODUCTION DATA

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

YUSUF ARSLAN

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
COMPUTER ENGINEERING

SEPTEMBER 2015

Approval of the thesis:

BASIN-BASED CLUSTERING OF HYDROELECTRIC POWER PLANTS IN TURKEY BY USE OF STREAM-FLOW AND HYDROELECTRIC ENERGY PRODUCTION DATA

submitted by **YUSUF ARSLAN** in partial fulfillment of the requirements for the degree of **Master of Science in Computer Engineering Department, Middle East Technical University** by,

Prof. Dr. Gülbin Dural Ünver _____
Dean, Graduate School of **Natural and Applied Sciences**

Prof. Dr. Adnan Yazıcı _____
Head of Department, **Computer Engineering**

Dr. Ayşenur Birtürk _____
Supervisor, **Computer Engineering Department, METU**

Examining Committee Members:

Prof. Dr. Nihan Kesim Çiçekli
Computer Engineering Department, METU _____

Dr. Ayşenur Birtürk
Computer Engineering Department, METU _____

Prof. Dr. İsmail Hakkı Toroslu
Computer Engineering Department, METU _____

Assoc. Prof. Dr. Pınar Karagöz
Computer Engineering Department, METU _____

Assoc. Prof. Dr. Osman Abul
Computer Engineering Department, TOBB ETÜ _____

Date: _____

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last Name: YUSUF ARSLAN

Signature :

ABSTRACT

BASIN-BASED CLUSTERING OF HYDROELECTRIC POWER PLANTS IN TURKEY BY USE OF STREAM-FLOW AND HYDROELECTRIC ENERGY PRODUCTION DATA

Arslan, Yusuf

M.S., Department of Computer Engineering

Supervisor : Dr. Ayşenur Birtürk

September 2015, 94 pages

A grouping approach may ease the process of supply prediction. It is shown that trend of the streams in the same basins have similar trend and it is also important to analyse that whether there are similarity between the trend of the neighbour basins or not. It is seen in the experiments that structure based hierarchical clustering makes the clustering based on the trend of the time series and this method reveals the connections between the basins which have similar trends in the flow of their streams. The aim of this thesis is to find the basin based clustering of the hydroelectric power plants and stream-flow and hydroelectric energy production datasets are used and the results of both are compared with each other.

In conclusion, it is shown that basin based clustering is done successfully by use of structure based hierarchical clustering on the stream-flow dataset and the result is visualized on the Turkey map.

Keywords: hydroelectric power production, stream-flow rate, structure based hierarchical clustering, longest common subsequence, basin based clustering

ÖZ

TÜRKİYE'DEKİ HİDROELEKTRİK SANTRALLERİN AKARSU AKIŞ HIZI VE HİDROELEKTRİK ENERJİ ÜRETİMİ VERİLERİ KULLANILARAK HAVZA BAZLI KÜMELENMESİ

Arslan, Yusuf
Yüksek Lisans, Bilgisayar Mühendisliği Bölümü
Tez Yöneticisi : Dr. Ayşenur Birtürk

Eylül 2015 , 94 sayfa

Bir gruplama yaklaşımı arz tahmini işlemini kolaylaştırabilir. Aynı havzadaki akarsuların benzer eğilime sahip olduğu gösterilmiştir ve komşu havzaların da benzer eğilimlere sahip olup olmadıklarının analiz edilmesi de aynı derecede önemlidir. Yapılan deneylerde görülmüştür ki yapısal tabanlı hiyerarşik kümeleme zaman serilerinin eğilimlerine göre kümelenmesini gerçekleştirmiştir ve bu yöntem veri kümesindeki benzer eğilimlere sahip havzalar arasındaki ilişkileri ortaya çıkarmıştır. Bu tezin amacı havza bazlı hidroelektrik santral sınıflandırılmasıdır ve akarsu akış hızı ve hidroelektrik enerji üretimi veri kümeleri kullanılmıştır ve her ikisine ait sonuçlar birbiriyle kıyaslanmıştır.

Sonuç olarak, havza bazlı sınıflandırma yapısal tabanlı hiyerarşik kümelemenin kullanılmasıyla başarılı bir şekilde gerçekleştirilmiştir ve sonuçlar Türkiye haritası üzerinde görselleştirilmiştir.

Anahtar Kelimeler: hidroelektrik enerji üretimi, akarsu akış hızı, yapısal tabanlı hiyerarşik kümeleme, en uzun ortak küme, havza bazlı kümeleme

To my father, my mother and my grandparents, *who spent all their lives just to illuminate the way of their children*

ACKNOWLEDGMENTS

I would like to express my sincere gratitude to my supervisor Dr. Ayşenur Birtürk for her guidance, support and positive attitude throughout this study.

I would like to thank to the examining committee members: Prof. Dr. Nihan Kesim Çiçekli, Prof. Dr. İsmail Hakkı Toroslu, Assoc. Prof. Dr. Pınar Karagöz and Assoc. Prof. Dr. Osman Abul for their very precious reviews in my thesis and very constructive comments during my thesis presentation.

I would like to express my great appreciation to Sinan Eren for suggesting me to work on this topic and for helping me in every stage of this study.

I am grateful to Dr. Turan Demirci, Dr. Dilek Küçük and Prof. Dr. Zuhal Akyürek for their support which helped me sort out the technical details of this study.

I would like to express my gratefulness to my parents, my brothers and my sister for their encouragement, support and understanding throughout this study.

A special thanks to my friends Mehmet Barış Özkan and Jeyhun Karimov, who are there whenever I need them.

TABLE OF CONTENTS

LIST OF TABLES

TABLES

LIST OF FIGURES

FIGURES

LIST OF ABBREVIATIONS

TÜBİTAK	The Scientific and Technological Research Council of Turkey
YTBS	Dispatcher Information System
PBS	Planning Information System
TEİAŞ	Turkish Electricity Transmission Company
DSİ	General Directorate of State Hydraulic Works
EİE	General Directorate of Electrical Power Resources Survey and Development Administration
HES	Hydroelectric Power Plant
PCA	Principal Component Analysis
FRIEND	Flow Regimes from International Experimental and Network Data
DTW	Dynamic Time Warping
LCSS	Longest Common Subsequence

CHAPTER 1

INTRODUCTION

1.1 Streams and Basins

A stream is a body of water with a current, confined within a bed and stream banks [1]. There are plenty of streams in Turkey. Turkey is divided into seven geographical regions and it has streams in each of them. Karadeniz region has Kızılırmak, Yeşilırmak, Bartın Çayı, Kelkit Çayı, Filyos, Doğankent Çayı, Çoruh, İyidere and Fırtına deresi. Akdeniz region has Dalaman Çayı, Eşen Çayı, Manavgat, Aksu, Köprü, Seyhan, Ceyhan and Asi rivers. Ege region has Bakırçay, Gediz, Büyük Menderes and Küçük Menderes rivers. Marmara region has Meriç (Ergene), Sakarya, Susurluk, Orhaneli Çayı, Nilüfer Çayı and Gönen Çayı. İç Anadolu region has Çarşamba Suyu, Porsuk, Sakarya, Kızılırmak and Samantı Çayı. Doğu Anadolu region has Fırat, Aras, Kura, Karasu, Murat, Dicle, Arpaçay and Zap Suyu. Güneydoğu Anadolu region has Fırat and Dicle rivers. These streams are used for various reasons, such as agricultural irrigation, drinking water supply and energy production. In energy area, for example, Turkey's streams have a huge capacity for energy production when compared with the capacity of the streams in European countries. Turkey has the highest hydroelectric power potential among all European countries as can be seen in Figure 1.1.

According to the installed hydropower capacities, Turkey is the seventh among 26 European countries as can be seen in Figure 1.2.

Over 500 hydroelectrical power plants have been built in Turkey's streams until now and some of them are still under construction. The streams and hydroelectric power plants in Turkey can be seen in Figure 1.3.

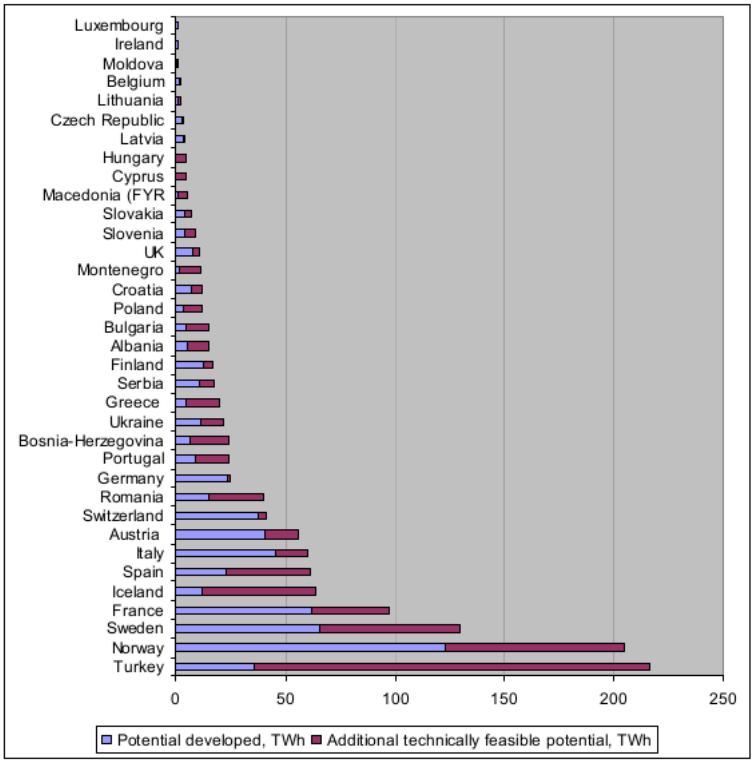

Figure 1.1: European countries hydropower potential [2]

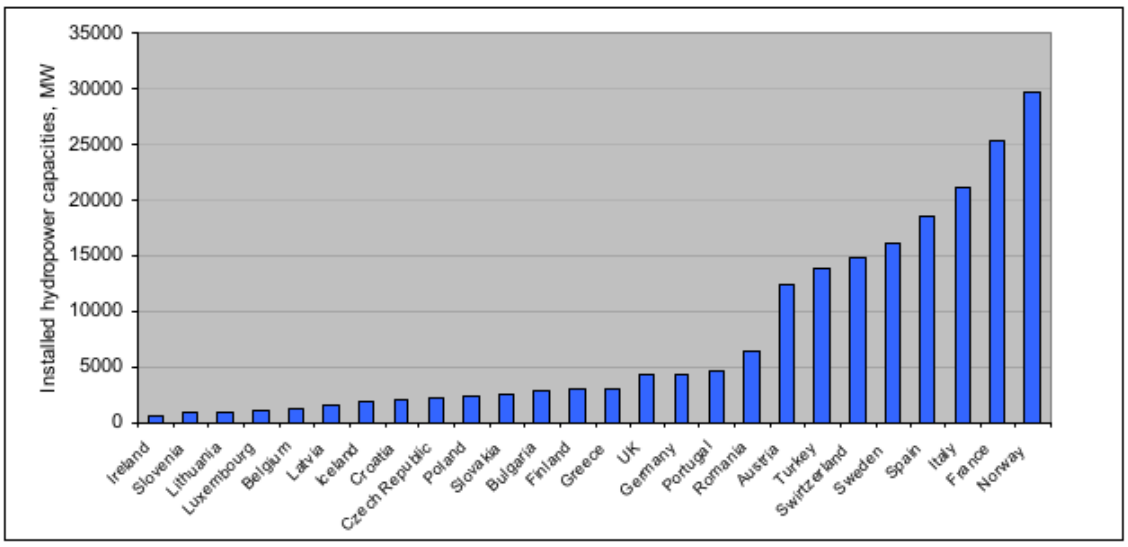

Figure 1.2: Installed hydropower capacities in MW in European countries, 2009 [2]

Figure 1.3: Streams and dams of Turkey

About 34% of the total electric energy production in Turkey is obtained from streams. The installed power capacity of electric energy of Turkey can be seen in Figure 1.4.

Basins are streams with all of the tributaries. River basin classification is important in ecology, hydrology and energy areas. With the help of river classification, researches are carried out on different topics in hydrology like flood and drought prediction, location of drinking water supply reserves and impact of global warming on the regions. Identification of hydrologic classes to increase the knowledge of flow variability in dispersion through streams and rivers, guidance of regionalisation analysis by using hydrologic classification, development of environmental flow guidance for water reserve management, identification and prioritization of protection attempts for freshwater ecosystems are some of the projects in ecology [3]. It is used for security of supply in energy domain. Turkey is divided into 26 basins by General Directorate of State Hydraulic Works (DSİ). The basin map of Turkey can be seen in Figure 1.5.

The annual flow and potential hydropower of each basin in Turkey can be seen in Table 1.1.

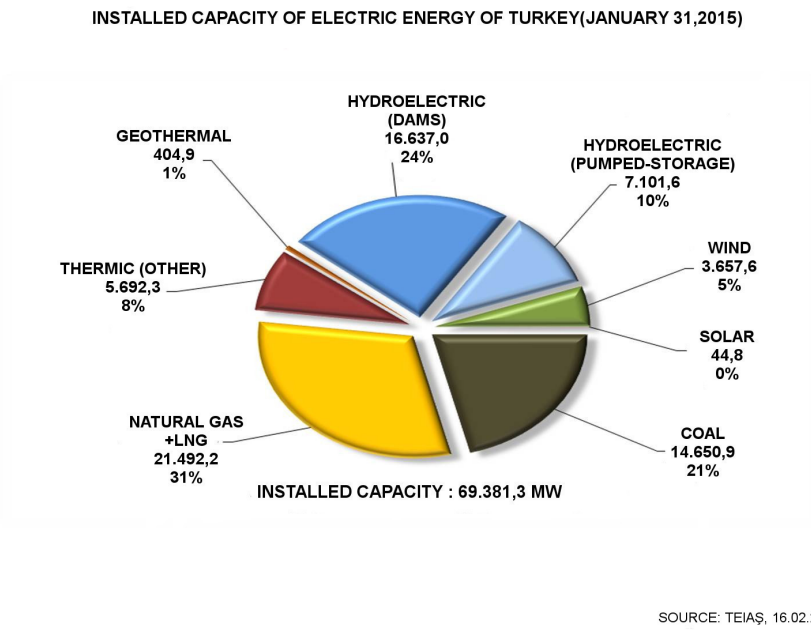

Figure 1.4: Installed power capacity of electric energy of Turkey

1.2 Motivation

There are a lot of streams in Turkey's basins. The amount and rate of the stream-flow in these rivers are recorded by hydroelectric power plants and the gauging stations. In some streams, these records go back to 1950s. Studies on these areas so far have used 31-years old stream-flow records (1964-1994) acquired from gauging stations of DSİ and General Directorate of Electrical Power Resources Survey and Development Administration (EİE) [5], [6], [7], [8]. These records originally collected from over 240 points on the streams. However, correctness of 60% of the measurement points data was found suspicious and discarded from the dataset. Records used in the studies contain 80 measurement points data. Moreover, the related dataset does not contain any information about 4 of the 26 basins of Turkey [9]. The datasets used in this project, on the other hand, have not been used in any project so far. One of these datasets contains stream-flow rate and amount information and it is collected from 26 dams with power plants. It contains stream-flow information of 14 out of 26 basins. It is mentioned as stream-flow dataset in our thesis. The other dataset contains 1 year electric energy production information of 75 hydroelectric power plants with dams and 311 run-of-river type hydroelectric power plants. It is mentioned as hydroelectric energy production dataset in our thesis. Our study aims to make an inference about

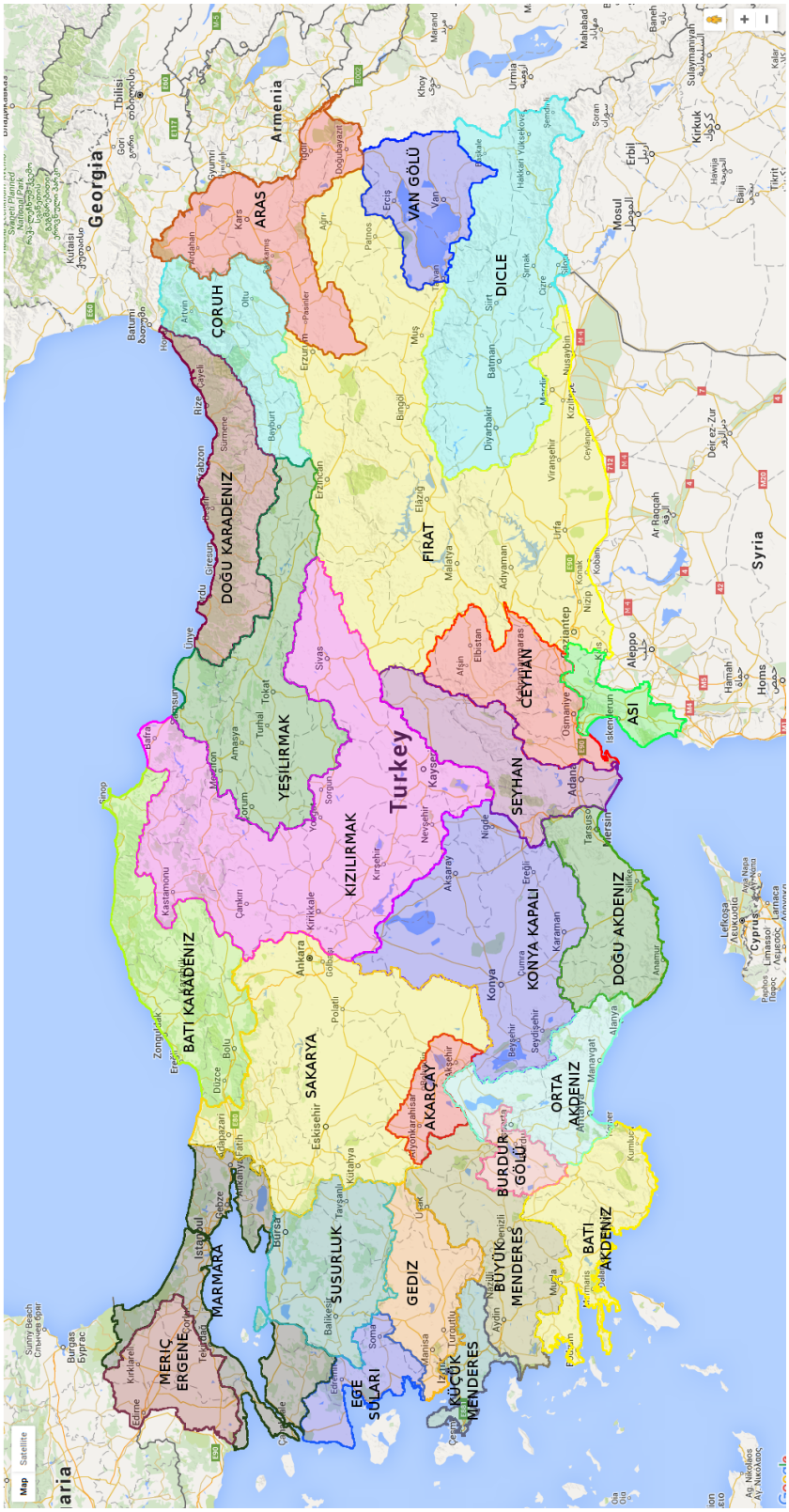

Figure 1.5: Basins of Turkey

Table 1.1: Distribution of Theoretic HPP Potential of Turkey based on Basins [4]

	Basin Name	**Average Streamflow (billion $m^3/year$)**	**Basin flow / $\sum Flow$**	**Theoratical HPP Potential (GWh/year)**	**Basin Potential / $\sum Potential$ (%)**
1	Fırat (Euphrates)	31,61	17,00	84,11	19,50
2	Dicle (Tigris)	21,33	11,50	48,71	11,30
3	Doğu Karadeniz(Eastern Black Sea)	14,90	8,00	48,48	11,20
4	Doğu Akdeniz(Eastern Mediterranean)	11,07	6,00	27,45	6,40
5	Antalya	10,06	5,40	23,08	5,30
6	Batı Karadeniz(Western Black Sea)	9,93	5,30	17,91	4,20
7	Batı Akdeniz(Western Mediterranean)	8,93	4,80	13,60	3,20
8	Marmara	8,33	4,50	5,18	1,20
9	Seyhan	8,01	4,30	20,88	4,80
10	Ceyhan	7,18	3,90	22,16	5,10
11	Kızılırmak	6,48	3,50	19,55	4,50
12	Sakarya	6,40	3,40	11,34	2,60
13	Çoruh (Chorokhi)	6,30	3,40	22,60	5,20
14	Yeşilırmak	5,80	3,10	18,69	4,30
15	Susurluk	5,43	2,90	10,57	2,40
16	Aras (Arax)	4,63	2,50	13,11	3,00
17	Konya Kapalı (Konya Closed)	4,52	2,40	1,22	0,30
18	Büyük Menderes (Big Menderes)	3,03	1,60	6,26	1,40
19	Kuzey Ege (Northern Aegean)	2,90	1,60	2,88	0,70
20	Van Göl. Kap. (Van Lake)	2,39	1,30	2,60	0,60
21	Gediz	1,95	1,10	3,92	0,90
22	Meriç Ergene	1,33	0,70	1,00	0,20
23	Küçük Menderes (Small Menderes)	1,19	0,60	1,38	0,30
24	Asi (Orontes)	1,17	0,60	4,90	1,10
25	Burdur Gölü (Burdur Lake)	0,50	0,30	0,89	0,20
26	Akarçay	0,49	0,30	0,54	0,10
	Turkey Total	**186,06**		**432,98**	

the characteristics of the streams by using stream-flow rate and energy production of the hydroelectric power plants information. It is planned to make a comparison of the results and discovery of the possible connections between the two. This study is planned to be used for security of the energy supply in long term. Our study differs from the studies done in this field in three points:

1. The related datasets have not been used in any study so far.

2. The studies which have been done so far are not related with the security of Turkey's hydroelectric energy supply topic.

3. There is no study in this field which is carried out by the computer engineers and scientists.

1.3 Thesis Organization

The rest of this thesis is organized as follows.

In Chapter 2, the literature review of basin clustering is presented. Related work in the literature is inspected in five sections, namely, datasets, aims, methods, problems, conclusions and remaining problems. Results are noted to highlight the remaining of this thesis.

In Chapter 3, background of the methods are explained. Algorithms for k-means clustering, hierarchical clustering, dynamic time warping and longest common subsequence are demonstrated and pseudo-codes of these methods are presented. K-means and hierarchical clustering are chosen since they are the most commonly used methods in the related work and dynamic time warping and longest common subsequence are chosen because they are the suggested methods for trend analysis.

In Chapter 4, experiments are done on the stream-flow and hydroelectric energy production dataset. Four methods from the background section are applied to the stream-flow dataset and only structure based hierarchical clustering and longest common subsequence are applied to the hydroelectric power plants with dams in the hydroelectric energy production dataset since these are the methods which provide promising results on stream-flow dataset. Only structure based hierarchical clustering is applied to the run-of-river type hydroelectric power plants in the hydroelectric energy production dataset since this method gives the best results on stream-flow dataset. The results of the experiments are explained in detail in this chapter. The visualization of the experiment results are presented.

In Chapter 5, thesis is concluded and possible future works are addressed for future research.

CHAPTER 2

LITERATURE SURVEY

Hydrologic classification and hydrologic regionalization are two crucial problems in hydrology and there is an important difference between them. Hydrologic classification is to assign most similar streams into same groups based on their flow regime. Hydrologic regionalization,on the other hand, is to group not only gauged streams but also ungauged streams. In hydrological classification, deductive and inductive reasoning are two approaches which are used based on the available data [10]. In case of scarcity of data, deductive reasoning are used and it uses geology, topography and climate data for hydrologic regionalization. It classifies the regions according to environmental similarity. Each class contains streams with similar environmental characteristics. Inductive reasoning approach, on the other hand, uses available or predicted discharge data for streamflow classification. It classifies the studied areas by using hydrologic measurements. Each class consists of streams with similar hydrologic attributes.

2.1 Related Work

2.1.1 Datasets

Review of literature reveals that streamflow attributes are used in many studies and because of available data, inductive reasoning approach is preferred. For example, Bower, Hannah, and McGregor [11] uses 35 UK river basin river flow information together with air temperature and rainfall time-series as dataset. Harris, Gurnell, Hannah, and Petts [12] uses 20-year record of flow and air temperature of 4 rivers

in UK. Hannah, Kansakar, Gerrard and Rees [13] uses monthly runoff data for 28 river basins in Nepal. Gottschalk [14] uses 20 years monthly runoff values from 139 stations and 15 years monthly runoff values from 89 stations. Lins [32] uses 48-year record form 182 gauging stations in USA. Jowett and Duncan [33] uses hydrologic data of 130 river sites records with average of 17.8 years in New Zealand. Kachroo, Mkhandi, and Prida [17] uses annual maximum flood data of 77 stations in Tanzania. Lundager [31] uses 165 stations from Denmark, Finland, Norway and Sweden in total. Krasovskaia, Arnell and Gottschalk [15] uses monthly flow data with at least 10 years of observations. Krasovskaia [16] uses monthly flow series of 49 stations with 66 years of observations. Mkhandi and Kachroo [35] uses data from 754 gauging stations of eleven countries in Southern Africa and average record lengths of stations are 24 years. Stahl [34] uses 612 stations covering most of northern, central and eastern Europe and Spain. Gubareva [37] uses 64 river basins of the Islands of Japan.

The dataset information and number of studies which are done worldwide other than Turkey are summarized in Table 2.1.

Table 2.1: Datasets (Worldwide)

Dataset Characteristics	**Studies**
Monthly River flow information Air temperature Rainfall time-series	[11]
River flow Air temperature	[12]
Monthly streamflow	[13] [14] [15] [16]
Annual maximum flood data	[17]

Similar to the studies from different countries as mentioned in previous paragraph, the studies about the hydrologic classification of Turkey also use streamflow information and prefer inductive reasoning approach. For instance, Kahya, Demirel, and Bég [5], Kahya, Kalaycı, and Piechota [9], Kahya and Demirel [18], Kahya and Kalayci [19], Demirel, Mariano, and Kahya [20], Işık and Singh [21], Işık, Turan, and Doğan [22] and Turan [23] use monthly streamflow information of Turkey. Dikbaş, Fırat, Koç, and Güngör [24], Kahya, Demirel, and Piechota [25] and Özfidaner [26] use annual streamflow data. Demirel [8] and Bayazıt, Cığızoğlu, and Önöz [27] uses both annual

and seasonal streamflow data. Yıldız and Saraç [28] and Cığızoğlu, Bayazıt, and Önöz [29] use daily streamflow information. Yanık [30] uses flow rate information for detection of the hydroelectric potential. The datasets of the studies which are done in Turkey and mentioned in this paragraph are acquired from DSİ and EİE. Moreover, Kahya et al. [5], Demirel [8], Kahya et al. [9], Kahya and Demirel [18], Kahya et al. [25] and Demirel et al. [20] use 80 stations information of 22 basins from total of 26 basins and are excluded 4 basins to satisfy homogenity condition. Yıldız and Saraç [28] uses information of 23 basins. Bayazıt et al. [27] and Cığızoğlu et al. [29] uses information of 24 basins. Kahya and Kalaycı [19], Işık and Singh [21], Işık et al. [22], Turan [23] and Özfidaner [26] use information acquired from all of the basins of Turkey. Besides, Dikbaş et al. [24] uses information from 117 stations but does not mention how many basins are covered in the study and Yanık [30] uses information of only one basin because of regional approach.

The characteristics of the dataset used in the studies which are conducted in Turkey and number of studies which used that information are summarized in Table 2.2.

Table 2.2: Datasets (Turkey)

Dataset Characteristics	**Studies**
Monthly stream-flow	[5] [9] [18] [19] [20] [21] [22] [23]
Annual stream-flow	[24] [25] [26]
Annual stream-flow Seasonal stream-flow	[8] [27]
Daily stream-flow	[28] [29]
Flow-rate	[30]

The source of the datasets included in the studies are General Directorate of State Hydraulic Works (DSİ) and General Directorate of Electrical Power Resources Survey and Development Administration (EİE).

The number of the regions used in the studies are also important together with the number of the studies used them and they are summarized in Table 2.3.

In this thesis, monthly stream-flow information from hydroelectric power plant and gauging stations and hydroelectric energy production information from hydroelec-

Table 2.3: Used regions

Covered Area	Studies
22 basins	[5] [8] [9] [18] [25]
23 basins	[28]
24 basins	[27] [29]
26 basins	[19] [21] [22] [23] [26]
1 basin	[30]
no information	[24]

tric power plants are used and because of available data, both inductive reasoning approach and deductive reasoning approach are preferred.

2.1.2 Aims

In the previous paragraphs, the used datasets in the related studies and their properties are introduced. At that point, it is important to know the aims of these studies to understand the field and the use of datasets in the projects.

Bower et al. [11] uses UK river flow information to generate a regime analysis method and to determine the climatic sensitivity of the river flow regimes. Harris et al. [12] and Hannah et al. [13] aim to classify the river regimes by use of river flow and monthly flow data respectively. Gottschalk [14], Kachroo et al. [17] and Lundager et al. [31] purpose is to hydrologic regionalization by use of mean monthly runoff data, annual maximum flood data and annual flow variability data respectively. Gubareva [37] target is both classification of river basins and hydrologic regionalization and it uses modulus of maximum annual flow data. Lins [32] intent is to identify similarity between streamflow and climatic variables by use of annual streamflow data. Mosley [36] desire is to identify basin with similar hydologic regime by use of annual flood data. Krasovskaia et al. [15] objective is to classify flow regime by use of monthly flow data. Jowett and Duncan [33] plan is to classify river regions and identify basin characteristics by use of daily mean flows. Krasovskaia [16] target is to use river flow regimes as a diagnostic device as the output of climate models by use of monthly flow series. Mkhandi and Kachroo [35] intent is flood frequency analysis by use of annual maximum instantaneous discharge series. Stahl [34] direction is to find the regional

effect of streamflow imperfection and shortage of water supply in Europe by use of daily flow data.

The aim of the studies are summarized in Table 2.4.

Table 2.4: Aims (Worldwide)

Aims	**Studies**
Generate regime analysis method & Determine climatic sensitivity of river flow regimes	[11]
Classify the river regimes	[12] [13] [15]
Hydrologic regionalization	[14] [17] [31]
Identify similarity between streamflow and climatic variables	[32]
Classify river regions Identify basin characteristics	[33]
Classify river regions Identify hydrologic regions	[37]
Identify basin with similar hydrologic regime	[36]
A diagnostic device for the output of climate models	[16]
Flood frequency analysis	[35]
Regional effect of streamflow imperfection & Shortage of water supply in Europe	[34]

Studies which are performed in Turkey have different purposes. Dikbaş et al. [24] aim is that the identification of the hydrologically homogeneous regions and classification of the annual maximum flow. Kahya et al. [25] plans to regionalize annual streamflow pattern of Turkey. Kahya et al. [9] aims to describe hydrologically homogenous regions. Demirel [8] plan is classification of regions which contain similar streamflow patterns. Kahya et al. [5] aims to decide stream-flow zones of Turkey. Işık and Singh [21] aims to associate the 3 regionalization style and figure out the streamflow at ungauged sites. Cığızoğlu et al. [29] plans to identify the trend in maximum, mean and low flows of rivers. Özfidaner [26] aims to analyse statistical trend of monthly and annual precipitation data which were collected from precipitation observation stations in Turkey between 1932 and 2002. Kahya and Demirel [18] aims to classify similar catchments. Işık et al. [22] plans to classify the river basins by use of cluster analysis

based on hydrological homogenity. Turan [23] aims to cluster basins on the basis of hydrometeorological homogeneity. Kahya and Kalaycı [19] aims to characterize the Turkish streamflow data for confirmation of climate change. Demirel et al. [20] plans to identify the regions with similar drought patterns. Bayazıt et al. [27] aims to detect the trends in Turkey's streamflow and precipitation. Yıldız and Saraç [28] aims to find effects of streamflow on HES energy production. Yanık [30] aims to make flow forecast for locations with non existing or incomplete flow data.

Studies which are done in Turkey are summarized in Table 2.5.

Table 2.5: Aims (Turkey)

Aims	**Studies**
Clustering of basins on the basis of hydrometeorological homogeneity	[22] [23]
Identification of hydrologically homogeneous regions & Classification of annual maximum flow	[24]
Identification of hydrologically homogeneous regions	[9]
Regionalization of annual streamflow patterns	[25]
Classification of regions with similar streamflow patterns	[8]
Determination of streamflow zones	[5]
Combination of 3 regionalization technique & Computation of streamflow at ungauged sites	[21]
Trend in maximum, mean and low flows of rivers	[29]
Statistical trend analysis of monthly and annual precipitation data	[26]
Classification of similar catchment areas	[18]
Characteristics of Turkish streamflow data	[19]
Zones with similar drought patterns	[20]
Trend detection in Turkey's streamflow and precipitation	[27]
Streamflow effects on HES energy production	[28]
Flow predictions for locations with non existing or incomplete/inadequate flow data	[30]

In this thesis, our aim is hydrological classification of streams by use of monthly

stream-flow rate information and hourly hydroelectric energy production information.

2.1.3 Methods

The aims and datasets of studies are given in the previous sections. In this section, used methods in the studies will be inspected and summarized to bear a torch for this study.

The most widely used method is hierarchical clustering which is used by Bower et al. [11], Hannah et al. [13], Stahl [34] and Mosley [36]. Bower et al. [11] and Stahl [34] applied Ward method to hierarchical clustering. Hannah et al. [13] applied agglomerative cluster analysis, which is a bottom-up approach, to hierarchical clustering. Mosley [36] benefited from the a cluster analysis program which is called BMDP2M for applying hierarchical clustering method. Bower et al. [11] not only took advantage of hierarchical cluster analysis but also non-hierarchical k-means cluster analysis. Gottschalk [14], Lins [32] and Gubareva [37] applied principal component analysis. Gottschalk [14] and Lins [32] supported the method by pairwise grouping method and visual assessment respectively. Gubareva [37] supported the method by use of the pair group average method, the pair group centroid method, the complete linkage method, the single linkage method and Ward's method. Harris et al. [12] applied average linkage technique. Jowett and Duncan [33] used a Fortran program which is called as TWINSPAN for organizing multivariate data in a structured two-way table as allocation of the aspect and individuals. Kachroo et al. [17] applied a homogeneity test to affirm the homogeneity of the described regions. Krasovskaia [16] applied entropy-based grouping. Olden et al. [10] introduced a methodological framework that illustrate important part of the classification process.

The methods which are used in the studies are summarized in Table 2.6.

Studies which are done in Turkey are also used hierarchical clustering widely. Kahya and Demirel [18], Demirel [8] and Işık et al. [22] applied hierarchical clustering. Kahya and Demirel [18] applied single, complete and Ward linkage criterion in hierarchical clustering. Demirel [8] used quantitative method in hierarchical clustering. Yanık [30] and Işık and Singh [21] applied both hierarchical clustering and

Table 2.6: Methods (Worldwide)

Methods	**Studies**
Hierarchical clustering	[13] [34] [36]
Hierarchical & Nonhierarchical	[11]
Principal component analysis	[14] [32] [37]
Average linkage	[12]
TWINSPAN	[33]
Homogeneity test	[17]
Entropy based grouping	[16]

non-hierarchical k-means clustering. Yanık [30] applied Ward method in hierarchical clustering. Işık and Singh [21] applied Euclidean distance and Ward's algorithm criterion in hierarchical clustering. Turan [23] applied Ward method in hierarchical clustering, k-means in non-hierarchical clustering and hard c-means methods. Dikbaş et al. [24] applied k-means based on L-moments. Kahya et al. [25] applied k-means algorithm using reallocation criteria. Demirel et al. [20] applied both PCA and k-means. Kahya et al. [5] applied Euclidian, squared euclidean and Ward's minimum distance. Kahya and Kalaycı [19] applied Van Belle and Hughes' tests for checking the homogeneity of trends. Kahya et al. [9] applied rotated PCA and annual cycle analysis. Bayazıt et al. [27], Cığızoğlu et al. [29] and Özfidaner [26] applied parametric T test and non-parametric Mann-Kendall test.

The methods used in the studies which are done in Turkey are summarized in Table 2.7.

In this thesis, k-means and hierarchical clustering are used since these two are the most commonly used methods in this chapter. Dynamic time warping and longest common subsequence are used since these two are the methods which are suggested for the trend analysis.

2.1.4 Problems

Problems encountered during the studies are summarized in this section.

Table 2.7: Methods (Turkey)

Methods	**Studies**
Hierarchical clustering	[18] [8] [22]
Hierarchical Nonhierarchical	[30] [21]
Hierarchical Nonhierarchical Hard c-means	[23]
Nonhierarchical (k-means)	[24] [25]
K-means PCA	[20]
Van Belle Hughes' tests	[19]
Rotated PCA Annual cycle analysis	[9]
Parametric T-test Non-parametric Mann-Kendall test	[27] [29] [26]

Hannah et al. [13] has a problem of differences in record length and time in gauging stations. This problem stems from recording of the measurements in different time intervals. For instance, one time series contains 30 years while the other one contains only 10 years measurement. It is mentioned that this problem may cause prejudice. This problem is overcome in our thesis by use of exactly the same time interval for all the time series in the dataset. Mosley et al. [36] pointed out that subjective conclusion is not eliminated by use of cluster analysis. Moreover, it mentioned that uncertainties of the regionalisation may counterbalance and cancel out the gain of statistical virtue of fit.

At that point, inspecting the problems of the studies which are done in Turkey may be useful. Kahya and Demirel [20] pointed out chaining problem. It is mentioned that chaining problem resulted from single linkage and it prevented the use of Cophenet coefficient efficiently. The chaining problem is related with the merging method of the single linkage [38]. Merging method of the single linkage clustering is local. In single linkage, two closest member are grouped without giving attention to overall shape of the clusters and it increases the distance between the cluster [39]. In our thesis, this problem is achieved by use of Ward method, which is known as having

the low tendency to the chaining problem [38]. It also remarked that raw data effect leads the failure of complete linkage. Yıldız and Saraç [28] mentioned that 2 out of 25 river basins are eliminated because of the problematic data. Işık and Singh [21] indicated that the result of the flow duration curves had been successful in small homogeneous regions but the result were not successful in large non-homogeneous regions. Moreover, Yanık [30] pointed one more problem about flow duration curves as single linkage, median linkage, centroid linkage, average linkage, and weighted average linkage methods within other nonhierarchical methods can not be applied the flow duration curves. Demirel [8] pointed out that current climate regions did not overlap with the streamflow regions which is generated by the study.

The problems faced in these studies are used in our thesis to understand and detect possible problems and find appropriate solutions to them. The prejudice of the results because of different record interval of the measurement are handled in our thesis by use of exactly same time interval for all the time series in the dataset. Chaining problem is handled by use of ward as a linkage method in our thesis.

2.1.5 Achievements in the Previous Studies

Datasets, aims, problems and methods of the previous studies are mentioned in the previous sections. It is important to explain the results of these studies to understand their success. Therefore, the achievements of the previous studies are summarized in this section.

Bower et al. [11] concluded that if the hydro-climatology of the environment is known then the regime classification and novel sensitivity index are adequate methods. Harris et al. [12] analysis shows that some special annual flow and temperature sequences controls the conflict between flow and temperature regimes. Hannah et al. [13] concluded that the applied classification method is convenient tool for identification of the basic spacial structure of annual flow regime shape, such as timing of peak, and magnitude in an extreme physical environment where regional hydrological patterns are complicated and not well known up to the present. Gottschalk [14] claimed that definition of the hydrological regions and determination of the spatial scales of variation can be done by the areal classification tool which is presented in the paper. Lins

[32] mentioned that the identification of five statistically significant modes of variation of annual stream-flow of United states is done. Mosley [36] concluded that four regions are found for South Island but any region are found for North Island in New Zealand. Jowett and Duncan [33] found that six groups were classified based on the flow variability by use of 130 sites in New Zealand. Kachroo et al. [17] claimed that the procedure of homogeneous regions definition is proved effective by use of geographical information constituting mean annual rainfall, major basin boundaries and topography. Lundager et al. [31] found that simple hydrological regionalization which is described in the paper can be used in practical hydrology such as within network planning and generalization of conclusions from representative and experimental basins. Krasovskaia et al. [15] concluded that 13 regime types are found in total and 4 of them are identified as transitive in northern and western Europe. Krasovskaia [16] claimed that utilization of river flow regimes as a diagnostic tool for climate model and also in river flow sensitivity researches can be done by use of the concept of the entropy. Mkhandi and Kachroo [35] concluded that the most of the recommended regions satisfied the homogeneity test which is applied in the study. Stahl [34] found that drought has effects on several regions in Europe. Gubareva [37] concluded that 2 hydrological regions are identified in Japan.

The achievements of the previous studies which are done in Turkey may be useful in identifying the success of the studies. Kahya and Demirel [18] had 3 results in the paper. First one is that climatology and river basin characteristics have to be used for better clustering. Second one is that standardization is necessary to acquire equally weighted clusters and third one is that Ward has a better performance than single and complete linkage. All these three suggestions and results are used in our thesis. Dikbaş et al. [24] found that classification of the annual maximum flows and description of hydrologically homogeneous regions can be successfully done by use of k-means method. Kahya et al. [5] found 6 cluster for each month as a result of the study. Kahya et al. [25] concluded that spatial variability of homogeneous streamflow regions can be showed by use of 8 cluster level. Kahya and Kalaycı [19] detected presence of linear trend in monthly mean streamflow data of Turkey. Demirel et al. [20] concluded that clustering strategy is failed when PCA is practised for describing drought zones of Turkey. Kahya et al. [9] had 2 findings in the paper.

First one is that the three approaches applied in the study for identifying homogeneous streamflow regions have similar results. Second one is that homogeneous streamflow regions and climate zones of Turkey are found similar with respect to geographical extent. Bayazıt et al. [27] concluded that mean streamflow, minimum streamflow and floods decrease in Trakya, West, South and Middle part of the Turkey. Yıldız and Saraç [28] detected trends in the most of the rivers in Marmara, Aegean, Inner Anatolia (Sakarya basin included) and Mediterranean regions. Işık and Singh [21] found 6 homogeneous regions and concluded that non-hierarchical k-means method is better than hierarchical method in homogeneous regions. Işık et al. [22] found 6 homogeneous regions as well and note the similarity between rainfall and water yield distribution. Cığızoğlu et al. [29] found a serious decrease in the mean and low flows in western, central and southern parts of Turkey. Demirel [8] concluded that climate zones of Turkey which is redefined by applying cluster analysis to total precipitation data is consistent with streamflow regions and homogeneous streamflow regions of Turkey is described by applying principal component analysis method. Result of Yanık [30] is that cluster analysis methods can be used to determine regional flow duration curves. Özfidaner [26] had 3 results in the paper. First one is that the precipitation data has a decreasing trend in winter in all seven regions of Turkey. The second one is that the precipitation has an increasing trend in the summer, spring and autumn in all seven regions. The third one is that the trend of precipitation data does not affect stream-flows except for South Eastern Anatolia zone. Turan [23] divided Turkey river basins into 6 homogeneous regions and found the yield and rainfall distribution very similar to these 6 homogeneous regions.

The results of the studies in this section will be used for comparison with our study.

2.1.6 Remaining Problems

The suggested future work by the previous studies are summarized below.

Harris et al. [12] suggests that proper time scale is necessary for illustrating benchmark regimes for inspecting ecosystem dynamics, for checking human effects, and for collecting practical tools for water resources management and this can be achieved by use of larger dataset, which climatologists typically use 30 years data, and more sta-

tions. Moreover, it advised that ecological and climatological records, which contains isolated flow and temperature respectively, can be combined in case of a research about the explicit ecological impacts and the hydroclimatic imposing mechanisms that generate the different types, sequences and combinations of flow and temperature regimes. Gottschalk [14] specified the difficulty of figuring out about the way physio-graphical and hydrological model regions relations. Its suggestions for the future studies is to identification of such relations. Furthermore, it is pointed out that systematic reasoning of the spatial variation arrangement of hydrological variables is necessary for developing such relations. Lundager et al. [31] indicated that crucial tasks of characterizing homogeneous hydrological regions remained unsolved and suggested that exploration of new methods was needed for more objective methods of classification and regionalization. Krasovskaia et al. [15] indicated that more analysis are needed both for the last formulation of the discriminating principle for both the different regime types and the advancement of the complete interpolation routines for the all 16 countries in northern and western Europe in FRIEND. Stahl [34] advised the researchers that the future studies should be focused on especially the interaction of seasonal changes for a better understanding and in order to elaborate acceptable strategies to prohibit and mitigate unwanted outcomes. Olden et al. [10] expected that researchers and managers would be more informed when having to make choices about the selection and accurate implementation of methods for hydrologic classification in future.

Recommendations of the studies which are done in Turkey are also important and some of them are explored in this thesis. Kahya et al. [25] suggested PCA and K-means coupling for future similar grouping projects. Kahya and Kalaycı [19] advises to check whether relations exists between trends in 3 climatologic variables of Turkey mentioned in the study, by using neural network model. Demirel et al. [20] explained the future plan of the study as the confirmation of cluster analysis of short-term intermittent flow prediction models.

The future work summarized in this section reveals the potential investigation questions and brightens the unsolved issues on the topic.

In our study, k-means is applied as suggested in this section.

CHAPTER 3

BACKGROUND

This chapter presents information about the methodologies which are used in the implementation of the clustering of the Turkey stream-flow and hydroelectric energy production data. K-means clustering, hierarchical clustering, dynamic time warping clustering and longest common subsequence clustering are explained in detail in this chapter.

3.1 K-means Clustering

K-means clustering is one of the popular methods used in the cluster analysis. K-means is an unsupervised clustering algorithm. "K" in the algorithm name defines the number of the intended clusters in the dataset. The observations in the dataset are clustered to k clusters as a result of algorithm. First of all, k sample from the dataset are chosen randomly. They are intended to be as much as far away from each other and they are marked as clusters' center. After that, remaining samples are assigned nearest cluster according to a distance function. Center of each clusters is recalculated. The new centers are identified and procedure repeats itself. The procedure continues until reaching a stable point in which cluster of the samples do not change any more. The biggest problem of this clustering approach is to decide the "k" value. Generally pre-processing is done on the dataset to decide an appropriate "k" number. The pseudo-code of the k-means algorithm can be seen in Figure 3.1.

input : $S = \{s_1, ..., s_n\}$ (n number of samples)
output : $C = \{c_1, ..., c_k\}$ (n number of k clustered samples)
choose k random sample as cluster centroids
 # loop continues until cluster centroids do not change and samples of the clusters do not change
while cluster centroids not change $\&\&$ clusters' samples not change **do**

for {i=0;i<n;n++} **do**
temporaryDistance = ∞
for {j=0;j<k;j++} **do**

if temporaryDistance > distanceFunction(s_i, c_j) **then**
temporaryDistance = distanceFunction(s_i, c_j)
cluster(s_i) = j
end if
end for
recalculate cluster centers by use of average distance of assigned samples to the each cluster
end for
end while

Figure 3.1: K-means Algorithm Pseudo-code

input : $S = \{s_1, ..., s_n\}$ (n number of samples) and $dist(c_i, c_j)$ (distance function)
output : $C = \{c_1 \cup ... \cup c_n\}$ (all samples are belong to same cluster)
for i=1 to n **do**
  # each sample is a cluster
  $c_i = \{s_i\}$
end for
 # there are n number of samples and n number of cluster
$C = \{c_1, ..., c_n\}$
 # loop continues until all the samples are grouped in the same cluster
while size(C) > 1 **do**
  Find most similar c_i and c_j according to $dist(c_i, c_j)$
  Remove c_i and c_j from C
  Add $c_i \cup c_j$ to C
end while

Figure 3.2: Hierarchical Agglomerative Algorithm Pseudo-code [41]

3.2 Hierarchical Clustering

Hierarchical clustering is one of the another widely used clustering technique. It is unsupervised clustering algorithm. The method works by grouping the samples into the tree of clusters [40]. It uses distance functions as a clustering criteria and it requires a termination condition. It can be applied by use of two strategies, namely, agglomerative and divisive. Agglomerative strategy is a bottom-up approach. Similar elements are clustered until all the samples are assigned one of the cluster. Divisive strategy, on the other hand, is a top-down approach. All samples are in the same cluster initially and dissimilar samples are assigned to different clusters. The results of the hierarchical clustering are visualized by dendograms. The pseudo-code of the agglomerative hierarchical clustering algorithm is given in the Figure 3.2.

Distance between clusters are generally described by three ways, namely, single linkage, complete linkage and average linkage.

3.2.1 Single Linkage

Single linkage is described as the distance between two closest member of the clusters. It is also known as nearest neighbour. Single linkage distance between two

cluster can be calculated as follows:

$$D(X,Y) = \min_{x \in X, y \in Y} d(x,y) \tag{3.1}$$

where

X is a cluster

Y is a cluster

$D(X,Y)$ is the distance between these two cluster

3.2.2 Complete Linkage

Complete linkage is described as the distance between two farthest member of the clusters. It is also known as farthest neighbour. Complete linkage distance between two cluster can be calculated as follows:

$$D(X,Y) = \max_{x \in X, y \in Y} d(x,y) \tag{3.2}$$

where

X is a cluster

Y is a cluster

$D(X,Y)$ is the distance between these two cluster

3.2.3 Average Linkage

Average linkage is described as the average distance between all member of two clusters. Average linkage distance between two cluster can be calculated as follows:

$$D(X,Y) = \frac{1}{|X||Y|} \sum_{x \in X} \sum_{y \in Y} d(x,y) \tag{3.3}$$

where

X is a cluster

Y is a cluster

$D(X,Y)$ is the distance between these two cluster

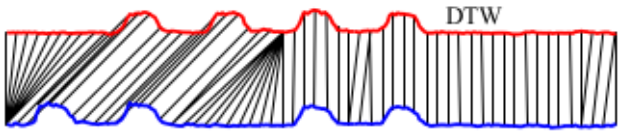

Figure 3.3: Dynamic Time Warping of two time series [50]

3.3 Dynamic Time Warping

In time series analysis, dynamic time warping is defined as an algorithm to identify the similarity between two time series [42]. Dynamic time warping is used in various areas for different aims. For instance, it is used in hand sign and gesture recognition [43] [44], in chemical engineering [45], in signal processing [46] and in data mining classification and clustering [47] [48]. One of the important feature of the dynamic time warping is that time series are warped nonlinearly in this method [49] and time shifts are handled. This feature can be seen in Figure 3.3.

Dynamic time warping algorithm is applied in two phases. The cost matrix of the time series is calculated in the first phase and optimal warping path is found in the second phase. Cost matrix calculation is described in Figure 3.4 and optimal time warping algorithm is explained in the Section 3.5.

3.4 Longest Common Subsequence

Longest common subsequence is another method which is especially used in biology and genetics for identification of the resemblance of the genetic sequences. The algorithm finds the common parts between two sequences. The pseudo-code of the algorithm can be seen in Figure 3.6.

In this study, k-means clustering, shape based and structure based hierarchical clustering, dynamic time warping clustering and longest common subsequence clustering are explained and these methods are used in the experiments presented in Chapter 4.

 # cost matrix is calculated by use of pairwise distances between X and Y
input : C (cost matrix), X (first time series) and Y (second time series)
output : dtw
$n \leftarrow |X|$
$m \leftarrow |Y|$
$dtw[] \leftarrow new[n \times m]$
#first element of the DTW matrix is set to 0
$dtw(0,0) \leftarrow 0$
 # calculate dtw values of first column by use of cost matrix
for i=1 to n **do**
  $dtw(i,1) \leftarrow dtw(i-1,1) + c(i,1)$
end for
 # calculate dtw values of first row by use of cost matrix
for j=1 to m **do**
  $dtw(1,j) \leftarrow dtw(1,j-1) + c(1,j)$
end for
 # calculate dtw values of all rows and columns by summing minimum of the left,upper and diagonal neighbour of the matrix element with its projection in the cost matrix
for i=1 to n **do**
  for j=1 to m **do**
    $dtw(i,j) \leftarrow c(i,j) + min\{dtw(i-1,j); dtw(i,j-1); dtw(i-1,j-1);\}$
  end for
end for
return dtw

Figure 3.4: Cost Matrix Algorithm Pseudo-code [51]

input : dtw
output : $path$
$path[] \leftarrow newarray$
 # assign row length to i
$i = rows(dtw)$
 # assign column length to j
$j = columns(dtw)$
 # loop until reaching dtw(1,1)
while $(i > 1)\&\&(j > 1)$ **do**
  if $i == 1$ **then**
    $j = j - 1$
  else if $j == 1$ **then**
    $i == i - 1$
  else

    # backtraking from last element to first element
    if $dtw(i-1, j) == min\{dtw(i-1, j); dtw(i, j-1); dtw(i-1, j-1);\}$ **then**
      $i == i - 1$
    else if $dtw(i, j-1) == min\{dtw(i-1, j); dtw(i, j-1); dtw(i-1, j-1);\}$
    then
      $j = j - 1$
    else
      $i = i - 1; j = j - 1$
    end if
    $path.add((i, j))$
  end if
end while
 # return optimal warping path matrix between X and Y
return $path$

Figure 3.5: Optimal Warping Path Algorithm Pseudo-code [51]

input : $X = \{x_1, ..., x_n\}$ (first sequence) and $Y = \{y_1, ..., y_m\}$ (second sequence)
output : $C[m, n]$ (length of the longest common subsequence of X and Y)
C = array(0..m,0..n)
 # set first column to 0
for i=0 to m **do**
  $C[i, 0] = 0$
end for
 # set first row to 0
for j=0 to n **do**
  $C[0, j] = 0$
end for
 # from beginning to end of the X and Y sequence
for i=1 to m **do**
  for j=1 to n **do**
    # if X and Y are matched then
    if X[i] == Y[j] **then**
      # add 1 to upper left diagonal value of the C matrix element and assign to C element
      $C[i, j] = C[i - 1, j - 1] + 1$
    else
      # assign C element maximum of its left and upper neighbour values
      $C[i, j] = max(C[i, j - 1], C[i - 1, j])$
    end if
  end for
end for
 # return length of the longest common subsequence of X and Y
return C[m,n]

Figure 3.6: Longest Common Subsequence Algorithm Pseudo-code [52]

CHAPTER 4

EXPERIMENTAL RESULTS

R [54] is used in the experiments. R is an interpreted programming language and it is a software platform for statistical computing and graphics. R objects can be directly manipulated by the use of different programming languages like Java, C, Python. Moreover, C, C++ and Fortran programs can be linked and called at run time. It is commonly used by data miners and statisticians for data analysis and statistics. R has an extensible object system which consists of time-series and its libraries carried out statistical and graphical techniques, including time series analysis [53].

The plot of the flow rate of the streams in 26 dams for 12 years period can be seen in Figure 4.1. This figure shows flow rate of 26 dams as time series from 2002 to 2014. It is mentioned in Chapter 2 that recording of the measurements in different time interval causes biased results. The same time interval is used in our thesis to get rid of this problem.

The comparison of the dams in Fırat basins can be seen in Figure 4.2.

In Figure 4.1, it is seen that different trends exist in the streams where the measurements are done. In Figure 4.2, dams are in the same basin but their stream-flows are different and it is seen that the trends are similar. The comparison of the other dams in the same basins are in Appendix A and the similarity of the trends can be seen in these figures. These figures prove that dams in the same basin but in different stream-flows has a similar trend. The aim of our thesis is to identify the similarity of the stream-flow trends between basins.

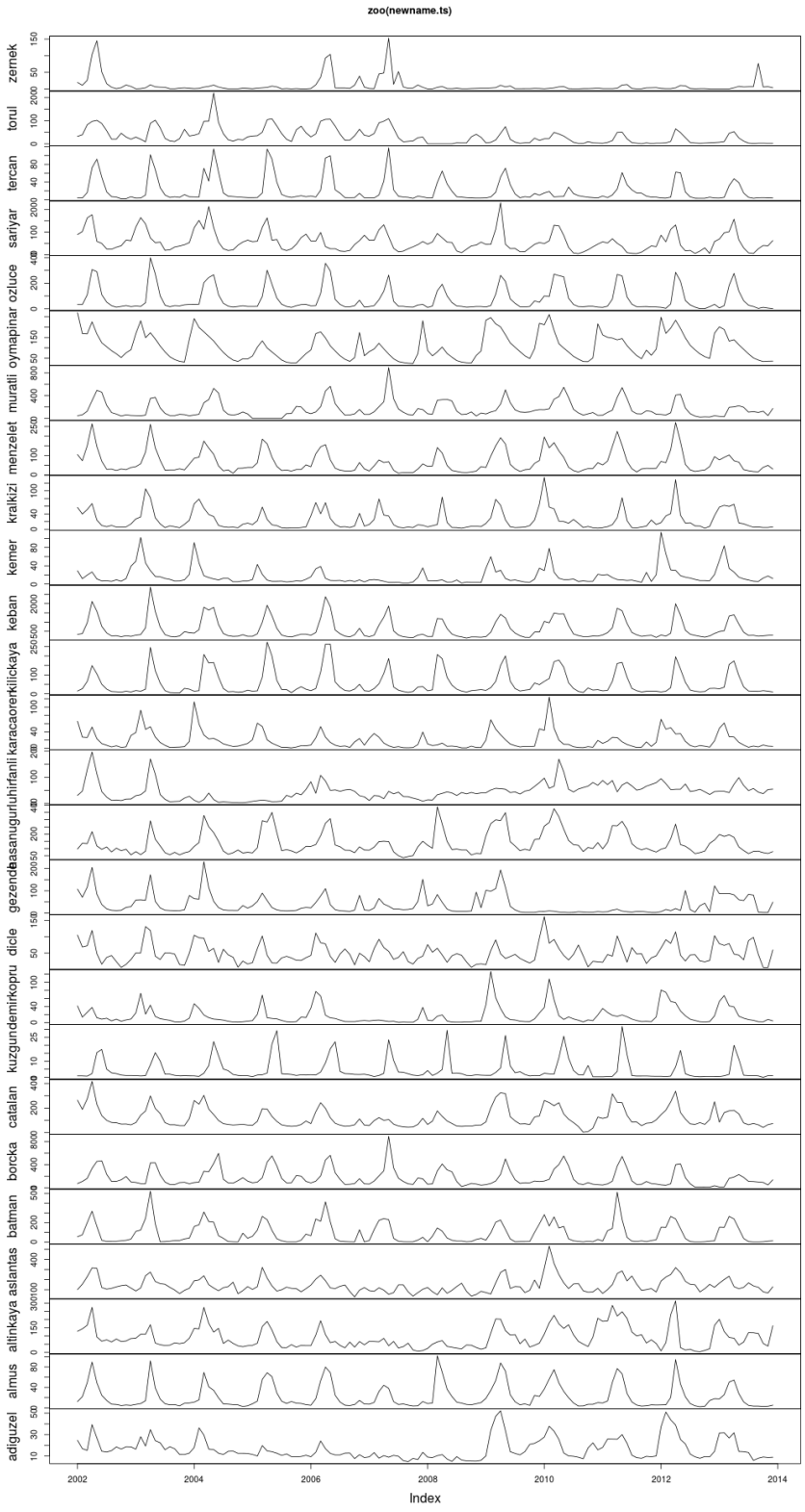

Figure 4.1: Plot of Flow Rate vs. Years

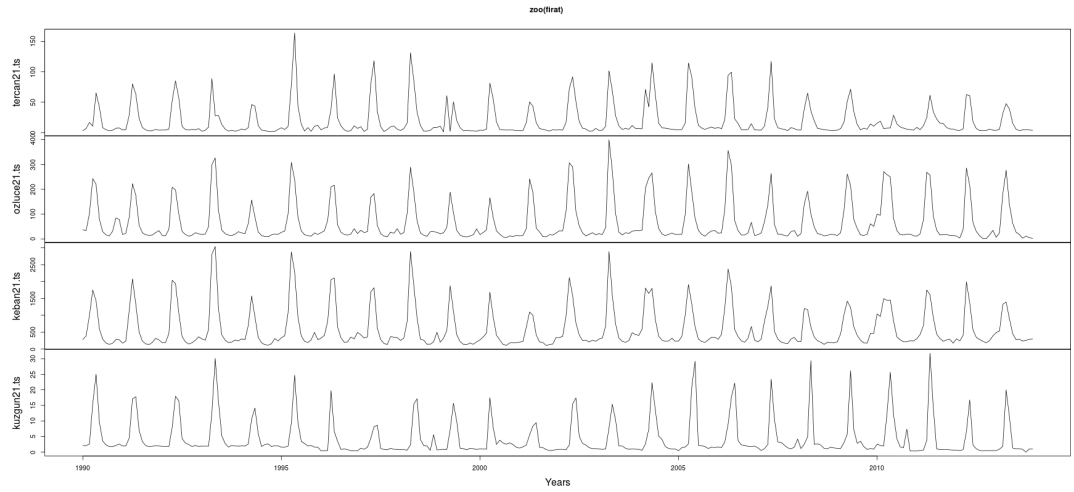

Figure 4.2: Plot of Flow Rate vs. Years in Fırat Basin

4.1 Clustering with K-means

12 years time series is shown in the Figure 4.1 since it is the shortest time series between all stream records and this shortest time series belongs to Torul dam. The reason of adding the flow-rate of the streams in the same basins is to show the similarity between them. The resemblance of the flow-rate of the streams in the same basins can be assessed visually.

K-means clustering is applied to the time series data which is used in this project. The result can be seen in Figure 4.3. K-means clustering is executed 20 times by setting k to 2, 3, 4, 5 and 6. Therefore, algorithm is executed 100 times. In each run, k centroids are chosen randomly. For each k value, distances between the centroids of the clusters are maximized and best solution is found according to this criteria. In other words, distances between the cluster centroids are calculated for each k value 20 times and best solution is marked as the solution in which cluster centroids are furtherest to each other. Best k value is chosen after running 100 times the function and is set to the best partition of the clusters according to cluster centroids.

K is set 4 as a result of the best partition. The best partition which is 4 is marked with a black dot in Figure 4.3. The clusters of each stream in the dataset by setting k to 4 can be seen in Table 4.1

Table 4.1 shows the k-means clustering results. Several conclusions can be drawn

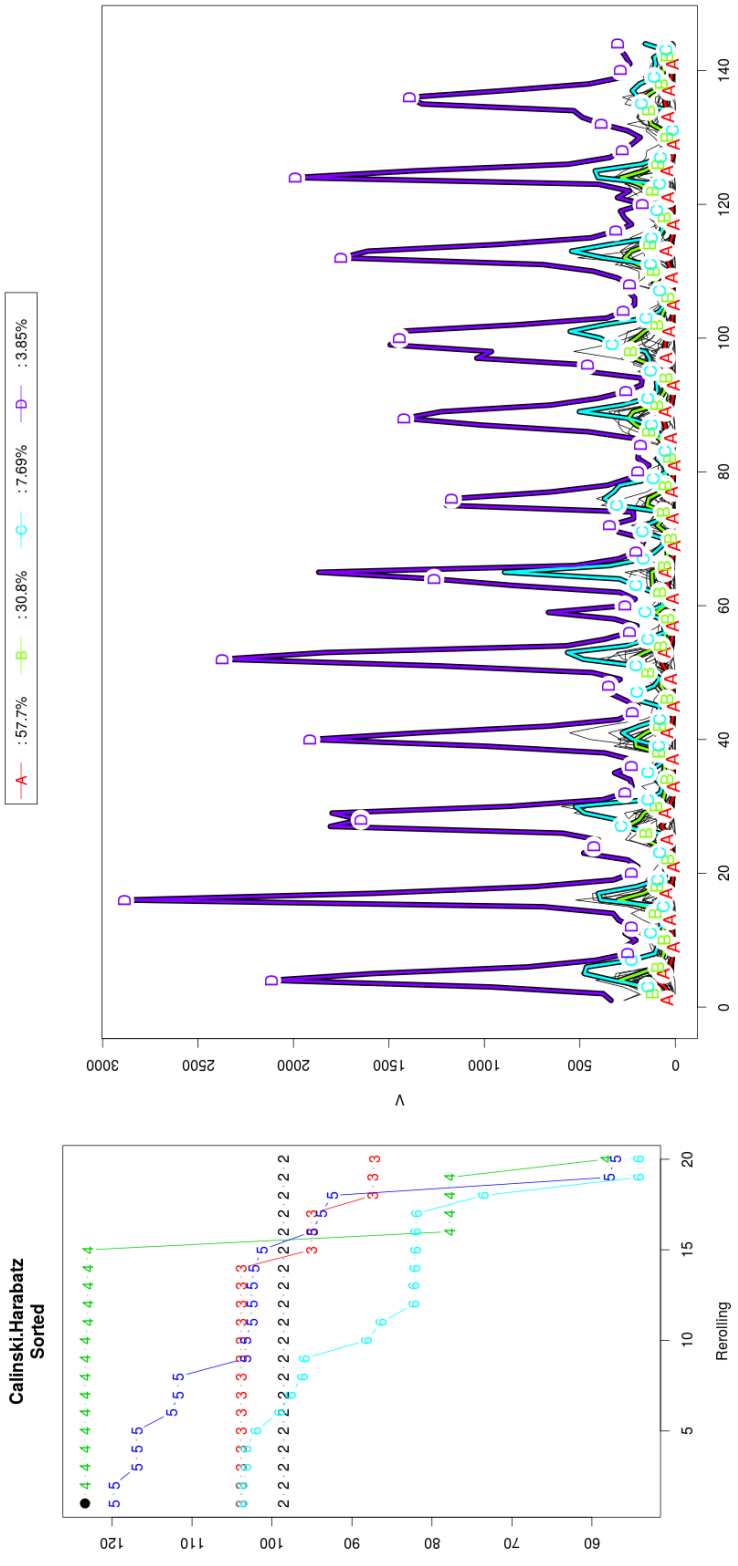

Figure 4.3: K-means clustering of stream-flow values

Table 4.1: K-means Clustering of the Stream-flows

Stream-flow Name	**Basin Name**	**Cluster**
Demirköprü	5	A
Adıgüzel	7	A
Kemer	7	A
Karacaören	9	A
Sarıyar	12	A
Almus	14	A
Kılıçkaya	14	A
Hirfanlı	15	A
Gezende	17	A
Tercan	21	A
Kuzgun	21	A
Torul	22	A
Zernek	25	A
Kralkızı	26	A
Dicle	26	A
Oymapınar	9	B
Hasan Uğurlu	14	B
Altınkaya	15	B
Çatalan	18	B
Menzelet	20	B
Aslantaş	20	B
Özlüce	21	B
Batman	26	B
Borçka	23	C
Muratlı	23	C
Keban	21	D

from the table. It can be seen that Muratlı and Borçka are clustered correctly as cluster C. Keban is assigned to cluster D because of the its high stream-flow rate. Clustering of Keban points a problem of the k-means clustering in time series. K-means method clusters time series based on the stream-flow rate values rather than the trend of the them. At that point, logarithms of the stream-flows are also grouped to eliminate the effect of the high stream-flows values.

The flow-rates of the streams depend on the geographic structure of the streams such as height. Therefore, logarithmic values of the flow-rates is also grouped by use of

k-means. The Figure 4.4 shows the result of the k-means clustering of the flow-rates of the streams.

K-means clustering is executed 20 times by setting k to 2, 3, 4, 5 and 6 for logarithmic values of the time series. Therefore, algorithm is executed 100 times also for logarithmic values. In each run, k centroids are chosen randomly. For each k value, distances between the centroids of the clusters are maximized and best solution is found according to this criteria. Best k value is chosen after running 100 times the function and is set to the best partition of the clusters according to cluster centroids. K is set 2 as a result of the best partition of the logarithmic values. The best partition which is 2 for the logarithmic values is marked with a black dot in Figure 4.4. The clusters of the each stream in the dataset can be seen in Table 4.2

Several conclusion can be drawn from the K-means clustering of the log values of stream-flow. It can be seen that some of the dams which are in the same basin are correctly clustered like 7,15 and 23 and some of the dams in the same basin are assigned different clusters like 14, 21 and 26. Keban is clustered as alone when using the raw values of stream-flow. By use of log values, Keban stream-flow are decreased and it is assigned to cluster A.

Stream-flow values are inspected in monthly resolution until now. In Chapter 2, it is pointed that some of the previous researches used dataset with yearly resolution. Therefore, the mean of the years are also inspected in this study. Trends of the yearly stream-flow rate can be seen in Figure 4.5.

K-means clustering of the streamflow in yearly resolution can be seen in Figure 4.6.

K is set to 6 for k-means clustering of yearly mean stream-flow values because of the best partition results of the k-means clustering function and it is marked with a black dot as can seen in Figure 4.6. The clusters of the each stream in the dataset can be seen in Table 4.3

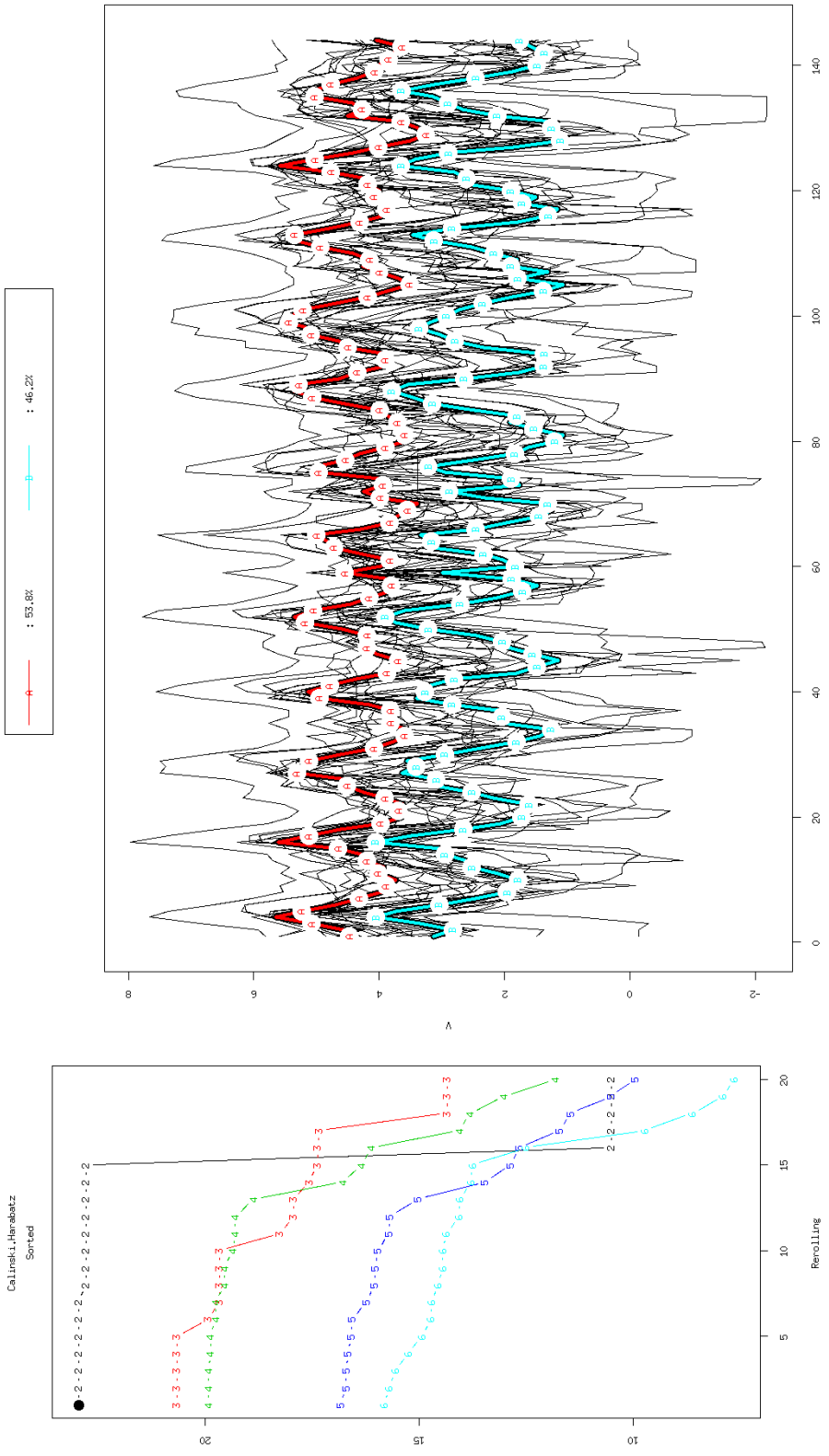

Figure 4.4: K-means clustering of stream-flow log values

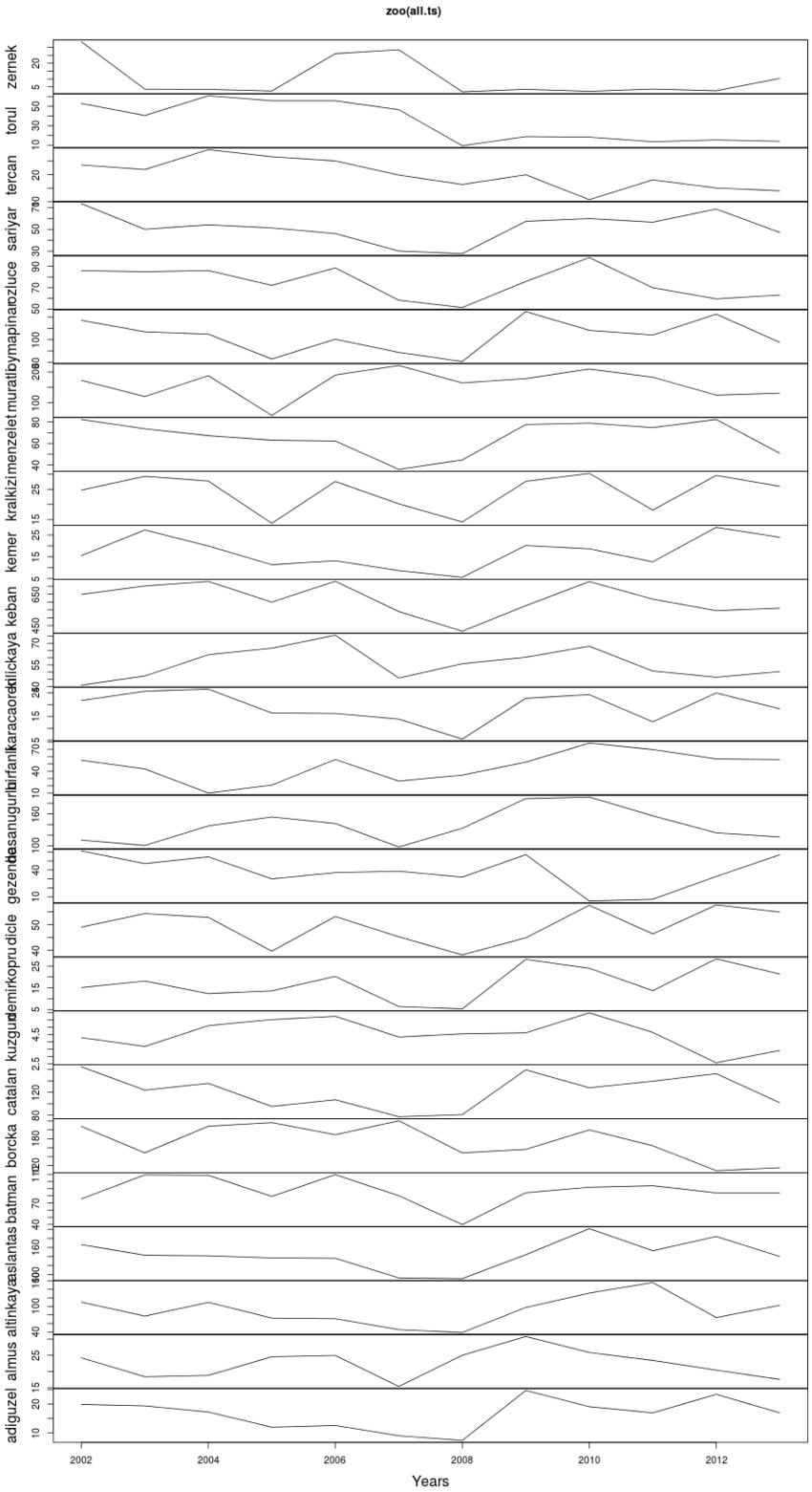

Figure 4.5: Plot of Flow Rate vs. Years in yearly resolution

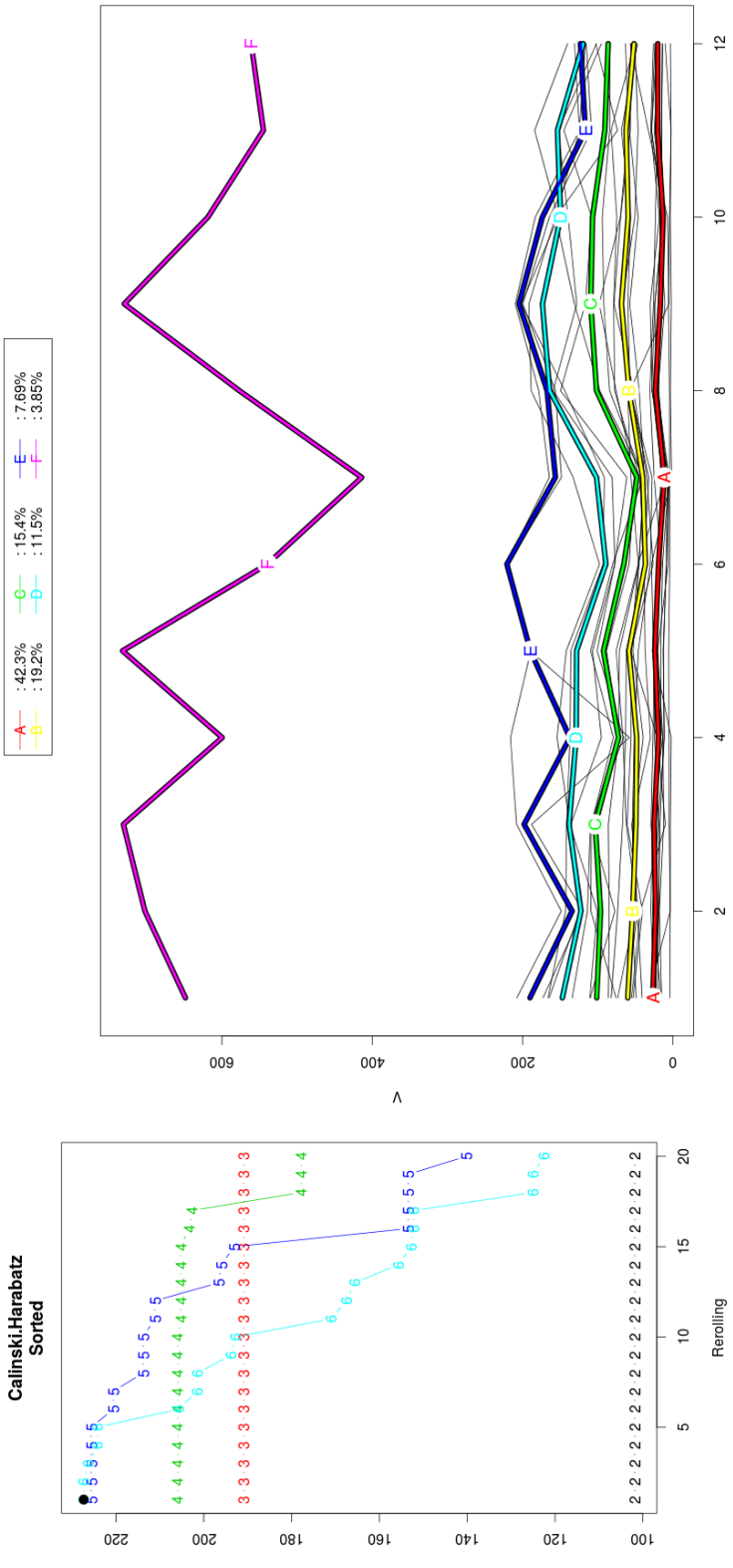

Figure 4.6: K-means clustering of yearly stream-flow values

Table 4.2: K-means Clustering of the Stream-flows log values for 2 cluster solution

Stream-flow Name	**Basin Name**	**Cluster**
Oymapınar	9	A
Sarıyar	12	A
Hasan Uğurlu	14	A
Kılıçkaya	14	A
Altınkaya	15	A
Hirfanlı	15	A
Çatalan	18	A
Aslantaş	20	A
Menzelet	20	A
Özlüce	21	A
Keban	21	A
Muratlı	23	A
Borçka	23	A
Dicle	26	A
Demirköprü	5	B
Kemer	7	B
Adıgüzel	7	B
Karacaören	9	B
Almus	14	B
Gezende	17	B
Kuzgun	21	B
Tercan	21	B
Torul	22	B
Zernek	25	B
Kralkızı	26	B
Batman	26	B

4.1.1 Discussion

In Figure 4.3, it can be seen that k is set to 4 because it gives the best partition result in which cluster centroids are furtherest to each other. However, Keban is assigned to D cluster as alone. This result shows that the high stream-flow rate values of the Keban dam causes such a partition.

At that point, it is important to get rid of this problem to make accurate inferences about the clustering results. For that reason, logarithmic values of the stream-flow

Table 4.3: K-means Clustering of the Stream-flows for 6 cluster solution

Stream-flow Name	**Basin Name**	**Cluster**
Demirköprü	5	A
Adıgüzel	7	A
Kemer	7	A
Karacaören	9	A
Almus	14	A
Gezende	17	A
Tercan	21	A
Kuzgun	21	A
Torul	22	A
Zernek	25	A
Kralkızı	26	A
Sarıyar	12	B
Kılıçkaya	14	B
Hirfanlı	15	B
Menzelet	20	B
Dicle	26	B
Oymapınar	9	C
Altınkaya	15	C
Özlüce	21	C
Batman	26	C
Hasan Uğurlu	14	D
Çatalan	18	D
Aslantaş	20	D
Muratlı	23	E
Borçka	23	E
Keban	21	F

dataset are also used and k is set to 2 as a result of the k-means clustering function. Keban is assigned to cluster A together with other 13 dams and B cluster has 12 dams. This result shows that clustering of the dams are done in a homogeneous way since the number of dams in the clusters are nearly same. However, it can be seen in Table 4.2 that dams in the same basins are assigned to different clusters in Orta Akdeniz (9), Yeşilırmak(14), Fırat(21) and Dicle(26). This result shows that although the problem of alone clustering of Keban is solved, the method failed when applied to the whole dataset.

In Figure 4.6, the result shows that Muratlı and Borçka are assigned to the same group in clustering of the yearly stream-flow data. Keban is still clustered as alone which shows the failure of the use of yearly stream-flow rate data. It can be concluded that an appropriate solution to the problem is not found by the use of k-means method.

4.2 Clustering with Hierarchical Clustering

4.2.1 Stream-Flow Dataset

One of the other popular method is hierarchical clustering. It is the most commonly used methods both in the national and international studies as mentioned in Chapter 2. Choosing appropriate dissimilarity concept is important to get better results in this method. There are two types of dissimilarity concepts, namely, shape-based and structure-based. The purpose of shape-based clustering is to analyse the geometric profile of the series and the purpose of the structure-based clustering is to detect underlying dependence structures [55]. These two concepts are very crucial and should be understood very well for choosing appropriate method for dissimilarity measure.

"Euclidean" is applied for shape-based clustering and "correlation" is applied for structure-based clustering. Two methods, namely, euclidean and correlation are used to calculate the distance matrix in this section to check the results of both the shape-based and structure-based clustering. After the generation of the distance matrix, each objects in the distance matrix is linked by use of a linkage method. Dendogram is generated after all the objects are linked. There are several different linkage methods in the literature and some of them are explained in Chapter 3 in detail. Figure 4.7 shows hierarchical clustering by use of euclidean as a dissimilarity metric and single method as a linkage method.

Keban is clustered alone by use of this method. The reason is the high values of Keban as shown also in k-means clustering. One of the possible methods to get rid of these high values is to use the logarithmic values of the samples in the dataset. Same method is applied to the log values of the dataset by the use of same dissimilarity metric and linkage method. The result can be seen in Figure 4.8.

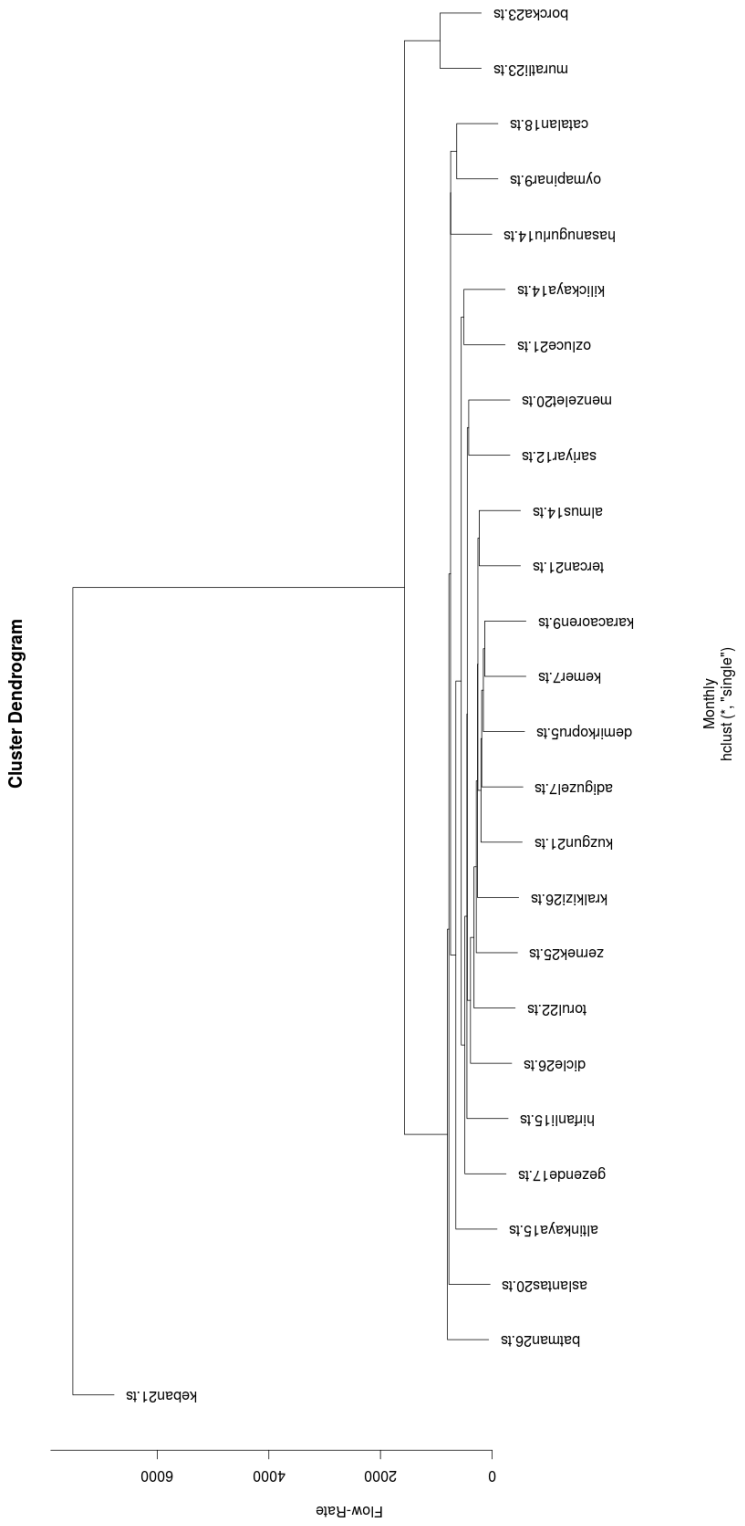

Figure 4.7: Hierarchical clustering (Euclidean) of stream-flow dataset (Monthly)

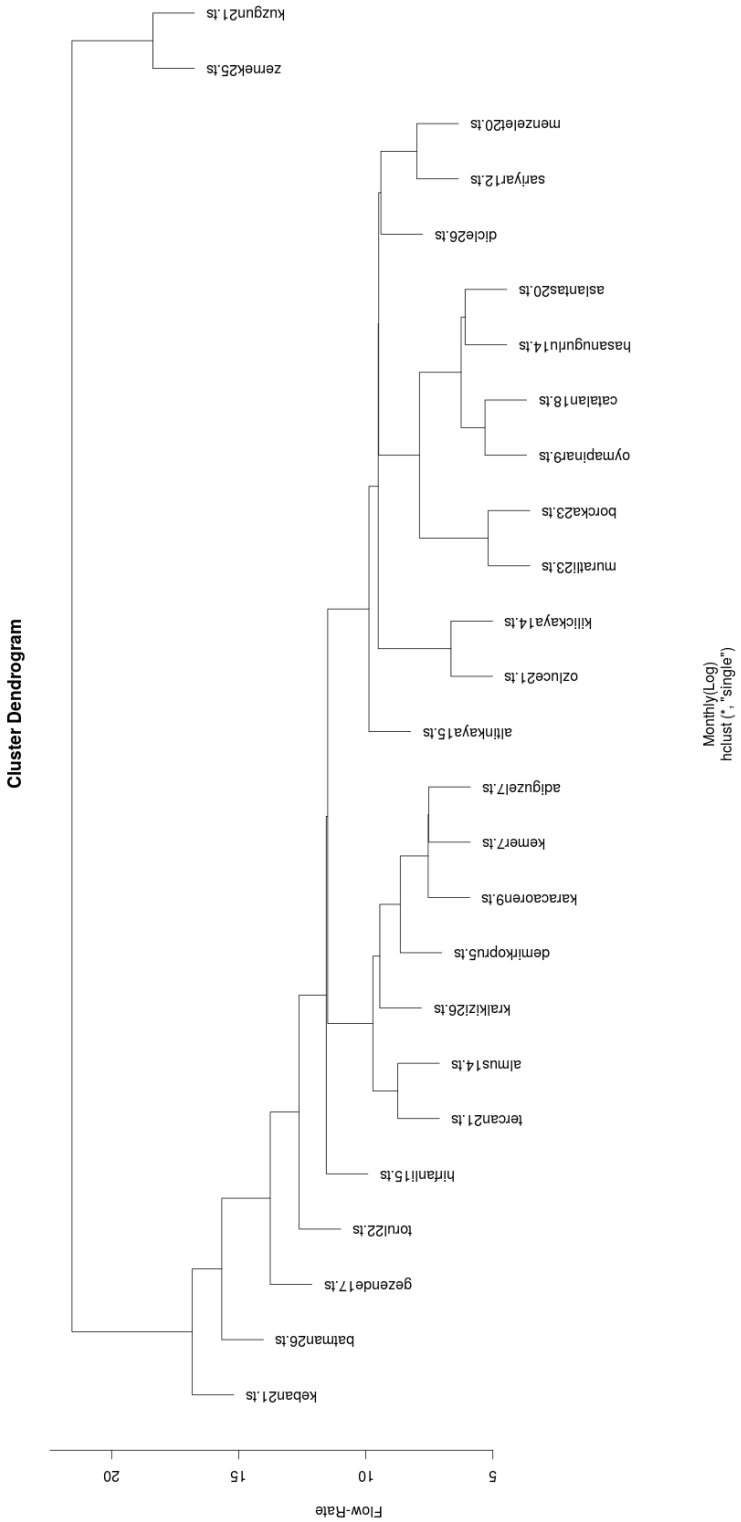

Figure 4.8: Hierarchical clustering (Euclidean) of stream-flow dataset (Monthly-Log)

Figure 4.8 shows that logarithmic values of the dataset decrease the high values of Keban. However, it is still linked to the other objects in the dendogram lastly in the leftmost of the dendogram.

In the previous section k-means clustering is applied to the dataset and in this section hierarchical clustering is applied to the same data set. Clustering results of these two methods are evaluated. Evaluation results can be seen in Table 4.4. Results can be between 1.0 and 0. Higher values indicate a high resemblance between two clustering results since 1.0 shows that clustering results of the two methods are identical.

Table 4.4: Cluster Evaluation

First Method	**Second Method**	**Evaluation Results**
K-means (Monthly)	Hierarchical Clustering (Euclidean)(Monthly)	0.9543651
K-means (Monthly-Log)	Hierarchical Clustering (Euclidean)(Monthly)	0.5555258
K-means (Yearly)	Hierarchical Clustering (Euclidean)(Monthly)	0.7249389

The evaluation result of the k-means and euclidean based hierarchical clustering is 0.95 and this result shows that 95% of the clustering results of the both methods are same and it can concluded that the results are nearly identical.

Figure 4.9 shows hierarchical clustering by use of correlation as a dissimilarity metric and single method as a linkage method.

Single linkage, which is also known as nearest neighbour, is used as the linkage method in the Figure 4.9 and the branches of the dendogram in Figure 4.9 is too close to each other and it makes hard to inspect it. The reason of this problem is grouping method of the single linkage. Single linkage merge principle is local [39] and two nearest samples are grouped together in this method. It does not take care of the dendogram output. For that reason, it can be hard to inspect the resulted dendogram. This problem is also explained in Chapter 2. Figure 4.10 shows hierarchical clustering by use of correlation as a dissimilarity metric and “ward” method as a linkage method.

As can be seen in Figure 4.10 that this dendogram is easier to inspect since the branches of the dendogram are not too close to each other. The visualization of the clustering on the map can be seen in Figure 4.11.

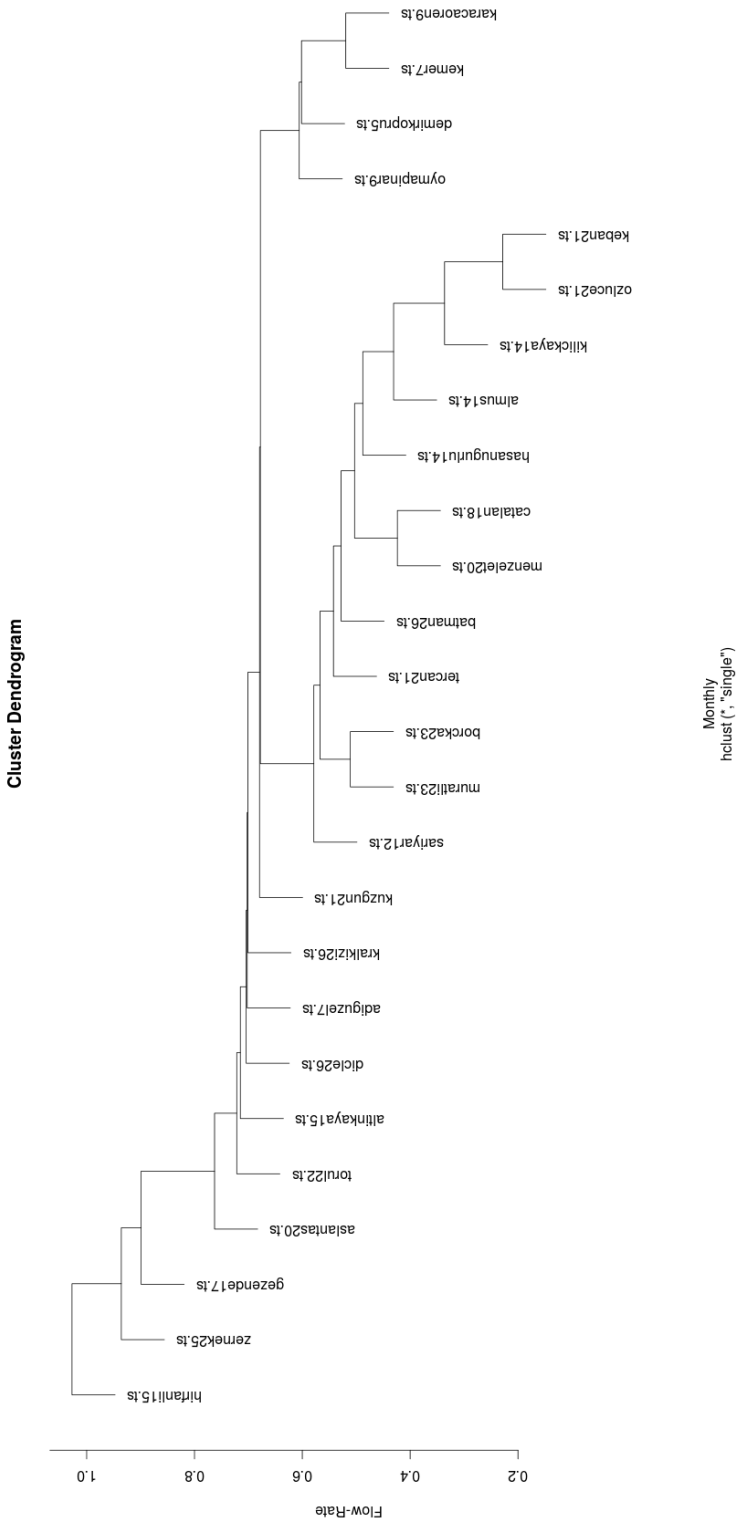

Figure 4.9: Hierarchical clustering (Correlation) of stream-flow dataset (Monthly)

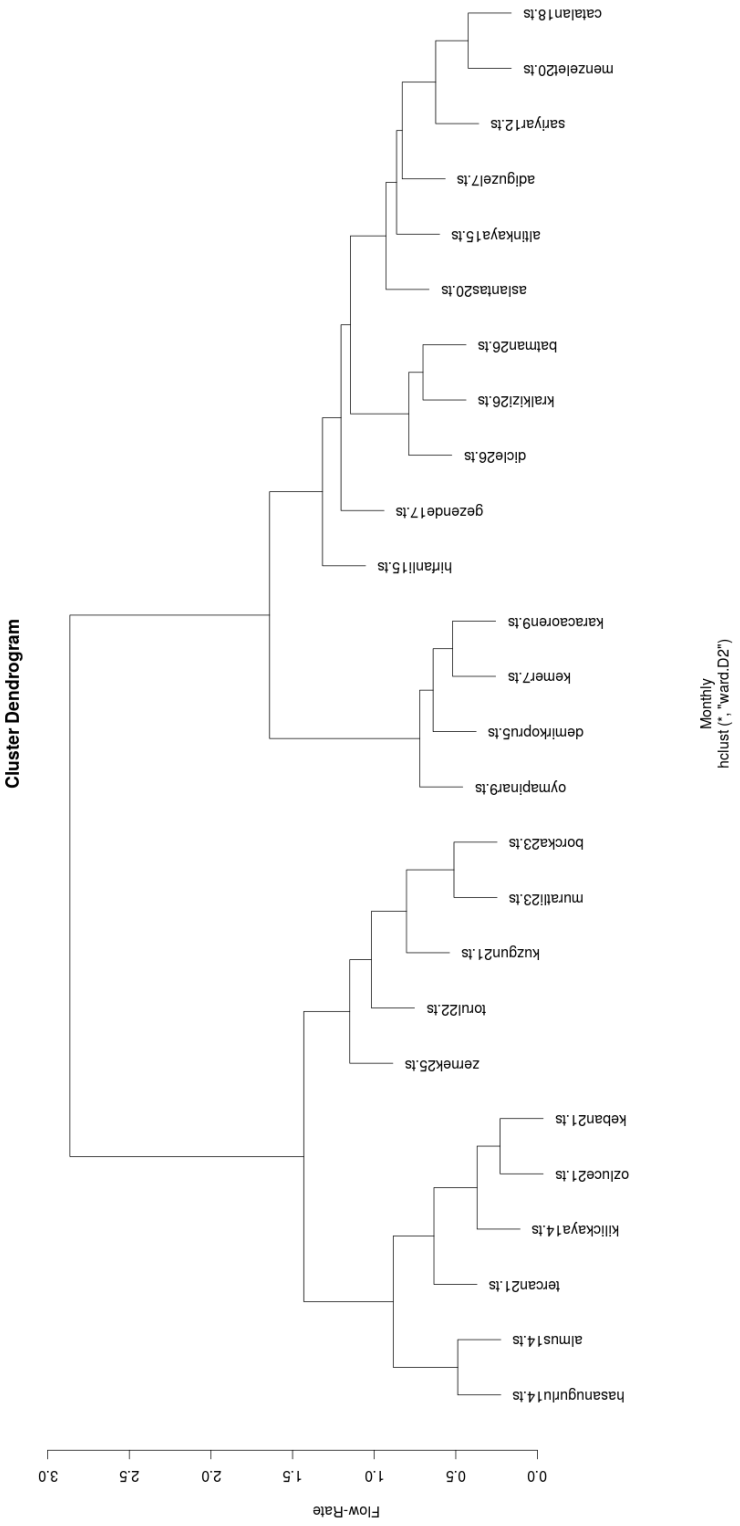

Figure 4.10: Hierarchical clustering (Correlation) of stream-flow dataset (Monthly)

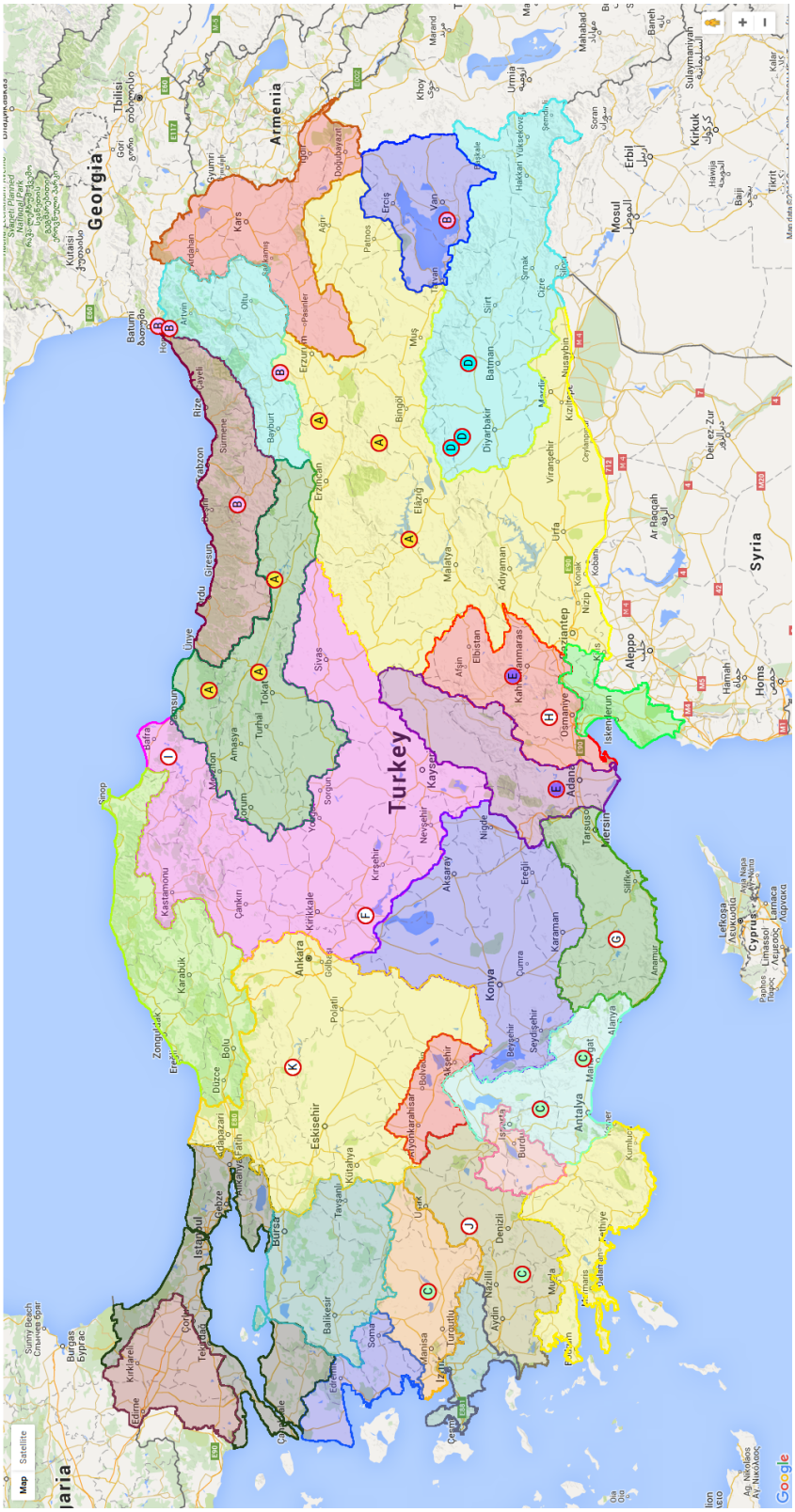

Figure 4.11: Visualization of correlation based hierarchical clustering results on the Turkey map

Figure 4.12 shows hierarchical clustering of logarithmic values of stream-flow by use of correlation as a dissimilarity metric and “ward” method as a linkage method.

Correlation based hierarchical clustering by use of “ward” method as a linkage method is also applied to the yearly averages of the stream-flow dataset. The corresponding dendogram can be seen in Figure 4.13.

4.2.1.1 Discussion

Hierarchical clustering is applied to the dataset by use of two different distance metrics. The first applied metric is euclidean distance. Results show that Keban is clustered alone in this clustering method. Logarithmic values of the samples in the dataset decrease the gap between Keban and the other objects, however, it is still linked lastly in the leftmost of the dendogram in Figure 4.8. This result shows that Keban has the highest dissimilarity value even after the use of logarithmic values. This result also shows that the use of logarithmic values failed. It can be concluded in Figure 4.7 that euclidean based hierarchical clustering clustered the time series dataset according to the magnitude of the values. However, correct way of clustering of the dataset is to cluster the streams based on the trend of the values. As a result both experiments, which are done by use of raw and logarithmic values, failed.

Table 4.4 reveals an important connection between k-means clustering and euclidean based hierarchical clustering. The result shows a resemblance of over 95% between two clustering methods. It can be concluded that these two methods bring about similar clustering results in stream-flow dataset.

The second applied metric is correlation distance.

Figure 4.10 and 4.12 show the dendograms of the “correlation” based hierarchical clustering. The dendograms of the Figure 4.10 and Figure 4.14 are same. Names of the streams in Figure 4.10 are removed for a better inspection in Figure 4.14.

It can be seen from the leftmost branch of the Figure 4.14 that basins 14 and 21, which are Yeşilırmak and Fırat, are clustered together. There are several important points in these clustering. Firstly, all three samples from Yeşilırmak basin are in this

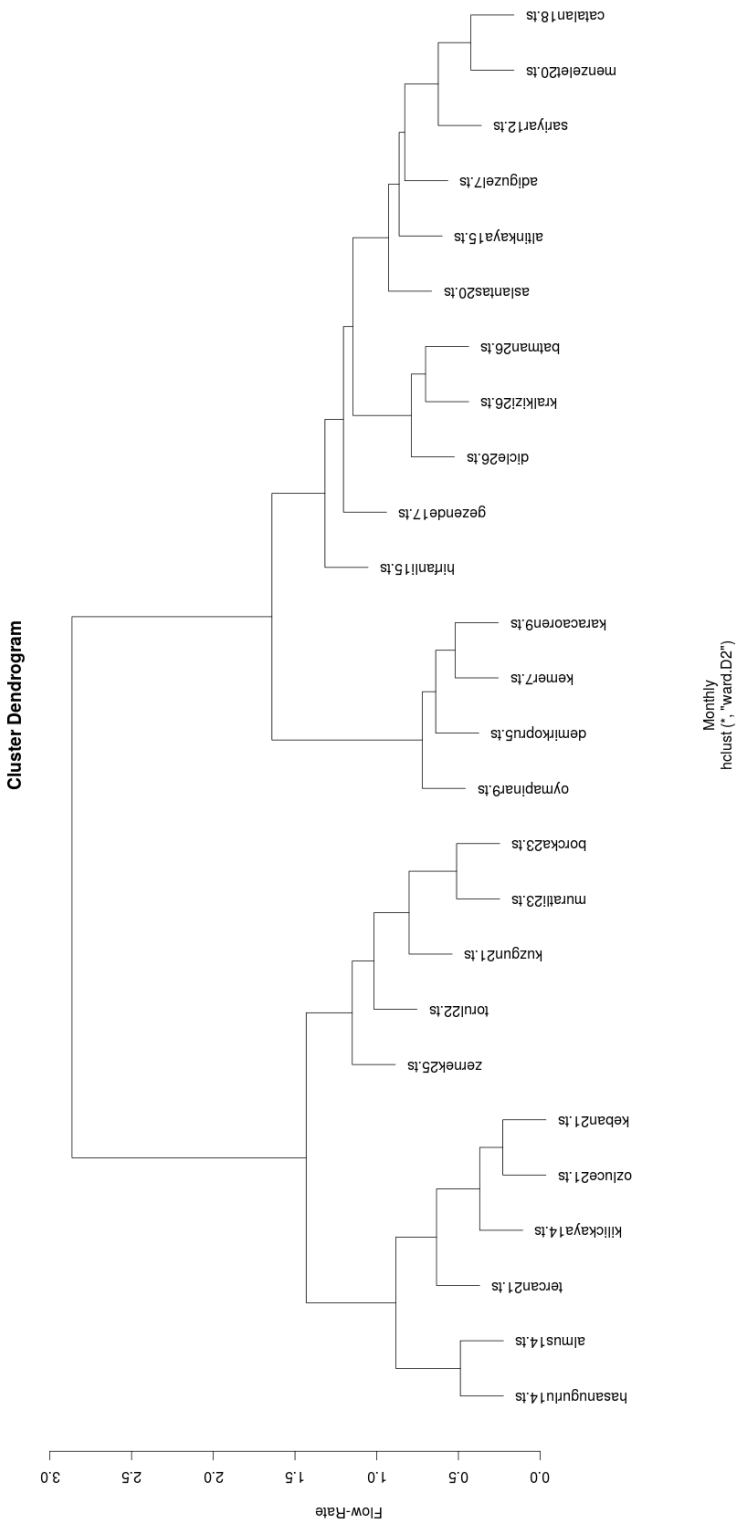

Figure 4.12: Hierarchical clustering (Correlation) of stream-flow dataset (Monthly-Log)

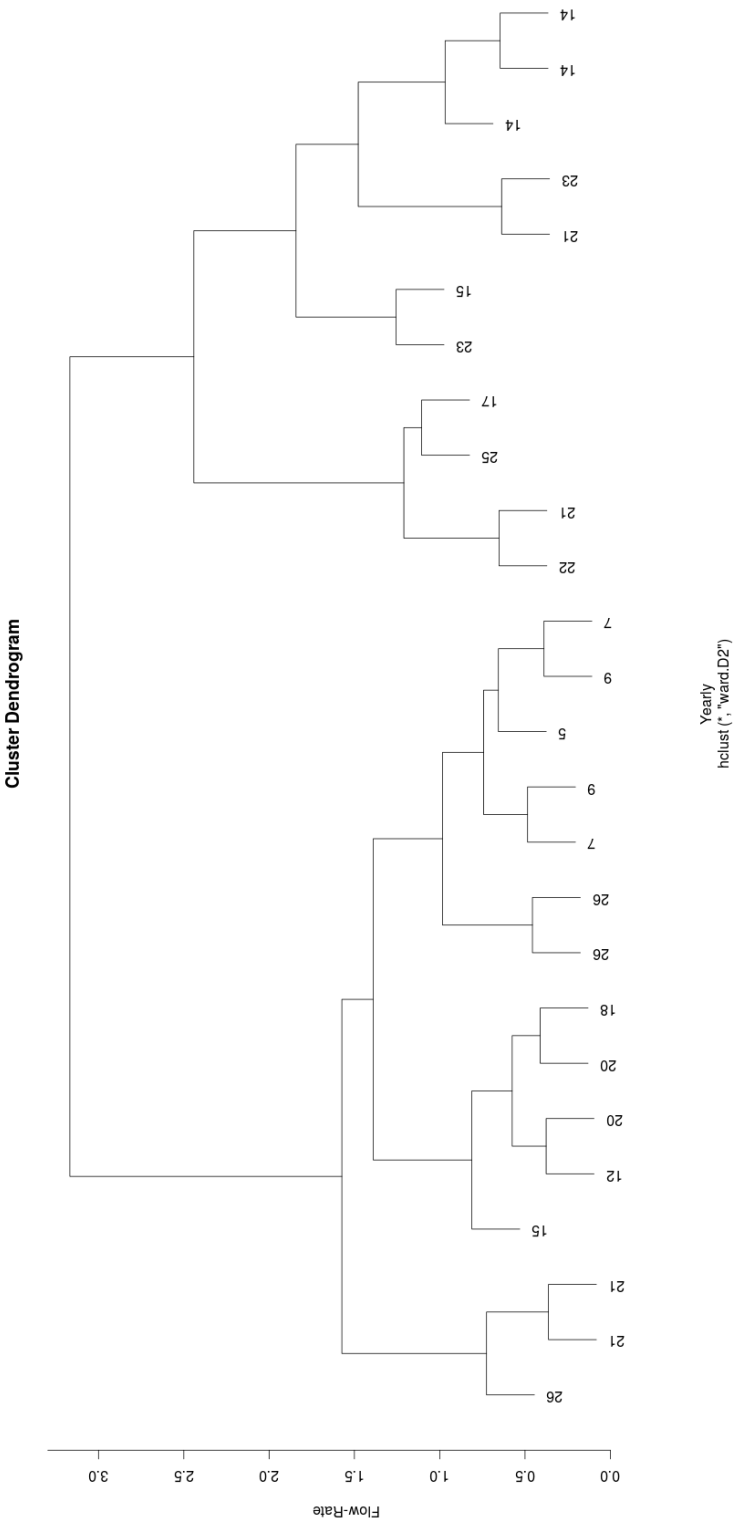

Figure 4.13: Hierarchical clustering (Correlation) of stream-flow dataset (Yearly)

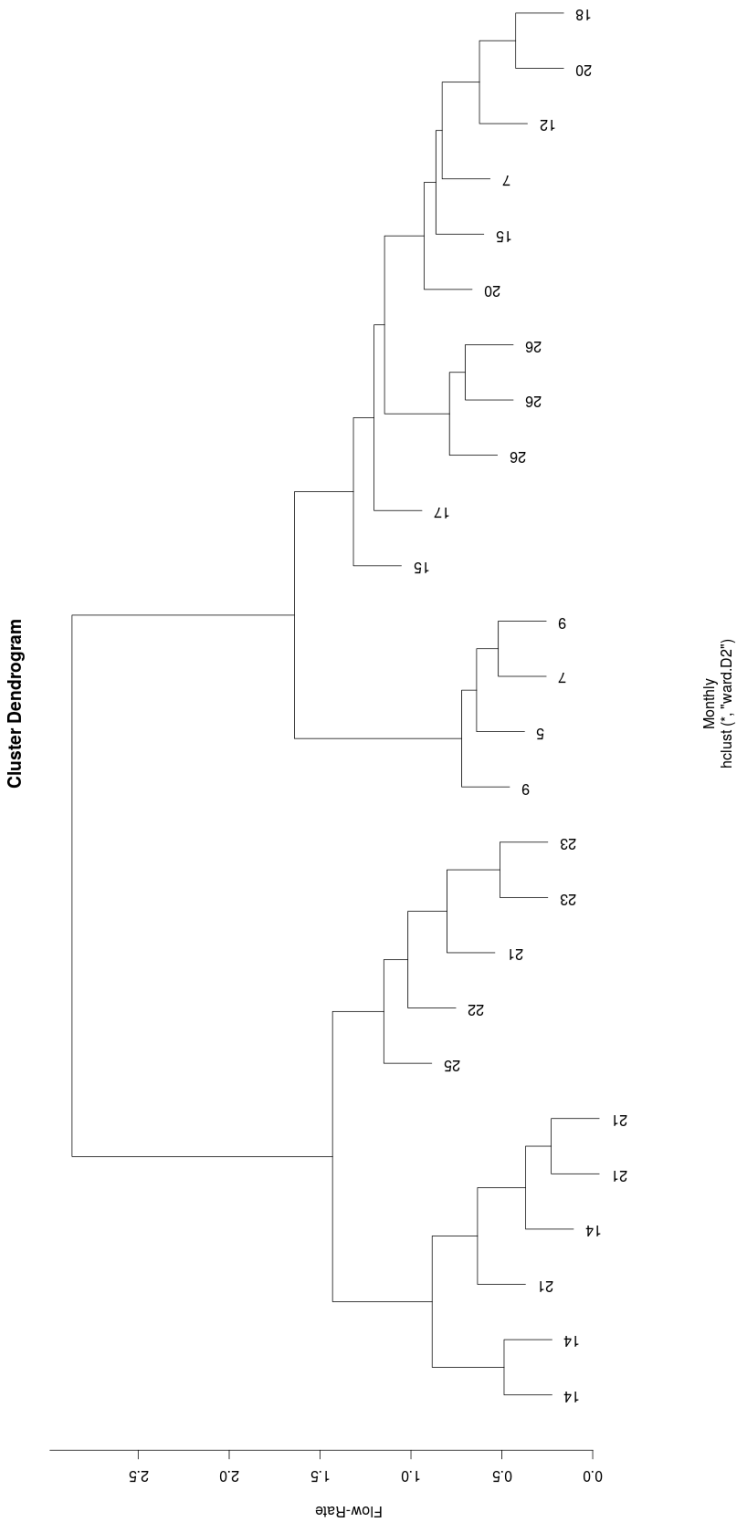

Figure 4.14: Hierarchical clustering (Correlation) of stream-flow dataset (Monthly)

branch of the dendogram. Yeşilırmak samples are clustered with Fırat and they are geographically neighbours. Secondly, only one of the samples of Fırat is missing from this branch of the dendogram and because of the size of Fırat, this result can be accepted normal. Better analysis can be made by use of exact location of the samples instead of basin location. Moreover, this sample of Fırat is in the left part of the dendogram and in the neighbour branch to the other samples of Fırat.

One of the other big branch just in right of the leftmost branch has samples from 21, 22, 23 and 25 basins, which are Fırat, Doğu Karadeniz, Çoruh and Van Gölü, respectively. There are several important points that can be noted about this branch. First of all, both samples of the Çoruh basin (23) in the dataset are clustered together. Moreover, Doğu Karadeniz and Van Gölü has only one samples in the dataset and these samples are clustered in the same branch. Secondly, all these 4 basins are geographically neighbour. Besides, one of the missing sample of the Fırat basin can be seen in this branch. As mentioned since Fırat basin has a biggest size between all 26 basins, if the clustering is done by use of coordinates of the samples more accurate conclusion can be drawn.

Until that point, inspection of the left part of the dendogram from top is done. The right part of the dendogram will be inspected at that point. The leftmost branch of the right of the dendogram has samples from 5, 7 and 9 basins, which are Gediz, Büyük Menderes and Orta Akdeniz, respectively. There are several important points that can be noted about that part of the branch. Firstly, both samples of the Orta Akdeniz basin are in this branch. Moreover, Gediz basin has only one sample in the dataset and this sample is in this branch, also. Only one of the sampled from Orta Akdeniz basin is missing and this sample is in the right of this branch and right part of the dendogram. Secondly, these three basins are not geographically neighbours, but they are close to each other. Moreover, it should be noted that, the stream-flow dataset does not contain samples from 4,6,8 and 10 basins which means by existence of these basins information more meaningful results can be drawn. In hydrological classification, deductive and inductive reasoning are two approaches which are used based on the available data[10]. In case of scarcity of data, deductive reasoning are used and it uses geology, topography and climate data for hydrologic regionalization. Inductive reasoning approach, on the other hand, uses available or predicted discharge data for

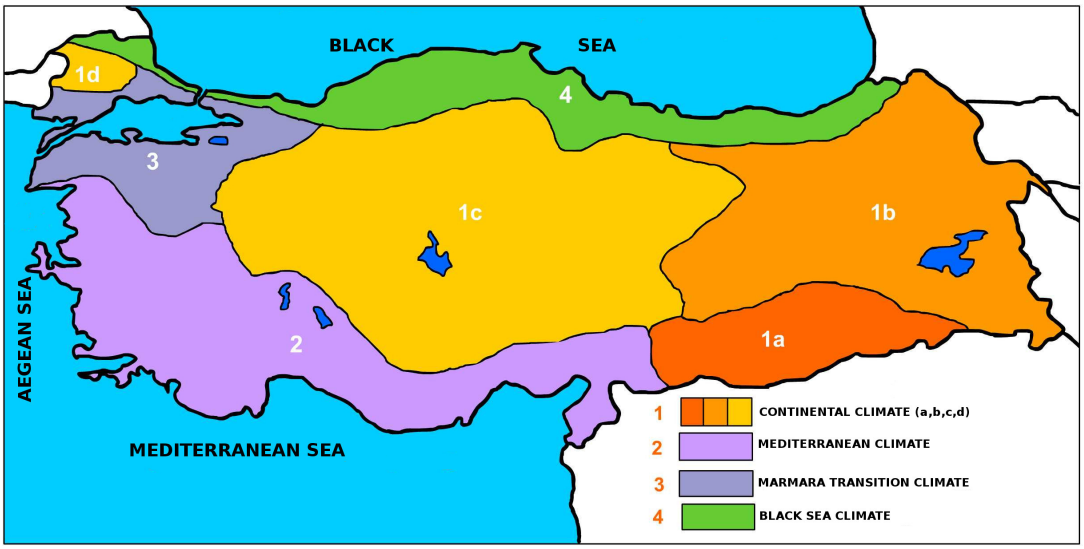

Figure 4.15: Climate Map of Turkey

stream-flow classification. Because of the scarcity of the data, climate condition can be used for samples from 4, 6, 8 and 10. It can be noted that because these four basins' sizes are small compared to other big basins, like Fırat and Dicle, and this part of the Turkey has a similar climatic condition, Mediterranean climate as can be seen in Figure 4.15 , that all these seven basins can be clustered together.

The rightmost of the dendogram has the biggest branch and several interesting conclusion can be drawn from this branch. First of all, all the samples of the Dicle basin (26) is clustered in one of the sub-branch and it can be concluded that Dicle basin has a specific condition that it is not clustered with samples form the other basins. One more important points is that samples from basins 18 and 20, which are Seyhan and Ceyhan, are clustered together in the rightmost part of the dendogram. One upper level of this branch contains samples from basins 7, 12, 15, 17 and 20, which are Büyük Menderes, Sakarya, Kızılırmak, Doğu Akdeniz and Ceyhan, respectively. Although these five basins are not related geographically, checking exact location of the samples in the dataset can lead to draw better conclusion from this branch.

All of the dendogram is inspected from left to right and it is seen that some logical conclusion can be drawn based on this dendogram. At that point, it can be acclaimed that correlation based hierarchical clustering is worked and the result of this method is visualized on the Figure 4.16 by use of hierarchical clustering results and climate information.

Figure 4.16: Clustered Turkey Basin Map

Figure 4.12 shows the logarithmic values of the stream-flow by use of same methods which are also used in Figure 4.10. It can be seen in Figure 4.12 that the left part of the dendogram from top is nearly same as Figure 4.10. There are some differences between Figure 4.10 and 4.12 in the right part of the dendograms from top, and it can be concluded that the dendogram of the raw values gives a better results than the dendogram of the logarithmic values. The reasons are that the samples of the Dicle basin are not clustered together in Figure 4.12 and branches contain samples that are not related geographically.

The cluster evaluation result between euclidean and correlation based hierarchical clustering is 0.3867355 and this value shows that the resemblance of the clustering of two methods is less than half and can be accepted low and this is why the results of euclidean based hierarchical clustering is accepted as unsuccessful and the results of correlation based hierarchical clustering is accepted as successful.

Correlation based hierarchical clustering in monthly stream-flow values supplies important results, therefore the same method is applied to the yearly averages in Figure 4.13. Several conclusion can be drawn from the dendogram inspection. First of all, the leftmost branch of the left of the dendogram from the top clusters Fırat and Dicle (21,26) samples. However, some of the other samples of these two basins are in the other branches. Secondly, the rightmost branch of the left of the dendogram from the

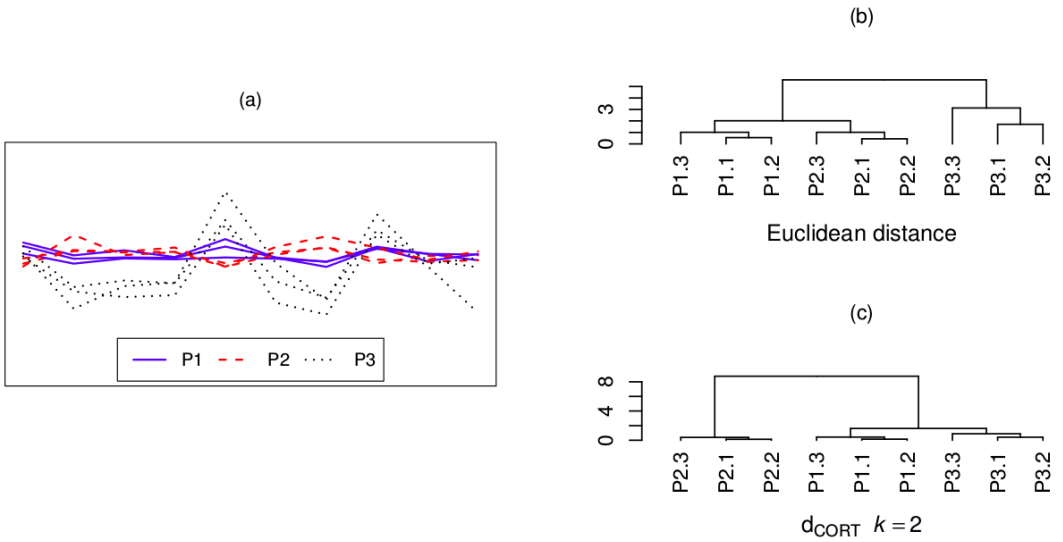

Figure 4.17: (a) Time-series dataset with 9 samples and 3 patterns (P1,P2,P3). (b) Dendogram of Euclidean based (shape-based) hierarchical clustering. (c) Dendogram of temporal correlation (structure-based) hierarchical clustering [55].

top clusters all the samples from Gediz (5), Büyük Menderes (7) and Orta Akdeniz (9) together. The other branches of the left of the dendogram can be accepted outlier since it is difficult to draw conclusions from them. The rightmost branch of the right of the tree from top clusters all the samples from Yeşilırmak (14) and left of this branch contains samples from Fırat and Çoruh (21,23), which are geographically neighbour of the Yeşilırmak basin. Besides, one of the samples of Fırat in this branch is far away from the other samples of the basin and exact location inspection may help to conclude more accurate results.

In this discussion section, the applied methods and their results are presented and evaluated until now. At this point, it is crucial to explain why correlation based hierarchical clustering is worked while euclidean based hierarchical clustering is failed. First of all, it is explained in the previous section that there are two types of the dissimilarity calculation concept, namely, shape-based and structure-based. Euclidean is used for shape-based clustering and correlation is used for structure-based clustering. Montero and Vilar [55] explains the result of these two different approaches on a dataset. Figure 4.17 shows this dataset(a) and the result of the euclidean(b) and temporal correlation(c) clustering.

A dataset with 9 samples are presented in Figure 4.17(a). This 9 samples has 3 differ-

ent patterns, namely, P1, P2 and P3. It can be seen in Figure 4.17(a) that samples of P1 and P2 have a similar values in magnitude. On the other hand, samples of the P1 and P3 follows a similar trend despite the big difference between their values. Figure 4.17(b) shows the result of the Euclidean based (shape-based) hierarchical clustering and P1 and P2 is clustered firstly and P3 is added to them lastly. This result shows that shape-based hierarchical clustering is done based on the closeness of the geometric profiles [55]. Figure 4.17(c) shows the result the temporal correlation based (structure-based) hierarchical clustering and P1 and P3 is clustered firstly and P2 is added to them lastly in this example. This result shows that structure-based hierarchical clustering is done based on the underlying dependence structure of the time series and since trend (increasing/decreasing) of the P1 and P3 are closer to each other than P1 and P2, they are grouped together [55]. This result shed light on why correlation based hierarchical clustering is successful in monthly stream-flow dataset. The connections of the trends between time-series samples are searched in these experiments and from Figure A.1 to Figure A.7 allow a visual verification of the resemblance of the stream-flow trends in the same basin. In conclusion, the result of the structure-based hierarchical clustering reveals trend similarity between basins.

4.2.2 Hydroelectric Energy Production Dataset

In this chapter, correlation based hierarchical clustering gives promising results. Therefore, this method is also applied to hydroelectric energy production dataset. There are 26 samples from 14 basins in stream-flow dataset. Hydroelectric energy production dataset contains 75 pumped-storage type hydroelectric power plants and 25 of these points are the points where stream-flow dataset values are acquired. Only hourly electric production information of Hoşap is missing which is inside the Van Gölü basins. These 25 points are also clustered by use of correlation based hierarchical clustering since it supplies successful results in former dataset and allows to make a comparison between the samples of these two datasets. Figure 4.18 shows the clustering results of hydroelectric energy production dataset.

In Figure 4.18, same method which is successful on stream-flow dataset is used. Distance matrix is calculated by use of correlation method and “ward” is used as a linkage

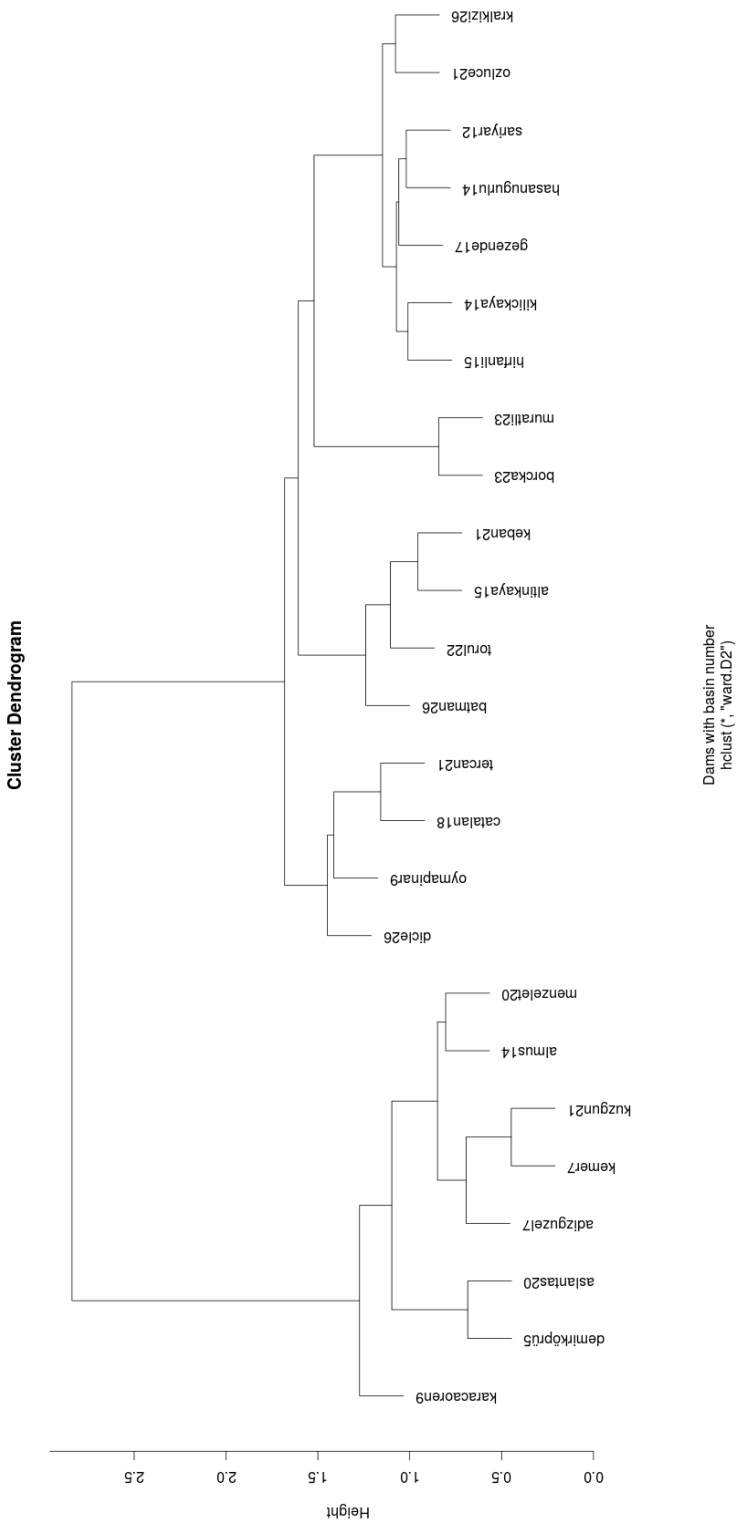

Figure 4.18: Correlation based hierarchical clustering of 25 hydroelectric energy production dataset(Hourly)

method. Only the rightmost branch contains samples from neighbour basins, namely, Fırat(21) and Dicle(26). It is not possible to connect the other branch with the same way.

In Figure 4.19, same method is applied to the log values of the hourly electric production data of the hydroelectric power plants with dams in the hydroelectric energy production dataset. The result shows that some of the samples of the Fırat(21) and both samples of the Çoruh(23) are clustered together and other than these branches it is not possible to make a connection between the samples in the branches.

The cluster evaluation result between raw and log values is 0.5285714 and this result shows that the outputs of two methods are not similar. Although the output of raw and log values are different, the results of the both methods are not useful.

Hydroelectric power plants dataset with 25 points does not supply meaningful information, and therefore the same method is applied to the whole hydroelectric power plants with dams data which contains 75 samples from 26 basins. The result of this clustering can be seen in Figure 4.20.

The rightmost branch of the Figure 4.20 shows that some of the samples of the Fırat(21) and Dicle(26) are clustered together. In the left of this rightmost branch Kızılırmak(15), Seyhan (18), Ceyhan(20) and Çoruh(23) are clustered. Kızılırmak(15), Seyhan (18) and Ceyhan(20) are neighbour but Çoruh(23) is not.

Left of this branch contains 4 samples in Fırat(21) basin. Then one of the branch contains samples of Sakarya(12) and Yeşilırmak(14). One of the other branch has samples from Doğu Karadeniz(22) and Çoruh(23). One of the other branch has Kızılırmak(15), Fırat(21) and Dicle(26) samples. Kızılırmak(15), Seyhan (18) and Ceyhan(20) are clustered together one of the other branch. Two samples from Yeşilırmak(14) and one sample from Kızılırmak(15) are clustered together in one of the other branch. These are branches with the neighbour samples.

Correlation based hierarchical clustering is applied to the run-of-river type hydroelectric power plants. There are 311 samples in the dataset and these samples contain at least 7000 observation for one year. The clustering results of the logarithmic values of the hydroelectric energy production data of run-of-river type hydroelectric power

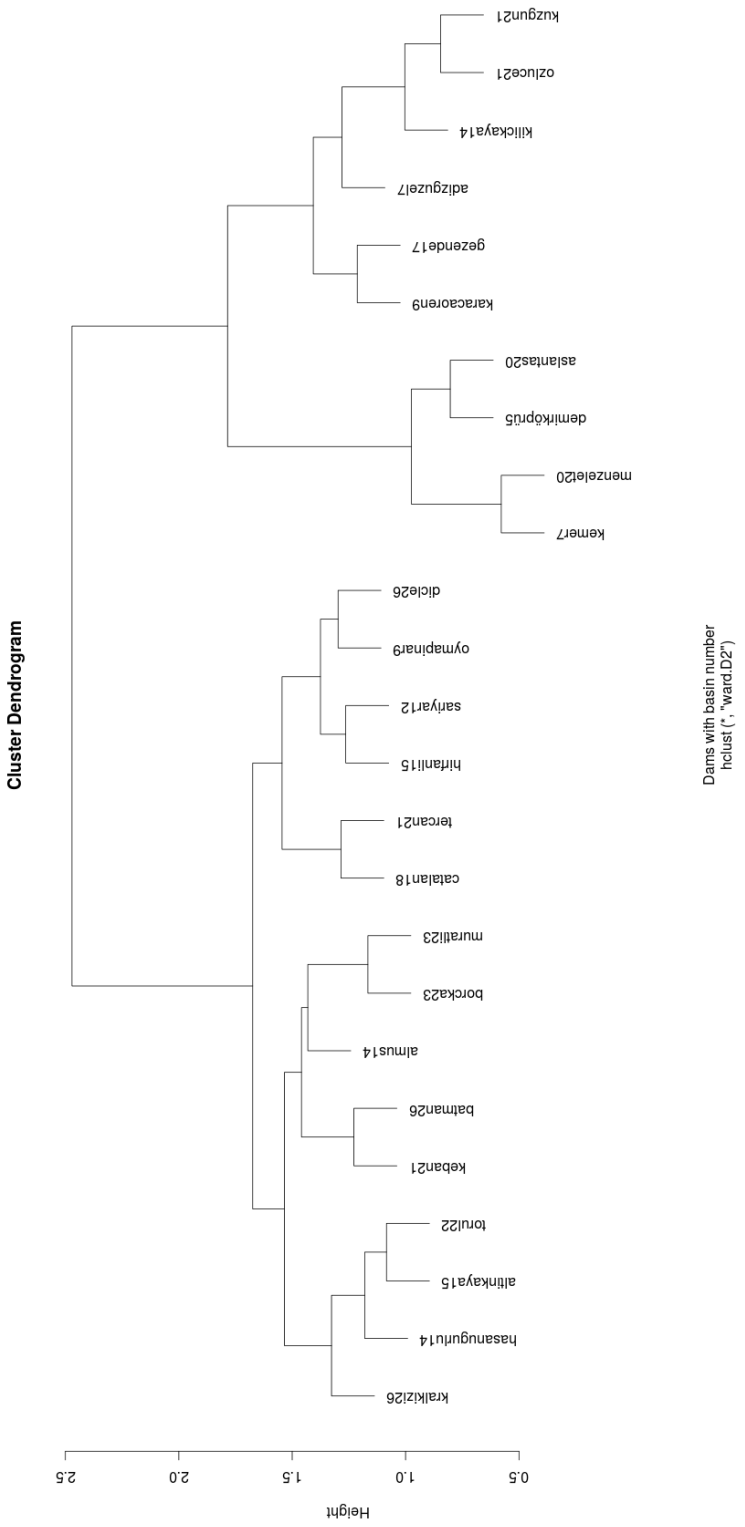

Figure 4.19: Correlation based hierarchical clustering of 25 hydroelectric energy production dataset(Hourly)

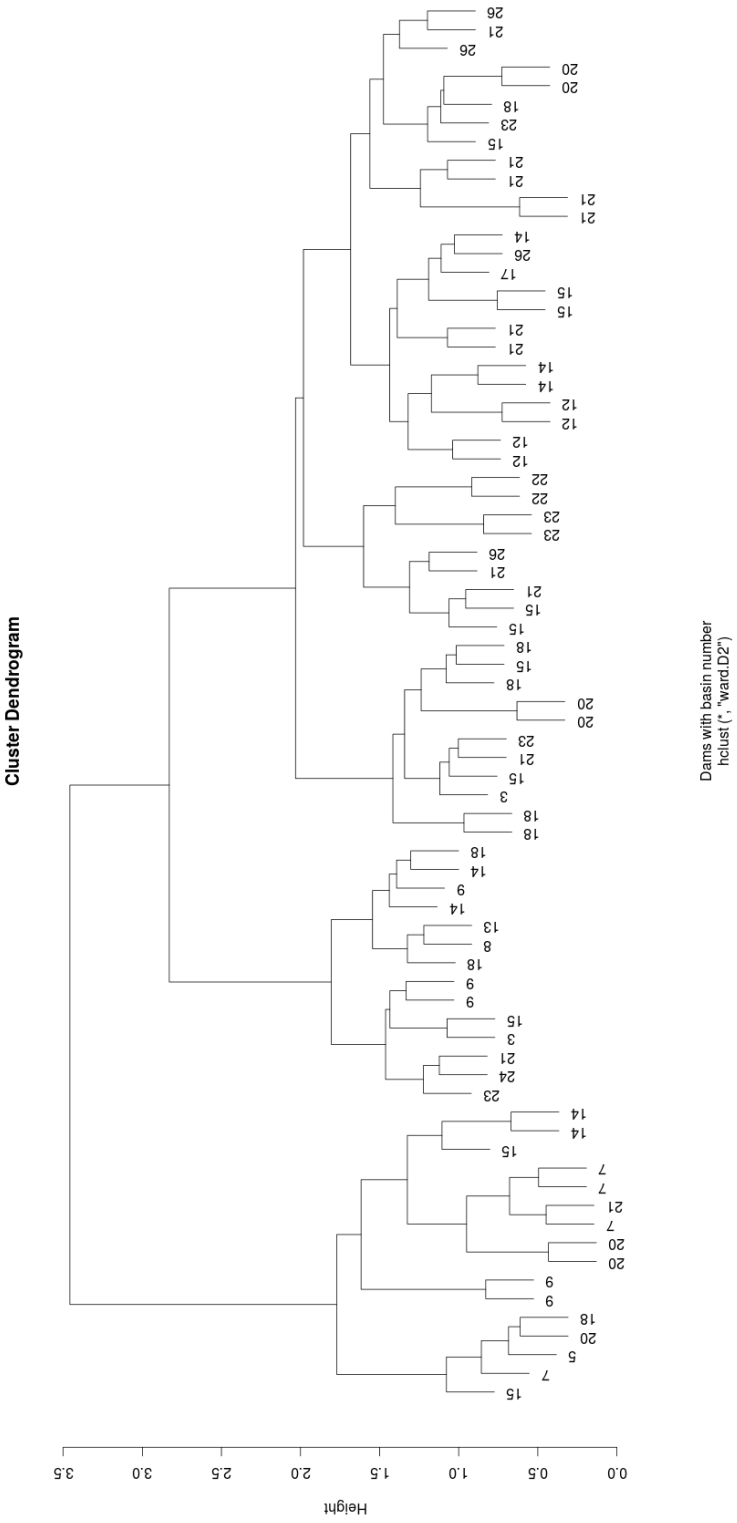

Figure 4.20: Correlation based hierarchical clustering of 75 hydroelectric energy production dataset(Hourly)

plants can be seen in Figure 4.21.

Dendogram in Figure 4.21 is cut in three levels. 4,6 and 12 clusters are acquired. Results can be seen in Figures 4.22, 4.24 and 4.26.

The results of the four, six and twelve cluster solutions are visualized on Turkey map and they can be seen in Figures 4.23, 4.25 and 4.27. In each of these figures, samples of the same clusters are shown with circles of same color and same letters. Therefore, both color and letter codes are used for identification of the samples of the clusters.

4.2.2.1 Discussion

75 points in the hydroelectric energy production dataset are belong to pumped-storage type hydroelectric power plants. This type of plants store the water and use it when needed. They work like batteries. Therefore, energy production is done when needed.

The results of the both clustering of 25 dams and 75 dams in Figure 4.18 and Figure 4.19 and Figure 4.20 do not supply significant results although correlation based hierarchical clustering is successful in stream-flow dataset. This 75 dams are used to identify a possible connection between the stream-flow rate and hydroelectric energy production and such a connection may help to make more accurate inferences about the basin based clustering of the stream-flow dataset since it has more samples. However, method fails on this dataset. The reason is that there is not much correlation between the electric production and stream-flow rate. Only in case of a long term decrease in the stream-flow, the water level in the dams can decrease as well and the stream-flow rate can affect the hydroelectric energy production.

The clustering results of the run-of-river type hydroelectric power plants also do not supply significant results. The dendogram is clustered to four, six and twelve groups but concentration of the same clusters in the same part of the map is not detected. Therefore, correlation based hierarchical clustering in the run-of-river type hydroelectric power plants dataset is failed.

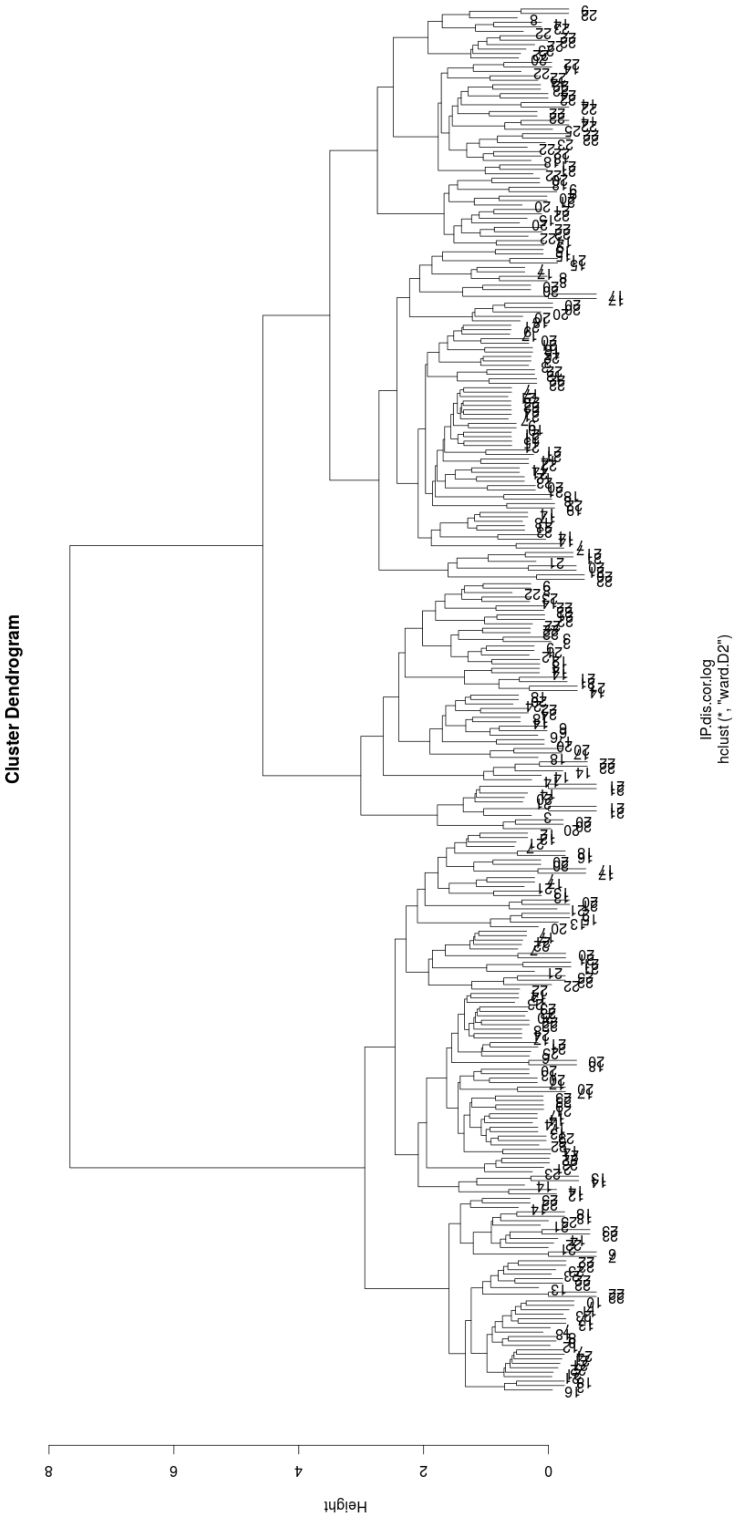

Figure 4.21: Correlation based hierarchical clustering of 311 run-of-river type hydroelectric power plants

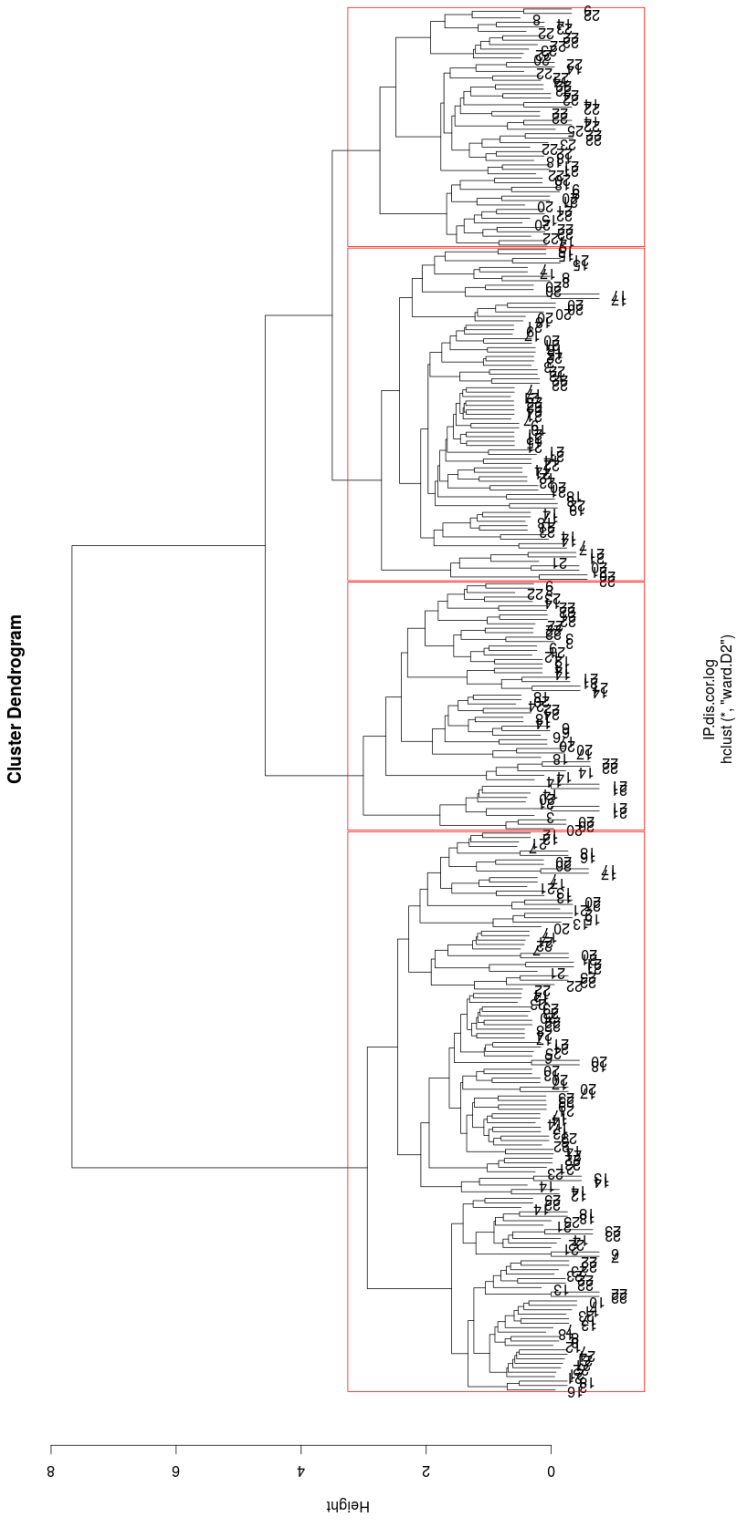

Figure 4.22: Correlation based hierarchical clustering of 311 run-of-river type hydro-electric power plants (Four Cluster)

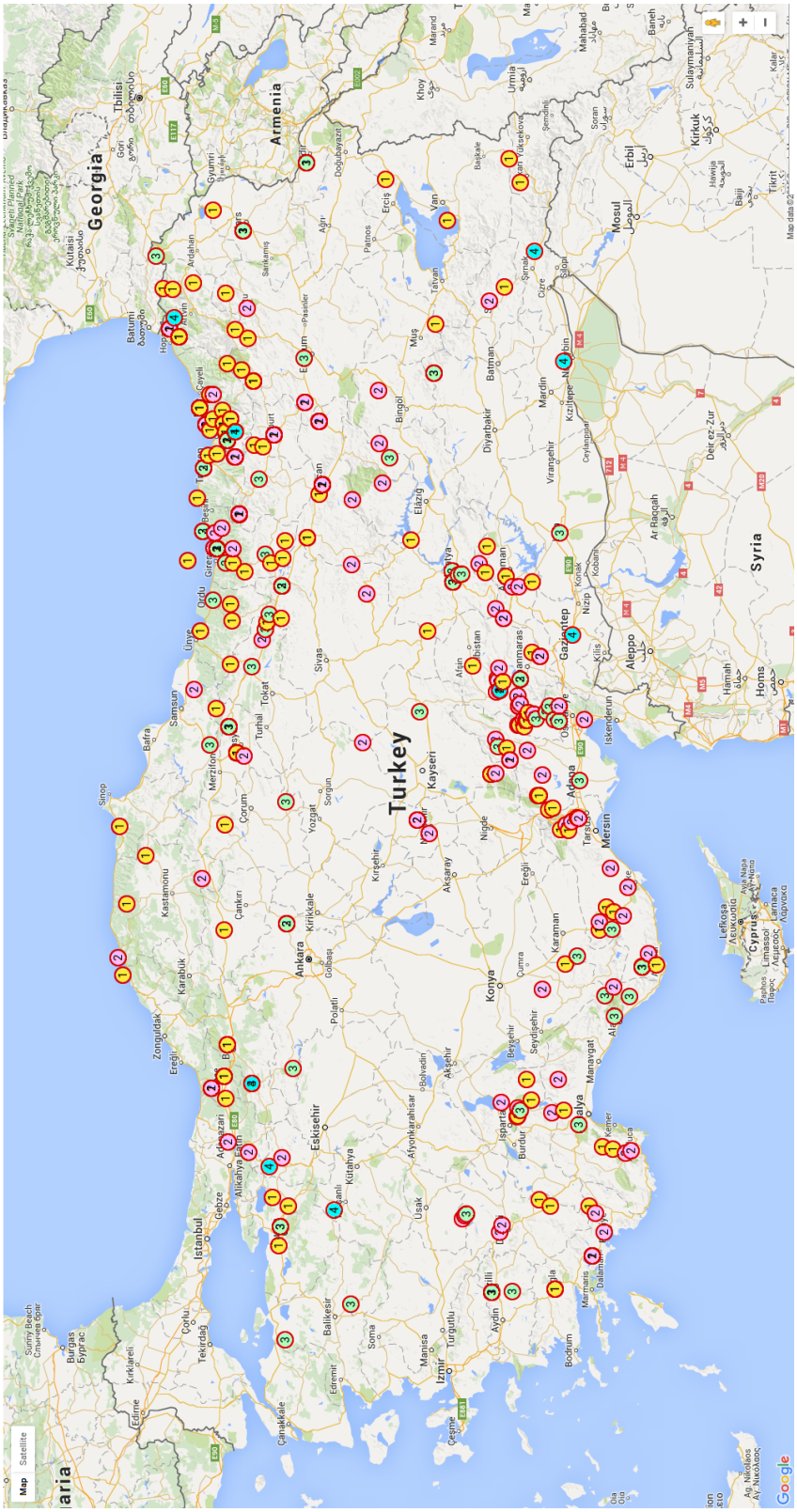

Figure 4.23: Four Cluster solution on Turkey map

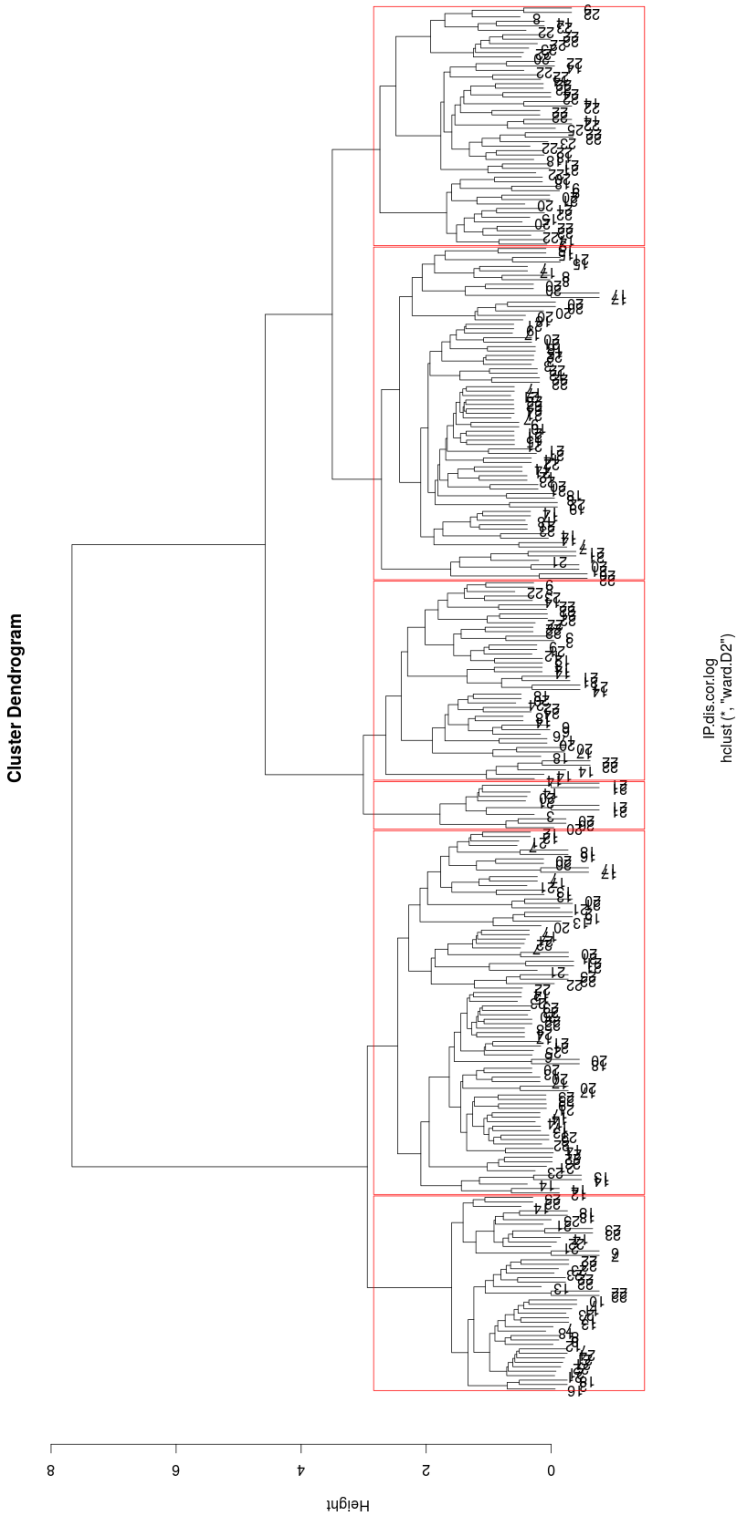

Figure 4.24: Correlation based hierarchical clustering of 311 run-of-river type hydro-electric power plants (Six Cluster)

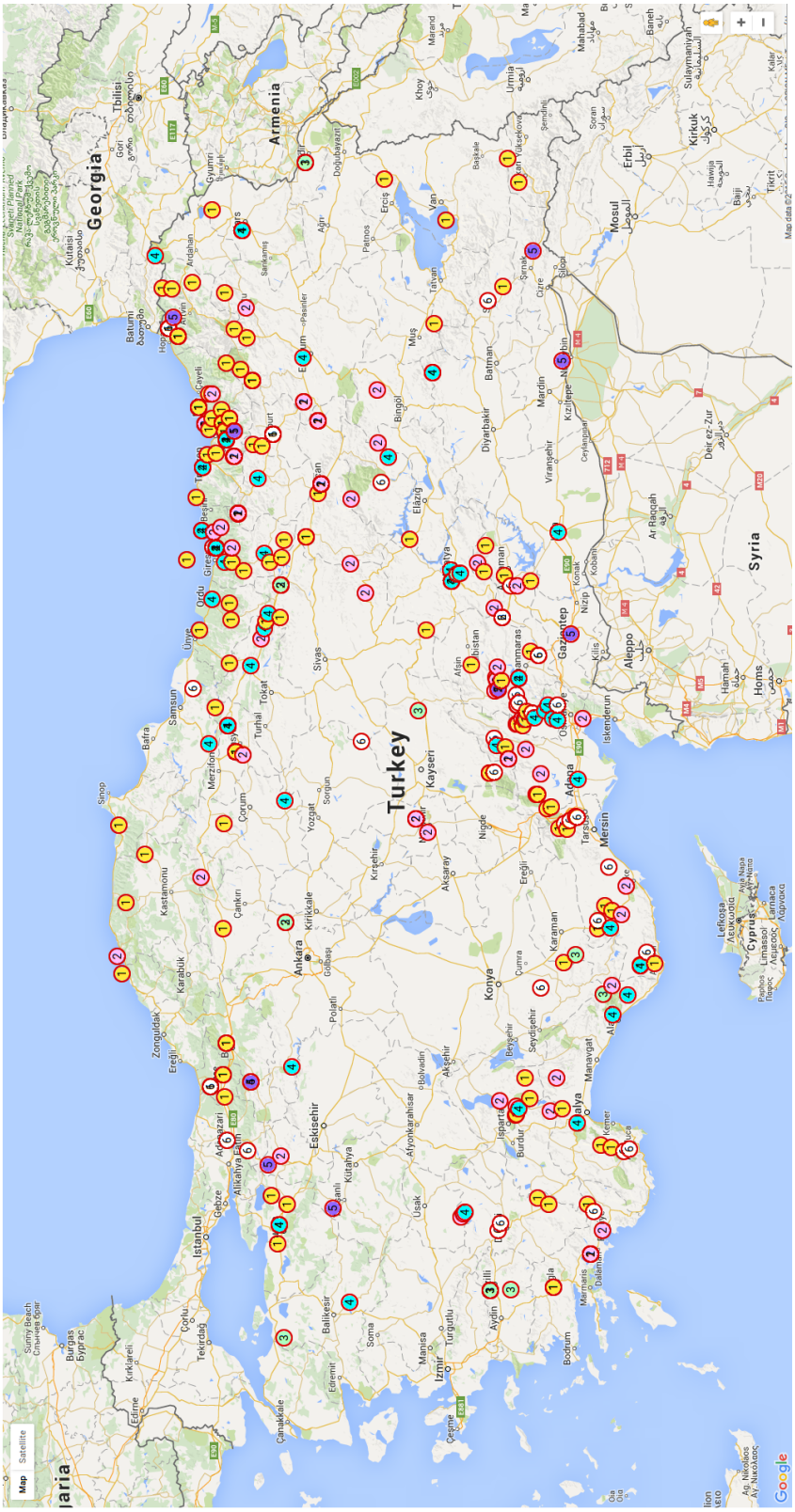

Figure 4.25: Six Cluster solution on Turkey map

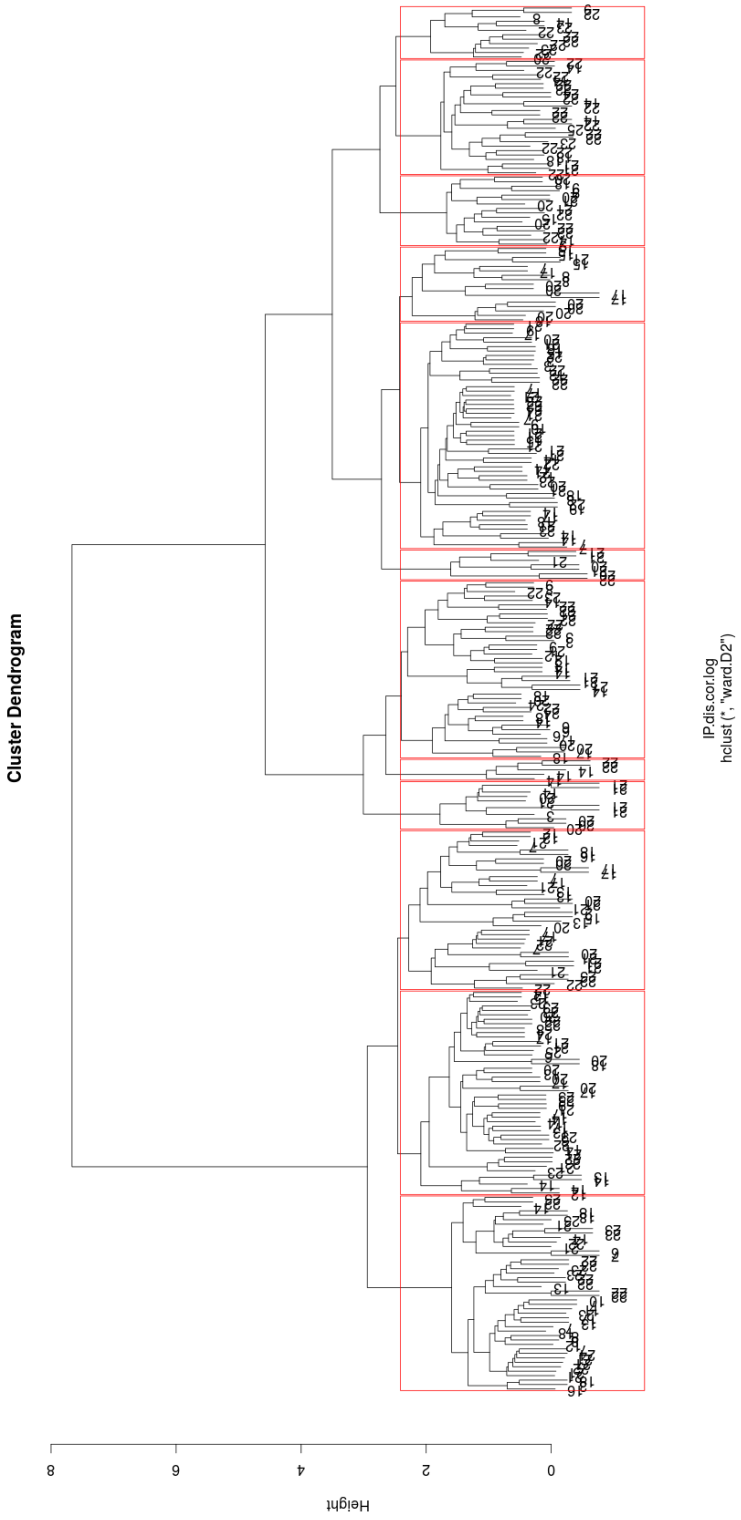

Figure 4.26: Correlation based hierarchical clustering of 311 run-of-river type hydro-electric power plants (Twelve Cluster)

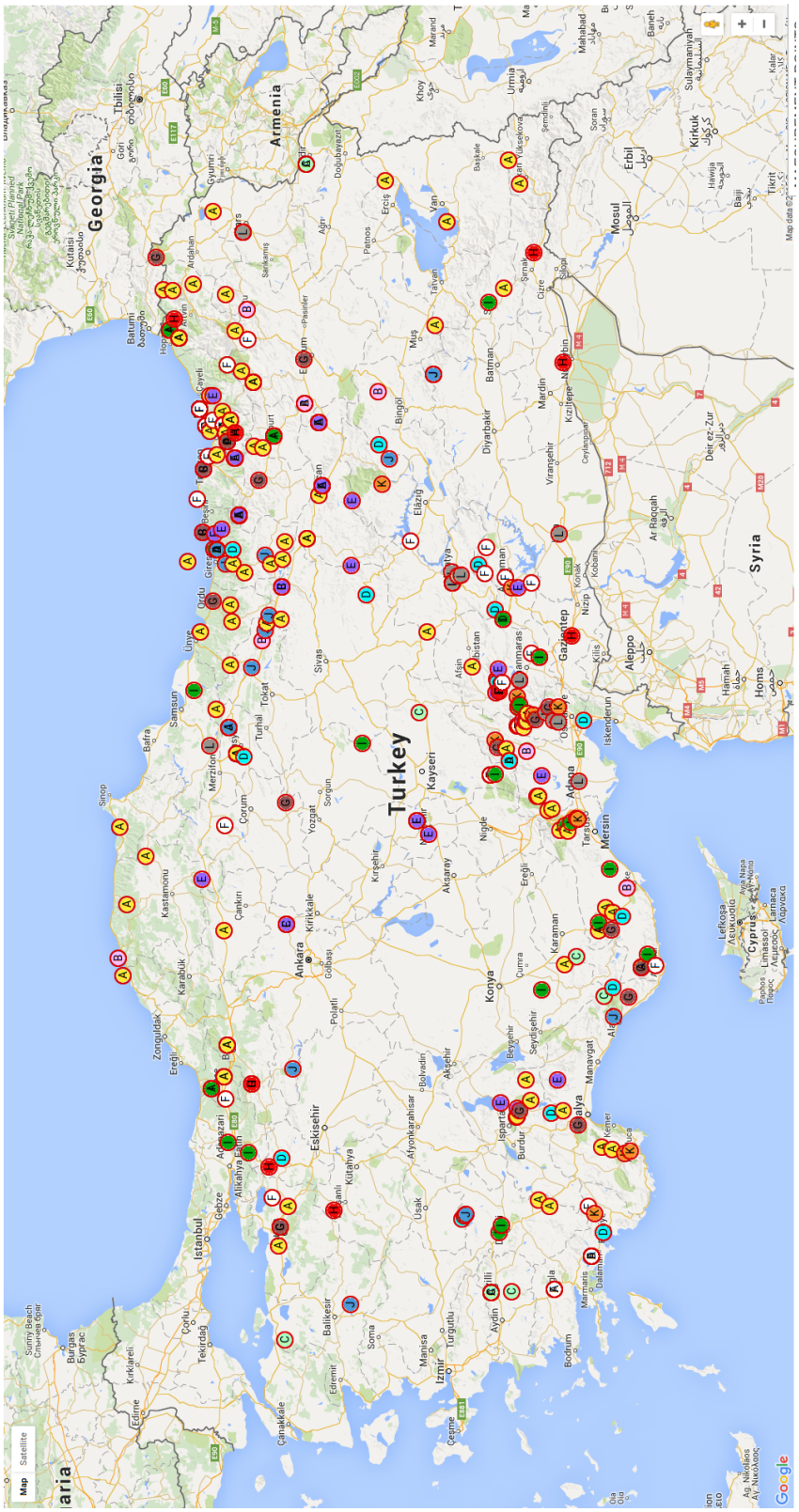

Figure 4.27: Twelve Cluster solution on Turkey map

4.3 Clustering with Dynamic Time Warping

Dynamic time warping is applied to the stream-flow dataset and results can be seen in Table 4.5.

Table 4.5: Dynamic Time Warping based Hierarchical Clustering

Stream-flow Name	**Basin Name**	**Cluster DTW**
Demirköprü	5	1
Adıgüzel	7	1
Kemer	7	1
Karacaören	9	1
Sarıyar	12	1
Kılıçkaya	14	1
Almus	14	1
Hirfanlı	15	1
Gezende	17	1
Menzelet	20	1
Özlüce	21	1
Kuzgun	21	1
Tercan	21	1
Torul	22	1
Zernek	25	1
Kralkızı	26	1
Dicle	26	1
Oymapınar	9	2
Altınkaya	15	2
Muratlı	23	3
Borçka	23	3
Keban	21	4
Hasan Uğurlu	14	5
Çatalan	18	5
Aslantaş	20	5
Batman	26	6

The evaluation of the clustering in monthly and yearly dataset can be seen in Table 4.6. The values of evaluation results change between 1 and 0. 1 shows that the result of the methods are identical whereas 0 shows that the result of the clustering between two methods is totally different.

Table 4.6: Cluster Evaluation

First Method	Second Method	Evaluation Results
Monthly DTW (Average)	Monthly DTW (Single)	0.7667895
Monthly DTW (Average-Log)	Monthly DTW (Single-Log)	0.8796296
Monthly DTW (Average-Log)	Monthly DTW (Average)	0.6266204
Monthly DTW (Single-Log)	Monthly DTW (Single)	0.5191638
Yearly DTW (Average)	Yearly DTW (Single)	0.8649425
Yearly DTW (Average-Log)	Yearly DTW (Single-Log)	0.8175926
Yearly DTW (Average-Log)	Yearly DTW (Average)	0.7100529
Yearly DTW (Single-Log)	Yearly DTW (Single)	0.3744658
Monthly DTW (Average)	Yearly DTW (Average)	0.6329806
Monthly DTW (Average-Log)	Yearly DTW (Average-Log)	0.8031215
Monthly DTW (Single)	Yearly DTW (Single)	0.6982906
Monthly DTW (Single-Log)	Yearly DTW (Single-Log)	0.6620370

Dynamic time warping clustering is applied both raw and logarithmic values of the dataset in monthly and yearly resolution by use of average and single agglomeration methods. The results show that the most similar clustering scheme is happened between average and single agglomeration methods of logarithmic monthly dataset since higher values indicate a high resemblance between two clustering results.

4.3.1 Discussion

It is presented in Table 4.5 that Keban is still clustered as alone. This result shows that problem of the clustering Keban continues and dynamic time warping method is failed.

4.4 Clustering with Longest Common Subsequence

4.4.1 Stream-Flow Dataset

Longest common subsequence is one of the another algorithm which can be used for trend detection in time series. This algorithm is basically compare two time series and

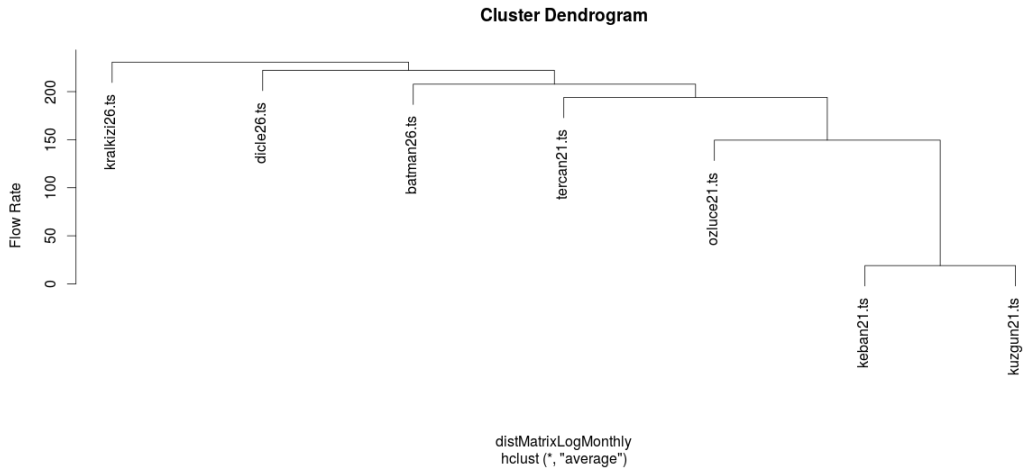

Figure 4.28: Plot of Fırat and Dicle basins clustering by use of LCSS method(Monthly-Log)

detect the longest common part between two time series. LCSS is applied to the log values of monthly stream-flow in Fırat and Dicle basins. The result of the clustering can be seen in Figure 4.28.

It can be seen that Fırat and Dicle streams are clustered correctly by use of LCSS method. This result shows that the problem of clustering of the Keban stream is overcame. Since the problem faced in the previous methods is solved, LCSS method is applied to whole dataset and the result can be seen in Figure 4.29.

4.4.1.1 Discussion

Fırat and Van Gölü basins are clustered with Dicle,Çoruh and Kızılırmak in the upper level in the leftmost part of the figure. These basins are also connected to each other geographically. However, the other branches are not connected geographically. For that reason, although the clustering of Fırat and Dicle is successful, the result of the method on the whole dataset is failed.

4.4.2 Hydroelectric Energy Production Dataset

In this chapter, different clustering methods are applied to the monthly stream-flow dataset and it is concluded that correlation based hierarchical clustering and longest

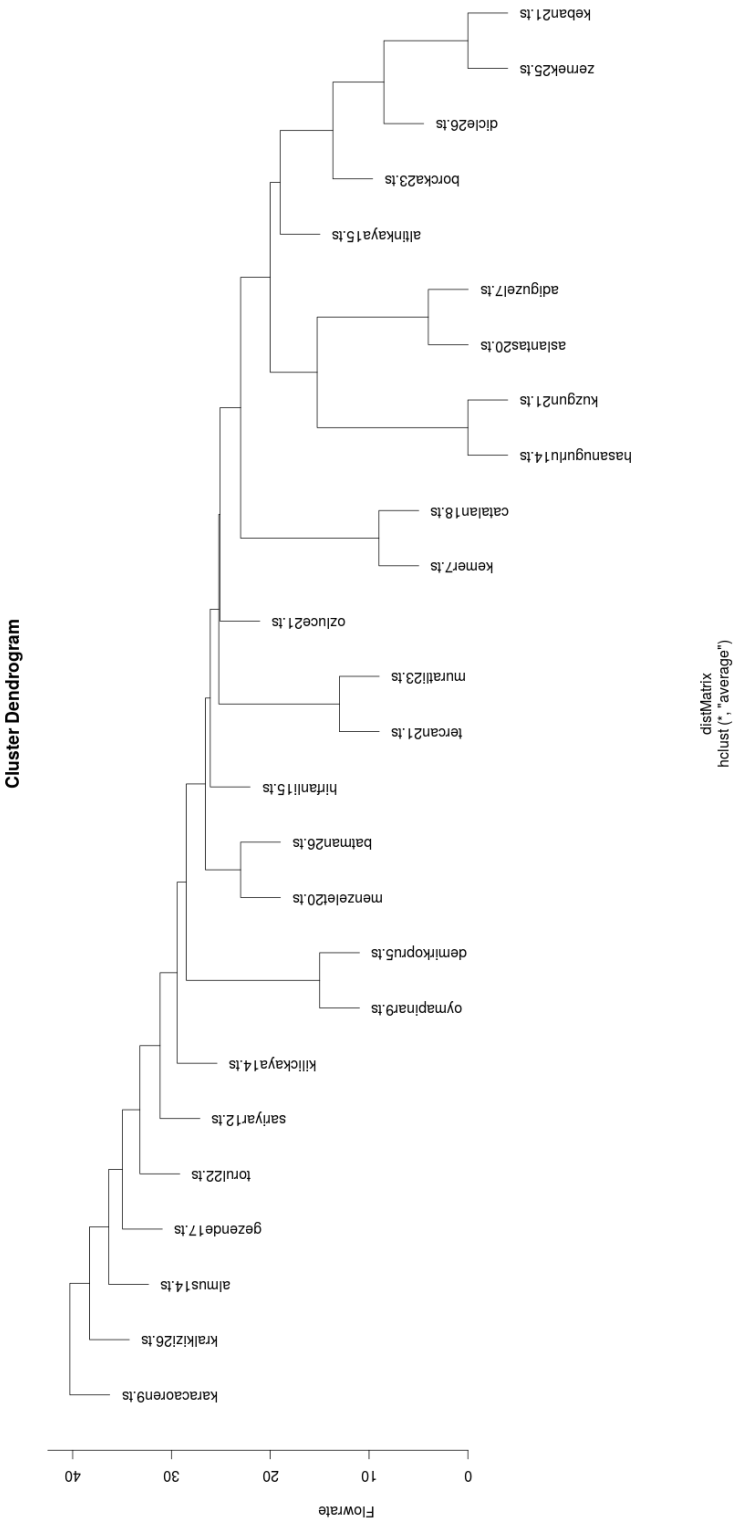

Figure 4.29: Plot of basins clustering by use of LCSS method

common subsequence method supply meaningful results. Therefore, LCSS is also applied to hydroelectric power plants with dams in the energy production dataset as clustering method. LCSS clustering result of the 25 out of 26 points, which also exist in the monthly stream-flow dataset, in hydroelectric energy production dataset can be seen in Figure 4.30.

25 points in this dataset are belong to pumped-storage type hydroelectric power plants. This type of plants store the water and use it when needed. It works like a battery. In this dataset, there are hourly electric production of 75 plants for one year. Figure 4.30 contains 25 of them to allow to make a comparison between stream-flow and electric production datasets. The LCSS clustering of 75 plants can be seen in Figure 4.31.

4.4.2.1 Discussion

In Figure 4.30, dendogram of the 25 pumped-storage type hydroelectric power plants are shown and it is hard to make a location based connection between the samples. It does not reveal an important connection between samples. In Figure 4.31, dendogram of the 75 pumped-storage type hydroelectric power plants are shown. In this figure, some branches have neighbour samples. For example, one of the branch contains samples from Yeşilırmak(14), Kızılırmak(15), Fırat(21). One of the branch contains samples from Ceyhan(20) and Fırat(21). Although the mentioned branches contain samples from neighbour basins, there are several other branches with samples which do not have neighbour connection. As a result, it is hard to make an inference by using these results and it can be concluded that method longest common subsequence clustering fails on the hydroelectric energy production dataset.

4.5 Validation

Calinski Harabatz criterion is used to find the optimal number of the cluster in the experiments. The cluster number is chosen accordingly before beginning the experiments. It is also important to validate the results of the clustering of the stream-flow dataset by using validation techniques after the experiments to evaluate the quality of

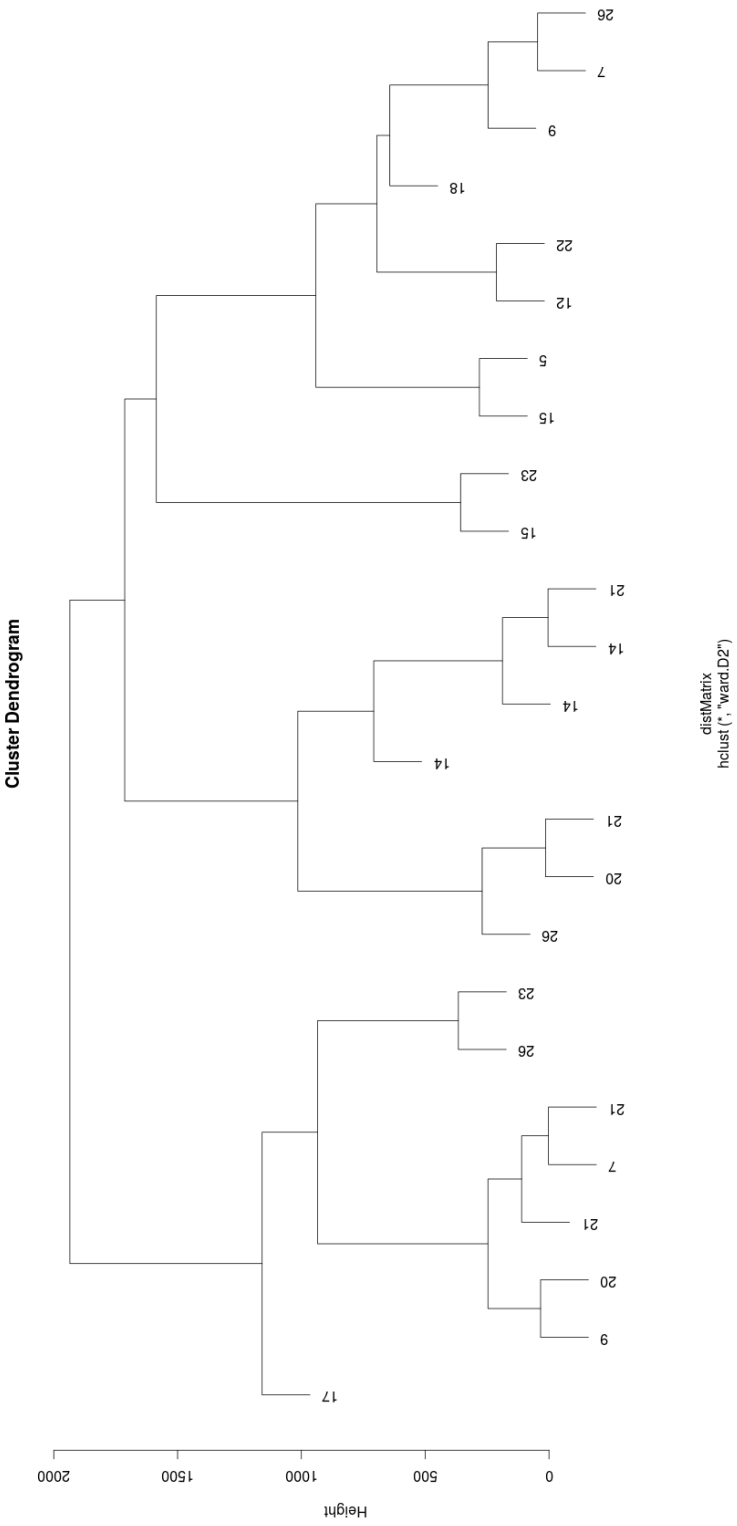

Figure 4.30: LCSS clustering of 25 hydroelectric energy production dataset(Hourly)

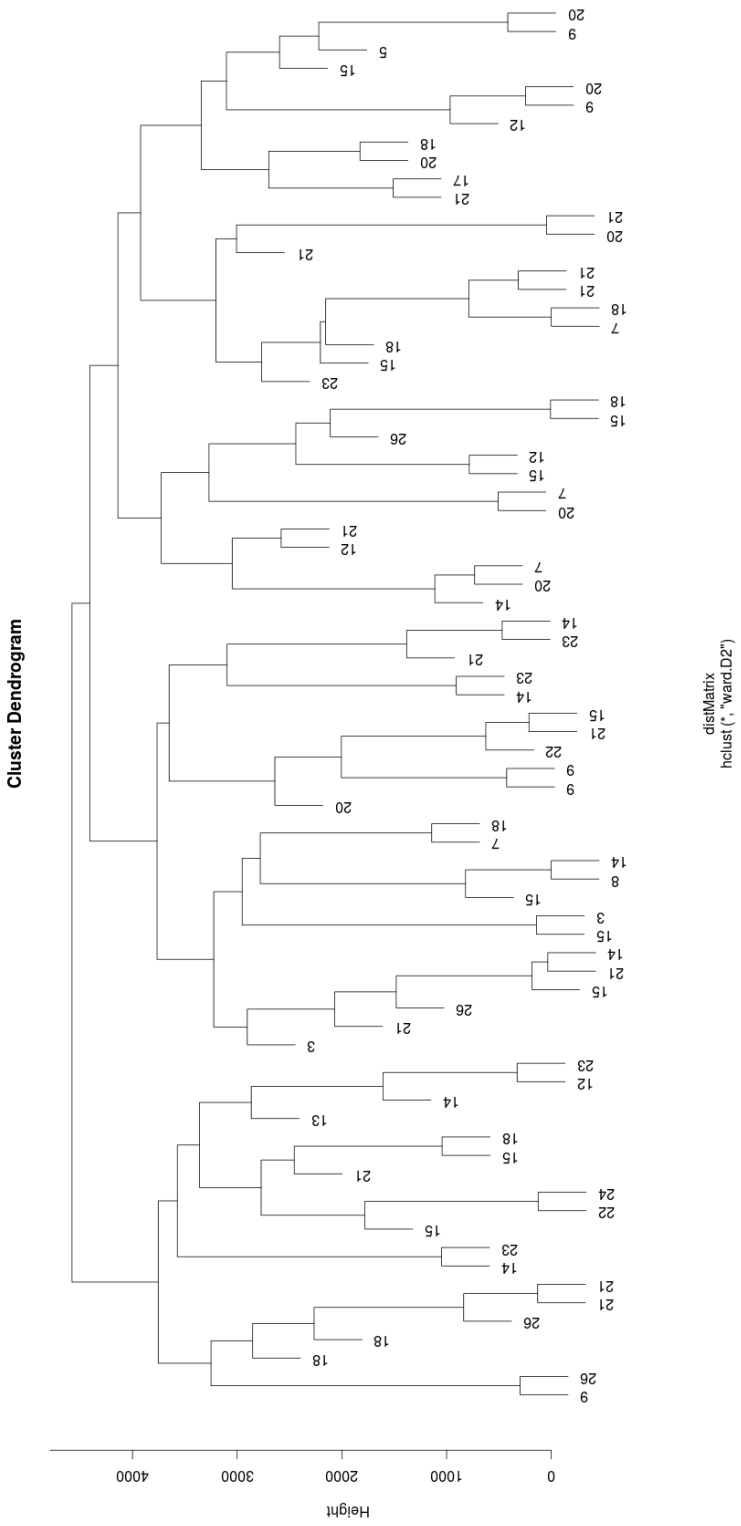

Figure 4.31: LCSS clustering of 75 hydroelectric energy production dataset(Hourly)

the results. Validation can be measured by using Dunn index, silhouette width and connectivity. Dunn index is calculated by dividing the smallest length between samples which are not in the same cluster to the biggest intra-cluster distance. Dunn index value may be between zero and infinity and it should be maximized. The silhouette value is the measurement of the confidence in clustering assignment of a specific sample. The silhouette value may vary between -1 and 1. 1 indicates well-clustered samples whereas -1 indicates the poorly clustered samples. The silhouette width is calculated by use of silhouette number. It is the average of each sample's silhouette value. It should be maximized. The degree of the connectedness between clusters is calculated by use of connectivity. Connectivity value may be between zero and infinity and it should be minimised [56]. In Figure 4.32, the optimal number of the cluster and values of the Dunn index, silhouette width and connectivity are shown.

In Figure 4.33, validation results of the logarithmic values of the stream-flow dataset are presented.

4.5.1 Discussion

Validation results show that both k-means and hierarchical clustering on stream-flow dataset have very similar Dunn index, silhouette and connectivity values. Moreover, optimal number of clusters is found as two. All optimum values belong to hierarchical clustering. Silhouette width is 0.8420 and this value is near 1, which indicates a well-clustered samples. Dunn index is 2.7895 and it has the highest value for two cluster as well. Connectivity is 2.9290 and it should be minimised. However, the connectivity value is high and it shows the connectivity of the samples are not good in the clusters.

Calinski Harabatz criterion is found the optimal number of cluster for the logarithmic values of the stream-flow dataset as 2. Best connectivity and silhouette width are acquired in 2 cluster solution as well while best Dunn index value is acquired in 3 cluster solution. Results show that hierarchical clustering gives the best results for connectivity and Dunn index while k-means gives the best results for the silhouette width. Silhouette width is 0.518 and it shows that silhouette width of the logarithmic values of the stream-flow dataset is worse than the silhouette width of the real values of the stream-flow dataset. Connectivity should be minimised but it is high like the

```
Clustering Methods:
 hierarchical kmeans

Cluster sizes:
 2 3 4 5 6 7 8

Validation Measures:
                                 2       3       4       5       6       7       8

hierarchical Connectivity   2.9290  6.7869 14.2817 19.3619 21.3619 24.3952 25.4786
             Dunn           2.7895  0.8281  0.4086  0.5379  0.5379  0.5552  0.5552
             Silhouette     0.8420  0.5474  0.4243  0.4096  0.3724  0.3413  0.3161
kmeans       Connectivity   2.9290  6.7869 16.9333 19.8187 21.8187 26.2032 28.2865
             Dunn           2.7895  0.8281  0.3191  0.4077  0.4077  0.4389  0.4701
             Silhouette     0.8420  0.5474  0.4447  0.4017  0.3640  0.3434  0.3126

Optimal Scores:

             Score  Method       Clusters
Connectivity 2.9290 hierarchical 2
Dunn         2.7895 hierarchical 2
Silhouette   0.8420 hierarchical 2
```

Figure 4.32: K-means and Hierarchical clustering validation results

```
Clustering Methods:
 hierarchical kmeans

Cluster sizes:
 2 3 4 5 6 7 8

Validation Measures:
                                 2       3       4       5       6       7       8

hierarchical Connectivity   2.9290  6.7869 10.3115 12.3115 17.6575 25.3881 26.5881
             Dunn           0.3565  0.5240  0.4842  0.4842  0.3740  0.3449  0.3449
             Silhouette     0.4443  0.3558  0.4765  0.4465  0.3754  0.3761  0.3520
kmeans       Connectivity   3.5246 13.0234 12.5444 14.5444 21.6647 25.3881 29.7786
             Dunn           0.2765  0.0849  0.4418  0.4418  0.1994  0.3449  0.3449
             Silhouette     0.5180  0.3839  0.4692  0.4368  0.3792  0.3761  0.3346

Optimal Scores:

             Score Method       Clusters
Connectivity 2.929 hierarchical 2
Dunn         0.524 hierarchical 3
Silhouette   0.518 kmeans       2
```

Figure 4.33: K-means and Hierarchical clustering validation results of logarithmic values

silhouette values of the real values of the stream-flow dataset. Dunn index should be minimised but it is high unlike the Dunn index of the real values of the stream-flow dataset. As a result, cluster validation of the logarithmic values of the stream-flow dataset gives bad results.

CHAPTER 5

CONCLUSION AND FUTURE WORK

Basin clustering is important for many fields and one of these fields is security of the energy supply. Supply is predicted by use of yearly, seasonal and regional information. Yearly and seasonal information exist in the stream-flow dataset and regional information is the only missing part for researches. The aim of this thesis is to find an appropriate basin clustering by use of data mining clustering techniques.

In this thesis, previous approaches to basin clustering problem are inspected. The problem of previous approaches are detected. The clustering techniques which are used in previous researches are repeated and the reasons of failures of these techniques explained. After that, it is shown that structure based hierarchical clustering can be used for trend detection of the time series. The clustering methods of shape based and structure based clustering are explained by Montero and Vilar [55] on a sample dataset and it is proved that shape based clustering can be used to detect closeness of the geometric profiles of time series like magnitude of the values in the time series and structure based clustering can be used to detect underlying dependence structure of the time series like trend of the values in the time series. With the help of this information, correlation based hierarchical clustering is suggested for an accurate clustering since it does structure based clustering. Basin grouping is done by using the output of this method and visualization of suggested basin grouping is presented in the related section.

Longest common subsequence clustering and structure based hierarchical clustering deliver good results, and therefore these two methods are also applied to the hydroelectric power plants with dams in the hydroelectric energy production dataset. This

dataset contains one year (between 01.06.2014 to 30.06.2015) hourly electric production data of 75 pumped-storage type hydroelectric power plants and 311 run-of-river type hydroelectric power plants of Turkey. 25 of these points are the same points where the measurements of stream-flow dataset are done. For this reason, clustering methods are applied to this sub-sample dataset. After that, the whole hydroelectric power plants with dams in the hydroelectric energy plant dataset is used for clustering. Both structure based hierarchical clustering and longest common subsequence clustering failed on the hydroelectric energy production dataset. The reason of failure is that there is not much correlation between the electric production and stream-flow rate. Only in case of a long term decrease in the stream-flow, the water level in the dams can decrease as well and the stream-flow rate can affect the hydroelectric energy production. Structure based clustering is also applied to the 311 run-of-river type hydroelectric power plants, but the method does not supply meaningful results.

In conclusion, basin clustering of Turkey is done by use of structure based clustering technique on stream-flow dataset and aim of this thesis is achieved.

For the future work, the results can be checked by using a bigger dataset. Several small basins are missing in stream-flow dataset used in this project. Neighbourhood connections can be identified more accurately in this way. In hydroelectric energy production dataset, electric production of one year is used. The increase of this time interval may help to enhance the clustering results.

REFERENCES

[1] Stream. n.d In Wikipedia. Retrieved April 4, 2015, from http://en.wikipedia.org/wiki/Stream.

[2] Schlosser, P., and Denina, A. (2011). Hydro in europe: Powering renewables.

[3] Olden, J.D., Reidy Liermann, C.A., Pusey, B.J., and M.J. Kennard. (2009). Protocols for hydrologic classification and a review of Australian applications. Chapter 2 in *Ecohydrological regionalisation of Australia: a tool for management and science*. 28 pages. Technical Report PN22591, Land and Water Australia

[4] World Energy Council Turkish National Committee(2014, January). *Energy Report* 2013, ISSN:1301-6318, 118.

[5] Kahya, E., Demirel, M.C., and Bég, A.O. (2008), Hydrologic Homogeneous Regions Using Monthly Streamflow in Turkey, *Earth Sciences Research Journal*, 12(2), 181-193.

[6] Karabörk, Ç., and Kahya, E. (1999).Multivariate Stochastic Modeling of Streamflows in the Sakarya Basin (in Turkish). *Turkish Journal of Engineering and Environmental Sciences*, 23(2), 133-147.

[7] Kahya, E., and Karabörk, M.Ç. (2001). The Analysis of El Nino and La Nina Signals in Streamflows of Turkey. *International Journal of Climatology*, 21(10), 1231-1250.

[8] Demirel, M.C., 2004: *Cluster Analysis of Streamflow Data over Turkey*. M.Sc. Thesis, Istanbul Technical University, Istanbul.

[9] Kahya, E., Kalaycı, S., and Piechota, T. C. (2008). Streamflow regionalization: Case study of Turkey. *Journal of Hydrologic Engineering*, 13(4), 205-214.

[10] Olden, J. D., Kennard, M. J., and Pusey, B. J. (2012). A framework for hydrologic classification with a review of methodologies and applications in ecohydrology. *Ecohydrology*, 5(4), 503-518.

[11] Bower, D., Hannah, D. M., and McGregor, G. R. (2004). Techniques for assessing the climatic sensitivity of river flow regimes. *Hydrological Processes*, 18(13), 2515-2543.

[12] Harris, N. M., Gurnell, A. M., Hannah, D. M., and Petts, G. E. (2000). Classification of river regimes: a context for hydroecology. *Hydrological Processes*, 14(16-17), 2831-2848.

[13] Hannah, D. M., Kansakar, S. R., Gerrard, A. J., and Rees, G. (2005). Flow regimes of Himalayan rivers of Nepal: nature and spatial patterns. *Journal of Hydrology*, 308(1), 18-32.

[14] Gottschalk, L. (1985). Hydrological regionalization of Sweden. *Hydrological Sciences Journal*, 30(1), 65-83.

[15] Krasovskaia, I., Arnell, N. W., and Gottschalk, L. (1994). Flow regimes in northern and western Europe: development and application of procedures for classifying flow regimes. *IAHS Publications-Series of Proceedings and Reports-Intern Assoc Hydrological Sciences*, 221, 185-192.

[16] Krasovskaia, I. (1997). Entropy-based grouping of river flow regimes. *Journal of Hydrology*, 202(1), 173-191.

[17] Kachroo, R. K., Mkhandi, S. H., and Parida, B. P. (2000). Flood frequency analysis of southern Africa: I. Delineation of homogeneous regions. *Hydrological sciences journal*, 45(3), 437-447.

[18] Kahya, E., and Demirel, M. C. (2007). A Comparison of low-flow clustering methods: streamflow grouping. *Journal of Engineering and Applied Sciences*, 2(3), 524-530.

[19] Kahya, E., and Kalaycı, S. (2004). Trend analysis of streamflow in Turkey. *Journal of Hydrology*, 289(1), 128-144.

[20] Demirel, M. C., Mariano, A. J., and Kahya, E. (2007). Performing k-means analysis to drought principal components of Turkish rivers. *Hydrology days*, 1, 145-151.

[21] Işık, S., and Singh, V. P. (2008). Hydrologic regionalization of watersheds in Turkey. *Journal of Hydrologic Engineering*, 13(9), 824-834.

[22] Işık, S., Turan, A., and Dogan, E. (2006). Classification of river yields in Turkey with cluster analysis. In *Proc., EWRI 2006 World Environmental and Water Resources Congress*.

[23] Turan, A. (2005). *Türkiye akarsu verimlerinin küme analizi ile sınıflandırılması*. Sakarya Üniversitesi. Fen Bilimleri Enstitüsü Yüksek Lisans Tezi, 155s.

[24] Dikbaş, F., Fırat, M., Koç, A. C., and Güngör, M. (2013). Defining homogeneous regions for streamflow processes in Turkey using a K-means clustering method. *Arabian Journal for Science and Engineering*, 38(6), 1313-1319.

[25] Kahya, E., Demirel, M. C., and Piechota, T. C. (2007). Spatial grouping of annual streamflow patterns in Turkey. *27th AGU Hydrology Days*, 169-176.

[26] Özfidaner, M. (2007). *Türkiye yağış verilerinin trend analizi ve nehir akımları üzerine etkisi*. ÇÜ Fen Bil. Enstitüsü Tarımsal Yapılar ve Sulama Anabilim Dalı, Yüksek Lisans Tezi, (3061).

[27] Bayazıt, M., Cığızoğlu, H. K., and Önöz, B. (2002). Türkiye akarsularında trend analizi, *Türkiye Mühendislik Haberleri* (pp. 420-421). Sayı 420-421-422, 4-6.

[28] Yıldız, M., Saraç, M. (2008). Türkiye Akarsularındaki Akımların Trendleri ve Bu trendlerin Hidroelektirk Enerji üretimine Etkileri, VII. Ulusal Temiz Enerji Sempozyumu, İstanbul, 17-19 Dec 2008

[29] Cığızoğlu, H. K., Bayazıt, M., and Önöz, B. (2005). Trends in the maximum, mean, and low flows of Turkish rivers. *Journal of Hydrometeorology*, 6(3), 280-290.

[30] Yanık, B. (2004). *Doğal akışlı hidroelektrik potansiyelin belirlenmesinde bölgesel analiz yaklaşımı* (Doctoral dissertation).

[31] Lundager, J. J., Dan, L., Reijo, S., and Arne, T. (1979). Hydrologic regions in the Nordic countries. *Nordic hydrology*, 10(5), 273-286.

[32] Lins, H. F. (1985). Streamflow variability in the United States: 1931-78. *Journal of Climate and Applied Meteorology*, 24(5), 463-471.

[33] Jowett, I. G., and Duncan, M. J. (1990). Flow variability in New Zealand rivers and its relationship to in-stream habitat and biota. *New Zealand journal of marine and freshwater research*, 24(3), 305-317.

[34] Stahl, K. (2001). *Hydrological drought-a study across Europe* (Doctoral dissertation, Universitätsbibliothek Freiburg).

[35] Mkhandi, S. and Kachroo, S. (1997). Regional flood frequency analysis for Southern Africa. *Southern African FRIEND, Technical Documents in Hydrology*, (15).

[36] Mosley, M. P. (1981). Delimitation of New Zealand hydrologic regions. *Journal of Hydrology*, 49(1), 173-192.

[37] Gubareva, T. S. (2012). Classification of river basins and hydrological regionalization (as exemplified by Japan). *Geography and Natural Resources*, 33(1), 74-82.

[38] Tibshirani, R. (2013) Hierarchical clustering [PDF document] Retrieved from Lecture Notes Online Web site: http://www.stat.cmu.edu/∼ryantibs/datamining/

[39] Manning, C. D., Raghavan, P., and Schütze, H. (2008). Introduction to information retrieval (Vol. 1, p. 496). Cambridge: Cambridge university press.

[40] Rani, Y., and Rohil, D. H. A. (2013). Study of Hierarchical Clustering Algorithm. *International Journal of Information and Computation Technology*. ISSN, 0974-2239.

[41] Hierarchical Clustering. n.d. Retrieved July 30, 2015, from http://www.saedsayad.com/clustering_hierarchical.htm.

[42] Dynamic Time Warping. n.d In Wikipedia. Retrieved August 1, 2015, from https://en.wikipedia.org/wiki/Dynamic_time_warping.

[43] Kuzmanić, A., and Zanchi, V. (2007). Hand shape classification using DTW and LCSS as similarity measures for vision-based gesture recognition system. In *EUROCON, 2007. The International Conference on "Computer as a Tool"* (pp. 264-269). IEEE.

[44] Corradini, A. (2001). Dynamic time warping for off-line recognition of a small gesture vocabulary. In *Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time Systems, 2001. Proceedings. IEEE ICCV Workshop on* (pp. 82-89). IEEE.

[45] Vial, J., Noçairi, H., Sassiat, P., Mallipatu, S., Cognon, G., Thiébaut, D., ... and Rutledge, D. N. (2009). Combination of dynamic time warping and multivariate analysis for the comparison of comprehensive two-dimensional gas chromatograms: application to plant extracts. *Journal of Chromatography A*, 1216(14), 2866-2872.

[46] Müller, M., Mattes, H., and Kurth, F. (2006). An Efficient Multiscale Approach to Audio Synchronization. In *ISMIR* (pp. 192-197).

[47] Gu, J., and Jin, X. (2006). A simple approximation for dynamic time warping search in large time series database. In *Intelligent Data Engineering and Automated Learning–IDEAL 2006* (pp. 841-848). Springer Berlin Heidelberg.

[48] Euachongprasit, W., and Ratanamahatana, C. A. (2008). Efficient multimedia time series data retrieval under uniform scaling and normalisation. In *Advances in Information Retrieval* (pp. 506-513). Springer Berlin Heidelberg.

[49] Müller, M. (2007). *Information retrieval for music and motion* (Vol. 2). Heidelberg: Springer.

[50] Keogh, E., and Ratanamahatana, C. A. (2005). Exact indexing of dynamic time warping. *Knowledge and information systems*, 7(3), 358-386.

[51] Senin, P. (2008). Dynamic time warping algorithm review. *University of Hawaii.*

[52] Longest Common Subsequence. n.d In Wikipedia. Retrieved August 1, 2015, from https://en.wikipedia.org/wiki/Longest_common_subsequence_problem

[53] R (programming language) n.d In Wikipedia. Retrieved August 21, 2015, from https://en.wikipedia.org/wiki/R_(programming_language)

[54] R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from http://www.R-project.org/.

[55] Montero, P., and Vilar, J. A. (2014). TSclust: An R Package for Time Series Clustering. *Journal of Statistical Software*, 62(1), 1-43. Retrieved from http://www.jstatsoft.org/v62/i01/.

[56] Brock, G., Pihur, V., Datta, S. [Susmita], and Datta, S. [Somnath] (2008). clValid: An R Package for Cluster Validation. *Journal of Statistical Software*, 25(4), 1-22. Retrieved from http://www.jstatsoft.org/v25/i04/.

[57] Achim Zeileis and Gabor Grothendieck (2005). zoo: S3 Infrastructure for Regular and Irregular Time Series. *Journal of Statistical Software, 14(6), 1-27*. Retrieved from http://www.jstatsoft.org/v14/i06/

[58] Christophe Genolini, Xavier Alacoque, Marianne Sentenac, Catherine Arnaud (2015). kml and kml3d: R Packages to Cluster Longitudinal Data. *Journal of Statistical Software, 65(4), 1-34*. Retrieved from http://www.jstatsoft.org/v65/i04/.

[59] Giorgino T (2009). "Computing and Visualizing Dynamic Time Warping Alignments in R: The dtw Package. *Journal of Statistical Software, 31(7), pp. 1-24*. Retrieved from http://www.jstatsoft.org/v31/i07/.

[60] Usue Mori, Alexander Mendiburu and J.A. Lozano (2015). TSdist: Distance Measures for Time Series Data. R package version 2.2. Retrieved from http://CRAN.R-project.org/package=TSdist

APPENDIX A

FLOW RATE OF STREAMS

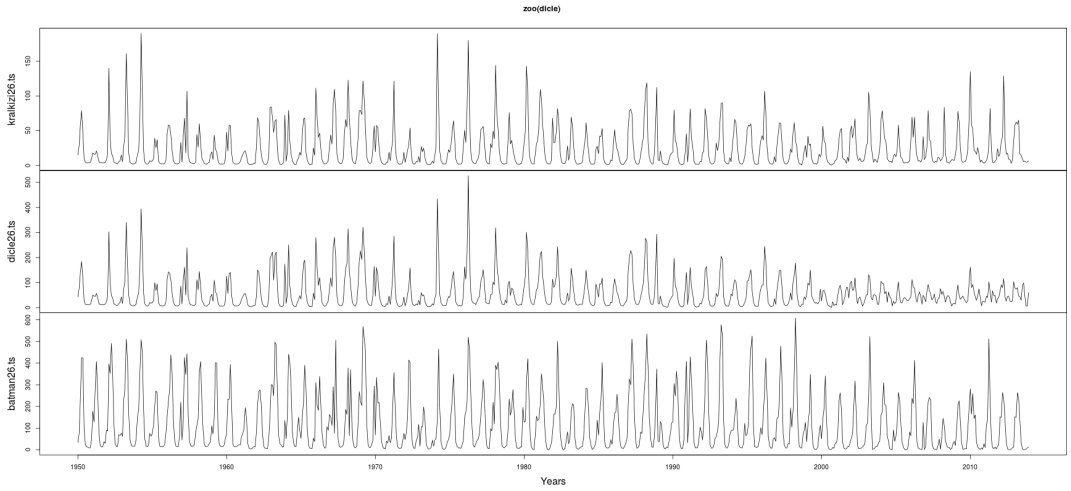

Figure A.1: Plot of Flow Rate vs. Years in Dicle Basin

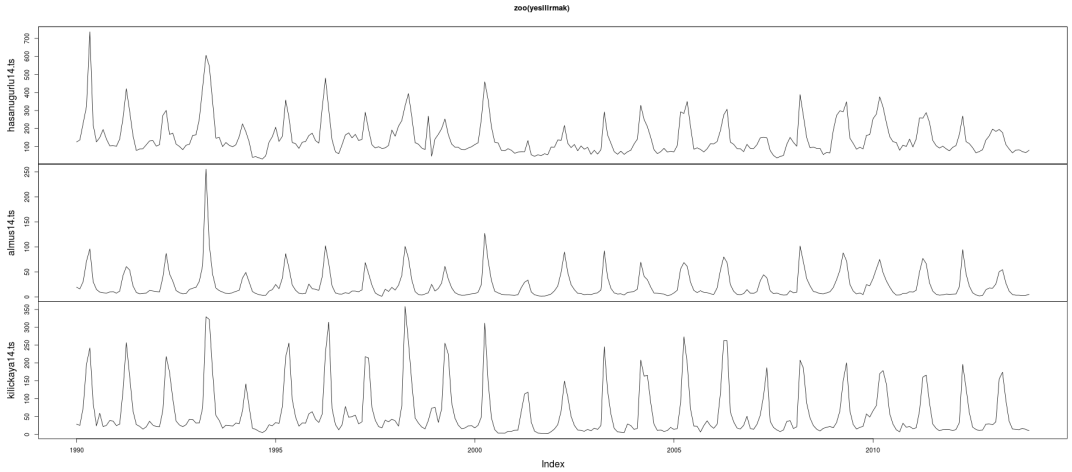

Figure A.2: Plot of Flow Rate vs. Years in Yeşilırmak Basin

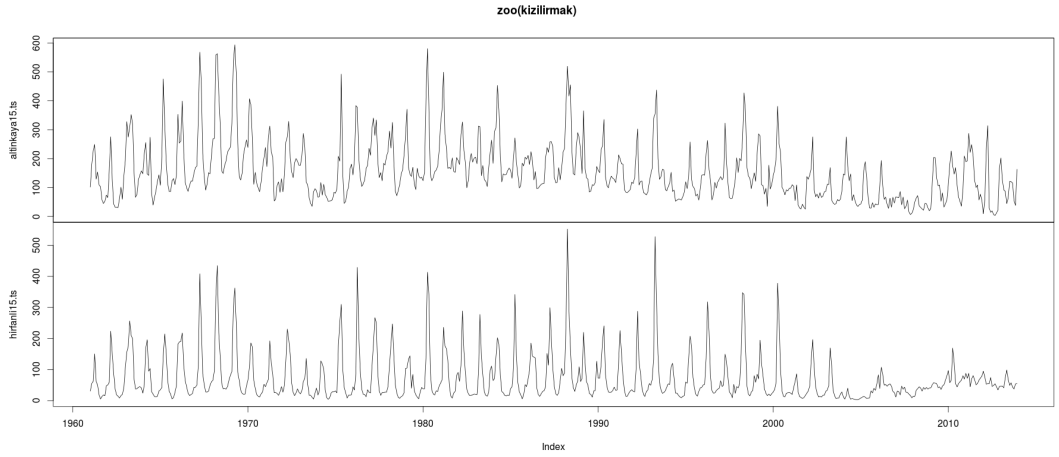

Figure A.3: Plot of Flow Rate vs. Years in Kızılırmak Basin

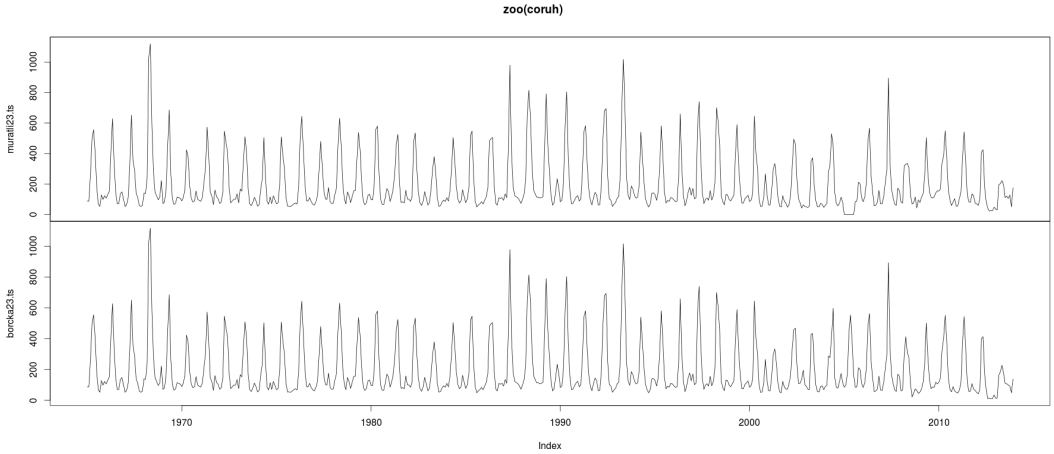

Figure A.4: Plot of Flow Rate vs. Years in Çoruh Basin

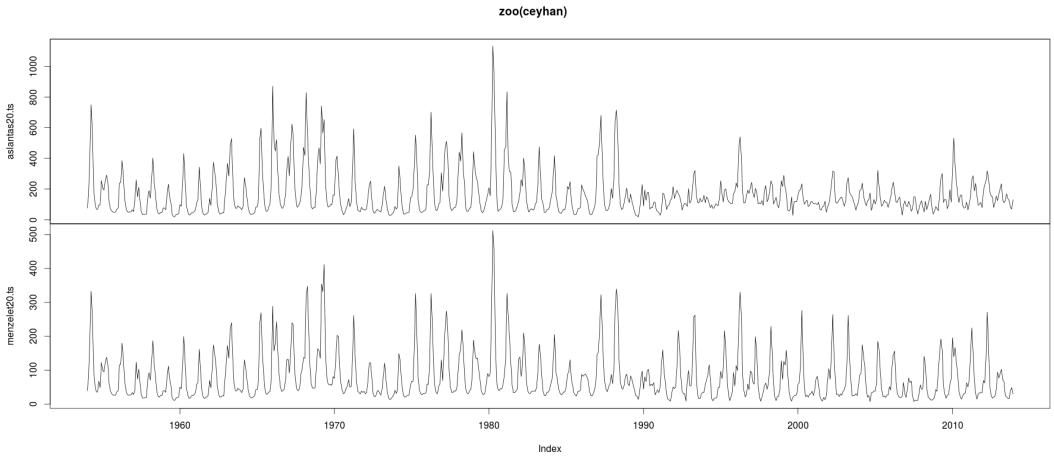

Figure A.5: Plot of Flow Rate vs. Years in Ceyhan Basin

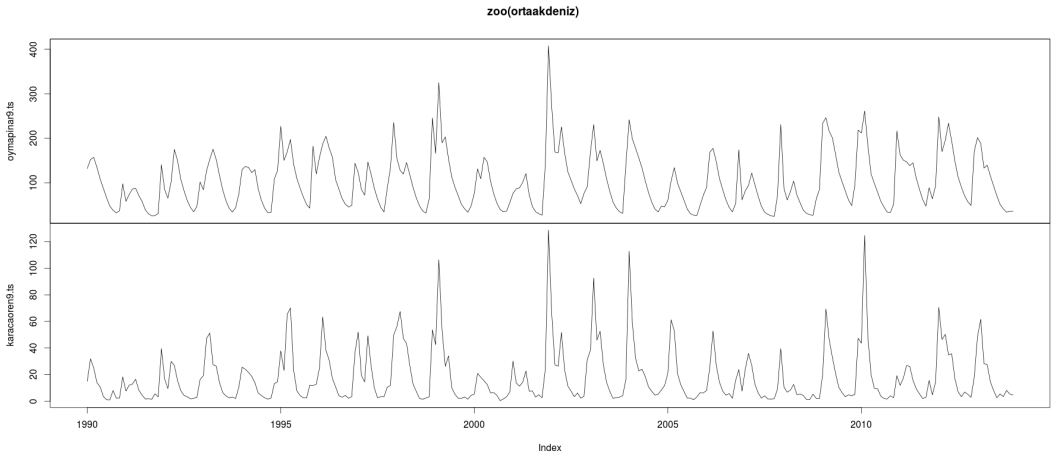

Figure A.6: Plot of Flow Rate vs. Years in Orta Akdeniz (Antalya) Basin

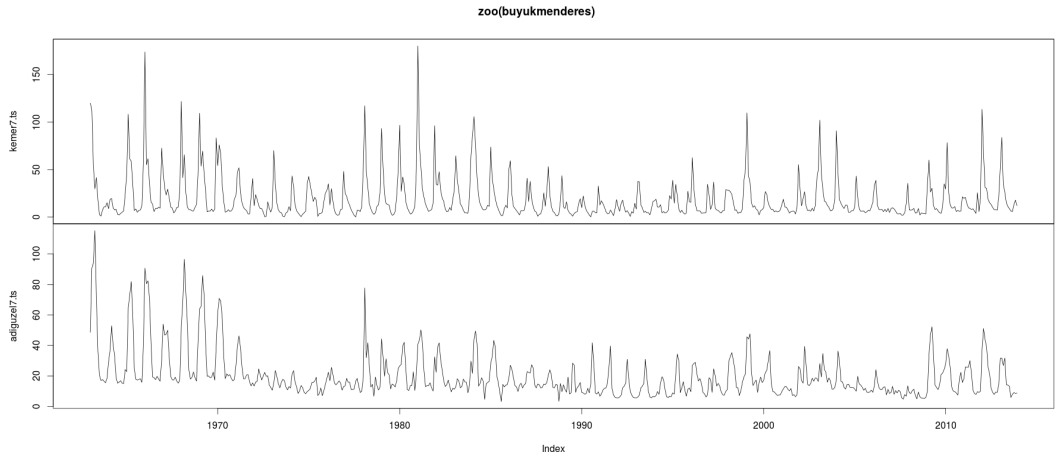

Figure A.7: Plot of Flow Rate vs. Years in Büyük Menderes Basins

APPENDIX B

INFORMATION ABOUT R PACKAGES

"Zoo" [57] package of the R is used to plot the time series of the streams. The package automatically arrange the time axis based on the shortest records between all the time series.

"kml" [58] package is used for clustering longitudinal data. "kml" function in the package is used for k-means clustering of data. Kml function takes the dataset, sets the k values 2, 3, 4, 5 and 6 in default and applies the clustering algorithm 20 times. The result is shown graphically and the suggested number of the cluster for the dataset is also shown with a black dot on the plot by use of "choice" function. The cluster of each longitudinal object for each k values are stated by use of "getCluster" function.

R has "TSclust" [55] package for hierarchical clustering of the time series. The evaluation of the cluster assignment can also be done by use of this package. It has several dissimilarity measures for time series clustering. It has model-free clustering approaches, model-based approaches, complexity-based approaches and prediction-based approaches and each approaches contains several methods.

"TSclust" package has a distance function and can be used to calculate it by use of dynamic time warping as a distance measure. Besides, R has another package called "DTW" [59]. This package also has a distance function and calculates the distance matrix of the time series dataset by use of dynamic time warping as a distance measure. Therefore, two different packages of the R can be used for dynamic time warping clustering.

The cluster assignment difference between two methods can be checked by use of "cluster.evaluation" function of the TSclust package. This function returns a value

between 1 and 0. Higher values indicate a high resemblance between two clustering results since 1.0 shows that clustering results of the two methods are identical.

“ward.D2” is one of the linkage method of TSclust package. TSclust package offers two ward methods. “ward.D” methods should be used with the square of the distance matrix and “ward.D2” function can be used directly with distance matrix in “hclust” function. Therefore, if distance matrix parameter is appropriately set then the result of two function will be same.

R has “TSdist” [60] package and this package has “tsDatabaseDistances” function which accepts a time series dataset and distance matrix (e.g. LCSS, Fourier, etc.) as argument and returns the distance matrix of the dataset.