# Clustering of Hydrological Stations

## A Cluster Analysis Based on Temporal Discharge and Temperature Data of the Swiss River Network

Bachelor Thesis

Faculty of Science, University of Bern

submitted by

**Jan Andrin Zurbrügg**

from Solothurn, Switzerland

Supervision:

PD Dr. Kaspar Riesen
Institute of Computer Science (INF)
University of Bern, Switzerland

# Abstract

The aim of this thesis is to divide hydrological stations into groups of similar discharge and temperature behaviour. The resulting groups can be used to compensate for missing values and to improve computational time in a deep learning project, that forecasts river temperatures. The groups are determined through an analysis of the discharge and temperature data of 75 hydrological stations in Switzerland. Each station is characterized by a feature vector derived from this data. The features are divided into subsets of different size and on each of them a k-means and a hierarchical clustering is performed. To determine the best cluster composition with the optimal number of clusters, the clusterings are evaluated with the silhouette and davies-bouldin score. With these scores the hierarchical clustering on the additional features from both data-sets and the k-means clustering on all features from both data-sets are recommended to be used in the deep learning project. Both of these clusterings result in seven clusters, but with a different composition. Therefore the results are not unequivocal but show a stable tendency in the composition of the groups and should both be tested.

# Contents

# Chapter 1

# Introduction

This chapter motivates the need for hydrological modelling and presents a brief overview of different modelling methods. It locates the work of B. Fankhauser [1] in the field and explains the connection it has to the presented thesis. The specific research questions are stated in section 1.2. Section 1.3 elaborates on how the questions are approached and gives an outline of the thesis.

## 1.1 Broad Perspective on Hydrological Modelling

Rivers are of great importance for the natural environment and for human societies all around the world. As human population we use rivers for freshwater supply, to generate hydro power, to cool nuclear power stations and to irrigate plants in agriculture and many more. From an ecological point of view, rivers are a complex and diverse ecosystem, serving as a habitat for many species and are essential for biodiversity. Some of these species are sensible to a change of water temperature. This water temperature is impacted by natural factors, such as snow or glacier melting, rainfall, ground water inflow and the rate of discharge. Through climate change this coomplex system has changed [2]. Especially the regime shift in the 1980's is significant and caused by anthropogenic and natural influences [3] .
Because the river temperature has a great impact on the ecosystem, the task of predicting water temperature is crucial. The research field that deals with these predictions is called hydrological modelling. There are many different approaches and models for different use cases. In the field of data driven models, B. Fankhauser [1] presents a novel approach, using graph based deep learning.

## 1.2 Specific Research Questions

During the project of B. Fankhauser [1] the question arose if the measurement stations could be clustered into groups of similar behaviour, with respect to water discharge and temperature. This clustering could be used to compensate for missing data of certain stations. This means that the model could be used in regions where less data is accessible than in Switzerland.
The Swiss river network is well suited for such a clustering task, because a large amount of data captures the behaviour of the river system very accurately, over a long period of time. Additionally due to the geographic diversity, with alpine regions and flat-lands, the Swiss river system consists of many different river types, from large streams to small mountain creeks. Another advantage of the clustering is that the neural network could be trained on the groups of stations. This would reduce computational cost significantly in comparison to training the network on every station separately.

The above described task states a classical machine learning clustering problem. These Problems fall into the field of unsupervised learning, because the outcome of the clustering is not predefined, in other words there is no ground truth. Specifically we have to search for an unknown pattern in the discharge and temperature data. In our case, in addition to not knowing how the clusters look like the number of clusters is not predefined as well. This leads to the following research questions:

- How is a station characterized ?
- What is the optimal number of clusters ?
- How are the clusters composed?

## 1.3 Structural Roadmap to Clustering Hydrological Stations

To answer the questions above, the following steps are taken. First the data is explored in terms of periodicity and general behaviour. This leads to a representation of a typical yearly discharge and temperature course for every station. These courses are then visually inspected and features are calculated based on this data, to find a n-dimensional representation for all stations. We use this representation to apply two widely used clustering algorithms to the problem. To determine what the optimal number of clusters and the more fitting algorithm is, two clustering metrics are calculated. After this the results are presented and two clusterings are recommended to be used in the deep learning project.

# Chapter 2

# Feature Engineering through a Visual Inspection

This chapter deals with the first research question: How is a station characterized? In section 2.1 the data is analyzed and prepared for feature calculation, which is explained in sections 2.2 and 2.3. In these sections the features are derived by visually inspecting the data. This results in a characterization of each station with 38 features.

The data used for this task consists of measurements taken at 75 hydrological stations in Switzerland, each characterized by a four digit unique ID number the associated stations can be found in tables A.1 ans A.2. In figure 2.1 the locations of all stations are shown on a map. The stations are distributed all around Switzerland and cover most areas with flowing water. For every station an average discharge rate and an average temperature is measured each day. These measurements were taken over the course of 20 to 40 years from 1980 up to 2021. This data serves as the foundation for the search for clusters.

The following practical part of the feature engineering was done using `python` as a programming language with the libraries `matplotlib.pyplot` [4] for plotting the data and `pandas` [5] for data analysis.

## 2.1 Data Preparation for Yearly Characteristics

### Discharge data

In figure 2.2 the discharge course of a typical station is illustrated. Every day the measured average discharge is denoted in $m^3/s$. A periodic pattern is visible with low water discharge in the winter months and increased discharge during the summer half year. This pattern occurs for many stations in the given data-set. To

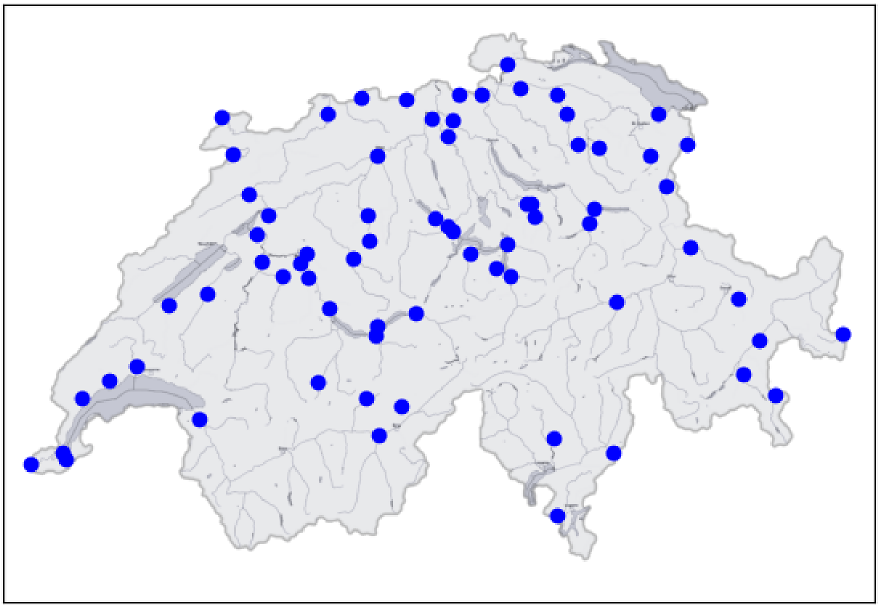


Figure 2.1: Locations of the measurement stations in Switzerland [6].

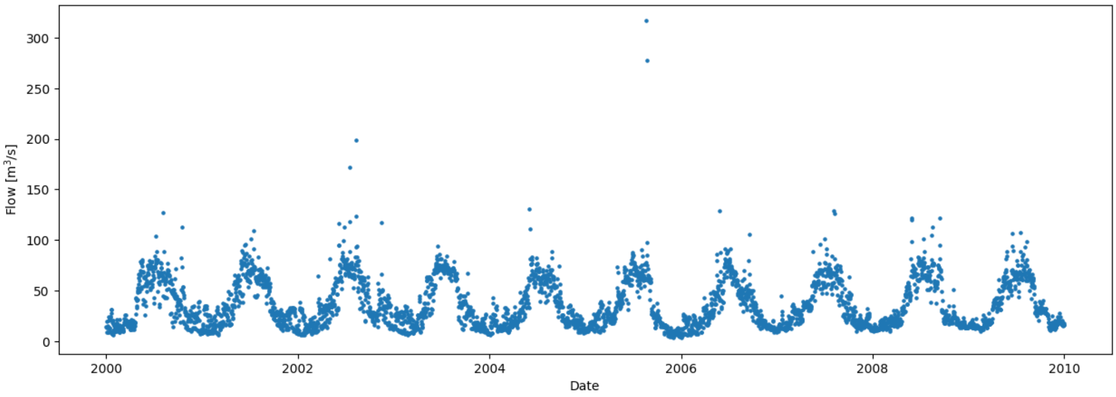
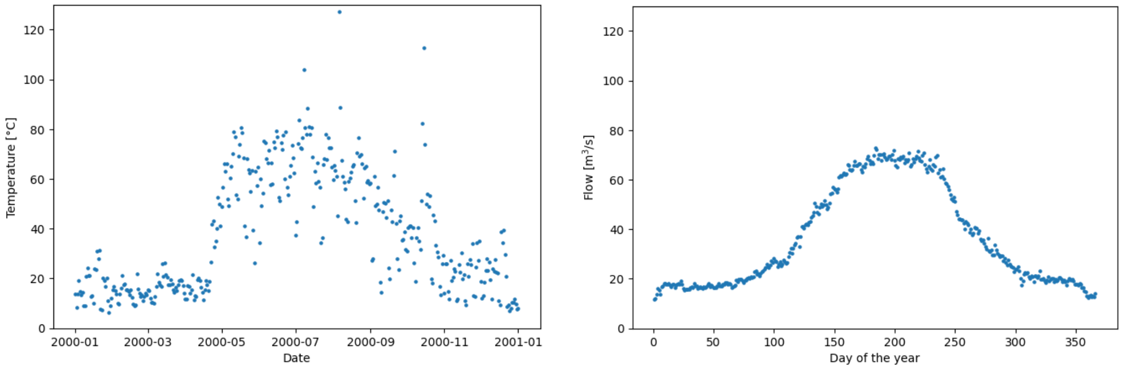


Figure 2.2: The course of average water discharge over a 10 years period for station no. 2019.



(a) Example year

(b) Averaged by day of the year

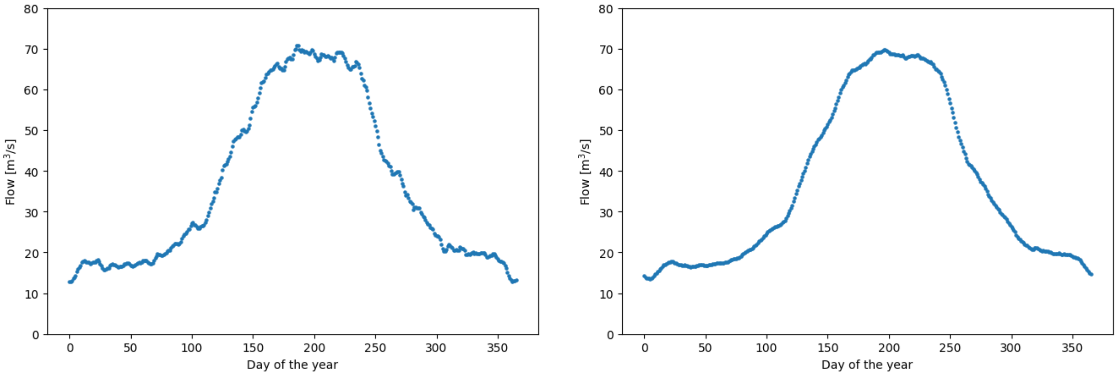
Figure 2.3: The course of water discharge for station no. 2009.

extract characteristic information of the given stations it makes sense to look at the discharge regime. To get the typical behaviour of the rivers discharge the data was averaged by the day of the year. The resulting discharge curve is shown in figure 2.3 on the right. On the left the original data of the year 2000 is plotted. In comparison to an arbitrary year the average years discharge course is smoother, this means extreme events do not effect the curve as much as in the original data. The resulting curve still has some fluctuations which could lead to problems when calculating sensitive features. For example if the day of the maximum is defined by an extreme event rather than the typical day of the year when the discharge is generally high.

To rule out such effects that can be triggered by outliers, the data runs through one more step of data preparation. In this step a window averaging function with the length of 15 days is applied to the averaged data [7]. The function used to calculate this window averaging function is displayed in the following code fragment.

```
def get_running_mean_df(station_number, window, flow_temp_data, Wert):
    daily_averaged_data = get_daily_averaged_df(station_number, flow_temp_data)
    # add the last days of the year to account for periodicity.
    expanded_data =
    ↪  daily_averaged_data.iloc[-window+1:].append(daily_averaged_data)
    expanded_data =
    ↪  expanded_data.rolling(window).mean().dropna().reset_index(drop=True)
    return expanded_data
```

Notice that the periodicity of the year is taken into account in the fourth line such that there still are 365 values per year. This is important as we will see in section 3.1 where the periodicity of the year has an effect on the normalization of the features.



(a) Window length of 5 days

(b) Window length of 15 days

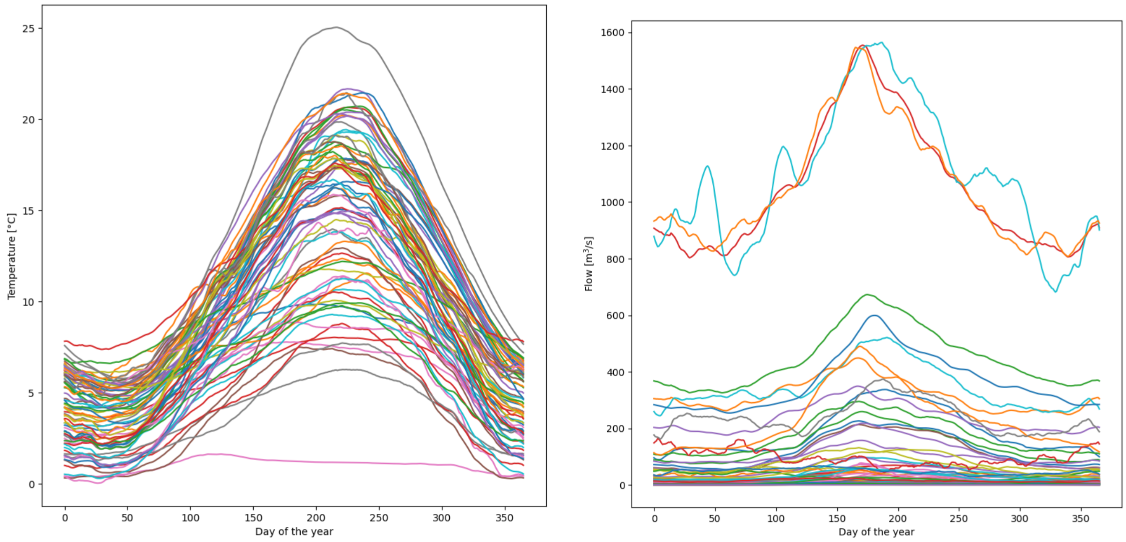
Figure 2.4: The course of the window averaged discharge of station no. 2019.

The result of the window averaging function is shown in figure 2.4. In figure 2.4a it is visible that with a window length if five days the curve still has small fluctuations. In contrast when the window length is set to 15 days the curve is smooth and better suited for the calculation of fluctuation sensitive features, as can be seen in figure 2.4b. Therefore we will use a window length of 15 days. This prepared discharge data for all stations is displayed in figure 2.5b.

Note that some stations do not follow this periodic pattern. This can be caused by effects such as water release from dams or a short-time increase in rainfall. Rivers with a lower average dicharge are far more sensitive to such events. We will take this behaviour into account in section 2.3 with a feature called similarity to mean. With this feature we will distinguish between periodic and aperiodic stations. In figure 2.9 an aperiodic discharge curve is shown.

### Temperature data

Regarding the temperature data the same steps for data preparation are applied as for the discharge data. For a review of the steps applied to the temperature data see Appendix A figures A.1, A.2 and A.3. It is to note that the temperature data is much less diverse than the discharge data as can be seen in figure 2.5. Almost all curves follow the typical temperature course with low temperature in the winter months and high temperature in the summer, with a continuous increase and decrease in between. This can be seen in figure 2.5a, where all the stations temperature data is shown. Additionally we see that no aperiodic temperature courses occur.



(a) Temperature

(b) Discharge

Figure 2.5: The prepared data of all 75 stations.

## 2.2 Definition of Basic Features

The so called basic features are basic statistical values calculated on the window averaged data described in section 2.1. In the following two sections those basic features are motivated by examples and explained.

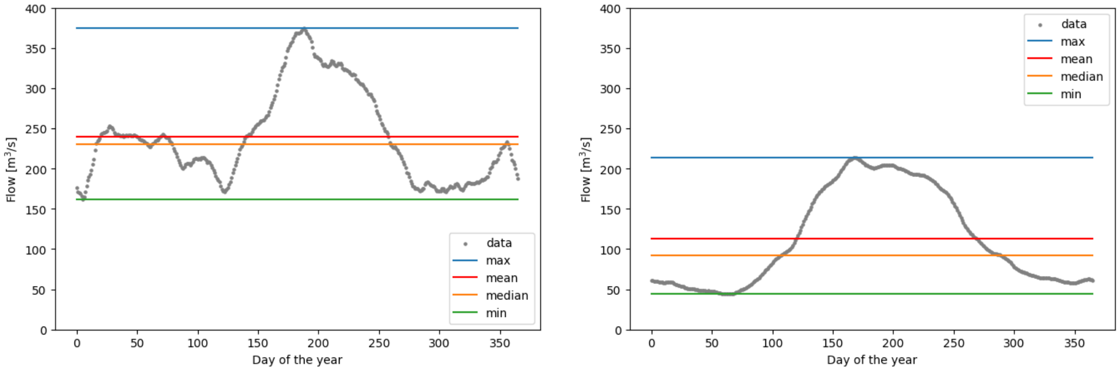
### 2.2.1 Discharge Data

Considering the discharge data, first we show how the basic features can contribute to a distinction between different stations. In figure 2.6 four of the basic features are illustrated on discharge curves of two stations. In figure 2.6 it is visible the maximum, minimum, mean and median make the two stations distinguishable.

In addition to the four displayed features two more features are counted to the basic features group. Namely the standard deviation and the range of the data. This results in six basic features on the discharge data. Table 2.1 shows an overview over the basic features.

### 2.2.2 Temperature Data

Because the course of the temperature looks similar to a typical discharge curve, as can be seen in figure 2.5, it makes sense to use the same basic features as are used for the discharge data-set. Therefore the six basic features are added on the temperature dataset as well.



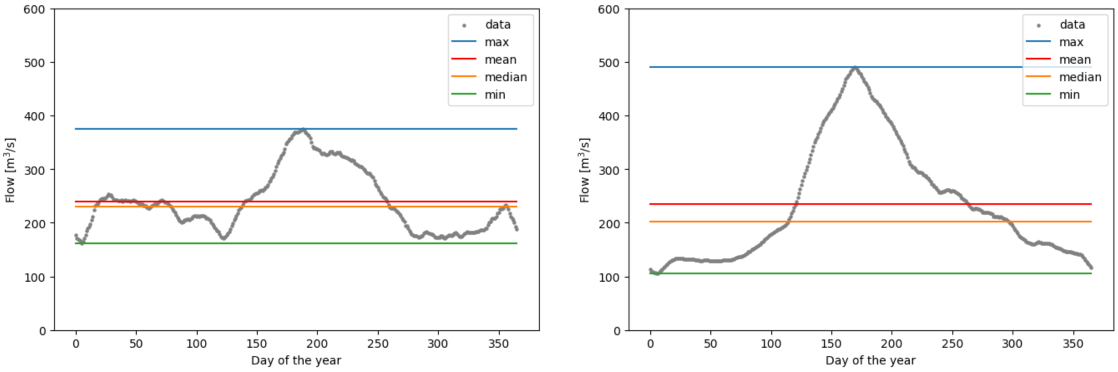
(a) Station no. 2606

(b) Station no. 2030

Figure 2.6: Prepared discharge data with four basic features plotted as indicated by the labels.

| Basic features | |
|---|---|
| Maximum | The maximum value of the prepared data |
| Minimum | The minimum value of the prepared data |
| Mean | The average value of the prepared data |
| Median | The median value of the prepared data |
| Standard Deviation | standard deviation of prepared data |
| Range | Range between max and min of the prepared data |

Table 2.1: Basic features for prepared data, calculated on the discharge and temperature data-set.



(a) Station 2606

(b) Station 2473

Figure 2.7: Prepared discharge data with four basic features plotted as indicated by the labels.

## 2.3 Definition of Additional Features

### 2.3.1 Discharge Data

Some stations have different characteristics but their values for the basic features are similar. In figure 2.7 we see that the two curves have three similar basic features although the discharge behaves different. For example the curve on the left has more than one peak and crosses the mean multiple times, where as the station plotted on the left has one very sharp peak and crosses the mean twice. To capture these and other characteristic behaviour, additional features are calculated. They are listed in table 2.2. For the ones that are not trivial a more detailed description is given in the following subsections.

#### Number of peaks

The feature number of peaks counts the number of local maxima of a discharge curve. To ignore small fluctuations a minimal distance between two peaks is set to be equal to 7 days. This means if two peaks are recorded within one week, only one of them is counted. This feature allows to make a clear distinction between curves, like the ones in figure 2.7. To implement this feature the method `signal.find_peaks()` from the `scipy` library [8] was used with the parameter `distance = 7`.

#### Slope

The two features: minimal slope and maximal slope are derived from the same calculations. Based on the prepared data the slope of the curve is calculated from two data points that are five days apart. From all these approximated slopes the smallest and the biggest slopes are taken as a feature. The difference of 5 days is

| Additional features | |
|---|---|
| Time above average | Number of days above the average of a stations data |
| Day of maximum | Day of the year the maximum is reached |
| Day of minimum | Day of the year the minimum is reached |
| Number of peaks | Number of peaks |
| Minimal slope | Slope measured for two days five days apart |
| Maximal slope | Slope measured for two days five days apart |
| Day upward crossing mean | Day of the year the mean is first crossed after the minimum |
| Day downward crossing mean | Day of the year the mean is first crossed after the maximum |
| Day upward crossing high quartile | Day of the year the upper quartile boundary is first crossed in after the minimum |
| Day downward crossing high quartile | Day of the year the upper quartile boundary is first crossed after the maximum |
| Day upward crossing low quartile | Day of the year the lower quartile boundary is first crossed after the minimum |
| Day downward crossing low quartile | Day of the year the lower quartile boundary is first crossed after the maximum |
| Similarity to mean | Added up differences from every years data and the mean year |

Table 2.2: Additional features for prepared data, calculated on the discharge and temperature data-set.

chosen to get a slope that is consistent over multiple days and therefore results in a trend, rather than a maximum steepness that rarely occurs.

### Day of Crossing a Given Value

To characterize the discharge regime more accurately another set of features called the "day of the year" features are added. Their values are the day of the year a given value is crossed in upward or downward direction. They are used o distinguish between curves, like the ones displayed in figure 2.7. The new added features consist of two features for crossing the mean, the high and low quartile. Resulting in six more features. They are enlisted in table 2.2.
The quartiles are calculated in the following way:

$$low\_quartile = min + \frac{max - min}{4}$$

and

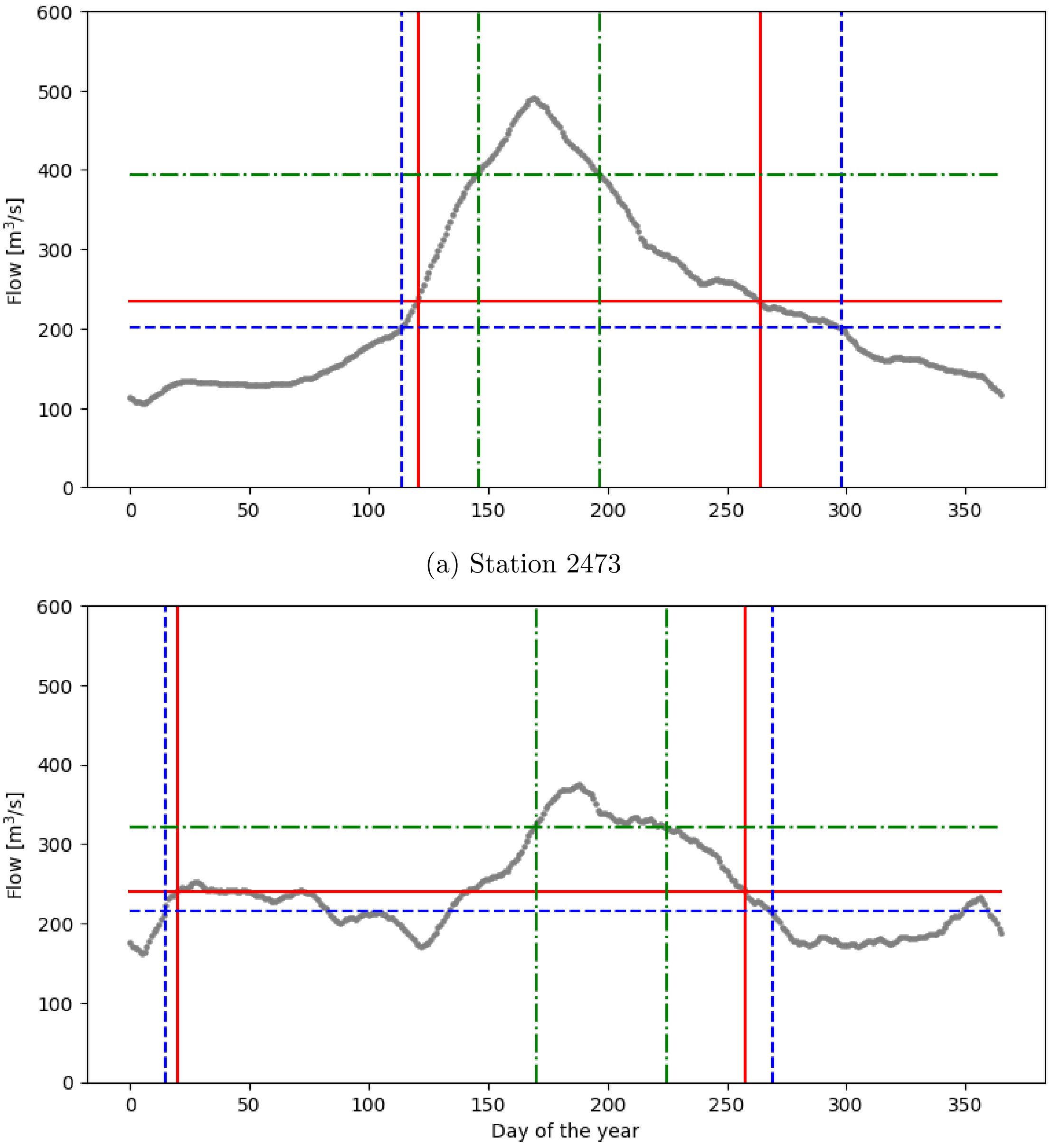
$$high\_quartile = max - \frac{max - min}{4}$$

To avoid ambiguity, due to multiple crossings of the given value a condition is formulated. To take the periodicity of the year into account and therefore eliminate an arbitrary cut in the year, the crossings were chosen with respect to the minimum or the maximum of the year. For the downward crossing feature this means that the first crossing of the value after the maximum is taken. For the upward crossing features the first crossing after the minimum is taken. Through this choice the features are clearly defined.

### Similarity to mean

One more additional feature is defined called the similarity to mean feature. This feature is introduced to take the variability of the data in comparison to the mean year into account. This means how similar the curve behaves in comparison to an average year. This feature is introduced especially to differentiate between periodic and aperiodic courses. In figure 2.9 such an aperiodic course is shown. The feature is calculated in the following way:

$$similarity\_to\_mean = \sum^{year} \sum^{day} mean\_discharge_{day} - discharge_{day}^{year}$$



(a) Station 2473

(b) Station 2606

Figure 2.8: The discharge curves are shown with the calculated ”day of the year” features. The solid red line indicates the day of crossing the mean, the dotted blue line the crossing of the low quartileand the dot-dashed green line the crossing of the high quartile.

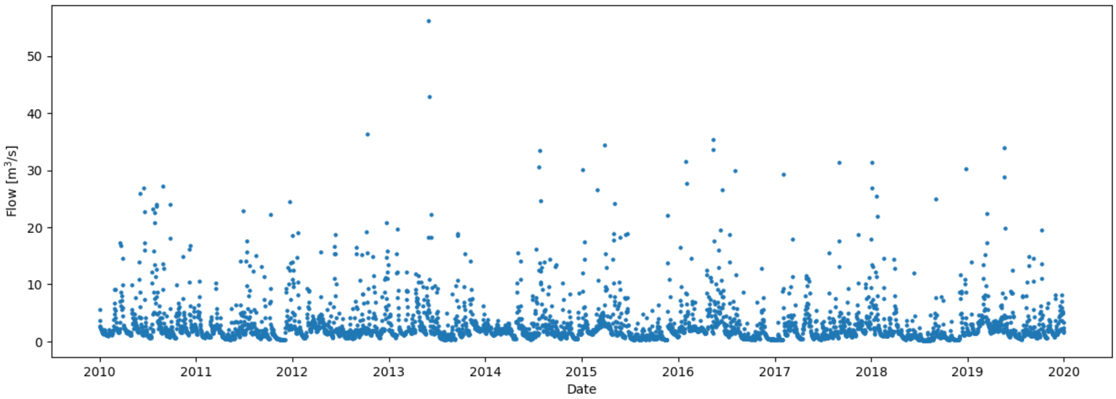


Figure 2.9: The course of average water discharge per second over a 10 years period of station no. 2374.

### 2.3.2 Temperature Data

The features are applied to the temperature data. This can be done because the temperature curves look similar to the discharge curves. This results in 19 features derived from the temperature data-set as well as 19 features derived from the discharge data-set. These 19 features are further divided into basic and additional features as is shown in tables 2.1 and 2.2. This results in an overall characterization of every station with 38 features.

# Chapter 3

# Clustering of Hydrological Stations

The goal of this chapter is to give a detailed explanation, on how the clustering is performed. In section 3.1 the specifics of how the features are normalized and preprocessed is described. Then in section 3.2 we elaborate on the used clustering algorithms and a description of the used evaluation metrics is presented in section 3.3. To visualize the clusters the features have to be mapped to two dimensions, therefore the multi dimensional scaling representation is introduced in section 4.1.

## 3.1 Preprocessing of the Features

Goal of the preprocessing is to prepare the features in a way such that they are most useful for the used machine learning algorithms. Due to the nature of the different features the preprocessing is done in two different ways. This so called featurewise normalization should lead to a better clustering performance in comparison to the use of one normalization method for all features [9]. In this case the "day of the year" features are normalized another way than the other statistical features.

### 3.1.1 Features Based on the Day of the Year

The "day of the year" features take on integer values between 1 and 365, namely the day of the year a given criteria is fulfilled. This representation of the data does not fully cover the reality, because it ignores the periodical repetition of the year and the predefined cut between December and January is irrelevant for our characterization. For example if Station A has its minimum at the 10th of January and station B has its minimum at the 20th of December they are considered very different by the calculated feature, but their behaviour is similar. To avoid this

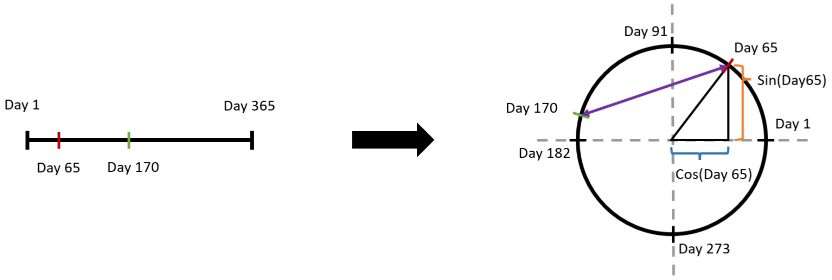


Figure 3.1: Illustration of the normalization conversion from linear to periodic year.

effect another representation of the data is introduced.

An good way in science to describe periodic patterns is to use the trigonometric functions sine and cosine. With them we will generate two new features that replace each already existing "day of the year" feature. The principle used to calculate the circular features is illustrated in figure 3.1.The illustration shows that instead of looking at a year as a straight line one can view a year as a circle in which each day is characterized by two parameters, specifically the sine and cosine. The exact calculations are shown in the following equations.

$$sin\_feat = sin(\frac{day\_fature \cdot 2\pi}{365})$$

$$cos\_feat = cos(\frac{day\_fature \cdot 2\pi}{365})$$

Through this calculations every "day of the year" feature is mapped to a sine and cosine value. This results in a circular representation of the year as is shown in figure 3.1. The distance between the points on the circle (denoted with the purple arrow) is a accurate measurement of their the similarity, independent of the years beginning. Because the sine and cosine only take on values in the range of 0 to 1 this also concludes the normalization step.

### 3.1.2 Statistical Features

All the other features have to be normalized as well because they are measured on different scales and are distributed differently. As a requirement we get that the resulting normalized features should land in the approximate range from -1 to 1, because this is the range that the "day of the year" features are mapped to and they should be evenly weighted. To achieve this the class `RobustScalar` from the `sklearn.preprocessing` is used to scale the data. This normalization method is chosen because it ignores outliers, which occured in some features. Therefore

| Feature subsets |
|---|
| Basic features temperature |
| Additional features temperature |
| All features temperature |
| Basic features discharge |
| Additional features discharge |
| All features discharge |
| Basic features both |
| Additional features both |
| All features both |

Table 3.1: List of all feature subsets that are used for the clustering basic additional and all refer to the used features. Discharge, temperature and both indicate which data-sets have been used for the calcualtions of the features.

through empirical evaluation it turned out that the best quartile range for the `RobustScalar()` constructor would be `quantile_range = (5,95)`.

## 3.2 Application of Clustering Algorithms

To cluster the stations we make use of two different machine learning clustering algorithms that operate on the given features. The clustering is done on different feature subsets. As described in section 2 the features are separated in basic and additional features for the discharge and temperature data. This results in four disjoint feature subsets. The clustering algorithms are applied to the feature subsets enlisted in table 3.1 separately.

On these subsets of features the clustering is performed with a k-means and a hierarchical clustering algorithm. Because the number of clusters is unknown, a clustering on each subset results in multiple clusterings with different numbers of clusters.

### 3.2.1 K-Means

The first of the two clustering methods is the k-means algorithm. It is a widely used clustering technique based on the concept of finding cluster centroids in an iterative manner. Because of this centroid based functionality the clusters boarders are spherical and not flexible at all. This method is well applicable to general purpose problems, where the distribution of feature values is unknown, as is the case for ur features.

Key characteristics of the algorithm are that the number of cluster has to be fixed

before the procedure is started and that the algorithm always converges to a local minima. This leads to an optimal solution in a given interval, but it does not have to be the best solution overall. This is due to the random initialisation of the cluster centroids [10].
To take this characteristics into account the implementation used for our problem, from the Class `sklearn.clustering.KMeans` [11] can be executed with certain parameters. To reduce the impact of the randomly chosen centroids the the parameter `init ='k-means++'` is set. This leads to the use of a slightly different version of the k-means algorithm which makes it less dependent on the initialization [12]. In addition to that the number of times the algorithm is run is set to `n_init = 10`. This means the algorithm is run ten times and the best result in terms of inertia is chosen [11]. In this way the influence the initialization has is reduced to a minimum. Another important parameter that has to be fixed is the number of clusters. As one does not know in advance what number is optimal for the given problem, the algorithm is run 40 times with `n_clusters = n` with $n = 1, ..., 40$.

### 3.2.2 Hierarchical

The second method is an agglomerative hierarchical clustering algorithm. It uses a bottom up approach, where each station starts in its own cluster and similar clusters are iteratively merged together. By the end resulting in one big cluster. This method can be stopped at any point and therefore the number of clusters can be chosen after performing the clustering. Because of this iterative merging approach the clusters can have an arbitrary shape. This is an important difference to the k-means algorithm. Additionally, when using the correct linkage parameter the mathematical goal of the algorithm is exactly the same as the one for the k-means. This makes the algorithms ideal to compare and to cross reference the clusterings [13].
We use the class `sklearn.cluster.AgglomerativeClustering` for implementation, which provides multiple agglomerative clustering options. the most important parameter is the linkage criteria that determines the metric that is chosen for merging two clusters together. For our application it is crucial to use the option `linkage = 'ward'` [14]. This linkage version minimizes the sum of squared differences in all clusters. This is the same goal the k-means algorithm pursues and therefore the results may be compared in a meaningful way [11].

## 3.3 Evaluation Metrics for Clustering

In section 3.2 we have defined how the clustering is performed and explained which feature subsets are used. This results in nine subsets, each clustered with the k-means and hierarcical approach. To evaluate these 18 clusterings the silhouette and davies-bouldin scores are calculated. These two evaluation metrics are presented in the following subsections. They both quantify how well the clusters are defined, based on the given feature subset.

### 3.3.1 Silhouette Score

The silhouette score is the mean silhouette coefficient of the clusters. The coefficient falls into a range between minus one and one, with one representing perfect clustering and minus one, very bad clustering. The coefficient depends on the intra-cluster mean distance, and the mean minimal distance to the next cluster. Therefore it takes into account how coherent the cluster is internally as well as how well separated it is from other clusters. Through the averaging of the inter-cluster distance the silhouette score takes an average scenario into account, this means that if for example only two clusters are not well separated the score would only be slightly worse and not very bad if the other clusters are separated well [15]. This is a key difference to the davies-bouldin score.

### 3.3.2 Davies-Bouldin Score

The davies-bouldin score measures the quality of clustering as well. It differs from the silhouette score in the characterization of the inter-cluster separation. The difference is that it not only considers the closest cluster but a combination of a close cluster with a bad intra-cluster mean distance. Therefore if clusters are mixed the davies-bouldin score detects this behaviour. The davies-bouldin score is limited to positive values without an upper limit. With a value close to zero indicating a good clustering [16].

## 3.4 Multidimensional Scaling

With the scores discussed in 3.3 we get a quantitative insight in the quality of clustering but no visualization of the results. Through the feature engineering the stations are represented in n-dimensional feature space. To illustrate the relations between the stations the dimensions have to be reduced. One way of doing that is multidimensional scaling (MDS). In this case the `sklearn.manifold.MDS` class is used to achieve this dimension reduction. It is important to note that to preserve the distances between the stations, metric MDS is used. For the implementation this concludes in using the default constructor with `mds = MDS(metric = True)`. The result of this representation for the different feature subsets is shown in figure 4.1. This representation is only used to illustrate the data and not used for any quantitative evaluation but rather to compare the visual image to the calculated values.

# Chapter 4

# Clustering Results

The results[1] of the clustering and the associated clustering scores are extensive. The number of clusters considered is reduced to the range of 4 to 15. A minimum of four clusters is chosen because two or three clusters would not represent the variety of different stations in an accurate way. The maximum of fifteen clusters was chosen because an average number of stations per cluster should at least be five stations. Otherwise the generalization that we want to achieve would not be of any use, because the clustering would consist of many special cases.
In section 4.1 the resulting MDS representations for the different feature subsets are presented. This representations is used to illustrate the clusterings in section 4.2.

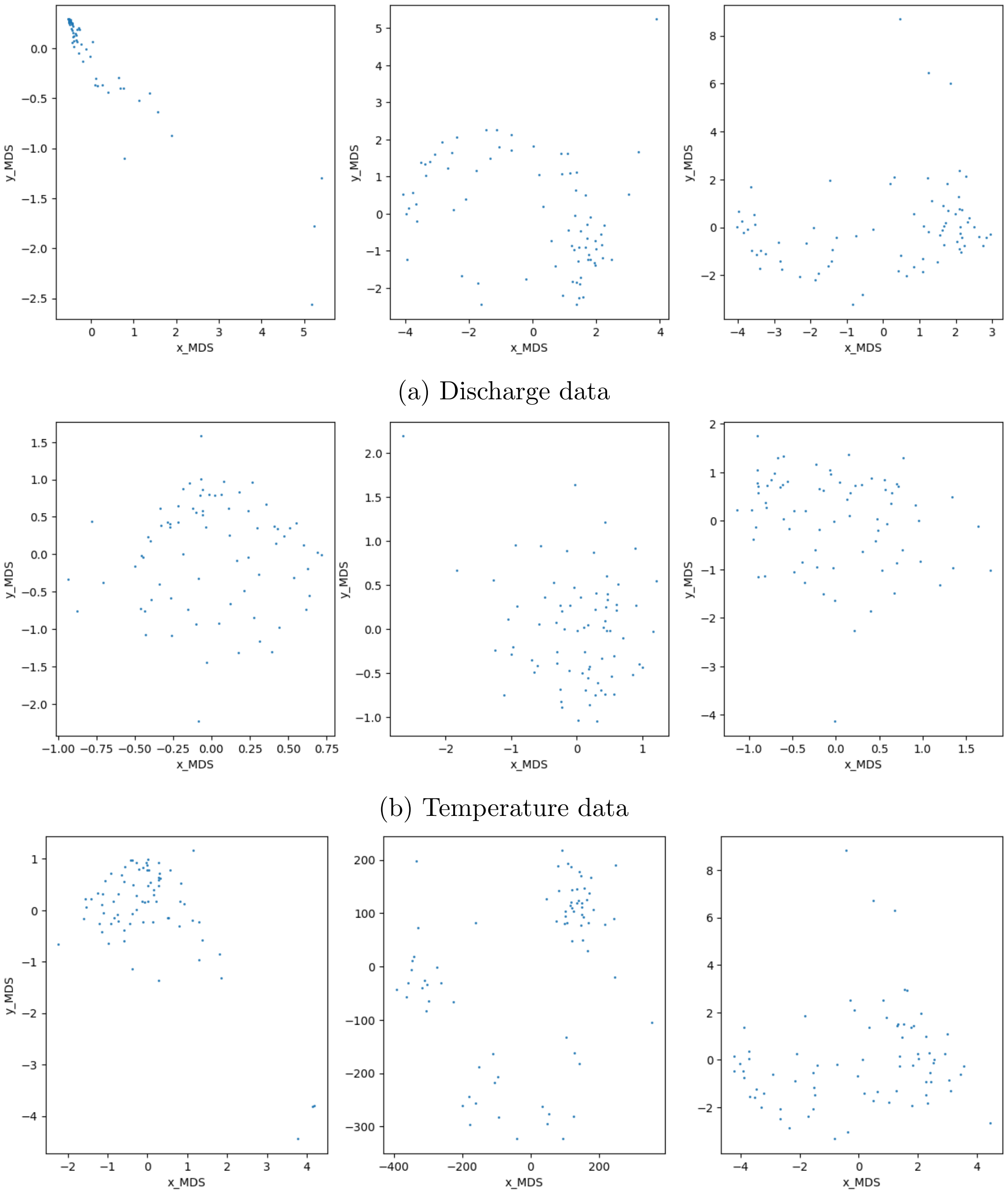
## 4.1 MDS-Representation of the Stations

For a broad overview of all the feature subsets from table 3.1 the MDS representations are shown in figure 4.1. In figure 4.1a the results for the discharge subsets, in figure 4.1b the results for the temperature subsets and in 4.1c the results for the subsets based on both data-sets are plotted.

## 4.2 Evaluation of Clustering Metrics

To get an overview over all the clustering scores, and to be able to extract valuable information, the data is presented in the following way: For every combination of clustering method, data-set and subset the best of the silhouette and the davies-bouldin score is selected and the according number of cluster is retrieved. These results are presented in table 4.1 for the davies-bouldin score and in table 4.2 for

[1] The code and the results are available at: `https://github.com/JanZurb/Clustering-of-Hydrological-Stations`



(a) Discharge data

(b) Temperature data

(c) Discharge and temperature data

Figure 4.1: The MDS representation of all feature subsets. From left to right basic, additional and all features.

| Clustering method | Data-set | Subset | DB score | Num. clusters |
|---|---|---|---|---|
| hierarchical | both | basic | 0.724 | 12 |
| hierarchical | both | additional | 0.453 | 7 |
| hierarchical | both | all | 1.047 | 10 |
| k-means | both | basic | 0.653 | 12 |
| k-means | both | additional | 0.501 | 6 |
| k-means | both | all | 1.009 | 7 |
| hierarchical | discharge | basic | 0.351 | 15 |
| hierarchical | discharge | additional | 0.473 | 6 |
| hierarchical | discharge | all | 0.868 | 10 |
| k-means | discharge | basic | 0.364 | 14 |
| k-means | discharge | additional | 0.535 | 5 |
| k-means | discharge | all | 0.938 | 14 |
| hierarchical | temp | basic | 0.697 | 15 |
| hierarchical | temp | additional | 0.517 | 4 |
| hierarchical | temp | all | 0.985 | 15 |
| k-means | temp | basic | 0.666 | 10 |
| k-means | temp | additional | 0.548 | 4 |
| k-means | temp | all | 0.548 | 4 |

Table 4.1: The best davies-bouldin scores with respective number of clusters, for each combination of data-set, feature subset and clustering method are shown. A lower davies-bouldin score indicates a better clustering. The grey values are best scores on a given subset.

the silhouette score.

Highlighted in grey are: The best score overall, the best score of a clustering that makes use of the temperature and the discharge data-set and the best score from the clusterings on all features from both datasets. These clusterings are discussed in sections 4.2.1,4.2.2 and 4.2.3.

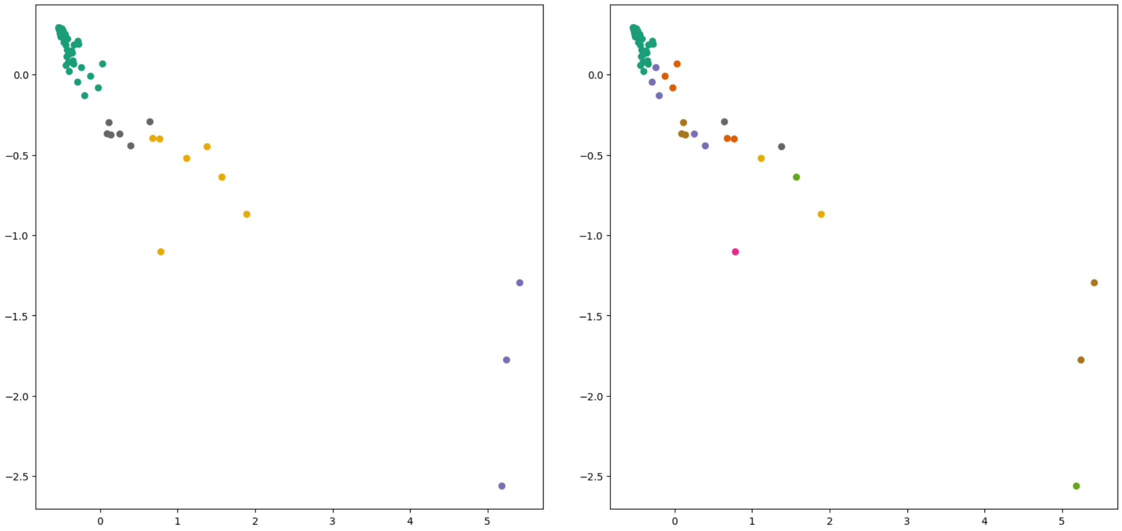
### 4.2.1 Best Scores on Arbitrary Feature Subset

The best silhouette score overall is achieved by performing the k-means clustering on the basic features of the discharge data. This clustering results in four clusters. In figure 4.2a the MDS representation of these features is shown with the clusters indicated by colour. It shows the data in four groups of different size. One very big group and three smaller groups. The smallest group consists of three members.

The best davies-bouldin score overall was achieved on the basic discharge features as well but clustered with a hierarchical approach and resulting in 15 clusters. In figure 4.2b the 15 groups are displayed. Fourteen of the groups consist of three or less stations and one group consists of more than half the stations.

| Clustering method | Data-set | Subset | Sil. score | Num. clusters |
|---|---|---|---|---|
| hierarchical | both | basic | 0.397 | 4 |
| hierarchical | both | additional | 0.602 | 7 |
| hierarchical | both | all | 0.297 | 4 |
| k-means | both | basic | 0.414 | 4 |
| k-means | both | additional | 0.601 | 6 |
| k-means | both | all | 0.288 | 4 |
| hierarchical | discharge | basic | 0.728 | 4 |
| hierarchical | discharge | additional | 0.643 | 6 |
| hierarchical | discharge | all | 0.374 | 4 |
| k-means | discharge | basic | 0.742 | 4 |
| k-means | discharge | additional | 0.642 | 5 |
| k-means | discharge | all | 0.414 | 4 |
| hierarchical | temp | basic | 0.36 | 15 |
| hierarchical | temp | additional | 0.37 | 5 |
| hierarchical | temp | all | 0.225 | 12 |
| k-means | temp | basic | 0.421 | 4 |
| k-means | temp | additional | 0.384 | 5 |
| k-means | temp | all | 0.384 | 5 |

Table 4.2: The best silhouette scores with the respective number of clusters, for each combination of data-set, feature subset and clustering method are shown. A silhouette score close to one indicates a good clustering. The grey values are best scores on a given subset.



(a) K-means clustering resulting in the best silhouette score.

(b) Hierarchical clustering resulting in the best davies-bouldin score.

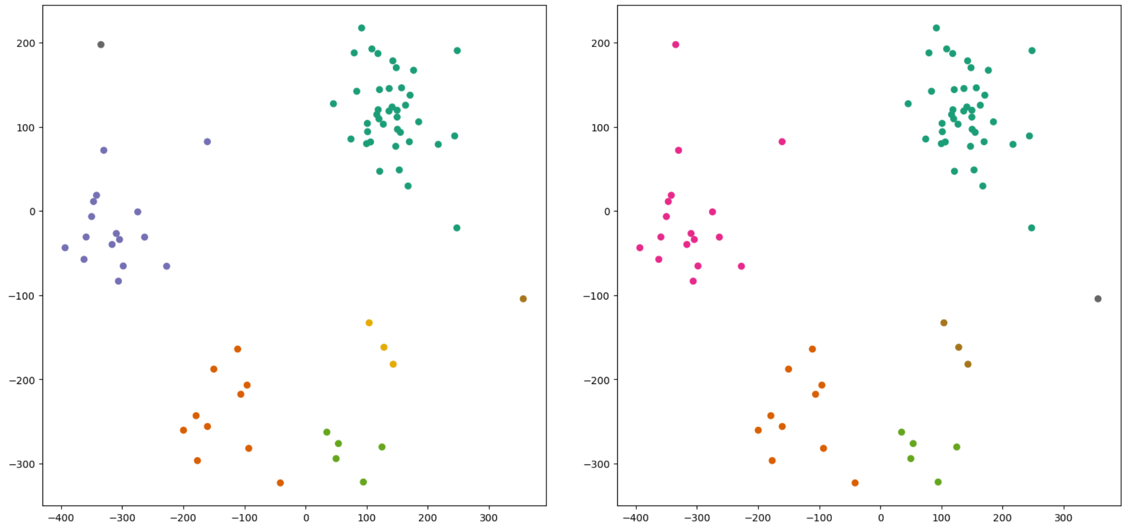
Figure 4.2: Illustration of the MDS representation of the clustering resulting from the basic features discharge data-set.

If we examine the MDS representation visually we see that approximately two thirds of the station belong to the same cluster independent of the clustering algorithm and the other stations are divided in three groups for the k-means algorithm or in 14 groups for the hierarchical method.

### 4.2.2 Best Scores on Combined Temperature and Discharge Subset

For this evaluation, we consider the feature subsets that contain discharge and temperature data. In the tables 4.2 and 4.1 these are the entries above the horizontal line. The best davies-bouldin and silhouette score are achieved by the hierarchical algorithm performed on the ”additional features both” feature subset. From figure 4.3a where the resulting clusters are shown it can be seen that, the clusters are of similar size, with the exception of two groups only consisting of one station and another group consisting of three stations.

From table 4.2 we derive that the silhouette score of the k-means algorithm on the same feature subset has only a 1% worse score than the considered hierarchical clustering, but results in six clusters. The result of this clustering is shown in figure 4.3b. It is visible that one station has changed the cluster affiliation. The one represented on the far upper left in figure 4.3a in the MDS representation. All the other clusters are exactly the same as with the hierarchical algorithm.



(a) Hierarchical algorithm resulting in the best davies-bouldin and silhouette score.

(b) K-means algorithm resulting in the second best silhouette score.

Figure 4.3: Illustration of the MDS representation of the clustering resulting from the additonal features both data-set.

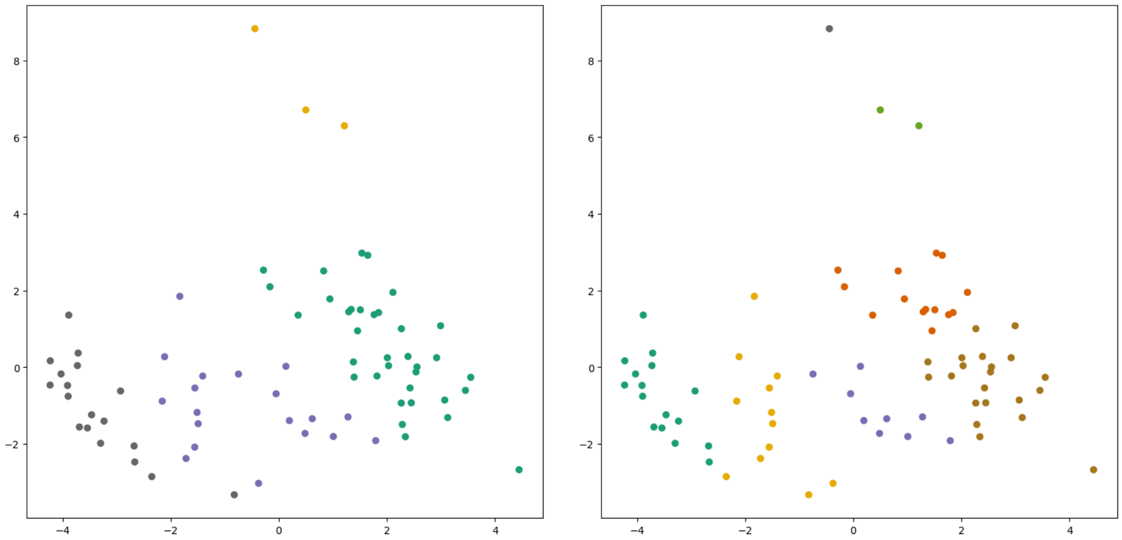
### 4.2.3 Best Scores on All Features

The scores on all features from both data-sets are among the worst three from all the subsets that are clustered. The best silhouette score is achieved by the hierarchical clustering and results in 4 clusters, see figure 4.4b. The associated clusters are of similar size except one cluster that consists of three stations. The best davies-bouldin score is achieved by the k-means clustering and results in seven clusters, see figure 4.4b. In this case the clusters are of similar size with the exception of the same three stations, now split in a cluster of one and two stations.

Because the two best scores do not result in the same number of clusters a direct comparison between the two best scores is difficult. Therefore in figure 4.5 the results of the clustering with the other algorithm is presented. If we compare the clusterings resulting in the best scores (figure 4.4)with the clusterings done on the same data-set but with the other method (figure 4.5)we see that for the version with 4 clusters we get different clusters, whereas for the one with 7 clusters the result is the same with the exception of four stations that change clusters.

## 4.3 Geographic Location of Clusters

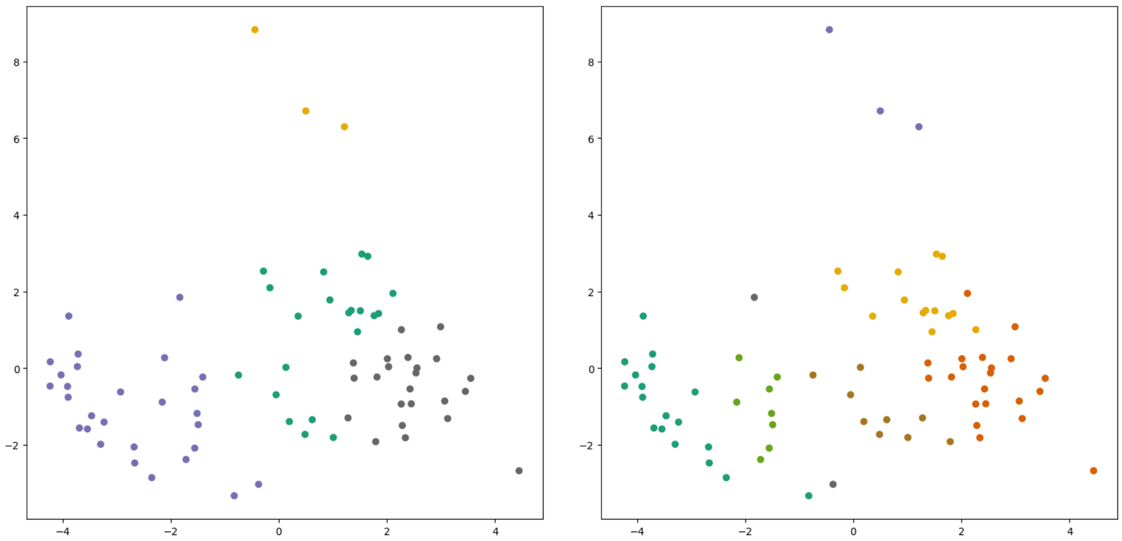
To further investigate the clustering, the clusters are plotted on a map with the Swiss river network. In figure 4.6a the clustering achieved on the additional features from both data-sets is shown and in figure 4.6b the clustering achieved on all features



(a) Hierarchical algorithm resulting in the best silhouette score.

(b) K-meanss algorithm resulting in the best davies-bouldin score.

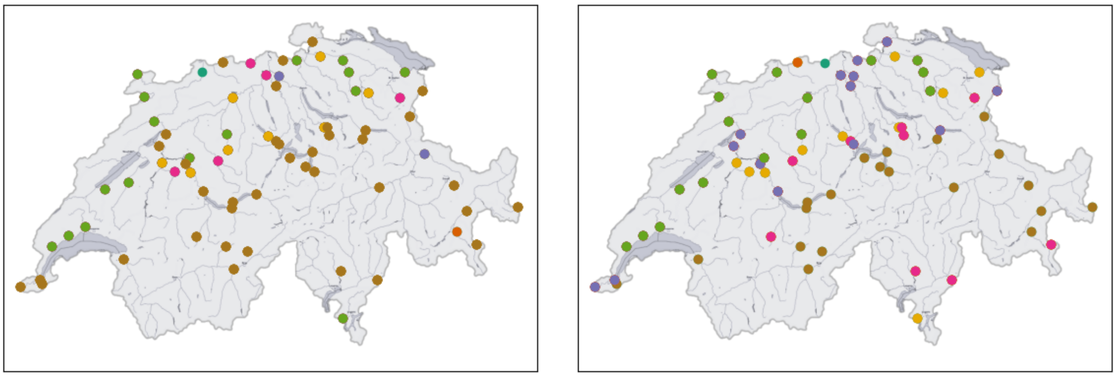
Figure 4.4: Illustration of the MDS representation of the clustering resulting from the all features both data-set.



(a) K-meanss clustering resulting in four clusters.

(b) Hierarchical clustering resulting in seven clusters.

Figure 4.5: Clusterings with other clustering algorithm but same number of clusters as in figure 4.4 on all features both data-set.



(a) Clustering with seven clusters on the additional both feature subset.

(b) Clustering with seven clusters on the all both feature subset.

Figure 4.6: Illustration of the clusters on the map.

from both data-sets is shown, both resulting in seven clusters.

# Chapter 5

# Conclusions and Future Work

In the following chapter the results presented in chapter 4are interpreted. The chapter is structured by focusing on the previously formulated questions in chapter 1. How do we characterize each station? What is the best number of clusters? And how are the clusters composed? In addition to the results further questions that arose during the thesis are stated in section 5.3.

## 5.1 Characterization of the Stations

The question how each station is characterized is answered in chapter 2. In general each station is characterized by 38 features. Composed of 6 basic features and 13 additional features, that are calculated on the discharge and temperature data-set. This set of features is divided into subsets, as described in table 3.1. To get some insight into the characterization we analyze the resulting MDS representations shown in figure 4.1. It is noticeable that in figure 4.1a 80% of the stations are close together and therefore considered similar and 20% of the stations are more spread out and therefore clearly different. From this we conclude that the basic features on the discharge data are able to characterize the 20% well but for 80% of the stations the basic features are unable to distinguish between them. This indicates that considering the basic flow features only, does not result in a good characterization.

The MDS representation for the ”additional feature both” subset, shown in figure 4.1c stands out as well. It is the only representation with visible clusters. Therefore this subset should lead to a well defined clustering which is evaluated in section 5.2. It is remarkable that these clusters occur because whether the additional features on the discharge data-set, nor the additional features on the temperature data-set resulted in such a formation. Only the combination of the two sets is able to reveal this pattern. From this we conclude that there is a dependency between the additional discharge and temperature features.

It is interesting to look at the resulting MDS representation for all features on both data-sets shown in figure 4.1c. In comparison to the clustering that resulted from the additional features the clear structure is not present. Therefore through consideration of the basic features the clear cluster structure is lost. What this means is difficult to say but in section 5.2 the resulting clustering is discussed further.
On a more general note it is important to keep in mind that the more features are taken into account the more detailed a station is described. On the other hand through too many features the resulting representation might be dominated by noisy features, although they are calculated carfully. In addition to that the results of this interpretation have to be taken with caution. Because it is a mapping from multidimensional data into two dimensions. Therefore much of the original information is lost and no definite conclusions can be made from this visual evaluation.

## 5.2 Recommended Clusterings

After the visual discussion of the subset we will discuss the numerical results from the clustering evaluation. Goal of the experiment is to find a clustering on a subset of features that characterizes the stations well and results in good scores. For this reason we further investigate the achieved best scores on the subsets and discuss whether the resulting clustering is reasonable to use further. In section 3.3 the best scores achieved on different subsets with different clustering algorithms are presented. We will have a look at three resulting clusterings in more detail.

### Best Clustering Scores overall

As seen in section 4.2.1 the best silhouette and davies-bouldin score is achieved by performing a k-means clustering on the basic features discharge subset. As can be seen in figure 4.2 the two best scores result 4 and 15 clusters. This deviation shows that the clusters are not well defined. Additionally to that, because only 6 features are considered the stations are characterized insufficiently, as already discussed in section 5.1. In both cases the clustering results in one very big cluster, this indicates as well that for many stations although they are different they are put in the same cluster. Therefore this clustering is not recommended for further use.

### Best Clustering Scores from Both Data-sets

To take more features from both data-sets into consideration the best scores on a combined temperature and discharge feature subset is considered. As can be seen in section 4.2.2, the best davies-bouldin and silhouette score are achieved by performing a hierarchical clustering on the additional feature set from both data-sets. The result was in both cases the same clustering with seven clusters. This can be verified in figure 4.3a. That the best scores result in the same clustering is an indicator for well defined clusters. To investigate this stability further the result of the k-means clustering, resulting in the second best silhouette score is shown in figure 4.3b. There we see that the number of clusters has changed to six but the cluster composition is the same apart from one station. For the details see section 4.2.2. This leads to the conclusion that the clustering is stable with respect to the clustering method. For these reasons the hierarchical clustering on this subset is recommended for further use. The cluster assignment are displayed in table B.2.

### Best Clustering Scores on all Features from Both Data-sets

As stated in section 5.1 the more features that are considered the more detailed the characterization of the stations is. Therefore the best scores on all features are evaluated as well. Generally speaking if we look at the four resulting scores of these clusterings their silhouette and davies-bouldin scores are among the five worst scores overall. This indicates, that the clusters found are not separated well. In table 4.1 we see that the best davies-bouldin score is achieved when using the k-means algorithm and results in seven clusters. On the other hand the best silhouette score is achieved when using the hierarchical clustering and results in four clusters as can be seen in table 4.2. To determine which of the clusterings is more meaningful we look at the resulting clusterings that were achieved with the other clustering methods, see figure 4.5. By comparing these results with the results from the best scores in figure 4.4 we see that for the clustering resulting in 4 groups the clusters composition changes significantly. In comparison the clustering with the seven groups is stable with about 10% of the stations changing clusters. Therefore the clustering with the seven stations can be recommended for further use as well, although not having a good score the additional consideration of the basic features can contribute to a better clustering. Because the stations are characterized in more detail as is described in section 5.1. The resulting cluster assignment are displayed in table B.1.

### Comparison of the Recommended Clusterings

Both recommended clusterings result in seven clusters. It is of interest to compare these clusterings to see if they result in the same clusters or if there are differences. In figure 4.6 the clusters are plotted in colour on a map of Switzerland. The green, brown and yellow group are very similar in both cases. The brown and green group follow a geographic separation along the north-east axis. The other four groups are composed differently and there is no direct mapping between them. Therefore it makes sense to consider both clusterings because the detailed composition differs and therefore might impact the performance of the model significantly. It can be concluded that for three of the seven groups the cluster assignments are stable and therefore these groups are characterized well.

## 5.3 Further Research

It remains to be stated that the task we confronted ourselves with is one with no guaranteed outcome, the work with real-life data, combined with an approach that is highly dependent on the used features, is difficult to evaluate because there is no ground truth. Only the testing of the clusters by applying them to the neural network will show if the clustering is of any use. Some further questions occurred during the thesis. For one, which features are responsible for a good clustering? The question was in some ways already dealt with by dividing the set of features into different subsets, but a more systematic approach would be useful to leave out unnecessary features and get an insight into which features contribute to a well defined clustering. An option to achieve that would be a principal component analysis.

Considering the features themselves, most of them are engineered for periodic data courses. But as already mentioned in section 2.3, some discharge curves have an aperiodic behaviour and do not follow this pattern. To characterize these curves better specific features should be engineered.

A more general question is, if the feature engineering approach is well suited for the analysis of time series data? Because the curves look very similar especially for the temperature data it is difficult to find features that result in clear groups and not only in a even distribution. An alternative option could be to construct a distance matrix with the help of dynamic time warping. This would lead to a less handcrafted solution for the problem.

# Appendix A

# Feature Engineering

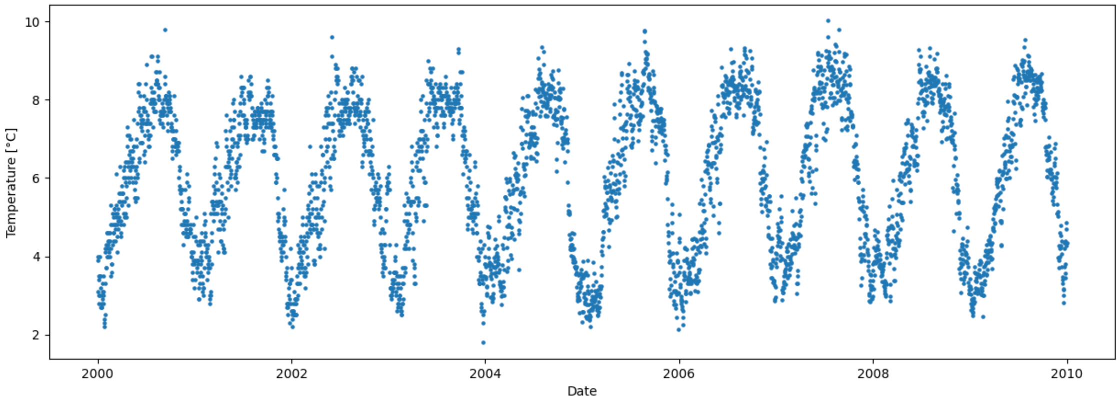
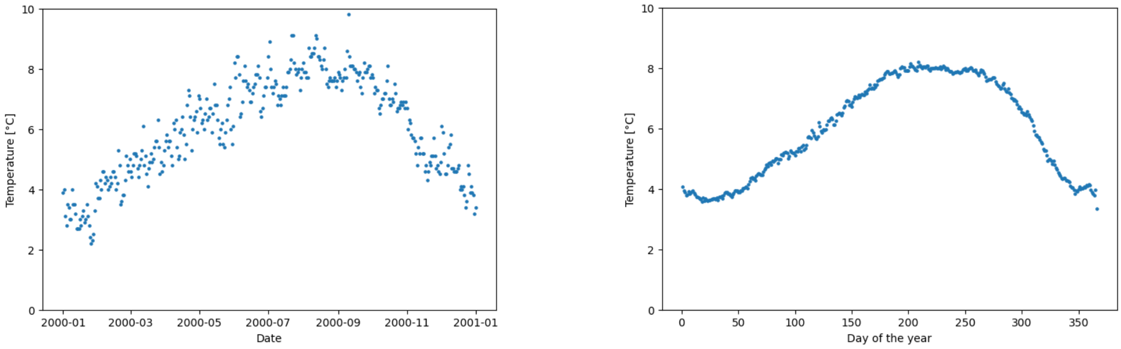


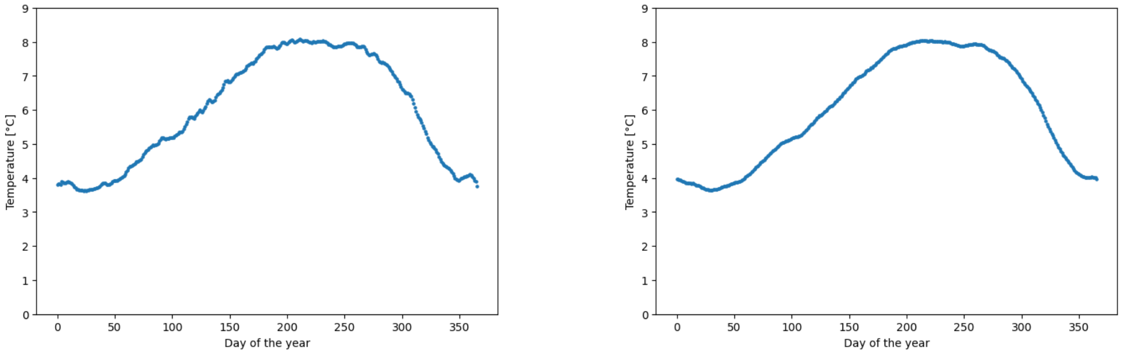
Figure A.1: The course of average temperature over a 10 years period for station no. 2019.



(a) Example year

(b) Averaged by the day of the year

Figure A.2: The course of water temperature for station no. 2019



(a) window length of 5 days

(b) window length of 15 days

Figure A.3: The course of the window averaged temperature course of station 2019.

| ID no. | Name |
|---|---|
| 2009 | Rhône-Porte du Scex |
| 2016 | Aare-Brugg |
| 2018 | Reuss-Mellingen |
| 2019 | Aare-Brienzwiler |
| 2029 | Aare-Brügg, Aegerten |
| 2030 | Aare-Thun |
| 2033 | Vorderrhein-Ilanz |
| 2034 | Broye-Payerne, Caserne d'aviation |
| 2044 | Thur-Andelfingen |
| 2056 | Reuss-Seedorf |
| 2070 | Emme-Emmenmatt |
| 2084 | Muota-Ingenbohl |
| 2085 | Aare-Hagneck |
| 2091 | Rhein-Rheinfelden, Messstation |
| 2104 | Linth-Weesen, Biäsche |
| 2106 | Birs-Münchenstein, Hofmatt |
| 2109 | Lütschine-Gsteig |
| 2112 | Sitter-Appenzell |
| 2113 | Aare-Felsenau, K.W. Klingnau |
| 2126 | Murg-Wängi |
| 2130 | Rhein (Oberwasser)-Laufenburg |
| 2135 | Aare-Bern, Schönau |
| 2143 | Rhein-Rekingen |
| 2150 | Landquart-Felsenbach |
| 2152 | Reuss-Luzern, Geissmattbrücke |
| 2159 | Gürbe-Belp, Mülimatt |
| 2161 | Massa-Blatten bei Naters |
| 2167 | Tresa-Ponte Tresa, Rocchetta |
| 2170 | Arve-Genève, Bout du Monde |
| 2174 | Rhône-Chancy, Aux Ripes |
| 2179 | Sense-Thörishaus, Sensematt |
| 2210 | Doubs-Ocourt |
| 2232 | Allenbach-Adelboden |
| 2243 | Limmat-Baden, Limmatpromenade |
| 2256 | Rosegbach-Pontresina |
| 2269 | Lonza-Blatten |
| 2276 | Grosstalbach-Isenthal |
| 2282 | Sperbelgraben-Wasen, Kurzeneialp |

Table A.1: Station ID numbers with the station names.

| ID no. | Name |
|---|---|
| 2288 | Rhein-Neuhausen, Fluringerbrücke |
| 2307 | Suze-Sonceboz |
| 2308 | Goldach-Goldach, Bleiche |
| 2327 | Dischmabach-Davos, Kriegsmatte |
| 2343 | Langeten-Huttwil, Häberenbad |
| 2347 | Riale di Roggiasca-Roveredo, Bacino di compenso |
| 2351 | Vispa-Visp |
| 2356 | Riale di Calneggia-Cavergno, Pontit |
| 2366 | Poschiavino-La Rösa |
| 2369 | Mentue-Yvonand, La Mauguettaz |
| 2372 | Linth-Mollis, Linthbrücke |
| 2374 | Necker-Mogelsberg, Aachsäge |
| 2386 | Murg-Frauenfeld |
| 2392 | Rhein (Oberwasser)-Rheinau |
| 2410 | Liechtensteiner Binnenkanal-Ruggell |
| 2414 | Rietholzbach-Mosnang, Rietholz |
| 2415 | Glatt-Rheinsfelden |
| 2432 | Venoge-Ecublens, Les Bois |
| 2433 | Aubonne-Allaman, Le Coulet |
| 2434 | Dünnern-Olten, Hammermühle |
| 2457 | Aare-Ringgenberg, Goldswil |
| 2462 | Inn-S-chanf |
| 2467 | Saane-Gümmenen |
| 2473 | Rhein-Diepoldsau, Rietbrücke |
| 2481 | Engelberger Aa-Buochs, Flugplatz |
| 2485 | Allaine-Boncourt, Frontière |
| 2493 | Promenthouse-Gland, Route Suisse |
| 2500 | Worble-Ittigen |
| 2604 | Biber-Biberbrugg |
| 2606 | Rhône-Genève, Halle de l'ìle |
| 2608 | Sellenbodenbach-Neuenkirch |
| 2609 | Alp-Einsiedeln |
| 2612 | Riale di Pincascia-Lavertezzo |
| 2617 | Rom-Müstair |
| 2623 | Rhone-Oberwald |
| 2634 | Kleine Emme-Emmen |
| 2635 | Grossbach-Einsiedeln, Gross |

Table A.2: Station ID numbers with the station names.

# Appendix B

# Clustering Results

| ID no. | Cluster assignment | ID no. | Cluster assignment |
|---|---|---|---|
| 2009 | 5 | 2276 | 5 |
| 2016 | 1 | 2282 | 4 |
| 2018 | 1 | 2288 | 1 |
| 2019 | 5 | 2307 | 0 |
| 2029 | 1 | 2308 | 4 |
| 2030 | 1 | 2327 | 5 |
| 2033 | 5 | 2343 | 0 |
| 2034 | 0 | 2347 | 2 |
| 2044 | 4 | 2351 | 5 |
| 2056 | 5 | 2366 | 2 |
| 2070 | 2 | 2369 | 0 |
| 2084 | 5 | 2372 | 5 |
| 2085 | 1 | 2374 | 4 |
| 2091 | 3 | 2386 | 0 |
| 2104 | 1 | 2410 | 5 |
| 2106 | 0 | 2414 | 0 |
| 2109 | 5 | 2415 | 0 |
| 2112 | 2 | 2432 | 0 |
| 2126 | 0 | 2433 | 0 |
| 2130 | 6 | 2434 | 0 |
| 2135 | 1 | 2457 | 5 |
| 2139 | 4 | 2462 | 5 |
| 2143 | 1 | 2467 | 4 |
| 2150 | 5 | 2473 | 1 |
| 2152 | 1 | 2481 | 5 |
| 2159 | 4 | 2485 | 0 |
| 2161 | 5 | 2493 | 0 |
| 2167 | 4 | 2500 | 0 |
| 2170 | 5 | 2604 | 4 |
| 2174 | 1 | 2606 | 1 |
| 2179 | 4 | 2608 | 4 |
| 2181 | 4 | 2609 | 2 |
| 2210 | 0 | 2612 | 2 |
| 2232 | 2 | 2613 | 3 |
| 2243 | 1 | 2617 | 5 |
| 2256 | 5 | 2634 | 2 |
| 2265 | 5 | 2635 | 2 |
| 2269 | 5 | | |

Table B.1: Resulting cluster assignments for the k-means clustering on all the features from both data-sets, resulting in seven clusters.

| ID no. | Cluster assignment | ID no. | Cluster assignment |
|---|---|---|---|
| 2009 | 0 | 2276 | 0 |
| 2016 | 3 | 2282 | 1 |
| 2018 | 0 | 2288 | 0 |
| 2019 | 0 | 2307 | 2 |
| 2029 | 0 | 2308 | 2 |
| 2030 | 0 | 2327 | 0 |
| 2033 | 0 | 2343 | 2 |
| 2034 | 2 | 2347 | 0 |
| 2044 | 1 | 2351 | 0 |
| 2056 | 0 | 2366 | 0 |
| 2070 | 3 | 2369 | 2 |
| 2084 | 0 | 2372 | 0 |
| 2085 | 0 | 2374 | 1 |
| 2091 | 0 | 2386 | 2 |
| 2104 | 0 | 2410 | 0 |
| 2106 | 6 | 2414 | 2 |
| 2109 | 0 | 2415 | 2 |
| 2112 | 3 | 2432 | 2 |
| 2126 | 2 | 2433 | 2 |
| 2130 | 3 | 2434 | 1 |
| 2135 | 0 | 2457 | 0 |
| 2139 | 1 | 2462 | 0 |
| 2143 | 0 | 2467 | 1 |
| 2150 | 4 | 2473 | 0 |
| 2152 | 0 | 2481 | 0 |
| 2159 | 1 | 2485 | 2 |
| 2161 | 0 | 2493 | 2 |
| 2167 | 2 | 2500 | 2 |
| 2170 | 0 | 2604 | 1 |
| 2174 | 0 | 2606 | 0 |
| 2179 | 3 | 2608 | 1 |
| 2181 | 1 | 2609 | 0 |
| 2210 | 2 | 2612 | 0 |
| 2232 | 0 | 2613 | 4 |
| 2243 | 4 | 2617 | 0 |
| 2256 | 5 | 2634 | 0 |
| 2265 | 0 | 2635 | 0 |
| 2269 | 0 | | |

Table B.2: Resulting cluster assignments for the hierarchical clustering on the additional features from both data-sets, resulting in seven clusters.

# Bibliography

[1] Benjamin Fankhauser and Kaspar Riesen. Graph-based deep learning on the swiss river network. (unpublished).

[2] A. Michel, T. Brauchli, M. Lehning, B. Schaefli, and H. Huwald. Stream temperature and discharge evolution in switzerland over the last 50 years: annual and seasonal behaviour. *Hydrology and Earth System Sciences*, 24(1):115–142, 2020.

[3] Philip C. Reid, Renata E. Hari, Grégory Beaugrand, David M. Livingstone, Christoph Marty, Dietmar Straile, Jonathan Barichivich, Eric Goberville, Rita Adrian, Yasuyuki Aono, Ross Brown, James Foster, Pavel Groisman, Pierre Hélaouët, Huang-Hsiung Hsu, Richard Kirby, Jeff Knight, Alexandra Kraberg, Jianping Li, Tzu-Ting Lo, Ranga B. Myneni, Ryan P. North, J. Alan Pounds, Tim Sparks, René Stübi, Yongjun Tian, Karen H. Wiltshire, Dong Xiao, and Zaichun Zhu. Global impacts of the 1980s regime shift. *Global Change Biology*, 22(2):682–703, 2016.

[4] Matplotlib documentation — Matplotlib 3.7.2 documentation. Available at `https://matplotlib.org/stable/index.html`. (visited on 12/08/2023).

[5] pandas documentation — pandas 2.0.3 documentation. Available at `http://pandas.pydata.org/docs/`. (visited on 12/08/2023).

[6] Federal Office of Topographie. Location of the hydrological measurement stations in switzerland. Available at `https://map.geo.admin.ch/?lang=en&amp;topic=gewiss&amp;bgLayer=ch.swisstopo.pixelkarte-grau&amp;E=2663102.38&amp;N=1188545.61&amp;zoom=1.3158869540534703&amp;layers=ch.bafu.hydrologie-hintergrundkarte,ch.bafu.hydrologie-wassertemperaturmessstationen`. (visited on 01/08/2023).

[7] Rob Hyndman. *International Encyclopedia of Statistical Science*, pages 866–869. Springer, 2010.

[8] SciPy documentation — SciPy v1.11.1 Manual. Available at `https://docs.scipy.org/doc/scipy/`. (visited on 12/08/2023).

[9] Dalwinder Singh and Birmohan Singh. Feature wise normalization: An effective way of normalizing data. *Pattern Recognition*, 122:108307, 2022.

[10] S. Lloyd. Least squares quantization in pcm. *IEEE Transactions on Information Theory*, 28(2):129–137, 1982.

[11] scikit-learn: machine learning in Python — scikit-learn 1.3.0 documentation. Available at `https://scikit-learn.org/stable/`. (visited on 12/08/2023).

[12] David Arthur and Sergei Vassilvitskii. K-means++: The advantages of careful seeding. *Proc. of the Annu. ACM-SIAM Symp. on Discrete Algorithms*, 8:1027–1035, 01 2007.

[13] Fionn Murtagh and Pedro Contreras. Algorithms for hierarchical clustering: an overview. *WIREs Data Mining and Knowledge Discovery*, 2(1):86–97, 2012.

[14] Joe H. Ward Jr. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301):236–244, 1963.

[15] Peter J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987.

[16] Maria Halkidi, Yannis Batistakis, and Michalis Vazirgiannis. On Clustering Validation Techniques. *Journal of Intelligent Information Systems*, 17(2):107–145, 2001.

# Erklärung

gemäss Art. 30 RSL Phil.-nat.18

Name/Vorname: Zurbrügg, Jan Andrin

Matrikelnummer: 19-917-012

Studiengang: Computer Science

Bachelor ☑ Master □ Dissertation □

Titel der Arbeit: Clustering of Hydrological Stations

LeiterIn der Arbeit: PD Dr. Kaspar Riesen

Ich erkläre hiermit, dass ich diese Arbeit selbständig verfasst und keine anderen als die angegebenen Quellen benutzt habe. Alle Stellen, die wörtlich oder sinngemäss aus Quellen entnommen wurden, habe ich als solche gekennzeichnet. Mir ist bekannt, dass andernfalls der Senat gemäss Artikel 36 Absatz 1 Buchstabe r des Gesetzes vom 5. September 1996 über die Universität zum Entzug des auf Grund dieser Arbeit verliehenen Titels berechtigt ist.
Für die Zwecke der Begutachtung und der Überprüfung der Einhaltung der Selbständigkeitserklärung bzw. der Reglemente betreffend Plagiate erteile ich der Universität Bern das Recht, die dazu erforderlichen Personendaten zu bearbeiten und Nutzungshandlungen vorzunehmen, insbesondere die schriftliche Arbeit zu vervielfältigen und dauerhaft in einer Datenbank zu speichern sowie diese zur Überprüfung von Arbeiten Dritter zu verwenden oder hierzu zur Verfügung zu stellen.

Biberist, 29.08.2023

Ort/Datum

Unterschrift J. Zurbrügg