Czech University of Life Sciences
Faculty of Environmental Sciences

Department of Water Resources and Environmental Modelling

Study Program: Landscape Engineering

Application of machine learning to satellite data for classification of the precipitation spatial characteristics

MASTER THESIS

Author: Mariia Kavalerova
Supervisor: doc. Mgr. Ing. Ioannis Markonis, Ph.D
Year: 2022

CZECH UNIVERSITY OF LIFE SCIENCES PRAGUE

Faculty of Environmental Sciences

DIPLOMA THESIS ASSIGNMENT

Mariia Kavalerova

Landscape Engineering
Environmental Modelling

Thesis title

Application of machine learning to satellite data for classification of the precipitation spatial characteristics

---

Objectives of thesis

The thesis uses a popular machine learning technique (Self-Organizing Map) to identify the dominant types of precipitation in space. The main objectives of the thesis are:

1. Classify the spatial patterns of precipitation over Czech Republic in a few meaningful categories.

2. Detect any changes in the spatial characteristics of precipitation during the last five years.

Methodology

To successfully achieve the thesis objectives the following steps will be taken:

1. Literature review regarding the Self-Organizing Map technique and the GPM satellite data product.

2. Presentation of the climatology of Czech Republic, with special focus in its climatic extremes.

3. Downloading and pre-processing of the GPM dataset.

4. Application of the classification algorithm.

5. Statistical analysis of the classification results.

The proposed extent of the thesis

50 pages

Keywords

Self-Organizing Map, clustering, precipitation in Czech Republic, GPM data

---

Recommended information sources

HUFFMAN, George J; BOLVIN, David T.; BRAITHWAITE, Dan; HSU, Kuolin; JOYCE, Robert; KIDD, Christopher; NELKIN, Eric J.; SOROOSHIAN, Soroosh; TAN, Jackson; XIE, Pingping. Algorithm Theoretical Basis Document (ATBD), NASA Global Precipitation Measurement (GPM) Integrated Multi-satellitE Retrievals for GPM (IMERG). 2019.

HUFFMAN, G.J.; STOCKER, E.F.; BOLVIN, D.T.; NELKIN, E.J.; JACKSON, Tan. GPM IMERG Final Precipitation L3 Half Hourly 0.1 degree x 0.1 degree V06, Greenbelt, MD. Goddard Earth Sciences Data and Information Services Center (GES DISC). 2019.

Kohonen, T. (2012). Self-organizing maps (Vol. 30). Springer Science & Business Media.

Lochbihler, K., Lenderink, G., & Siebesma, A. P. (2017). The spatial extent of rainfall events and its relation to precipitation scaling. Geophysical Research Letters, 44(16), 8629-8636.

Precipitation processing system. GPM IMERG Final Precipitation L3 Half Hourly 0.1 degree x 0.1 degree V06, Greenbelt, MD. Global precipitation measurement (GPM), Prelimitary file Specification for GPM products. 2021.

TRMM instruments. NASA, [n.d.]. Available also from: https://gpm.nasa.gov/ missions/TRMM/satellite.

Zveryaev, Igor I. ”Seasonality in precipitation variability over Europe.” Journal of Geophysical Research: Atmospheres 109, no. D5 (2004).

Expected date of thesis defence

2021/22 SS – FES

The Diploma Thesis Supervisor

doc. Mgr. Ing. Ioannis Markonis, Ph.D.

Supervising department

Department of Water Resources and Environmental Modeling

Advisor of thesis

Rajani Kumar Pradhan
MSc. Rajani Kumar Pradhan, MSc

Electronic approval: 29. 3. 2022

prof. Ing. Martin Hanel, Ph.D.

Head of department

Electronic approval: 29. 3. 2022

prof. RNDr. Vladimír Bejček, CSc.

Dean

Prague on 29. 03. 2022

Author's Statement

I hereby declare that I have independently elaborated the diploma/final thesis with the topic of: Development of a research support tool for literature review facilitating data mining techniques and that I have cited all the information sources that I used in the thesis and that are also listed at the end of the thesis in the list of used information sources. I am aware that my diploma/final thesis is subject to Act No. 121/2000 Coll., on copyright, on rights related to copyright and on amendment of some acts, as amended by later regulations, particularly the provisions of Section 35(3) of the act on the use of the thesis. I am aware that by submitting the diploma/final thesis I agree with its publication under Act No. 111/1998 Coll., on universities and on the change and amendments of some acts, as amended, regardless of the result of its defence. With my own signature, I also declare that the electronic version is identical to the printed version and the data stated in the thesis has been processed in relation to the GDPR.

In Prague, 31^{th} March 2022 Mariia Kavalerova

Acknowledgment

I would like to express my gratitude to my supervisor, doc. Mgr. Ing. Ioannis Markonis, Ph.D, for all his guidance and for his inspiring suggestions. I wish to extend my special thanks to my advisor, MSc. Rajani Kumar Pradhan, for his support and help.

Mariia Kavalerova

Thesis Title:

Application of machine learning to satellite data for classification of the precipitation spatial characteristics

Author: Mariia Kavalerova

Study Field: Landscape Engineering
Study Program: Environmental Modelling
Type of thesis: Master Thesis

Supervisor: doc. Mgr. Ing. Ioannis Markonis, Ph.D
Department of Water Resources and Environmental Modelling
Consultant: MSc. Rajani Kumar Pradhan
Department of Water Resources and Environmental Modelling

Abstract: Precipitation analysis of the trends and changes are taking part in climate change understanding. In addition, it plays a vital role in predicting and investigating extreme weather events, such as floods and droughts. Extreme events threaten climate-sensitive economic sectors, including agriculture, forestry, energy, insurance and others. In precipitation analysis, temporal changes and spatial variability are taken into account. Therefore, it is crucial to identify areas with the same and different precipitation patterns to improve precipitation analysis. Only ground-based observations were available and used in the studies for a long time. However, spatial data with high resolution became available, and many researchers attempted to use satellite data in their analysis. Spatial data includes an enormous amount of data compared to ground-based observation; thus, analysis involving the application of the unsupervised machine learning approaches is becoming more popular and is considered more efficient. In this research, the self-organising algorithm merged with hierarchical clustering was used on GPM data to classify precipitation of the Czech Republic spatially. GPM data is stored in the structured compressed "nc4" format; thus, it was pre-processed and converted to the tidy table format. As a result, homogeneous clusters with different precipitation patterns were received. Different grid sizes and numbers of clusters were applied; however, most of the clusters remain the same; this confirms the homogeneity of the regions. Interestingly, the mountains regions and the South Moravia region formed separate and homogeneous clusters. This approach showed a convenient, low computational and time cost method to classify precipitation patterns spatially. One of the main benefits is clear visualisation of the clusters on the map, which can benefit the analysis of the precipitation patterns and improve spatial precipitation variance analysis. Furthermore, spatial classification of the precipitation patterns can improve flood and drought modelling and predictions. However, this study did not include a comprehensive analysis of the frequency and intensity of the precipitations and their relation to the atmospheric circulation, which can be considered future development and an extension of the analysis. In addition, the GPM data quality should be considered while interpreting the results of this study.

Key words: Self-Organizing Map, clustering, precipitation in Czech Republic, GPM data

Název práce:

Aplikace strojového učení na družicová data pro klasifikaci prostorových charakteristik srážek

Autor: Mariia Kavalerova

Studijní obor: Krajinné inženýrství
Studijní program: Environmentální modelování
Typ práce: Diplomová práce

Vedoucí: doc. Mgr. Ing. Ioannis Markonis, Ph.D
Katedra vodních zdrojů a environmentálního modelování
Konzultant: MSc. Rajani Kumar Pradhan
Katedra vodních zdrojů a environmentálního modelování

Abstrakt: Srážková analýza trendů a změn se podílí na pochopení změny klimatu. Navíc hraje zásadní roli při předpovídání a zkoumání extrémních jevů počasí, jako jsou povodně a sucha. Extrémní klimatické události ohrožují ekonomicky citlivá odvětví na klima jako zemědělství, lesnictví, energetiky, pojišťovnictví a dalších. Při analýze srážek se berou v úvahu časové změny a prostorová variabilita. Proto je klíčové identifikovat oblasti se stejnými a odlišnými srážkovými vzory, aby se zlepšila analýza srážek. K dispozici byla pouze pozemní pozorování, která se ve studiích používala dlouhou dobu. Zpřístupnila se ale prostorová data s vysokým rozlišením a mnoho výzkumníků se pokusilo při své analýze použít satelitní data. Prostorová data zahrnují obrovské množství dat ve srovnání s pozemním pozorováním; proto se analýza zahrnující aplikaci přístupů strojového učení bez učitele stává populárnější a je považována za efektivnější. V tomto výzkumu byl použit samo-organizující se algoritmus spojený s hierarchickým shlukováním na datech GPM k prostorové klasifikaci srážek České republiky. Data GPM jsou uložena ve strukturovaném komprimovaném formátu „nc4“; a proto byla předzpracována a převedena do formátu uspořádané tabulky. Díky tomu byly získány homogenní shluky s různými vzory srážek. Byly použity různé velikosti mřížky a počty shluků; většina shluků zůstala stejná; to potvrzuje homogenitu regionů. Zajímavé je, že horské oblasti a jižní Morava tvořily samostatné a homogenní shluky. Tento přístup ukázal pohodlnou, výpočetně a časově nenáročnou metodu pro prostorovou klasifikaci vzorců srážek. Jednou z hlavních výhod je jasná vizualizace shluků na mapě, která může být přínosem pro analýzu vzorců srážek a zlepšit analýzu rozptylu prostorových srážek. Prostorová klasifikace vzorců srážek může navíc zlepšit modelování a předpovědi povodní a sucha. Tato studie nezahrnovala komplexní analýzu četnosti a intenzity srážek a jejich vztahu k atmosférické cirkulaci, a proto tyto kroky lze doporučit pro budoucí vývoj a rozšíření analýzy. Při interpretaci výsledků této studie je navíc třeba vzít v úvahu kvalitu dat GPM.

Klíčová slova: samo-organizující mapa, shlukování, dešťové srážky v České Republice, GPM data

Contents

1 Introduction

Extreme precipitation events, especially long-lasting ones, significantly impact social-economical areas. It is expected that the future changes in precipitation extremes will bring new challenges: heatwaves and droughts frequency increase and influence human health, agriculture and water availability. [1]

Many researchers analyse precipitation using ground-based observations. Weather stations can provide high-quality data; however, their distribution is limited. Analysing precipitation requires considering the orthography and atmospheric circulation patterns, which are defined spatially. Thus, the availability of ground-based stations is essential in precipitation analysis. Different interpolation techniques exist to cover the gap in observation spatial distribution. However, the results may differ considering the chosen technique. Therefore, different satellite data collection missions were implemented to improve availability and precipitation analysis.

Satellite data varies in different spatially, time-scale resolutions. Moreover, it can include different variables, for example, precipitation estimates, temperature, altitude and coordinates. All variables' estimates are calculated using data from multiple sources, including the constellation of satellites and its calibration based on the ground station observation. The more detailed process of calibration and estimation processes for GPM[1] data, which was used in this study, is explained in Chapter 4, Data and Methodology. The high spatial and temporal resolution of the satellite data increases the amount of data that can be analysed. Traditional analysis approaches might not provide sufficient analysis results. Thus, the demand and interest in more efficient analysing techniques are growing.

One of the advanced analysing techniques is the unsupervised machine learning approach. This method is widely used in different fields, such as palaeoceanography, climatology, data cleaning and pre-processing, customer segmentation, population analysis, fraud detection and others. The dimensionality reduction is one of the tasks unsupervised learning can perform. In climatology studies, this methodology is essential because it can contribute to understanding the comprehensive non-linear relationship between variables. The main aim of the dimensionality reduction algorithms is present high dimensional data

[1] Global Precipitation Measurement Mission

to low dimensional space. SOM[2] was used in this study. The algorithm is described in Chapter 2.2.2, Self-organising maps algorithm. The most interesting unique property of the SOM is the saving of the topology of analysed data. It is reached by applying the neighbourhood function.

This study aims to classify precipitations spatially using the SOM algorithm. After applying SOM, low-dimensional prototypes of the original data were used to explore the effect of different clustering schemes using hierarchical clustering. Finally, one scheme was chosen to analyse clusters using the traditional statistical approach.

[2] Self-organizing map

2 Objectives

The thesis aims to analyse the spatial and temporal distribution of the precipitation across the Czech Republic using a machine learning approach. Due to increased volumes of the data, comprehensive precipitation patterns, an unsupervised machine learning algorithm SOM was applied for analysis. The first objective is to receive meaningful categories of the spatial precipitation patterns across the Czech Republic. In order to fulfil this objective, the GPM satellite product data was downloaded and pre-processed. The literature review regarding GPM satellite product data, the SOM algorithm and the Czech Republic climatology was conducted. R programming language was chosen to implement the data manipulating and SOM algorithm and analyse data. The second objective is to detect any changes in the spatial characteristics of precipitation during the last five years. Statistical analysis of received clusters was applied in order to fulfil this objective.

3 Literature review

In order to conduct the study, the Czech Republic climate with the focus on precipitation and extreme events is presented in the following sub-chapters. Moreover, the concepts of the unsupervised machine learning approaches, in particular SOM algorithms, are investigated and described.

3.1 Czech Republic: climate and landscape

A temperate climate in Central Europe characterises the Czech Republic. The main rivers of the western part are Elbe and Vltava, on the eastern - Morava River and Oder river. The Czech Republic is divided into the 14 districts presented in the Figure3.1. The

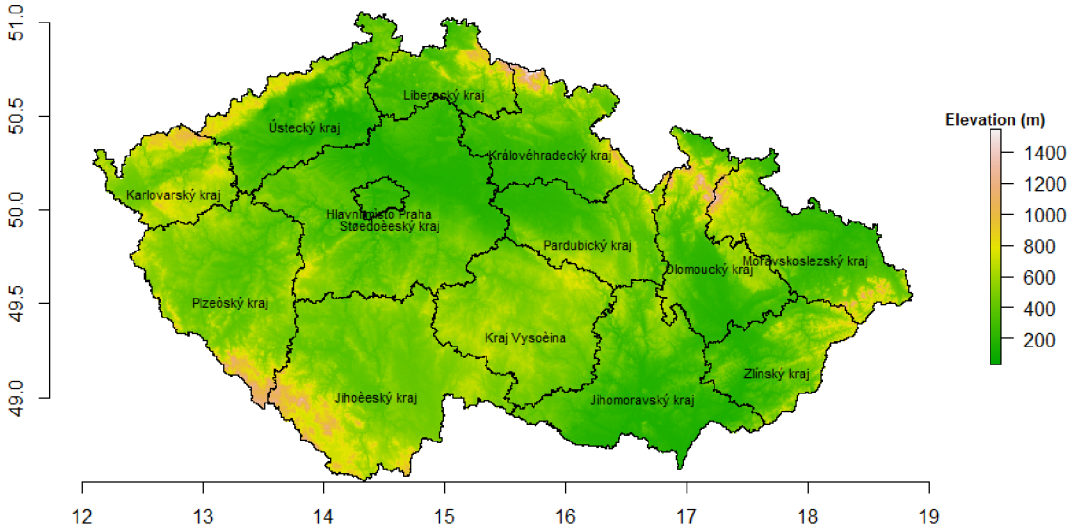

Figure 3.1: Czech Republic regions

Czech Republic's substantial spatial and temporal variability includes Atlantic, Mediterranean, and continental influence. Precipitation in central Europe is characterised by the

significant difference in precipitation between winter and summer time [2]. The maximum precipitation is the most common for July and the minimum for January. July is the warmest month, and January and February are the coldest. Precipitation during winter is usually characterised by lower intensity and longer duration. Otherwise, precipitation is higher intensity during summer and has a shorter duration. 37% of annual precipitation accounts for the summer season (July - August) and around 18% for the winter season (December - February). The northern mountain region of the Czech Republic is an exception to this precipitation pattern; for this region, winter precipitation totals are slightly higher than in summer. [3]

Beranová R. and Kyselý J. investigated the trends of precipitation characteristics in the Czech Republic. The results showed a significant difference between western and eastern regions. For the west part of the country, Atlantic influence is stronger for the east region - the continental and Mediterranean. Furthermore, the research aligns the dependence of precipitation trends in winter and altitude with NAO indexes. [4]

3.1.1 Extreme events CR

The Czech Republic has experienced extreme precipitation events such as droughts and floods for the past decades. The lack of long-lasting snowpacks, glaciers, natural lakes and aquifers makes the Czech Republic vulnerable to long drought events, which probability expects to increase [5]. Therefore, estimating the drought risk is essential for strategic planning, as it affects natural ecosystems, the economy and society [3]. South Moravia precipitation pattern showed a drier trajectory than Bohemia for precipitation in August 2011- May 2012 [5]. Zahradníček P. discovered that in 2011 - 2012 Moravia showed weak hydrological drought, which can be explained by the spring of the Morava river in the north [5]. Furthermore, Kyselý J. suggested that the eastern part of Europe has a more complex pattern in changes in precipitation than the eastern, which also can explain the rainfall in the Czech Republic [6]. Vojtech Bliznak et al. l. research confirmed the dependency between the spatial distribution of the mean and maximum precipitation in summer with altitude [7].

Droughts

Drought is one of the extreme events related to precipitation extreme. Droughts affect many socio-economic sectors, including agriculture, forestry, water management, the ecosystem's health, and people. Drought can occur from a precipitation deficit compared with the expected or normal amount. Droughts events can develop over the years and persist for a long time. Droughts can be divided by their spatial distribution to regional,

national and continental scales. Recently, lower intensities droughts that occur for days or weeks were described and called ”flash” droughts.

Czech Hydrometeorological Institute introduces a drought monitoring system. However, the European Observatory system with temporal resolution from multiple days to 1 month and spatial 1*1 km CHMI[1] monitoring system can benefit operational management at the local level. Considering the primary usage - agriculture, the system has a high resolution - 250 m. CzechDM[2] collects soil moisture from radar measurements based on the Central European level, soil moisture at 500 m, drought reports provided by farmers and foresters, remotely sensed vegetation conditions. [8]

In the 21st century, there have been several severe drought events in the Czech Republic. The drought during spring in 2000 resulted in crops damage and loss of the harvest. The drought of the year 2012 affected a relatively small region, whereas the event of 2015 affected the whole country from July to August. Additionally, CzechDM analysed the past drought events and drought forecast.

One of the most severe droughts in the Czech Republic was an extreme event in 1947. It lasted from April to October. Additionally, it is the most severely influenced social-political field [9]. Brazdil’s (2009) research confirms the significant tendency towards more intensive dry events in the Czech Republic. One of the driving reasons is temperature increase and precipitation decrease. However, changes in precipitation totals are not as crucial as changes in temperature. The main reason is global warming. It has been confirmed that for the long term drought cycles, the increase in air temperature leads to higher potential evapotranspiration estimates, which results in a higher probability of soil moisture deficit. Whereas for the shorter cycles, soil moisture drives mainly by precipitation. Projections of the global circulation models for the Czech Republic show an increase in evapotranspiration caused by temperature increase. Additionally, it shows that precipitation will not increase enough to balance the rise in temperature, which results in more frequent drought events. [10]

Floods

In addition to the increasing droughts frequency events, floods occurrence also increased in the past ten years. The most extensive flood took place in 2002, 2004, 2005 and 2007 [11]. Analysis of rainfall data in the Czech Republic showed the share of the heavy, short-term precipitation to the precipitation total significantly increased at about half of 17 weather stations spread over the country [12].

Floods like droughts threaten social and economic fields. Floods might destroy or

[1] Czech Hydrometeorological Institute
[2] Czech Drought Monitoring

damage cities' housing, transport infrastructure, engineering works and utilities, water structures and streams; it is dangerous for people and animals' life. In addition, like a drought, floods events also affect the agriculture and forestry sector. In the past decades, many flood events occurred: Moravia - July 1997 and May to June 2010, Bohemia - August 2002 and June 2013, spring floods - March 2006, North Bohemia - August 2010. Floods of the June 2003 and August 2002 were caused by large-scale precipitation events, so the lower part of the Elbe and Vltava rivers reached their maximum flow. [13]

3.2 Methods to analyse precipitation

The most common approach to investigate and predict drought is to use drought indices. The main purpose of the drought indices is to predict the duration, intensity and spatial distribution of drought. The commonly used indices can be found on the world meteorological organisation resources. One of the most common is SPI.

Some indices can be calculated based on the ground collected data, whereas others require spatial data. Ground-based require temperature and precipitation variables. Remotely based may include spatial data such as precipitation, temperature, evapotranspiration, and vegetation information. Remotely based indices allow better coverage of spatial distribution, although the quality of spatial data is limited to atmospheric conditions and algorithms [14]. As a result, the accuracy of remotely sensed indices is questionable [15]. At the same time, ground-based indexes are limited to the availability of the stations; thus, spatially limited [16]. Researches use spatial interpolation techniques to cover gaps in spatial distribution, although interpolation may cause uncertainties due to the specifics of the algorithm and topography [17]. Although the remotely based indices benefit in spatial distribution, they can not replace ground-based indices due to the short period of records [18].

Consequently, many researchers tried to reproduce ground-based indices using ANN[3] [19] and autoregressive integrated MA[4] models [20]. However, these models have limitations to work with non-linear or non-stationary processes. More advanced machine learning methods can be used to cover these limitations. Machine learning methods can recognise co-linear variables and non-linear relationships between variables. These capabilities result in better performance in droughts recognition[16]. Dimensionality reduction is one of the methods capable of working with independent variables. Dimensionality reduction methods are described in the following chapters in more detail.

[3] artificial neural networks
[4] moving average

3.2.1 Unsupervised machine learning

Unsupervised machine learning is used for analysis to discover an informative way to visualise or group high dimensional data. One of the unsupervised methods is principal component analysis. PCA[5] allows explaining original data with a smaller number of representative variables. PCA seeks for the low dimensional representation of the data, which preserves the most extensive possible variability.

PCA allows mapping data set into two dimensions. PCA projects correlated variables into a smaller number of uncorrelated variables and keeps as much variation in the original dataset as possible [21]. However, PCA has its limitations:

1. only two dimensions, for some problems, a higher dimensional visualisation might require,
2. uses only Euclidean distance as a dissimilarity measure.

Unsupervised machine learning techniques which aim to find subgroups in the data set are called clustering. The main purpose of clustering is to collect similar observations in one group, whereas observations from the different groups will have different characteristics. For example, clustering as a PCA aims to explain data through a smaller number of variables. However, whereas PCA is a low-dimensional representation of the data, clustering aims to divide data into homogeneous subgroups [22]. There are many clustering techniques, among them are K-mean, hierarchical clustering and SOM, which were used in this work.

K-mean clustering

In K-mean clustering, the number of clusters should be predefined. Then, each observation is assigned to the cluster. Within-cluster variation measure $W(C_k)$, where C_k is a denotation for the number of observations in the kth cluster, is used to evaluate the performance of the K-mean clustering. The smaller C_k means, the better performance. The within-cluster variation estimates the value showing how observation differs from each other within a cluster. So, to find the best performance, the within-cluster variation should be summed over all K clusters. One of the approaches to estimating $W(C_k)$ involves the squared Euclidean distance. Therefore, within-cluster variation for the particular cluster is equal to the sum of all squared Euclidean distances between the observations in this cluster divided by the total number of observations in this cluster.

The initial step is randomly assign the cluster number from 1 to K to each observation. Then, the cluster centroid is estimated for each cluster. When centroids are computed, each

[5]Principal component analysis

observation is assigned to the closest cluster. The closest cluster is defined by squared Euclidean distance. Steps of centroid estimation and assignation of the observation are repeated until all observations belong to the closest cluster. Considering that the initial assignation of observations to the clusters is a random process, it is necessary to run the algorithm several times and select the best result. For instance, the smaller within-cluster variation sum can justify the best result. [22]

Hierarchical clustering

Another clustering method is hierarchical clustering. The main difference from the K-mean is that the number of clusters should not be predefined. The hierarchical clustering approach creates a tree-based visualisation of the observations called a dendrogram. All observation is visualised at the bottom of the dendrogram; they are grouped by clusters on the next level. On the next level, these clusters are grouped again and they are grouped until two clusters are formed. Thus, the number of clusters can be selected by choosing the rational number of clusters and often, the choice is not clear. The limitation of this method to choose clusters is a nested approach. Dataset might consist of characters that cannot be nested. They can present different strategies on how to create subgroups. The building of the dendrogram starts from the bottom, and each observation is treated as its cluster. In the next step, the two clusters with the smaller Euclidean distance are merged. The new number of clusters is n-1, where n is the number of observations. This algorithm repeats until all observations belong to one cluster. [22]

Both methods have their limitations. As mentioned before, the most significant limitation is a nested structure for hierarchical clustering. In addition, the visualisation of the dendrogram might be too heavy and confusing for big data sets. K-means is limited by topography; it assumes only spherical clusters and tries to assign an equal number of observations to each cluster [23]. In addition, K-means clustering shows a decrease in efficiency with high dimensional data [24].

3.2.2 Self-organising maps algorithm

SOM is similar to K-mean clustering in some way. A cluster of k-mean can be compared to the SOM's nodes. The number of clusters in k-means is defined by the number of centroids, whereas in SOM, by the size of the grid [25].

At the begging of the SOM algorithm, each node is assigned randomly with a codebook vector. This vector will be a typical pattern for this node. Each object used for training is compared with the training node codebook vector. The "winning unit" will be updated with a weighted average to become more similar. This weight is used as a training parameter and called a learning rate α. With each iteration, the learning rate decreases. As

mentioned before SOM preserves topology and keeps neighbours close to each other. It means that neighbouring nodes should have similar codebook vectors. For this purpose, the ”winning units” are being updated. The algorithm stops after a predefined number of iterations.[26]

Despite the randomness in the initialisation, it has been noticed that conclusions based on the different maps stayed consistent. Therefore, it is still recommended to train several maps [25]. When the classes for the data are unknown, however, the relation between samples is assumed, the clusters can be found by unsupervised classification. There are self-organising maps, among other unsupervised learning algorithms; their main feature is to project and visualise high-dimensional signal spaces on a two-dimensional display. First, decoders or detectors are developed from the various neurons according to their signal domains. Then the network is formed, consisting of these decoders in a meaningful order. [26] SOM allows the two-dimensional grid of nodes for high-dimensional data with statistical relationships.

General learning process

The set of input variables are real vector:$x = [\epsilon_1, \epsilon_2, \epsilon_3..., \epsilon_n]^T \in \Re^n$. SOM creates a model where a parametric real vector $m^i = [\mu_i 1, \mu_i 2, ..., \mu_i n]^T \in \Re^n$ is created for each element. The distance between x and m_i is $d(x, m_i)$. The main goal is to find the m_i, so the mapping is ordered and descriptive of the x distribution. The initial values of the mi can be random. Each x(t), where t is an integer index, is compared with m_i and copied into the sublist of the most similar node to x(t). After all the x(t) have been sorted, we consider neighbourhood set N_i around model m_i. The neighbourhood set consists of nodes, which are distanced from node i up to a certain radius. The next task is to find the sample median of all x_i, so that it has the smallest sum of distances from samples x(t), where t is within neighbourhood N_i. $\overline{x}_i$ is a generalised set median and is restricted to being one of the x(t). However, x(t) might not cover the whole input domain, and thus the aim is to find the $\overline{x}'_i$ (generalised median) with an even smaller sum of distances from the x(t). The generalised median can be equal to the arithmetic mean for Euclidean vectors or the sum of squares of Euclidean distance for arbitrary Euclidean vectors from all the samples of the neighbourhood set. For example, it can be defined by the euclidean distance

$$d(x, x') = \sqrt{\sum_{i=1}^{n}(x_i - x'_i)^2} \tag{3.1}$$

where n is the dimension of the input vector. The next step is to form $\overline{x}_i$ and $\overline{x}'_i$ for each node and replace m_i respectively to the $\overline{x}_i$ and $\overline{x}'_i$. The original x(t) are again distributed into the nodes and the new $\overline{x}_i$ and $\overline{x}'_i$ are generated again and replace m_i. This regression process can be done until a satisfying sum of distances will not be reached. [26]

Neighbourhood and learning rate functions

The topology can be rectangular, hexagonal or even irregular. Input vector x(t) is connected to all nodes in parallel via scalar weights μ_{ij}, which is unique for each node. Vector x is compared with all m_i, in a simple case, Euclidean distance is used to define the best matching node for the vector x. Nodes that are topologically close to each other over to certain distance will learn something from the input vector x. This will affect the nodes' weight vectors and lead to global ordering. Initial values of the $m_i(0)$ are assigned randomly:

$$m_i(t+1) = m_i(t) + h_{ci}(t)[x(t) - m_i] \tag{3.2}$$

$h_{ci(t)}$ is the neighborhood function. This function should seek to zero whereas t to the endless to reach the convergence. The neighbourhood function is:

$$h_{ci}(t) = h(||r_c - r_i||, t) \tag{3.3}$$

r_c and r_i are the location vectors of the nodes. Different algorithms can be chosen for the neighbourhood function depending on the size and complexity of the SOM network. Another example of the algorithm which can be used for the neighbouring function is a smoother neighbourhood kernel written in terms of the Gaussian function:

$$h_{ci}(t) = \alpha(t) * exp(-\frac{||r_c - r_i||^2}{2\omega^2(t)}) \tag{3.4}$$

$\alpha(t)$ is a learning rate factor and ω(t) is the width of the kernel. Learning rate might be the following function:

$$\alpha(t) = \alpha_0 * (\alpha_{end}/\alpha_0)^t/t_{max} \tag{3.5}$$

The learning rate will have high values at the beginning of the learning process, in case of initial values are assigned randomly, and will decrease monotonically. The learning rate function can be linear, exponential or inversely proportional to t. The most important is that by the end of the learning process, α should assign to small values. There are online and batch learning modes. In online learning mode, the weight vector updates using the following equation:

$$w_i(t+1) = w_i(t) + \alpha h_{ij}(t)(x(t) - w_i(t)) \tag{3.6}$$

In batch learning:

$$w_i(t+1) = \frac{\sum_{j=1}^{N} h_{c,i}(t) * x_j}{\sum_{j=1}^{N} h_{c,i}(t)} \tag{3.7}$$

where $h_{c,i}(t)$ - neighborhood function of the node i at iteraction t. In this recursive process, we shall set up some parameters. First, the lattice type of the two-dimensional array of

nodes can be defined as rectangular, hexagonal or irregular. Hexagonal is considered the most effective for visualisation. The learning-rate factor should be followed to find the best-suited number of interactions. According to Teuvo Kohonen: ”the ”rule of thumb” is that, for good statistical accuracy, the number of steps must be at least 500 times the number of network units.” Teuvo Kohonen and his team have used up to 100000 steps; however, it is considered that 10000 steps and even less may be enough in some cases. [26]

3.3 Applications of machine learning in precipitation analysis

Bochenek B. and Ustrnul Z. conducted comprehensive research on the application of machine learning approaches in weather prediction and climate analysis. They highlighted the significance of the machine learning application in meteorology and climatology due to the fact that these fields deal with circulation-related issues, and thus there is a lack of quantitative definitions or criteria to perform objective analysis. Authors consider that the interest in machine learning applications will continue to grow. [27]

Purposes for application of the machine learning approaches for analysing precipitation vary. In the recent studies different ML[6] approaches were used to find the relation between the occurrence of regional precipitation and discharge extremes to synoptic-scale climate fields [28], to cluster catchments and classify them (SOM algorithm) [29, 30], prediction of the intensity, duration and frequency in assessing the extreme precipitation and floods [31], an estimation of the trends and seasonality components [32, 33, 34].

Guntu R. investigated the seasonal and temporal variability in regionalisation for the Indian subcontinent. The authors used SOM and a standardised variability index. The validation of the cluster was done with the Silhouette coefficient (SC). The cluster was defined by precipitation magnitude and its temporal variability analysis. Although the spatial constraints were not used for cluster classification, the regions were geographically cohesive in almost all cases. To further investigate the temporal variability of clusters, SOM was applied to the data split into three time periods. This approach revealed a significant change in different precipitation intensities and events across the same cluster using this approach. [35]

Another recently published research, conducted by Zeng P. et al., is aimed to target the most vulnerable to climate change area in Chine. SOMs K-means was used in this research, which helped distinguish clusters based on droughts characteristics and provided understandable and visualised results, which traditional methods can not easily reach. [36]

[6]Machine Learning

The recent study conducted by Markonis Y. and Strnad F. applied a self-organising map algorithm to classify the European continent and describe its spatial properties. The classification was performed using ”someplace” and ”Kohonen” R packages. Hexagonal grids of different sizes were used in this study. The within-cluster and between-cluster heterogeneity was applied to identify a sufficient number of clusters. The results showed that the resulting schema does not strongly depend on the number of nodes. Moreover, the received regions compile with known climatic processes. [37]

4 Data and Methodology

R programming language was chosen to perform data manipulation and application of the SOM algorithm. ”kohonen” R package is a convenient, simple to use the package to implement SOM. Sattelite GPM product data storage structure is reviewed, and pre-processing is performed to prepare data for the convenient format, which can be used for analysis. Moreover, methods to define a number of clusters for the model and clustering techniques are explained in the following sub-chapter.

4.1 GPM data

Precipitation studies are limited in some areas by lacking gauge stations and data. In helping cover these data gaps, the Tropical Rainfall Measuring Mission was launched in 1997 and operated until 2015. The TRMM[1] was operated by NASA[2] (National Aeronautics and Space Administration) and the JAXA[3] (Japanese space agency). The TRMM data of tropical precipitation aimed to improve understanding of precipitation and weather.[38]

Due to the success of the TRMM, NASA and JAXA launched Global Precipitation Measurement Mission in 2014. GPM is extended to improve near real-time monitoring of hurricanes and rainfall. TRMM primarily caught heavy to moderate precipitation over tropical and subtropical oceans, whereas GPM can detect the light (< 0.5 mm hr-1), solid, and the microphysical properties of precipitating particles.[39]

The GPM Core Observatory satellite includes the dual-frequence precipitation radar and a multi-channel GPM Microwave Imager. DPR[4] is responsible for the three-dimensional measurements of precipitation structure and characteristics, and it is more sensitive to light rain rates and snowfall. The GMI[5] can catch heavy, moderate and light precipitations.[39]

The data collected for more than 22 years from the TRMM and GPM missions can be reached pre-processed by the Integrated Multi-satellite Retrievals for GPM. This al-

[1] Tropical Rainfall Measuring Mission
[2] National Aeronautics and Space Administration
[3] Japanese space agency
[4] Dual-frequence precipitation radar
[5] GPM Microwave Imager

gorithm processes data from the constellation of satellites and estimates near-real-time precipitation over the Earth. [39]

GPM data contains metadata, including name, date and time of the observations, and data. Depending on the processing level of the data, it is stored in swath or grid structures. Levels 1 and 2 are stored in the swath format; they resemble the instrument scanning pattern. In this research, level three was used. It is stored in a grid structure and grided to the regular global grid. Grid structure contains grid boxes, which store data arrays.[40]

There are different products retrieved from the GPM data. This research uses an "IMERG[6] 30-minute" product. Integrated Multi-satellitE Retrievals data is pre-processed output. Data from the satellite is being inter-calibrated, merged and interpolated with satellite microwave precipitation estimates and precipitation gauge analyses. Figures 4.1 and 4.2 from the "PRELIMINARY File Specification for GPM Products" document show the structure of the "IMERG 30-minute" product [41]. IMERG retrievals are divided into

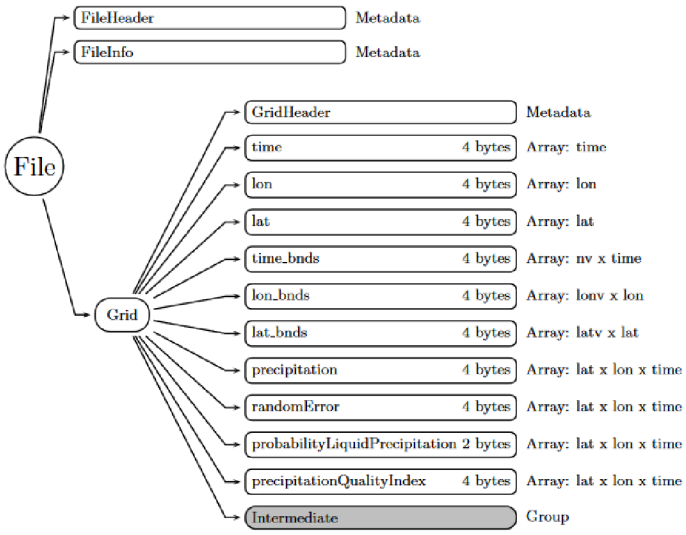

Figure 4.1: Data Format Structure for 3IMERGHH, IMERG 30-minute

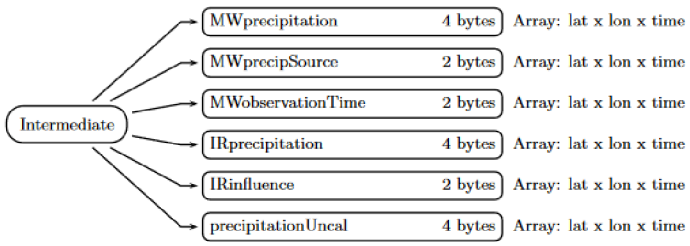

Figure 4.2: Data Format Structure for 3IMERGHH, IMERG 30-minute

[6]Integrated Multi-satellitE Retrievals

three types depending on the pre-processing stage. ”Early” multi-satellite product can be calculated after 4,5 hours from observation time, ”late” multi-satellite product - 14,5 hours and ”final” satellite-gauge product - 3.5 months after observation month. The Final-run product precipitation estimates are calibrated so that their monthly totals equal the monthly satellite-gauge combination computed in IMERG.[42] There are several variables available to download from GPM. In this research, ”PrecipitationCal” will be used. ”PrecipitationCal” is a variable of the final run precipitation estimates calibrated retrospectively by gauge calibration. IMERG Algorithm Theoretical Basis Document recommends this variable for general use.[42]

The overview of the GPM data application by Kirschbaum D et al. shows that GPM data is widely applicable [43]. It is valuable in the following areas: extreme events and disasters, water resources and agriculture, weather and climate modelling and public health and ecology. Whereas the data had significant input in several fields, it is crucial to continuously evaluate and validate it by the users' community. Pradhan R. et al. conducted another overview of GPM IMERG performance. The research reviews IMERG product usage across different climatic conditions and geographical locations. There are potential validation limitations of the IMERG data. For example, if data is compared with gauge stations precipitation observation, the uncertainty of interpolation techniques can influence the validation results. Based on the studies analysed in this overview, IMERG estimates the trend to have poor performance over complex terrains and mountain regions. [44]

4.2 Pre-processing of the GPM dataset

The IMERG dataset is used for this research. The Earth data feature is used to download the GPM dataset because it can subset data and generate all necessary links for downloading. Data consist of nc4 files for every 0.5 hours. The resolution is 0.1 degree * 0.1 degree. R is used to analyse the data. Data is stored as a grid, and it was formatted in tidy format, and all non-zero nc4 files were merged into one Table. Variable PrecipitationCal is chosen; the unit of the variable is mm/h.

Data can be downloaded as a rectangular shape with coordinates: 11.95, 48.55, 18.95, 51.15. As this research aims to analyse the Czech Republic, coordinates were cut to the country shapefile. Czech Republic shapefile is downloaded from the Geoportal CUZK with the name - ”Topographic database of the Czech Republic (Data200) - layer Settlements” [45].

The first step is to reshape data to the long table format with four columns: ”lat”, ”lon”, ”precipitation”, and ”date”. Table 1 shows the first ten rows of the reshaped data. Considering units of the precipitation variable - mm/hr and the fact that the variable is estimated for every half an hour, the average intensity was taken for each hour. There-

fore, average hourly intensities were summed to get the daily amount of precipitation. For convenience, columns containing year, month, name of the season were added, time was removed from the date column. Table 2 shows the first ten rows of the daily precipitation data. The daily precipitation amount was further aggregated to get:

1. Average annual number of wet days
2. Average number of wet days seasonally (4 columns)
3. Average annual sum of precipitations
4. Average sum of precipitation seasonally (4 columns)

A number of wet days were calculated for each year to get an average number of wet days, then the mean number of wet days was taken for each grid. The day is considered wet if its precipitation amount exceeds 0.1mm. To get an average number of wet days by season, first, the number of wet days was calculated for each season and year, then the average number of days was calculated for each season. All results were merged in one data frame. The first ten rows of the resulted data frame are shown in Table 3.

In addition, wet days were divided by precipitation amount; the classes were created from quantiles. For the values less than 25% quantile were classified as 'light', values between 25% and 50% - 'medium', 50% - 75% - 'high' and values which are greater than 75% quantile - 'very heavy'. The number of wet days was calculated regarding type and season; the following columns were received:

1. Light/moderate/high/very heavy average annual number of wet days (4 columns)
2. Annual number of wet days
3. Average number of wet days by seasons (4 columns)
4. Average number of wet days by season and type (16 columns)

In addition, the amount of precipitation totals was calculated annually, seasonally and by type. As a result, new columns were received:

1. Light/moderate/high/very heavy annual precipitation totals (4 columns)
2. Annual precipitation totals
3. Average precipitation by seasons (4 columns)
4. Average precipitation by season and type (16 columns)

The resulting Table contains 52 columns.

4.3 R package Kohonen

A two-level approach to clustering can be implemented when analysing high dimensional data. First, SOM model is created to receive lower-dimensional prototypes of the data. The grid size should be larger than the expected number of clusters. The next step is to cluster prototypes using hierarchical clustering to receive actual clusters. This approach allows for decreasing computational costs. In addition, using clustering methods on prototypes avoids hierarchical and k-means clustering limitations, such as nested dependency and lower efficiency for high dimensional data. [23] The R package is the set of functions to build and visualise self-organising maps. The somgrid function should be called in order to initialise the grid. It defines the size, and the topology of the grid [25]. After the grid is created, the model can be created using the function som. Several parameters can be changed or used default ones.

1. the input data should be standardised and formatted as a matrix
2. grid - the object created by the ”somgrid” function
3. rlen - number of iterations
4. mode - ”online”, ”batch” or ”pbatch”. ”pbatch” is used in this research
5. cores - number of cores for parallel computing

Parallelism and batch mode

The time required for SOM training increases with the size of the dataset. ”pbatch” mode allows parallel implementations. Batch mode updates weights based on the formula3.7. The training process is distributed in a memory system. The best matching unit independently finds each input variable, resembling the parallel version. The equally sized parts distribute data to each node. Communication between the primary and supplementary nodes is required to update the codebook. After all weights are calculated in the supplemental node, they are sent to the primary node, updating the codebook. Finally, the new codebook is forwarded to all supplementary nodes. [46]

4.4 Clustering of the SOM model

The ”Elbow” method is used to define the relevant number of clusters for each model. The method involves computing the within-cluster sum of squares with K-mean clustering.

The function ”kmeans()” is used to initialise K-means clustering [22]. K-mean within-cluster sum-of-squares can be obtained using ”k-mean.model$withinss”. K-mean within-cluster sum of squares values are computed for a different number of clusters and are stored in the array. Then, the values are plotted against the number of clusters to visualise the changes. The within-cluster sum of squares decreases with the increasing number of clusters. The idea is to select the number of the cluster after which the changes will not be so dramatic as in the beginning [47].

When SOM codebooks are computed, hierarchical clustering can be applied to have better visualisation of clusters. First, the Euclidean distance matrix of the codebook is created using the ”dist” method from the ”stats” package. Then, ”hclust()” is used to perform the hierarchical clustering function. Then, ”cutree()” with a selected number of clusters is used to group observation to a selected number of subgroups.

5 Results

The research includes results on different experiments with SOM algorithm and clustering. Then the results of the exploratory data analysis of the chosen classification scheme are presented.

5.1 Application of SOM

SOM was trained on three sets of variables: annual, seasonal and all. ”pbatch” was chosen as a learning method for faster calculation, as it allows to perform computation in parallel, using a several cores. In addition, the experiment included different values for the following parameters:

1. number of iteration: 100, 1000, 10000
2. hexagonal grid: different grid sizes - 3*3, 4*4, 5*5
3. number of clusters: from 3 to 6 clusters

The k-mean matrix was created to find the optimal number of clusters, and a sum of the squared Euclidean distance was plotted for each number of clusters. Then the hierarchical clustering was performed with a chosen number of clusters. Finally, each point was clustered and plotted grouped by its cluster.

5.1.1 Number of iterations

The number of iterations depends on the nature of the data; it can differ. Teuvo Kohonen and his team used 100000 interations [26], however they consider that 10000 or less might be enough. Yannis Markonis and Filip Strnad used 10000 interactions for a paleoclimatic dataset to explore homogeneous of regions [37]. Generally, the distance between variables within-cluster should decrease with a higher number of iterations, and clusters become more homogeneous. The size grid 3*3 (9 nodes) was used to compare the effect of

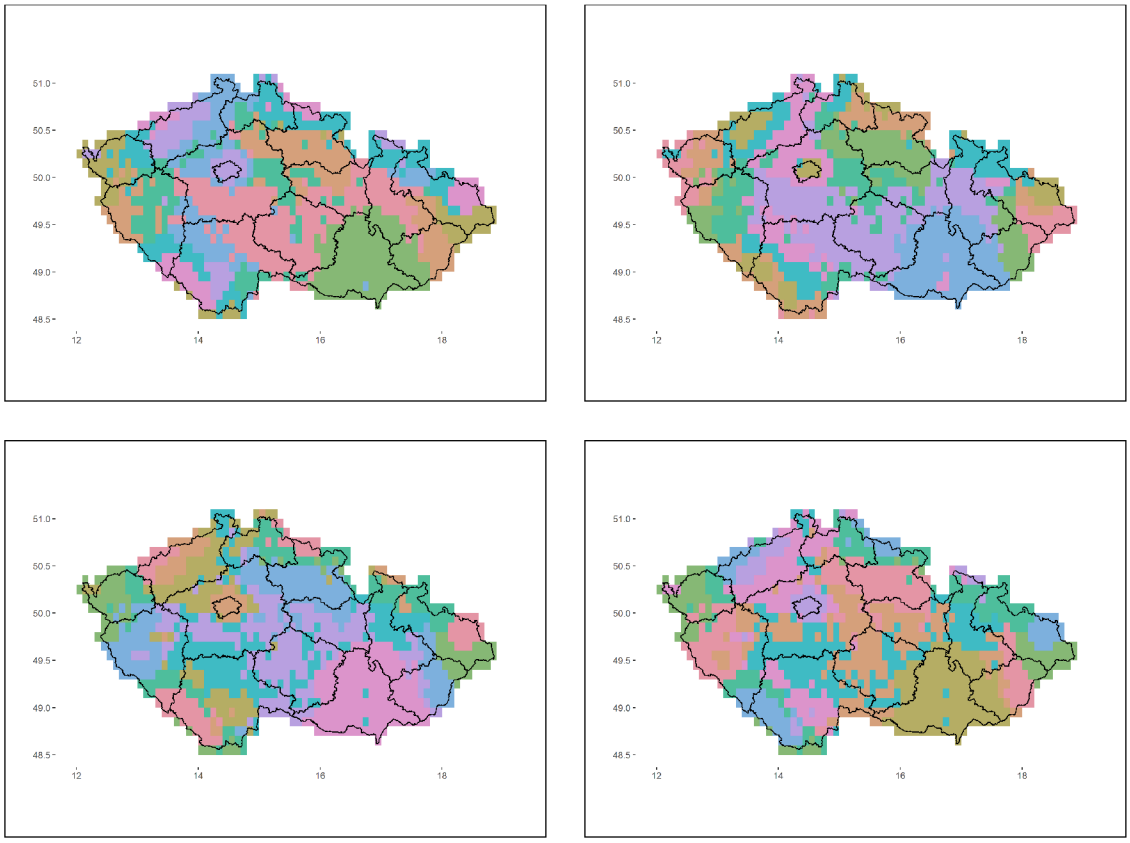

Figure 5.1: SOM's clusters for 100, 1000, 10000 and 100000 iterations (left to right)

a different number of interactions. Figure 5.1 shows the resulted node for each coordinate with different number of iterations.

All four results look similar; however, a slight change is with 100000 iterations - some coordinates appeared alone with fewer iterations surrounded mostly by another cluster, now form bigger groups. Therefore, for further experiments, 100000 interactions were chosen.

5.1.2 Different grids

Grid size should be chosen higher than the expected number of clusters. The experiment included grids with sizes: 3*3, 4*4, 5*5. Figure 5.2 shows the resulting clusters for each coordinate.

Similar classifying patterns characterise all three maps. For example, grids 3*3 and 4*4 define almost all coordinates in the South Moravia region (Jihomoravský kraj) to a unique node, although a new pattern among this cluster is detected on the grid 5*5. Additionally, almost all coordinates in Prague belong to one cluster for all grid sizes. Similar patterns can be noticed on the same experiment with seasonal variables in Figure 1 and all variables - Figure 2.

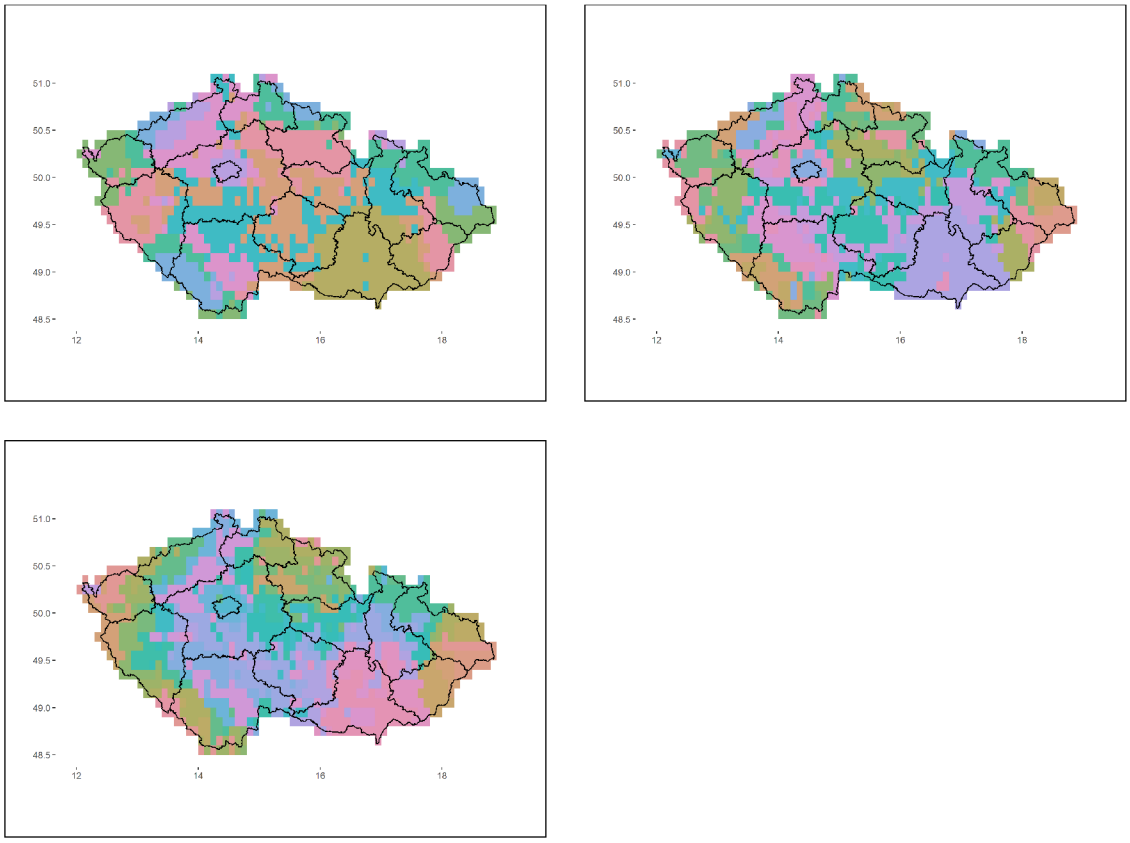

Figure 5.2: SOM's clusters for different grid sizes - 3*3, 4*4, 5*5 (left to right)

5.1.3 Different number of clusters

When coordinates are assigned to the nodes, hierarchical clustering can be applied to find the optimal number of clusters. All variables from the dataset were used in this experiment to create a SOM model. The plot of the within-cluster sum of squares by the k-mean algorithm was used to explore suitable numbers for clusters for each grid. Figure 5.3 shows the results for different numbers of clusters. Examining the "elbow point" revealed that the number of clusters can be chosen between 2 and 6. The experiment with only annual and season variables showed similar results (Figure 3). It is notable from these three figures that with fewer variables used in the input into SOM model, the within sum of squares between clusters is smaller with a smaller number of clusters.

Hierarchical clustering was performed for each chosen grid size and different variations of the dataset variables. Figure 5.4 shows the results for the model based on all variables from the dataset and represent results for 2-6 clusters for each grid size. Figure 4 and figure 5 presents the same experiment for only annual and seasonal variables of the same dataset.

In most cases, the south of Středočeský kraj, Vysočina, all/south part of the Pardubický kraj, the western part of the Olomoucký kraj, the north-eastern part of the Jihočeský

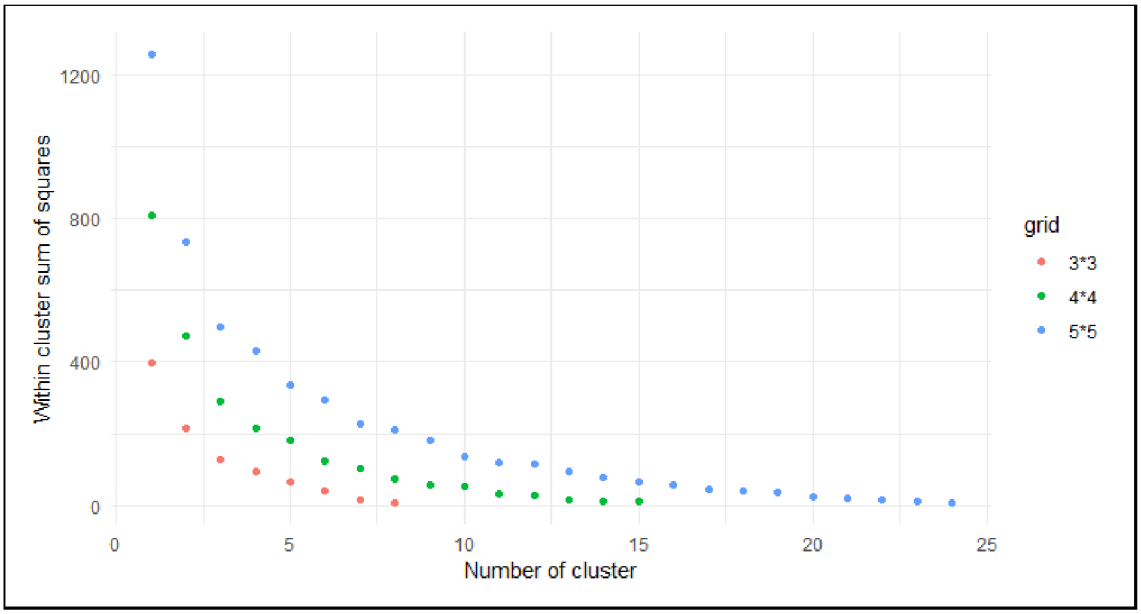

Figure 5.3: Within cluster sum of squared for different sized grids

kraj, the eastern part of the Plezenský kraj and Jihomoravský kraj belongs to one cluster. Jihomoravský kraj forms unique clusters for models: all variables 3*3 grid 4-6 clusters, all variables 4*4 and 5*5 grid 6 clusters, annual variables 4*4 6 clusters, seasonal 3*3 for all variance of cluster number, seasonal 4*4 4-6 clusters and seasonal 5*5 5-6 clusters. Moreover, mountain regions, such as Ore mountains, Sudetes mountains, Orlice mountains and Sumava mountains, belongs to one cluster in most cases. Additionally, Prague tends to have different clusters from the surrounding areas; it usually belongs to the cluster mainly consisting of the mountain regions, if the number of clusters is higher than with regions close by mountains. The above findings also are reflected in figures 5.2, 2, 1. Clusters show similar trends through all figures; thus, the model based on seasonal variables with a grid size of 4*4 and 6 clusters were chosen for further analysis.

5.2 Analysis of the received clusters

Figure 5.5 shows clusters distribution on the elevation map. The first and forth clusters are mainly distributed among the middle range elevation, second - close to mountains region and Prague region, third and fifth - mostly mountains regions, sixth - mostly cover the South Moravian region. The table 5.1 shows the summary statistic across all clusters and the Czech Republic. Cluster 4 is closer to the whole Czech Republic by mean annual precipitation totals. Cluster 5 has the highest mean precipitation annually and the highest variability, whereas cluster 6 has the lowest precipitation totals and lowest variability. Figure 5.6 shows a similar seasonality trend: clusters 5 and 6 show higher variability for April and August than other clusters. Cluster 5 shows the highest variability in June com-

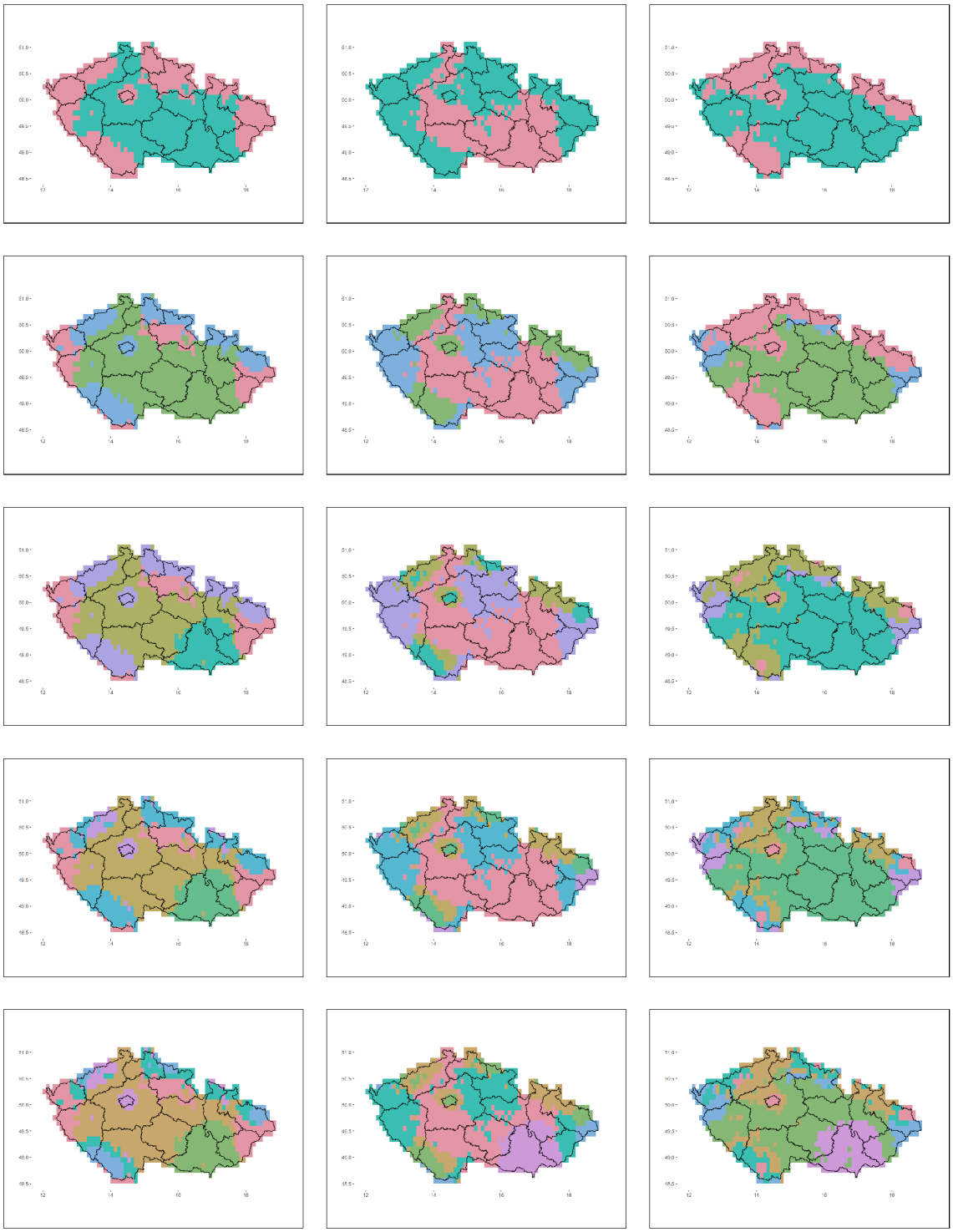

Figure 5.4: SOM's 2-6 clusters for grids 3*3, 4*4, 5*5

	cluster	mean	sd	min	max	CV
1	1	853.82	110.97	632.78	1059.56	0.13
2	2	887.23	119.45	642.79	1146.22	0.13
3	3	992.58	123.79	731.24	1223.33	0.12
4	4	936.17	113.61	709.33	1153.93	0.12
5	5	1071.72	135.06	820.06	1332.64	0.13
6	6	796.90	108.75	610.61	1052.32	0.14
7	CR	923.07	147.66	610.61	1332.64	0.16

Table 5.1: Annual precipitation statistic summary

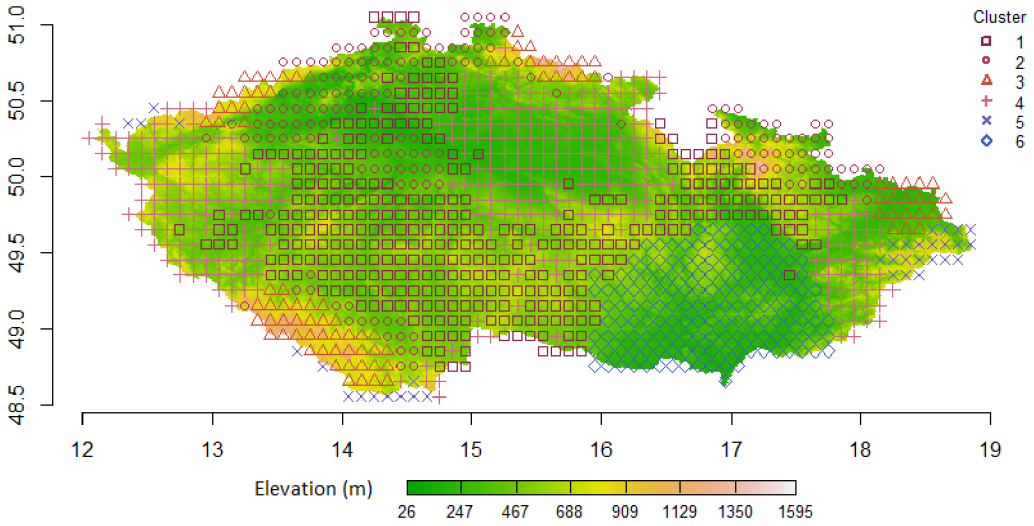

Figure 5.5: Clusters on the elevation map

pared with other clusters. Cluster 6 has the smallest variability in November and December compared to other clusters.

Table 4 shows outliers for monthly precipitation among each cluster. Cluster 4 has the highest number of outliers; cluster 6 has the smallest. The highest number of outliers belong to 2020, the smallest 2005,2018,2009, 2011 (by one outlier for each year). The highest number of outliers happened in August, smallest in March, October and November. The minimum monthly precipitation amount was in November 2011 (cluster 6) and the highest in August 2002 (cluster 1). Figure 5.7 shows the comparison between clusters' annual variations in monthly precipitation totals.

Figure 5.8 shows the annual variation in seasonal precipitation totals. The noticeable difference among clusters is in the winter (DJF) season. Cluster 6 has smaller precipitation totals, followed by cluster 1. The summer season shows higher precipitation among all clusters.

Figure 5.9 shows the precipitation totals for each cluster and year through 2001 to

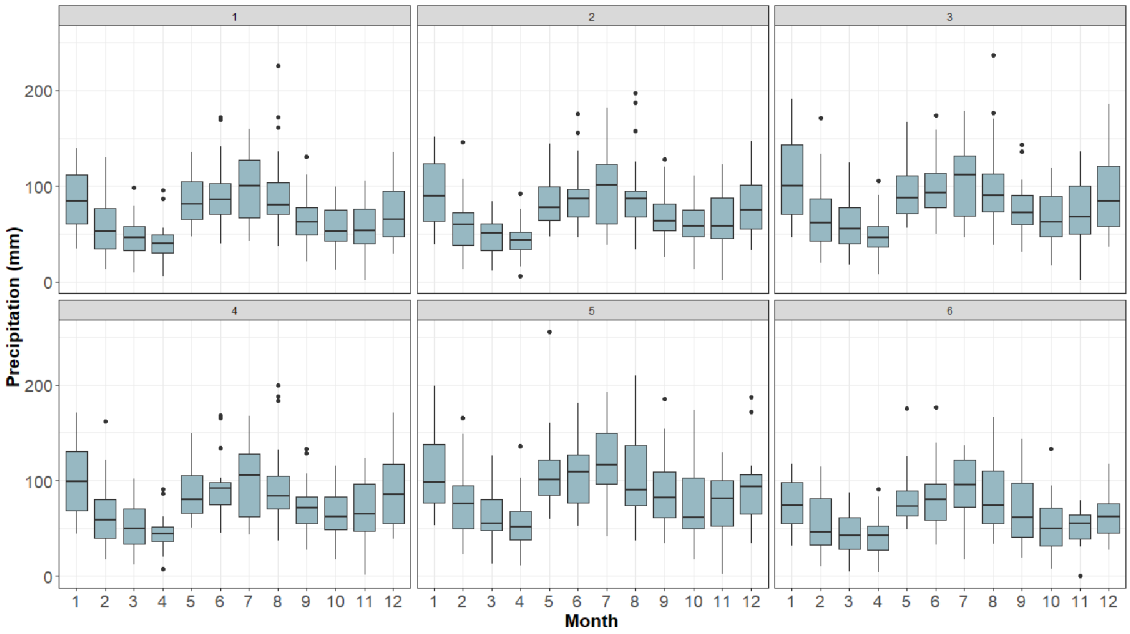

Figure 5.6: Annual variation in monthly precipitation totals (seasonality)

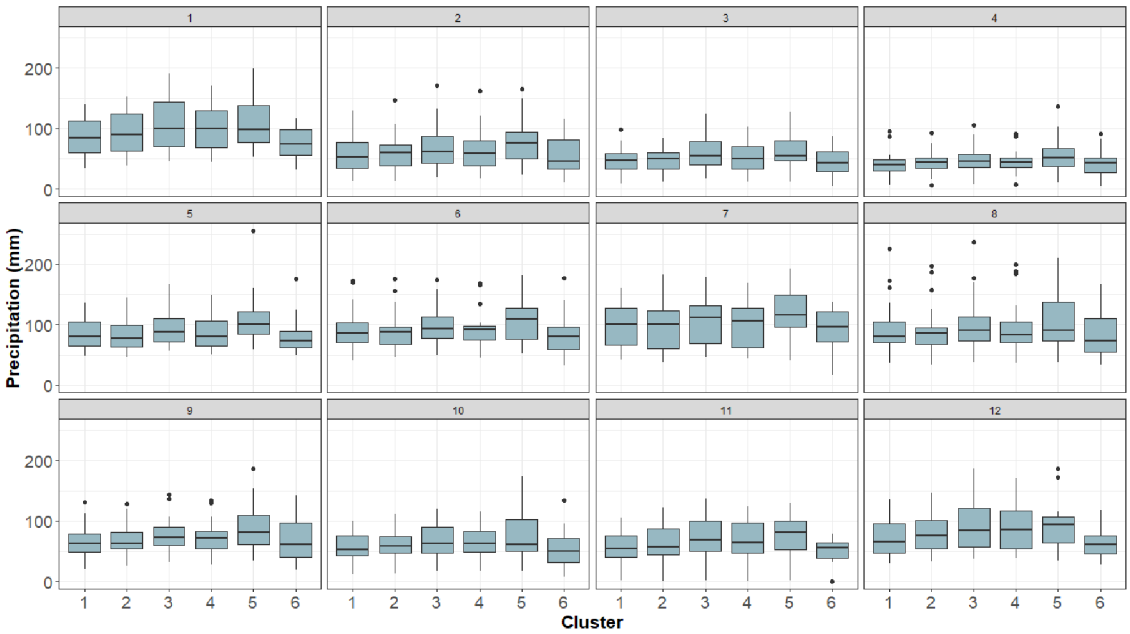

Figure 5.7: Annual variation in monthly precipitation totals (monthly comparison)

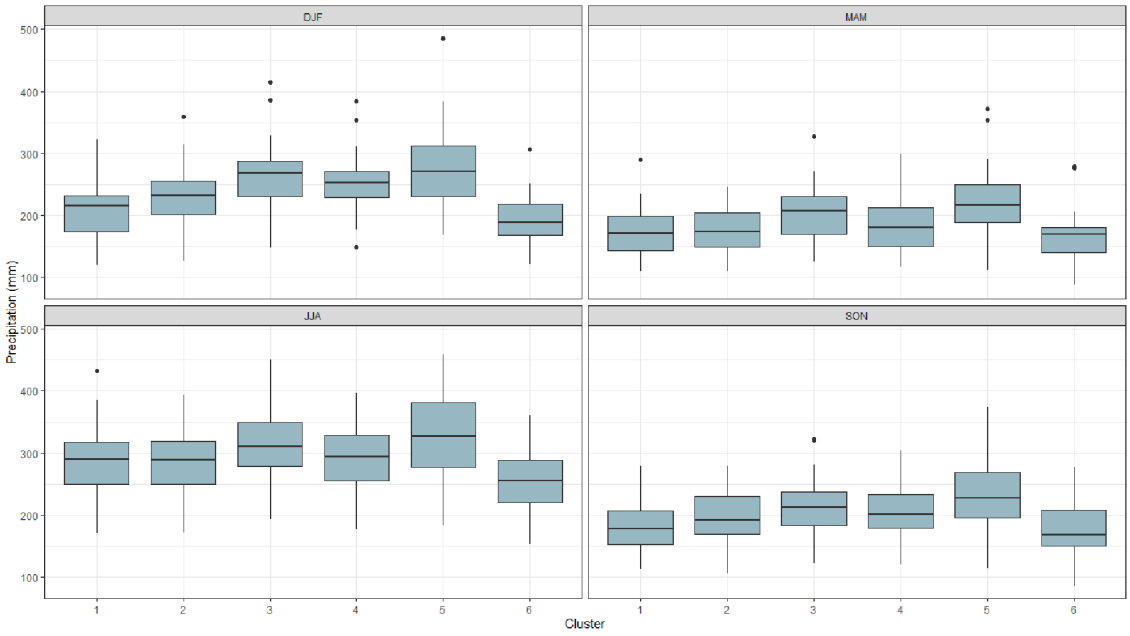

Figure 5.8: Annual precipitation variation in seasonal precipitation

2020. There are similar trends in annual precipitation behaviour for all clusters. Although, 2006 does not have a significant change in precipitations compared to 2005 and 2007, only for clusters 1 and 6. In addition, clusters have different behaviour through the years 2012-2014 (especially cluster 6), and cluster 6 had different trends through the years 2016-2018. Moreover, cluster 6 has lower annual precipitation totals through all years, whereas clusters 3 and 5 are the highest. In addition, two regression models were applied to discover the trend: ”lm” on the graph stands for linear regression model and ”loess” stands for locally estimated scatterplot smoothing, which combines polynomial regression and slopes between observations. Based on loess method behaviour 2001-2010 years trend seems similar to the 2011-2020 trend. Figure 5.10 shows the box plot comparison between these two periods for each cluster. Period 2011-2020 shifted to smaller annual precipitation across all clusters.

The figure 5.11 shows the seasonal variation in precipitation comparison between periods. Winter (DJF) remains the same for all clusters; the most considerable increase in variation of seasonal precipitation observes in cluster 5 followed by 6. Spring season (MAM) remains similar for clusters 1 and 4, followed by an increase for clusters 2 and 3, for cluster 5 variation increase, whereas for cluster 6 decreased. Among clusters 1,2,3 the variation in seasonal precipitation amount in summer (JJA) increased; however, the precipitation became less abundant. Even higher decrease towards lighter precipitation was observed among clusters 5 and 6. Autumn (SON) precipitation variation for clusters 5 and 6 becomes more skewed, increasing frequency among lighter precipitation days. Autumn precipitation variation is much lower for the period 2011-2020 among all clusters except 6. The figure 5.12 shows the comparison between monthly variability for each

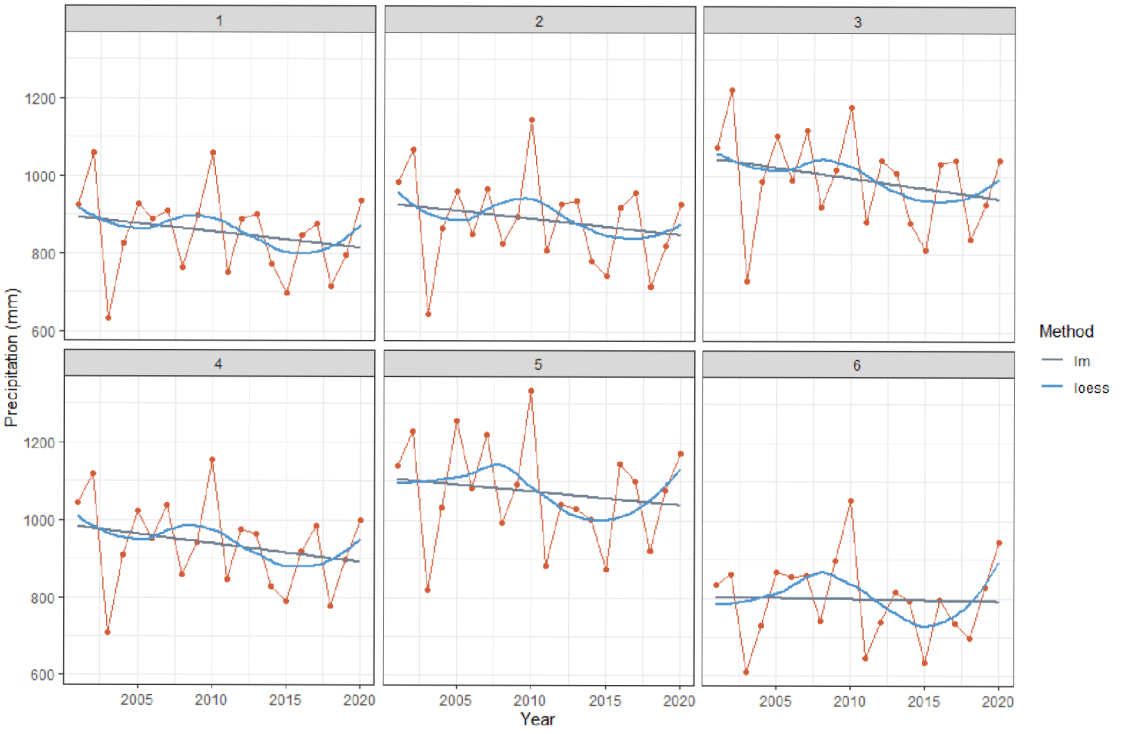

Figure 5.9: Annual precipitation totals

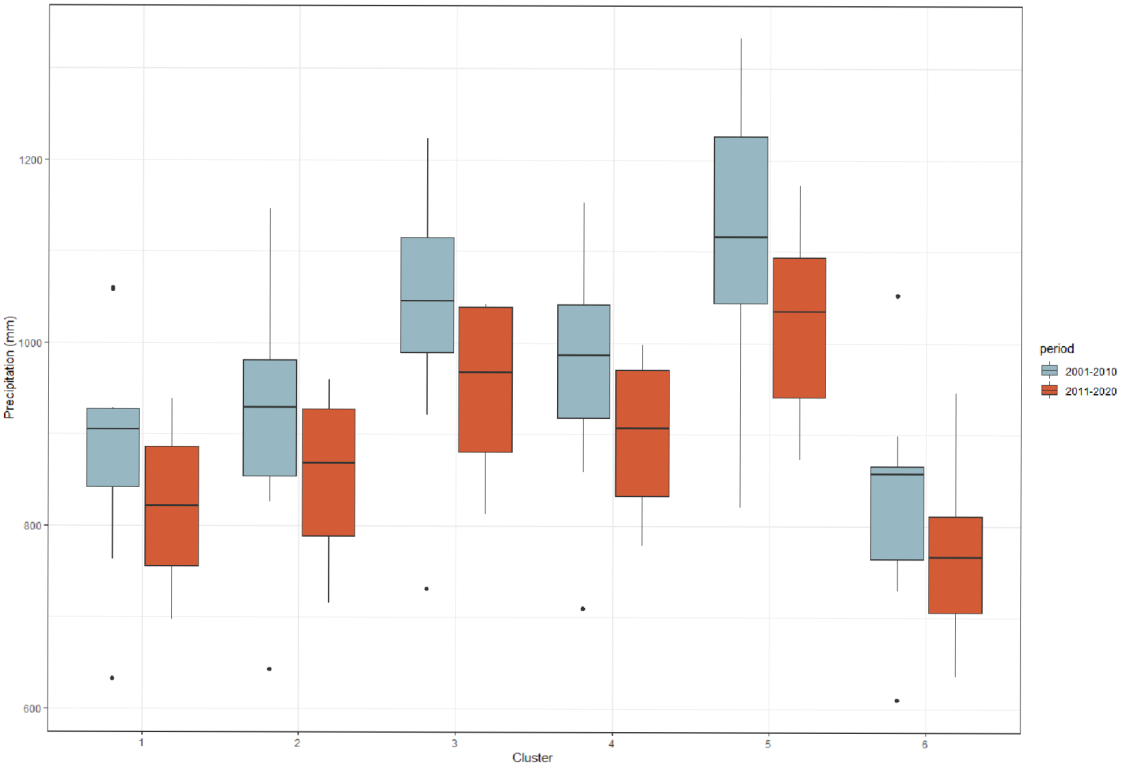

Figure 5.10: Annual precipitation variation comparison for the periods 2001-2010 and 2011-2020

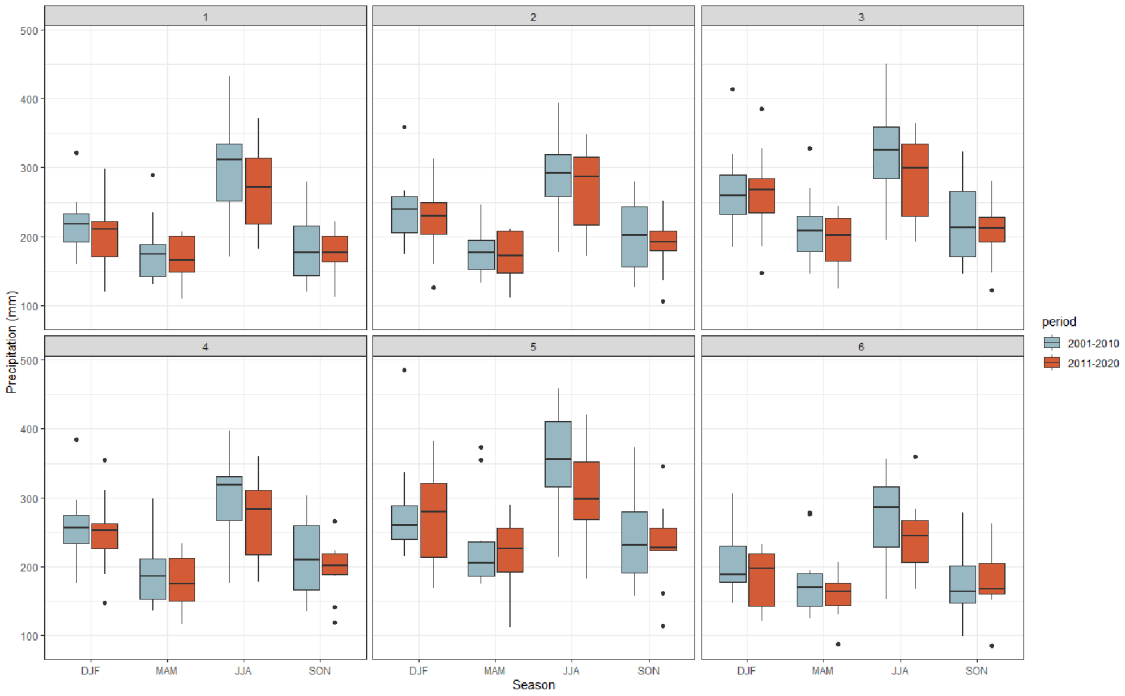

Figure 5.11: Seasonal precipitation variation comparison for the periods 2001-2010 and 2011-2020

cluster between two periods. There is a shift in precipitation variation for all clusters from August to July.

Moreover, the Figure 5.9 slope for 2001-2005 and 2011-2015 is negative, whereas slopes for the periods 2006-2010 and 2016-2020 are positive. In the following figures, these four periods are analysed. The figure 5.13 shows that 2011-2015 is the driest period in annual precipitation across all clusters, especially dramatically different for clusters 5 and 6. On the other hand, the period 2006-2010 has the lowest annual precipitation variation for clusters 1 and 6. The figure 5.14 shows the decrease in mean summer precipitation for the period 2016-2020 for all clusters and increase in mean winter precipitation totals. The figure 5.15 shows the shift from the amount in precipitation in January for the period 2011-2015 to February for the period 2016-2020. The annual variability in monthly totals is high in August for the period 2006-2010 and in July for the period 2011-2020.

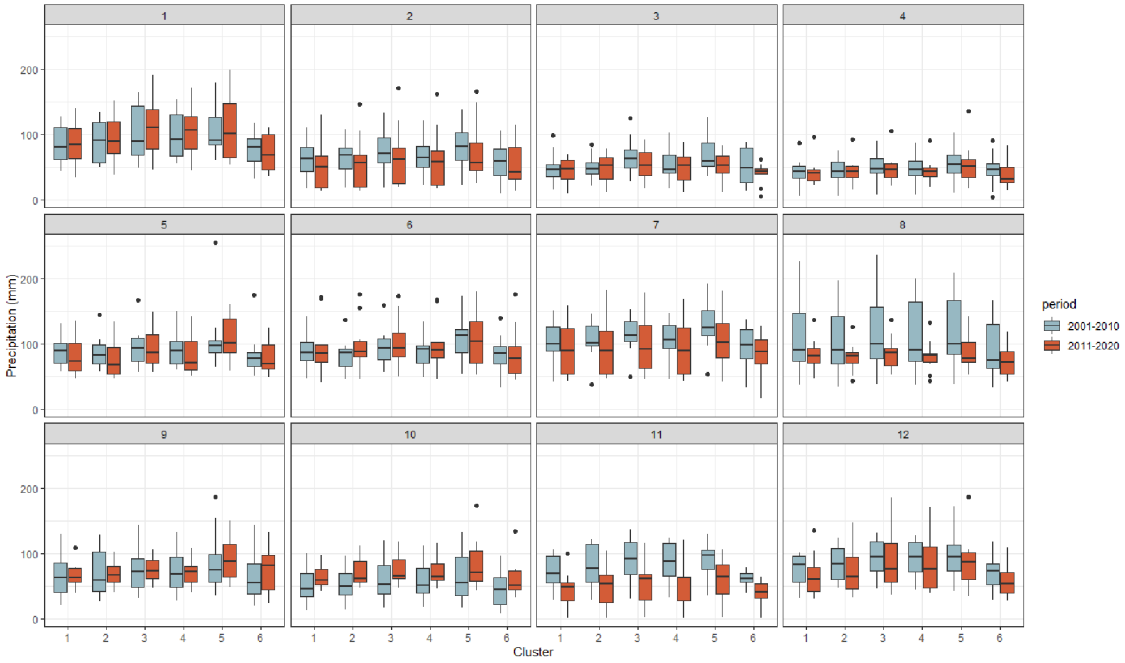

Figure 5.12: Monthly precipitation variation comparison for the periods 2001-2010 and 2011-2020

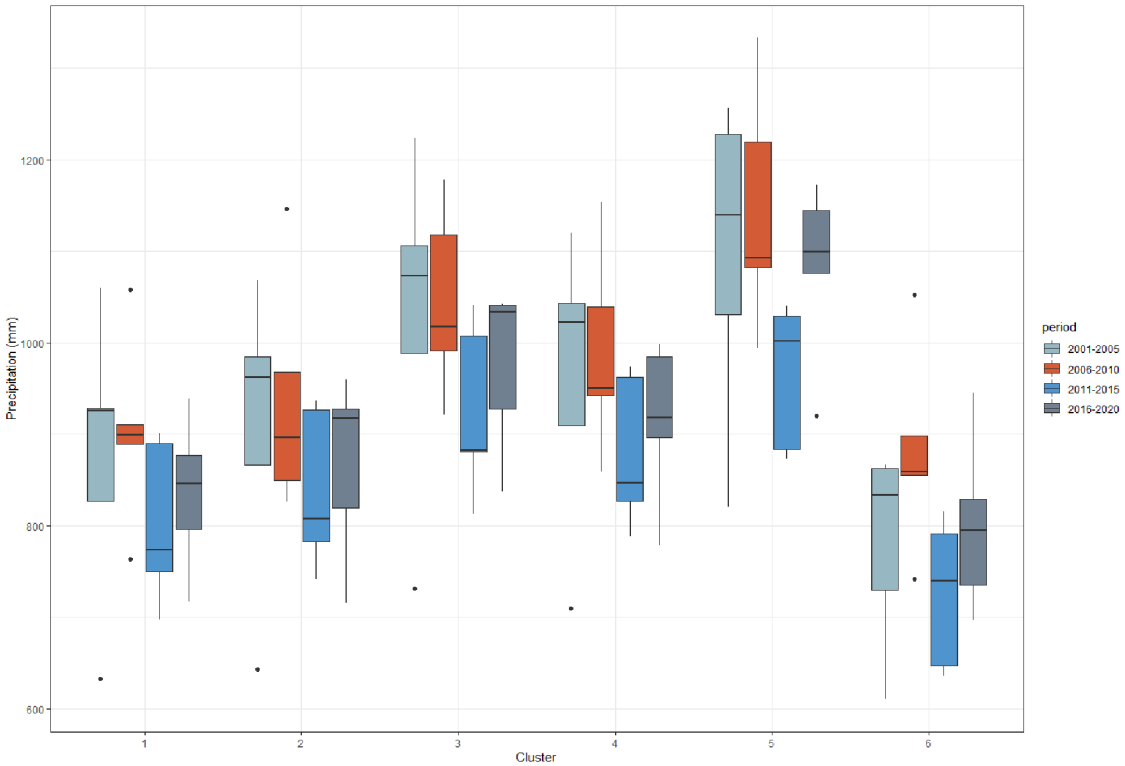

Figure 5.13: Annual precipitation variations for periods 2001-2005, 2006-2010, 2011-2015 and 2016-2020

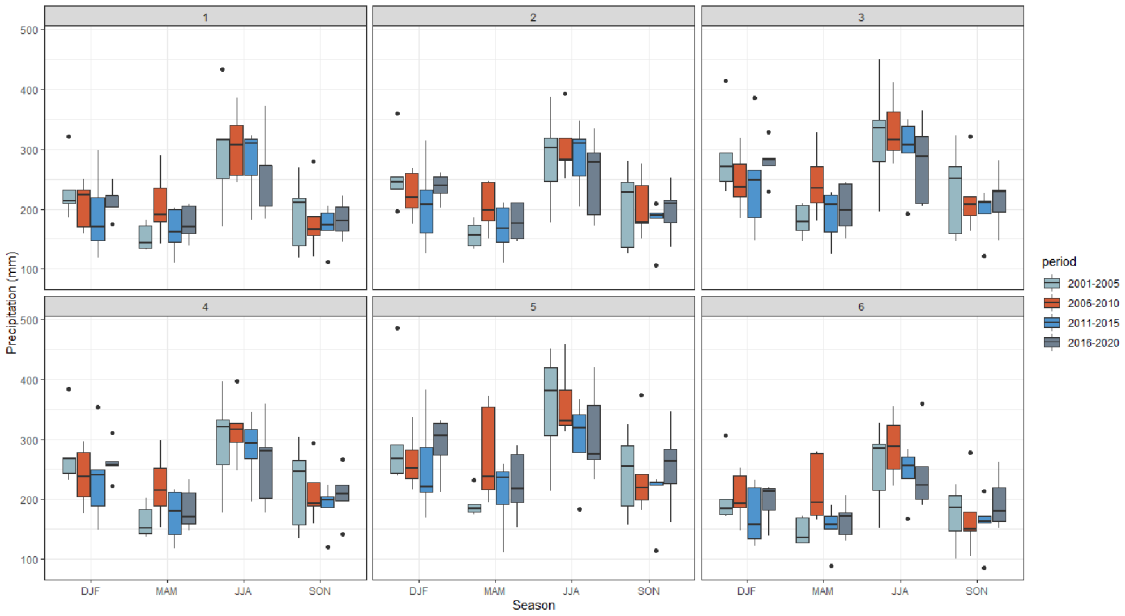

Figure 5.14: Seasonal precipitation variation comparison for the periods 2001-2005, 2006-2010, 2011-2015 and 2016-2020

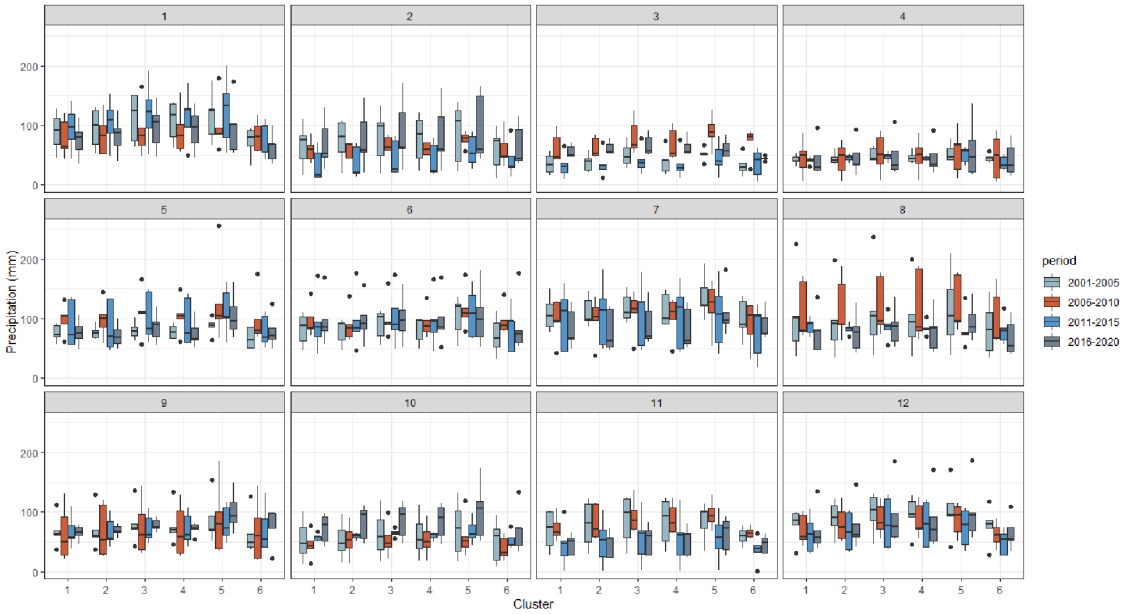

Figure 5.15: Monthly precipitation variation comparison for the periods 2001-2005, 2006-2010, 2011-2015 and 2016-2020

6 Discussion

This research performed a different number of iterations, sizes of grid and number of clusters with SOM application. The number of iterations did not influence significantly on cluster formation; however, the homogeneity of clusters was slightly improved with the increasing number of iterations (Figure 5.1). Different sizes of grids and the number of clusters also did not strongly influence on the formation of the clusters. Markonis Y. and Strnad F. also observed that there is no strong dependency between the grid size and clustering scheme [37]. In different variations of grid sizes and the number of clusters, similar patterns were observed: mountains regions Ore mountains, Sudetes mountains, Orlice mountains and Sumava mountains belonged to one cluster, South Moravia region formed a separate cluster, Prague region differs from the surrounding area. to one cluster

6.1 Clusters description

Cluster 1 is mainly formed through the middle elevation grid points, including the north part of the South Bohemia region, the Central Bohemia region except the north-eastern part and parts of the Vysocina, Pardubice and Olomouc regions closer to the South Moravia region border. The total amount of the annual precipitation is lower than the average across the Czech Republic, a little higher than for cluster 6. Mild temporal changes occurred in the precipitation of this cluster; however, they lie with changes in other clusters.

Prague and regions close to mountains belong to cluster 2. It has a slightly higher annual mean precipitation than cluster 1. DJF precipitation is slightly higher for this cluster than cluster 1; however, summer precipitation totals are almost the same (Figure 5.8). A comparison of the period 2001-2010 and 2011-2020 for seasonal precipitation showed no significant change in the mean precipitation totals; however, the variance increased for the seasons - MAM and JJA and decreased for SON. The frequency of the light events increased for summer precipitations. The Prague region belongs to a different cluster from the surrounding areas. Urban heat island effects are increasing in Prague [48], which might lead to the different precipitation patterns compared to the surrounding areas and thus the

different clusters.

Mountains regions formed homogeneous clusters 5 and 3 (Figure 5.5). Trnka M. et al. research based on Ward's regionalization of drought characteristics also defined mountain regions in one homogeneous cluster with low drought probability [3]. These clusters shows the highest mean in annual and seasonal precipitation totals (Figure 5.8, Table 5.1). The same trend was described by the Brazdil et al. research [49]. It was expected that regions with higher altitudes might belong to the same cluster. Interestingly, Trnka M et al. [3] results showed that the mountain regions belonging to clusters 3 and 5 in this research were grouped into homogeneous clusters characterized by rarely occurring dry events.

Cluster 4 mainly includes the Karlovy Vary region, south and west part of the Plzen, Hradec Kralove region, north part of the Pardubice region, north part of the Central Bohemia region, south part of the Oloumouc and north part of the Zlin region. The altitude varies from the lower to the middle range (Figure 5.5). The annual mean precipitation is closer to the mean annual precipitation of the Czech Republic.

Cluster 6 includes most of the area of the South Moravia region and partially the bordered regions. This cluster is characterized by the lowest amount of mean annual precipitations. Therefore, this cluster might represent the area in the Czech Republic that is more vulnerable to droughts and characterized by a lack of precipitation. Trnka M et al. [3] obtained that the highest number of dry events occurs in the north, central and southeastern regions of the country.

6.2 Temporal variability and changes

It is common for the Czech Republic climate that the maximum amount of precipitation accounts for July, whereas the minimum for January. This research result accounts for the minimum precipitation in April (except for cluster 6 - March) and the maximum in July. Brazdil et al. [49] showed that for the period 2001-2010, the minimum precipitation amount in mean annual variation occurs in April (or February for one altitude group) and maximum in August (or February for one altitude group), for the period 2011-2019 the minimum in February and the maximum in July. The minimum amount of precipitation in mean annual variability found in this research complies with Brazdil et al. results for the period 2001-2010, and an exception is cluster 4, for which the minimum occurs in March (Figure 5.12). Maximum precipitation occurs in July except for 2011-2020 across clusters 2, 3, 4, with maximum precipitation in January. This might be related to the tendency of the GPM data to overestimate precipitation for some regions and across some climate types [44].

In addition, it is expected that in the Prague region, winter precipitation totals will

slightly increase, whereas summer will decrease [50]. This study's results did not show the same pattern for the ten years periods in cluster 2 (Figure 5.11). However, a five-year analysis showed a decrease in the period 2001-2015 and an increase for the period 2016-2020. Winter precipitation totals for the period 2016-2020 are lower than 2001-2005. Summer precipitation totals decreased for the past years compared to 2001-2005.

The increase in winter precipitation and decline in summer precipitation is more significant in cluster 6 (mainly the South Moravia region) than in cluster 2 for ten years and five-year analysis (Figure 5.11 and Figure 5.14). It is expected that this effect is more significant for the South Moravia region [50].

Annual precipitation totals comparison between periods 2001-2010 and 2011-2020 showed a decrease in the mean value for all clusters. Seasonality for all clusters is similar. Comparison between the periods 2001-2010 and 2011-2020 showed an increase in seasonal mean and variability for the winter in the mountains (cluster 5) and a decrease in summer. The decrease in summer precipitation totals was observed through all clusters except cluster 2.

6.3 Limitations and Future Research

One of the limitations of this study is the small number of previous studies, especially for the Czech Republic. According to the articles mining performed by the Bochenek B. and Ustrnul Z. [27] only 150 articles were found with the application of the machine learning approaches in precipitation analysis. China, Australia, India and Germany are the top countries examining the application of the machine learning approaches in climate-related articles. Thus, it is not easy to compare this study's results with previous findings. In addition, it would be interesting to evaluate the performance of the used methodology with other similar studies.

This research was more focused on the possibility to apply an unsupervised machine learning approach to classify precipitation data in a line usage of the satellite data. Precipitation analysis is an extensive area involving different interdisciplinary fields. For example, precipitation intensity patterns might be investigated more deeply. In precipitation changes, the changes in intensity play a vital role. It is essential in drought analysis because, for example, the same amount of monthly precipitation can result in different soil moisture; less frequent showers will result in worse soil moisture than more frequent light rainfalls. Further studies using precipitation scaling methods merged with clustering can be conducted to investigate deeply the relationship between clusters and their significance.

It is worth mentioning that the SOM algorithm from the R package "kohonen" performed very well on the computer with RAM 8 Gb and 8 cores. The parallel method

”pbatch” was used on 6 out of 8 cores. The computing time of the biggest sized model took approximately 1.5 minutes. Parallel execution (”foreach” method) was also performed for the data pre-processing and merging files in one tidy table format. However, the amount of RAM was not enough to allocate all merged data. Switching to the computer with 4 cores but the increased amount of RAM - 28 GB, solved the issue.

7 Conclusion

One of the objectives of this study was to receive meaningful clusters to classify the spatial patterns of precipitation. Different setups to classification schemes were conducted, and clusters formation showed similar patterns. Annual, seasonal and monthly totals were compared and showed different values and trends among clusters. In addition, clustering revealed the Prague region's different patterns from the surrounding regions, which might direct to the change in the precipitation pattern caused by the urban heat island effect. Although, it is not clear why the same cluster includes the area of the lee side of the Ore Mountains and north parts of the Olomouc and Moravia-Silesia regions. More broadly, research is also needed to determine the similarity of this regions' precipitation pattern with the Prague region.

The dependency of precipitation behaviour on altitude is broadly investigated in many studies. Thus, it was suggested that the mountain regions would form one separate homogeneous cluster. Indeed, Ore mountains, Sudetes part, where pick Snezka is located, and the Bohemian Forest formed one homogeneous cluster. However, Beskids mountains formed different clusters.

The second objective was to analyse the change in precipitation during the last five years. Compared to the previous (2011-2015) five years, annual precipitation totals increased for all clusters; however, it is still lower than the period 2001-2005. For the last five years, all clusters tended to increase winter precipitation totals and decrease in summer precipitation totals. The most significant changes are observed in the mountains and South Moravia regions.

This research implemented an unsupervised machine learning method - SOM merged with hierarchical clustering on satellite GPM product data to classify precipitation patterns. The "kohonen" R package showed an outstanding performance and required low computational power. Opposite, the pre-processing of the GPM data required a larger amount of computational memory; this should be considered while using this dataset in the research. There is a growing demand for the usage of machine learning approaches in climate-related studies. Especially it might significantly benefit meteorological studies because this field involves comprehensive non-linear relationships between variables.

However, the debates about the future, profitability and contribution of machine learning in climate-related studies are still ongoing. This study showed that unsupervised machine learning might reveal an interesting pattern in spatial precipitation classification that future studies can deeply investigate.

Bibliography

1. PRETTEL, Leanne E. *Impact of Weather and Climate Extremes*. Hauppauge, UNITED STATES: Nova Science Publishers, Incorporated, 2011. ISBN 978-1-61122-374-3. Available also from: http://ebookcentral.proquest.com/lib/czup/detail.action?docID=3018775.
2. ZVERYAEV, Igor I. Seasonality in precipitation variability over Europe. *Journal of Geophysical Research: Atmospheres*. 2004, roč. 109, č. D5. Available from DOI: https://doi.org/10.1029/2003JD003668.
3. TRNKA, M.; DUBROVSKÝ, M.; SVOBODA, M.; SEMERÁDOVÁ, D.; HAYES, M.; ŽALUD, Z.; WILHITE, D. Developing a regional drought climatology for the Czech Republic. *International Journal of Climatology*. 2009, roč. 29, č. 6, pp. 863–883. Available from DOI: https://doi.org/10.1002/joc.1745.
4. BERANOVÁ, Romana; KYSELÝ, Jan. Trends of precipitation characteristics in the Czech Republic over 1961–2012, their spatial patterns and links to temperature and the North Atlantic Oscillation. *International Journal of Climatology*. 2018, roč. 38, č. S1, e596–e606. Available from DOI: https://doi.org/10.1002/joc.5392.
5. ZAHRADNÍČEK, Pavel; TRNKA, Miroslav; BRÁZDIL, Rudolf; MOŽNÝ, Martin; ŠTĚPÁNEK, Petr; HLAVINKA, Petr; ŽALUD, Zdeněk; MALÝ, Antonín; SEMERÁDOVÁ, Daniela; DOBROVOLNÝ, Petr; DUBROVSKÝ, Martin; ŘEZNÍČKOVÁ, Ladislava. The extreme drought episode of August 2011–May 2012 in the Czech Republic. *International Journal of Climatology*. 2015, roč. 35, č. 11, pp. 3335–3352. Available from DOI: https://doi.org/10.1002/joc.4211.
6. KYSELÝ, Jan. Trends in heavy precipitation in the Czech Republic over 1961–2005. *International Journal of Climatology*. 2009, roč. 29, č. 12, pp. 1745–1758. Available from DOI: https://doi.org/10.1002/joc.1784.
7. BLIŽŇÁK, Vojtěch; KAŠPAR, Marek; MÜLLER, Miloslav. Radar-based summer precipitation climatology of the Czech Republic. *International Journal of Climatology*. 2018, roč. 38, č. 2, pp. 677–691. Available from DOI: https://doi.org/10.1002/joc.5202.

8. TRNKA, Miroslav; HLAVINKA, Petr; MOŽNÝ, Martin; SEMERÁDOVÁ, Daniela; ŠTĚPÁNEK, Petr; BALEK, Jan; BARTOŠOVÁ, Lenka; ZAHRADNÍČEK, Pavel; BLÁHOVÁ, Monika; SKALÁK, Petr; FARDA, Aleš; HAYES, Michael; SVOBODA, Mark; WAGNER, Wolfgang; EITZINGER, Josef; FISCHER, Milan; ŽALUD, Zdeněk. Czech Drought Monitor System for monitoring and forecasting agricultural drought and drought impacts. *International Journal of Climatology*. 2020, roč. 40, č. 14, pp. 5941–5958. Available from DOI: https://doi.org/10.1002/joc.6557.

9. BRÁZDIL, R.; RAŠKA, P.; TRNKA, M.; ZAHRADNÍČEK, P. et al. The Central European drought of 1947: causes and consequences, with particular reference to the Czech Lands. *Clim Res*. 2016, roč. 70, pp. 161–178. Available from DOI: https://doi.org/10.3354/cr01387.

10. BRÁZDIL, R.; TRNKA, M.; DOBROVOLNÝ, P.; CHROMÁ, K.; HLAVINKA, P.; ŽALUD, Z. Variability of droughts in the Czech Republic, 1881–2006. *Theoretical and Applied Climatology*. 2009, roč. 97, č. 3, pp. 297–315. ISSN 1434-4483. Available from DOI: 10.1007/s00704-008-0065-x.

11. Fifth National Communication of the Czech Republic on the UN Framework on Climate change including Supplementary Information Pursuant to Article 7.2 of the Kyoto Protocol. *Ministry of the Environment of the Czech Republic*. 2009.

12. HANEL, Martin; PAVLÁSKOVÁ, Alena; KYSELÝ, Jan. Trends in characteristics of sub-daily heavy precipitation and rainfall erosivity in the Czech Republic. *International Journal of Climatology*. 2016, roč. 36, č. 4, pp. 1833–1845. Available from DOI: https://doi.org/10.1002/joc.4463.

13. JAN, Daňhelka; JAN, Kubát; PETR, Šercl; RADEK, Čekal. Floods in the Czech Republic in June 2013. *Ministry of the Environment of the Czech Republic, Czech Hydrometerological Institute*. 2014.

14. ZHANG, Anzhi; JIA, Gensuo. Monitoring meteorological drought in semiarid regions using multi-sensor microwave remote sensing data. *Remote sensing of Environment*. 2013, roč. 134, pp. 12–23.

15. ALIZADEH, Mohammad Reza; NIKOO, Mohammad Reza. A fusion-based methodology for meteorological drought estimation using remote sensing data. *Remote sensing of environment*. 2018, roč. 211, pp. 229–247.

16. FENG, Puyu; WANG, Bin; LIU, De Li; YU, Qiang. Machine learning-based integration of remotely-sensed drought factors can improve the estimation of agricultural drought in South-Eastern Australia. *Agricultural Systems*. 2019, roč. 173, pp. 303–316. ISSN 0308-521X. Available from DOI: https://doi.org/10.1016/j.agsy.2019.03.015.

17. SWAIN, Sharmistha; WARDLOW, Brian D; NARUMALANI, Sunil; TADESSE, Tsegaye; CALLAHAN, Karin. Assessment of vegetation response to drought in Nebraska using Terra-MODIS land surface temperature and normalized difference vegetation index. *GIScience & Remote Sensing*. 2011, roč. 48, č. 3, pp. 432–455.

18. LAI, Chengguang; ZHONG, Ruida; WANG, Zhaoli; WU, Xiaoqing; CHEN, Xiaohong; WANG, Peng; LIAN, Yanqing. Monitoring hydrological drought using long-term satellite-based precipitation data. *Science of the total environment*. 2019, roč. 649, pp. 1198–1208.

19. MORID, Saeid; SMAKHTIN, Vladimir; BAGHERZADEH, K. Drought forecasting using artificial neural networks and time series of drought indices. *International Journal of Climatology: A Journal of the Royal Meteorological Society*. 2007, roč. 27, č. 15, pp. 2103–2111.

20. BELAYNEH, A; ADAMOWSKI, J; KHALIL, B; OZGA-ZIELINSKI, B. Long-term SPI drought forecasting in the Awash River Basin in Ethiopia using wavelet neural network and wavelet support vector regression models. *Journal of Hydrology*. 2014, roč. 508, pp. 418–429.

21. *11 Dimensionality reduction techniques you should know in 2021*. Rukshan Pramoditha, 2021. Available also from: `https://towardsdatascience.com/11-dimensionality-reduction-techniques-you-should-know-in-2021-dcb9500d388b`. Last accessed 13 March 2022.

22. GARETH, James; DANIELA, Witten; TREVOR, Hastie; ROBERT, Tibshirani. *An introduction to statistical learning: with applications in R*. Spinger, 2013.

23. VESANTO, Juha; ALHONIEMI, Esa. Clustering of the self-organizing map. *IEEE transactions on neural networks*. 2000, roč. 11 3, pp. 586–600.

24. *k-Means Advantages and Disadvantages*. Google, 2021. Available also from: `https://developers.google.com/machine-learning/clustering/algorithm/advantages-disadvantages#advantages-of-k-means`. Last accessed 13 March 2022.

25. WEHRENS, Ron; BUYDENS, Lutgarde M. C. Self- and Super-organizing Maps in R: The kohonen Package. *Journal of Statistical Software*. 2007, roč. 21, č. 5, pp. 1–19. Available from DOI: `10.18637/jss.v021.i05`.

26. KOHONEN, Teuvo. *Self-Organizing Maps*. Berlin; Heidelberg; New York; Barcelona; Hong Kong; London; Milan; Paris; Singapore; Tokyo: Springer-Verlag Berlin Heidelberg, 2001.

27. BOCHENEK, Bogdan; USTRNUL, Zbigniew. Machine Learning in Weather Prediction and Climate Analyses - Applications and Perspectives. *Atmosphere*. 2022, roč. 13, č. 2. ISSN 2073-4433. Available from DOI: `10.3390/atmos13020180`.

28. KNIGHTON, James; PLEISS, Geoff; CARTER, Elizabeth; LYON, Steven; WALTER, M. Todd; STEINSCHNEIDER, Scott. Potential Predictability of Regional Precipitation and Discharge Extremes Using Synoptic-Scale Climate Information via Machine Learning: An Evaluation for the Eastern Continental United States. *Journal of Hydrometeorology*. 2019, roč. 20, č. 5, pp. 883–900. Available from DOI: 10.1175/JHM-D-18-0196.1.

29. LEY, R.; CASPER, M. C.; HELLEBRAND, H.; MERZ, R. Catchment classification by runoff behaviour with self-organizing maps (SOM). *Hydrology and Earth System Sciences*. 2011, roč. 15, č. 9, pp. 2947–2962. Available from DOI: 10.5194/hess-15-2947-2011.

30. DI PRINZIO, M.; CASTELLARIN, A.; TOTH, E. Data-driven catchment classification: application to the pub problem. *Hydrology and Earth System Sciences*. 2011, roč. 15, č. 6, pp. 1921–1935. Available from DOI: 10.5194/hess-15-1921-2011.

31. HU, Huiling; AYYUB, Bilal M. Machine Learning for Projecting Extreme Precipitation Intensity for Short Durations in a Changing Climate. *Geosciences*. 2019, roč. 9, č. 5. ISSN 2076-3263. Available from DOI: 10.3390/geosciences9050209.

32. ELSANABARY, Mohamed Helmy; GAN, Thian Yew. Wavelet Analysis of Seasonal Rainfall Variability of the Upper Blue Nile Basin, Its Teleconnection to Global Sea Surface Temperature, and Its Forecasting by an Artificial Neural Network. *Monthly Weather Review*. 2014, roč. 142, č. 5, pp. 1771–1791. Available from DOI: 10.1175/MWR-D-13-00085.1.

33. GHADERPOUR, Ebrahim; VUJADINOVIC, Tijana; HASSAN, Quazi K. Application of the Least-Squares Wavelet software in hydrology: Athabasca River Basin. *Journal of Hydrology: Regional Studies*. 2021, roč. 36, p. 100847. ISSN 2214-5818. Available from DOI: https://doi.org/10.1016/j.ejrh.2021.100847.

34. COULIBALY, Paulin; BURN, Donald H. Wavelet analysis of variability in annual Canadian streamflows. *Water Resources Research*. 2004, roč. 40, č. 3. Available from DOI: https://doi.org/10.1029/2003WR002667.

35. GUNTU, Ravi Kumar; MAHESWARAN, Rathinasamy; AGARWAL, Ankit; SINGH, Vijay P. Accounting for temporal variability for improved precipitation regionalization based on self-organizing map coupled with information theory. *Journal of Hydrology*. 2020, roč. 590, p. 125236. ISSN 0022-1694. Available from DOI: https://doi.org/10.1016/j.jhydrol.2020.125236.

36. ZENG, Peng; SUN, Fengyun; LIU, Yaoyi; WANG, Yukun; LI, Gen; CHE, Yue. Mapping future droughts under global warming across China: A combined multi-timescale meteorological drought index and SOM-Kmeans approach. *Weather and Climate Extremes*. 2021, p. 100304. ISSN 2212-0947. Available from DOI: https://doi.org/10.1016/j.wace.2021.100304.

37. MARKONIS, Yannis; STRNAD, Filip. Representation of European hydroclimatic patterns with self-organizing maps. *The Holocene*. 2020, roč. 30, č. 8, pp. 1155–1162. Available from DOI: 10.1177/0959683620913924.

38. *TRMM instruments*. NASA, [n.d.]. Available also from: https://gpm.nasa.gov/missions/TRMM/satellite. Last accessed 21 February 2022.

39. *The global precipitation measurement mission (GPM)*. NASA, [n.d.]. Available also from: https://gpm.nasa.gov/missions/GPM. Last accessed 21 February 2022.

40. HUFFMAN, G.J.; STOCKER, E.F.; BOLVIN, D.T.; NELKIN, E.J.; JACKSON, Tan. GPM IMERG Final Precipitation L3 Half Hourly 0.1 degree x 0.1 degree V06, Greenbelt, MD. *Goddard Earth Sciences Data and Information Services Center (GES DISC)*. 2019. Last accessed 21 February 2022,10.5067/GPM/IMERG/3B-HH/06.

41. GPM IMERG Final Precipitation L3 Half Hourly 0.1 degree x 0.1 degree V06, Greenbelt, MD. *Precipitation processing system*. 2021. Last accessed 21 February 2022,10.5067/GPM/IMERG/3B-HH/06.

42. HUFFMAN, George J; BOLVIN, David T.; BRAITHWAITE, Dan; HSU, Kuolin; JOYCE, Robert; KIDD, Christopher; NELKIN, Eric J.; SOROOSHIAN, Soroosh; TAN, Jackson; XIE, Pingping. Algorithm Theoretical Basis Document (ATBD), NASA Global Precipitation Measurement (GPM) Integrated Multi-satellitE Retrievals for GPM (IMERG). 2019. Last accessed 21 February 2022.

43. KIRSCHBAUM, Dalia B.; HUFFMAN, George J.; ADLER, Robert F.; BRAUN, Scott; GARRETT, Kevin; JONES, Erin; MCNALLY, Amy; SKOFRONICK-JACKSON, Gail; STOCKER, Erich; WU, Huan; ZAITCHIK, Benjamin F. NASA's Remotely Sensed Precipitation: A Reservoir for Applications Users. *Bulletin of the American Meteorological Society*. 2017, roč. 98, č. 6, pp. 1169–1184. Available from DOI: 10.1175/BAMS-D-15-00296.1.

44. PRADHAN, Rajani K.; MARKONIS, Yannis; VARGAS GODOY, Mijael Rodrigo; VILLALBA-PRADAS, Anahí; ANDREADIS, Konstantinos M.; NIKOLOPOULOS, Efthymios I.; PAPALEXIOU, Simon Michael; RAHIM, Akif; TAPIADOR, Francisco J.; HANEL, Martin. Review of GPM IMERG performance: A global perspective. *Remote Sensing of Environment*. 2022, roč. 268, p. 112754. ISSN 0034-4257. Available from DOI: https://doi.org/10.1016/j.rse.2021.112754.

45. Topographic database of the Czech Republic (Data200) - layer Settlements. *Geoportal CUZK*. 2021. Available also from: https://geoportal.cuzk.cz/(S(cjnjk5c0eo5vxejp4kgwejg5))/Default.aspx?mode=TextMeta&side=mapy_data200&metadataID=CZ-CUZK-DATA200-SIDLA-V&head_tab=sekce-02-gp&menu=2295.

46. WITTEK, Peter; GAO, Shi Chao; LIM, Ik Soo; ZHAO, Li. somoclu: An Efficient Parallel Library for Self-Organizing Maps. *Journal of Statistical Software*. 2017, roč. 78, č. 9, pp. 1–21. Available from DOI: 10.18637/jss.v078.i09.

47. YUAN, Chunhui; YANG, Haitao. Research on K-Value Selection Method of K-Means Clustering Algorithm. *J.* 2019, roč. 2, č. 2, pp. 226–235. ISSN 2571-8800. Available from DOI: 10.3390/j2020016.

48. Capital city of Prague climate change adaptation strategy. *Prague City Hall*. 2020. Available also from: https://www.prague convention.cz/file/7601/prague_adaptation_strategy_eng_web_82020.pdf.

49. BRÁZDIL, Rudolf; ZAHRADNÍČEK, Pavel; DOBROVOLNÝ, Petr; ŠTĚPÁNEK, Petr; TRNKA, Miroslav. Observed changes in precipitation during recent warming: The Czech Republic, 1961–2019. *International Journal of Climatology*. 2021, roč. 41, č. 7, pp. 3881–3902. Available from DOI: https://doi.org/10.1002/joc.7048.

50. Strategy on Adaptation to Climate Change in the Czech Republic. *Ministries of Environment, Agriculture, Industry and Trade, Regional Development, Health and Interior*. 2015. Available also from: https://www.mzp.cz/C125750E003B698B/en/strategy_adaptation_climate_change/%5C$FILE/OEOK_Adaptation_strategy_20171003.pdf.

List of abbreviations used

ANN Artificial Neural Network

ML Machine Learning

TRMM Tropical Rainfall Measuring Mission

GPM Global Precipitation Measurement Mission

NASA National Aeronautics and Space Administration

JAXA Japanese space agency

DPR Dual-frequence precipitation radar

GMI GPM Microwave Imager

IMERG Integrated Multi-satellitE Retrievals

SOM Self-organizing map

PCA Principal component analysis

CHMI Czech Hydrometeorological Institute

CzechDM Czech Drought Monitoring

ANN artificial neural networks

MA moving average

List of Figures

List of Tables

Appendices

A Appendix. Figures

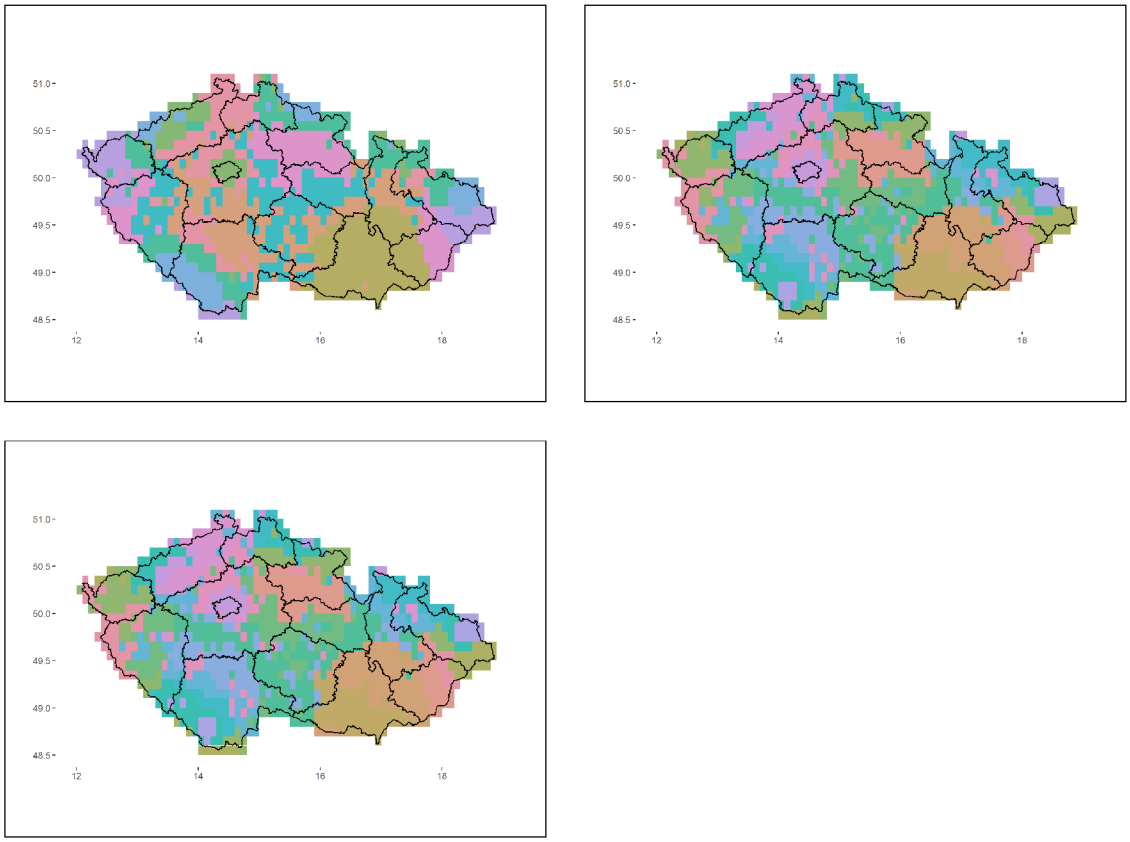

Figure 1: Seasonal variables: SOM's clusters for different grid sizes - 3*3, 4*4, 5*5 (left to right)

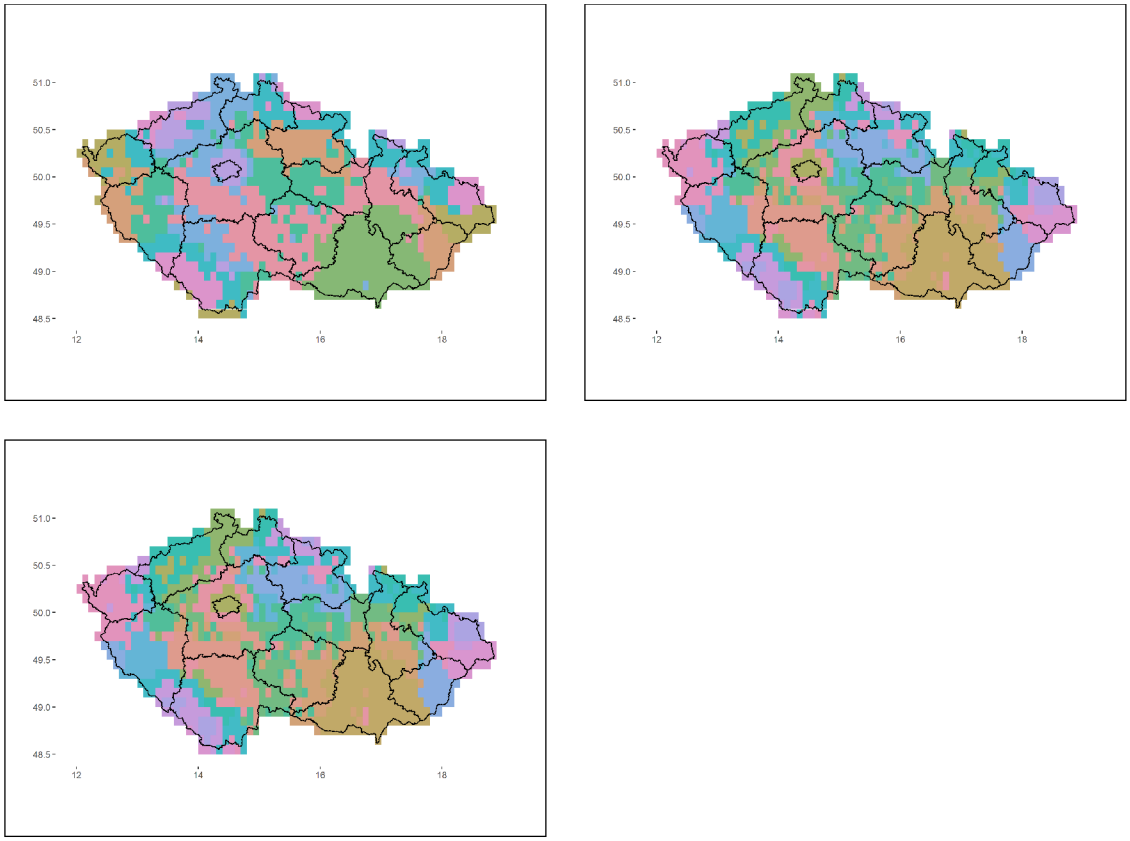

Figure 2: All variables: SOM's clusters for different grid sizes - 3*3, 4*4, 5*5 (left to right)

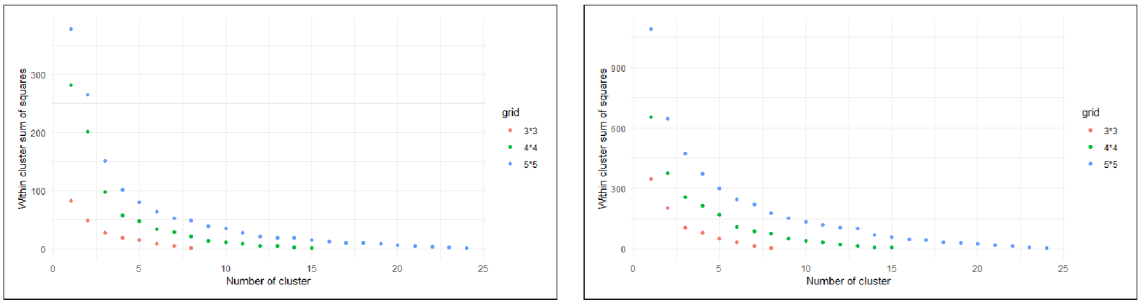

Figure 3: Annual and seasonal variables within cluster sum of squares (left to right)

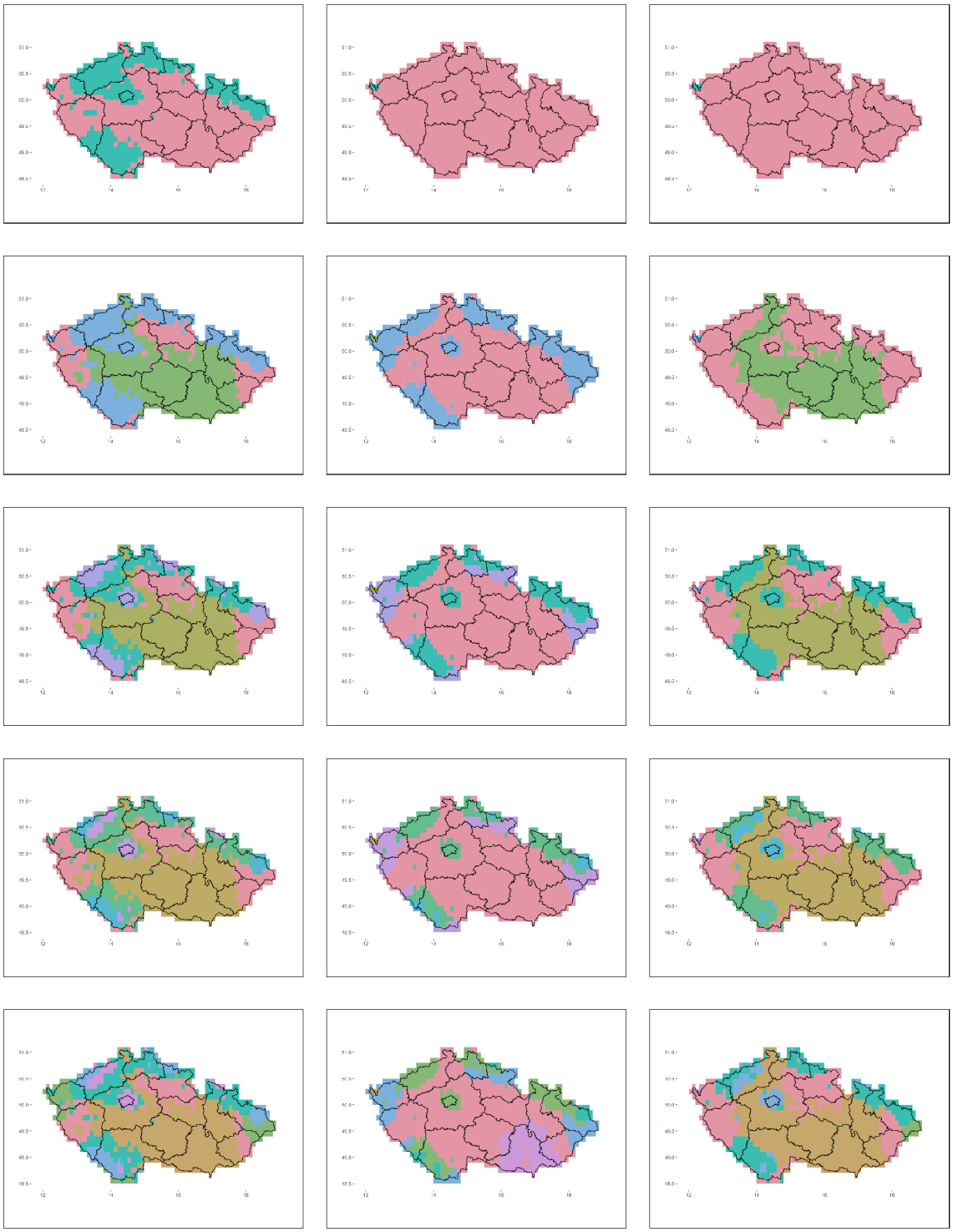

Figure 4: SOM's 2-6 clusters for grids 3*3, 4*4, 5*5 (annual variables)

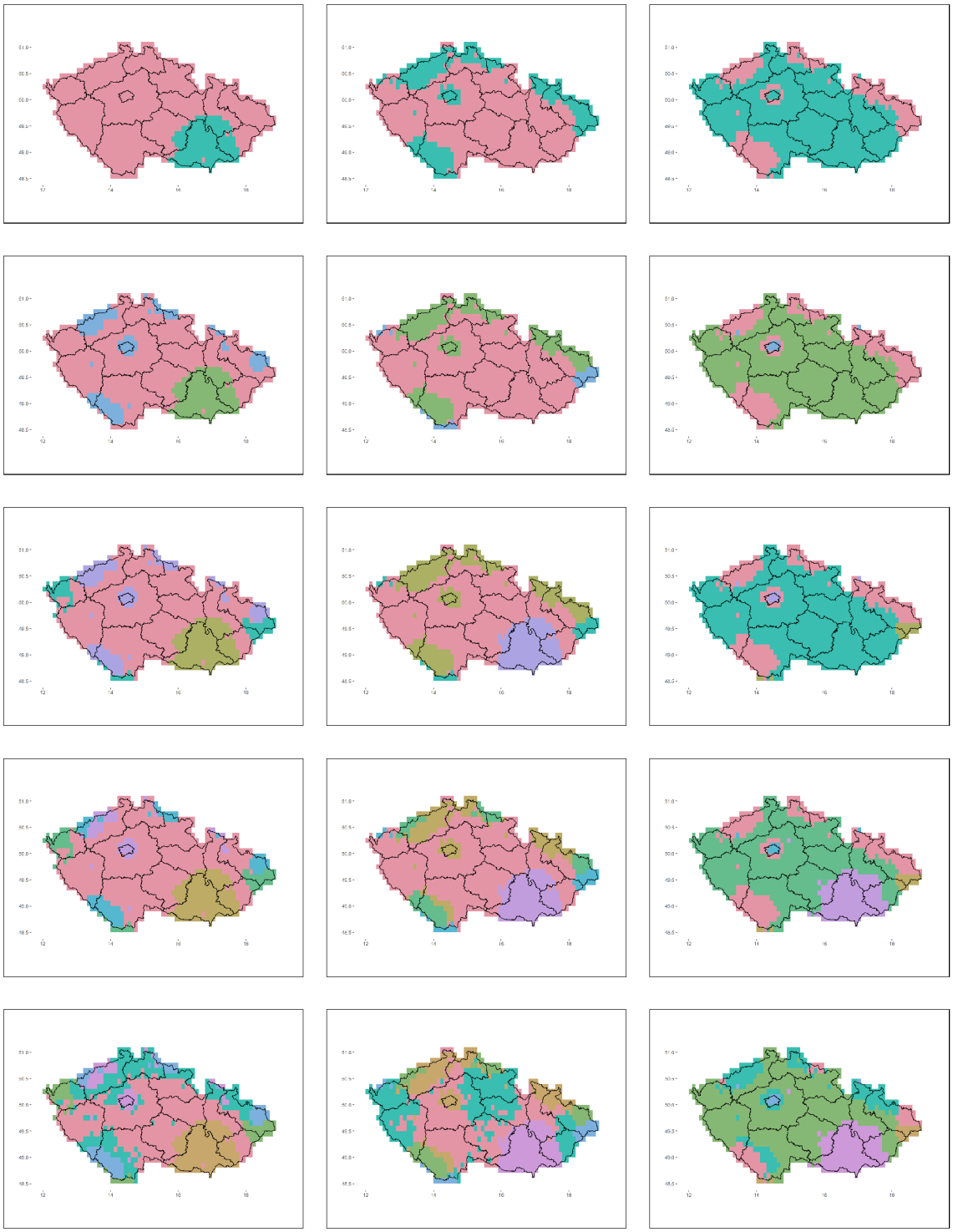

Figure 5: SOM's 2-6 clusters for grids 3*3, 4*4, 5*5 (seasonal variables)

B Appendix. Tables

	lat	lon	precipitation	date
1	50.25	12.15	0.01	2001-01-01 08:30:00
2	50.25	12.25	0.08	2001-01-01 08:30:00
3	50.25	12.35	0.27	2001-01-01 08:30:00
4	50.35	12.35	1.78	2001-01-01 08:30:00
5	50.25	12.45	0.31	2001-01-01 08:30:00
6	50.25	12.55	0.13	2001-01-01 08:30:00
7	50.45	12.55	1.53	2001-01-01 08:30:00
8	50.25	12.65	0.01	2001-01-01 08:30:00
9	50.35	12.65	0.18	2001-01-01 08:30:00
10	50.45	12.65	0.60	2001-01-01 08:30:00

Table 1: First 10 rows of precipitation data (mm/hr)

	lat	lon	year	month	season	date	precipitation
1	50.25	12.25	2017	4	MAM	2017-04-12	8.34
2	50.55	14.05	2002	9	SON	2002-09-21	2.53
3	50.25	16.85	2003	12	DJF	2003-12-11	4.67
4	50.45	15.35	2001	6	JJA	2001-06-11	5.90
5	50.05	12.45	2016	1	DJF	2016-01-06	2.55
6	49.15	14.55	2003	3	MAM	2003-03-18	0.01
7	49.75	13.85	2002	7	JJA	2002-07-07	0.98
8	50.15	15.65	2010	8	JJA	2010-08-25	0.08
9	50.25	17.35	2012	7	JJA	2012-07-21	2.04
10	49.55	16.35	2018	7	JJA	2018-07-27	11.47

Table 2: First 10 rows of daily precipitation amount

	lat	lon	ann_wetdays	wet_days_summer	wet_days_winter	wet_days_autumn	wet_days_spring
1	49.75	14.75	184.35	51.65	45.45	44.35	42.90
2	49.95	13.75	177.80	49.40	45.10	41.65	41.65
3	49.05	16.95	168.35	46.45	43.90	38.60	39.40
4	49.45	14.75	180.20	50.75	45.70	41.90	41.85
5	50.25	13.75	186.50	51.00	48.60	44.15	42.75
6	49.55	15.85	179.10	49.25	45.50	42.65	41.70
7	49.65	16.35	176.70	48.80	46.50	39.90	41.50
8	49.45	17.15	170.65	46.25	45.25	38.70	40.45
9	49.75	12.95	184.05	50.15	47.35	43.05	43.50
10	49.05	16.85	170.35	46.50	43.65	39.30	40.90

Table 3: First 10 rows of wet days count

	cluster	month	year	precipitation
1	1	8	2002	225.63
2	1	3	2006	98.61
3	1	4	2006	86.91
4	1	8	2006	161.72
5	1	9	2007	130.44
6	1	8	2010	172.44
7	1	6	2013	172.27
8	1	4	2017	95.70
9	1	6	2020	169.29
10	2	9	2001	128.48
11	2	8	2002	197.50
12	2	8	2006	157.51
13	2	4	2007	5.59
14	2	8	2010	187.54
15	2	6	2013	176.08
16	2	4	2017	92.29
17	2	2	2020	146.49
18	2	6	2020	155.57
19	3	9	2001	136.13
20	3	8	2002	236.64
21	3	9	2007	143.09
22	3	8	2010	176.66
23	3	6	2013	173.68
24	3	4	2017	105.58
25	3	2	2020	170.99
26	4	9	2001	133.12
27	4	8	2002	199.79
28	4	4	2006	86.55
29	4	8	2006	184.01
30	4	4	2007	7.78
31	4	9	2007	129.17
32	4	6	2009	134.49
33	4	8	2010	187.87
34	4	6	2013	165.88
35	4	4	2017	91.48
36	4	2	2020	161.76
37	4	6	2020	168.63
38	5	12	2005	171.54
39	5	9	2007	185.82
40	5	5	2010	255.06
41	5	4	2017	136.09
42	5	12	2018	186.95
43	5	2	2020	165.22
44	6	4	2006	90.74
45	6	5	2010	175.21
46	6	11	2011	0.59
47	6	6	2020	176.45
48	6	10	2020	133.31

Table 4: Monthly boxplot outliers