CHARLES UNIVERSITY

Faculty of science

Department of Applied Geoinformatics and Cartography

Bc. Daniel BÍCÁK

GEOGRAPHICAL RANDOM FOREST MODEL EVALUATION IN AGRICULTURAL DROUGHT ASSESSMENT

Evaluace geografického Random Forest algoritmu v posouzení sucha.

Master thesis

supervisor: Ing. Lukáš Brodský, Ph.D.

PRAGUE, 2021

UNIVERZITA KARLOVA PŘÍRODOVĚDECKÁ FAKULTA

Aplikovaná geoinformatika a kartografie 2020/2021

ZADÁNÍ DIPLOMOVÉ PRÁCE

pre Bc. Daniel Bicák

Kartografie a geoinformatika

Název tématu:: Evaluace geografického Random Forest algoritmu v posouzení sucha
Geographical Random Forest model evaluation in agricultural drought assessment

Zásady pro vypracování

Cílem práce je evaluace geografického Random Forest při hodnocení faktorů zemědělského sucha. Na základě rešerše literatury bude navržena metodika modelování sucha s využitím tohoto rozšířeného algoritmu strojového učení a při využití volně dostupných dat. Bude vytvořen model predikce umožňující výběr a hodnocení faktorů sucha. Proces strojového učení bude navržen s cílem kvantifikovat významnost faktorů zemědělského sucha v lokálních a globálních podmínkách.

Dílčími cíli práce jsou:

1. evaluace geografického rozšíření algoritmu Random Forest pro predikci zemědělského sucha, včetně ladění parametrů modelu,
2. ověření lokálních a globálních faktorů zemědělského sucha,
3. vytvoření mapy prostorové varibility zranitelnosti vůči suchu v Česku s využitím navrženého a parametrizovného modelu.

Rozsah grafických prací: cca 5 stran
Rozsah průvodní zprávy: cca 50 stran

Seznam odborné literatury:

Christopher M Bishop (2006), *Pattern recognition and machine learning*, springer

Leo Breiman (2001), "Random forests", *Machine learning*, 45, 1, pp. 5-32

Mehryar Mohri et al. (2018), *Foundations of machine learning*, MIT press

Omid Rahmati, Fatemeh Falah, et al. (2020), "Machine learning approaches for spatial modeling of agricultural droughts in the south-east region of Queensland Australia", *Science of The Total Environment*, 699, p. 134230

Vedoucí diplomové práce:: Ing. Lukáš Brodský, Ph.D.

Konzultant diplomové práce: -

Datum zadání diplomové práce: 16. ledna 2020

Termín odevzdání diplomové práce: červenec 2021

Platnost tohoto zadání je po dobu jednoho akademického roku.

Vedoucí diplomové práce **Garant studijního oboru**

V Praze, dne

DECLARATION

Hereby I declare that I have written this thesis myself and that I quoted all references accordingly. Neither this thesis nor its part was used to obtain other academic degrees of the same or other levels.

I hereby agree with using this thesis for academic purposes and I agree with its inclusion in the evidence of academic works.

Prague, August 2021 Bc. Daniel Bicák

ACKNOWLEDGEMENTS

I would like to thank my supervisor, Ing. Lukáš Brodský PhD. for his continuous guidance, shared knowledge in consultations and lessons and constructive feedback, without which the thesis would not have been written.

ABSTRACT

Drought is a natural disaster, which negatively affects millions of people and causes huge economic losses. This thesis investigates agricultural drought in Czechia using machine learning algorithms. The statistical models utilised were Random Forest (RF), Geographical Random Forest (GRF) and Locally Tuned Geographical Random Forest (LT GRF). GRF consists of several RF models trained on a subset of original data. The final prediction is a weighted sum of the prediction of a local and global model. The size of the subset is determined by the tunable parameter. LT GRF addresses spatial variability of subset size and local weight. During the tuning process, optimal parameters are found for every location and then interpolated for unknown regions. The thesis aims to evaluate the performance of each model and compare GRF feature importance output with the global model. The best model features meteorological importances are used to create a drought vulnerability map of Czechia. Produced assessment is compared to existing drought vulnerability projects.

Key words: drought, vulnerability assessment, Geographical Random Forest

ABSTRAKT

Sucho je přírodní katastrofa, která negativně ovlivňuje miliony lidí a způsobuje obrovské ekonomické ztráty. Tato práce zkoumá zemědělské sucho v Česku pomocí algoritmů strojového učení. Použité statistické modely byly Random Forest (RF), Geographic Random Forest (GRF) a Locally Tuned Geographic Random Forest (LT GRF). GRF se skládá z několika RF modelů vytrénovaných na podmnožinu původních dat. Konečná predikce je váženým součtem predikce lokálního a globálního modelu. Velikost podmnožiny je určena laditelným parametrem. LT GRF řeší prostorovou variabilitu velikosti podmnožiny a lokální váhu. Během procesu ladění jsou pro každé místo nalezeny optimální parametry a poté interpolovány pro neznámé oblasti. Tato práce si klade za cíl vyhodnotit přesnost každého modelu a porovnat výstup důležitosti faktorů GRF s globálním modelem. K vytvoření mapy zranitelnosti vůči suchu v Česku se využije důležitost meteorologických faktorů. Vytvořené hodnocení je porovnáno se stávajícími projekty zranitelnosti suchem.

Klíčová slova: sucho, posouzení zranitelnosti, Geographical Random Forest

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

AHP Analytical Hierarchy Process.
AWC Available Water Capacity.

CDI Combined Drought Indicator.
CLC Corine Land Cover.

EU-DEM European Digital Elevation Model.

fAPAR fraction of Absorbed Photosynthetically Active Radiation.

GRF Geographical Random Forest.
GWR Geographic Weighted Regression.

LHS Latin Hypercube Sampling.
LOO CV Leave-One-Out Cross-Validation.
LT GRF Locally Tuned Geographical Random Forest.
LUCAS Land Use and Cover Area frame Statistical Survey.

MAE Mean Absolute Error.
MAPE Mean Absolute Percentage Error.
MSE Mean Square Error.

OOB Out-of-Bag.

PDSI Palmer Drought Severity Index.

RF Random Forest.
RMSE Root Mean Squared Error.

SOM Soil Organic Matter.
SPEI Standardised Precipitation-Evapotranspiration Index.
SPI Standardized Precipitation Index.
SPLEI Standardized Precipitation Latent Heat Evapotranspiration Index.
SWI Soil Water Index.

TPPE Topsoil Physical Properties for the Europe.
TWI Topographical Wetness Index.

CONTENTS

1 INTRODUCTION

Environmental hazards are natural phenomena, which negatively affect human society in terms of economic and social losses. Drought hazard belongs to most damaging and widespread causes of huge economic and human losses. The severity of the drought depends on the environment's (or society's) ability to cope with hazards. For example, in developed countries, drought's direct impact is almost invisible, and indirect impact projects to higher consumption of water to irrigate agricultural plants. In developing countries, drought might cause crop failure and subsequent instability. However, climate change will worsen many aspects of drought, its recurrence, severity and timespan (Mukherjee et al., 2018). Therefore, the need for mechanisms and tools, which make the environment and society more resilient towards drought is pressing. The first step is identifying vulnerabilities in the agricultural system and describing natural coping strategies. One such tool is vulnerability assessment, which helps to identify vulnerable regions. Researchers can then find underlying causes of vulnerability. On the other hand, less vulnerable regions can point out factors if assessed, which increase resilience. Vulnerability assessments are usually conducted by subjective assigning weight to factors.

The use of machine learning models expanded into geographical topics. However, these algorithms are aspatial, thus cannot identify underlying spatial patterns in data. It is analysing spatial data, not spatial analysis (Fotheringham, Charlton, et al., 1996). The spatial analysis utilises spatial aspects existing within to create a more accurate conclusion and describe patterns from a spatial perspective. Models incorporating spatial aspects should perform better with spatial data than their aspatial counterparts. Spatial analysis can in addition explain spatial non-linearity and change in the dependent variable with space.

In this thesis variants of spatially explicit machine learning algorithm Geographical Random Forest (Georganos et al., 2019) (GRF), which builds on Random Forest (RF) algorithm, will be tested on environmental data. Best variant will be chosen to create a vulnerability map of study area, which is defined by bounding box of Czechia boundaries.

1.1 RESEARCH QUESTIONS

Three research tasks were formulated, centred around the problem of spatial non-linearity in vulnerability analysis.

- Build and evaluate GRF and Locally Tuned Geographical Random Forest (LT GRF) for prediction of drought indicator.
- Explore spatial variability of feature importance from local models in comparison to the global model.
- Create vulnerability assessment to drought hazard with the most accurate algorithm.

First task comprises of preparing a dataset of environmental data and building several regression prediction models - RF, GRF and LT GRF. Finally, these models are evaluated in terms of prediction accuracy. In addition, topics concerning performance of models including ability to capture spatial patterns, computational complexity, effectiveness of tuning process will be discussed. The main question, which the topic will also seek to answer is: "Is it worth it to build spatially explicit machine learning model such as GRF (possibly with local tuning) compare to regular aspatial RF ?"

The second task involves an examination of RF unique feature, ability to return importance of variables used during the training phase. RF returns importance for the whole dataset, on the other hand, GRF opens the possibility to study variable importance from each location. Differences between local and global variables will be studied. As a tool of analysis, only visual exploration will be employed.

Lastly, vulnerability assessment of drought hazards is created in form of a map. The task includes developing a new methodology, which processes existing data and returns the output. The validation process is done by comparing existing vulnerability assessments from different sources. The question is, whether it is possible to accurately assess vulnerability in comparison to the ;other methods.

1.2 THESIS OUTLINE

The thesis is structured into chapters:

- Chapter 2 reviews studies related to drought hazard and machine learning, random forest in particular. The chapter also describes studies and projects which monitor drought hazards in Czechia.
- Chapter 3 focuses on theoretical background. The chapter describes drought identification and quantification with drought indices, a framework for drought vulnerability assessment and factors, which influence the severity of drought. Next, the chapter focuses on machine learning, mainly on algorithm random forest. Besides that, spatial extensions to machine learning and tuning of parameters are discussed. Lastly, some sampling approaches are described.
- Chapter 4 includes a description of datasets, tools and study area.

- Chapter 5 describes the methodology. Firstly dataset preparation, visual exploration and sampling is addressed. Later, the model building of random forest, geographical random forest and locally tuned geographical random forest are described. At last, a method of assessing vulnerability is presented.
- Chapter 6 presents the results of model testing, variable importance and vulnerability assessment.
- Chapter 7 discuss the results and evaluates, if the research tasks were fulfilled.

2 REVIEW OF LITERATURE

This chapter aims to provide a comprehensive review of relevant scientific publications. Publications mentioned are a primary source for the theoretical part of the thesis, which provides knowledge to construct the methodological part. Review of literature consists of several subsections, first one lists publications dealing with drought hazards and the concept of vulnerability towards drought, second review publications about machine learning, RF algorithm, spatial aspect in machine learning and use of machine learning in drought modelling.

2.1 DROUGHT AND DROUGHT ASSESSMENT

Drought as one of the most severe natural hazards has been exhaustively described in many publications. Therefore, only a small fraction of studies, which are subjectively most beneficial, will be mentioned. Firstly, it is needed to bring up studies, which define natural hazards and provides a framework for vulnerability assessment. Several authors discussed the concept and definition of drought and its usability, for example, D. A. Wilhite and Glantz (1985), D. Wilhite A. (2000) argue the importance of the definition in case of drought. Lloyd-Hughes, 2014 stress the impracticality of a single, uniform definition of phenomena that is dependent on human needs and intervention. Despite the inconsistency in determining a uniform definition, the interpretation from Palmer (1965) is often used; "interval of time, generally of the order of months or years in duration, during which the actual moisture supply at a given place rather consistently falls short of the climatically expected or climatically appropriate moisture supply". However, this definition is suitable for meteorological drought. The thesis analyses agricultural drought, which definition is altered to focus on plants conditions. Mishra and Singh (2010) define agricultural drought as "period with declining soil moisture and consequent crop failure without any reference to surface water resources.". Crop failure refers to absent or diminished crop yield relative to expectations. The meteorological definition describes relative deficiency in precipitation, on the other hand, agricultural drought is a physical manifestation of such deficiency. For research purposes, drought needs to be quantified using an indicator.

Drought manifests itself differently across the world and can be researched from different perspectives. It is relative in space, time, scale and research field. The relative nature of drought gave rise to a high number of drought indicators and indices. Several studies summarise and list used indices and indicators. Mishra and Singh (2010) describe some of the most popular. A comprehensive list of approximately 50 indicators and indices provides pub-

lication from World Meteorological Organisation ((Svoboda, Fuchs, et al., 2016)). A lot of researchers summarise and point out aspects of indices, for example, Heim Jr (2002) for indices used in 20. century, Niemeyer et al. (2008) describe indices by categories or Zargar et al. (2011), whom review advantages and disadvantages of various indices. Many authors evaluate performance of indices in specific region for example, Adnan et al. (2018) in Pakistan, Teweldebirhan Tsige et al. (2019), Tian et al. (2018) in United States, Raible et al. (2017) in all of Europe or Szalai, Szinell, and Zoboki (2000) in Hungary. Keyantash and Dracup (2002) evaluated indices by qualitative factors. X. Liu et al. (2016) reviewed agricultural indices in particular and outlined challenges in detecting agricultural drought. From the research aims perspective the most important are studies, which evaluates agricultural indicators in climatic conditions similar to Central Europe.

Natural hazards and their mitigation is a topic often discussed by United Nations. In their publication, Office for Disaster Risk Reduction (2004) discussed vulnerability and risk mitigation initiatives around the globe. According to De Stefano et al. (2015) vulnerability is understood and conceptualized in different ways from diverse scientific domains. Despite that, two major epistemological approaches can be identified. Firstly, vulnerability as "relationship between the severity of the hazard and the degree of damage caused" (Füssel, 2007) and secondly, vulnerability as a condition of socio-economical background. The first approach is more relevant to this thesis. Various vulnerability models were constructed, for example the Pressure and Release Model (Blaikie et al., 2014), Hazard and Risk model (Office for Disaster Risk Reduction, 2004) or "Methods for the Improvement of Vulnerability Assessment in Europe" model by (Joern Birkmann et al., 2013).

Drought as natural phenomena and risk describes many authors, for example Blaikie et al., 2014 or Bryant, 2005. Comprehensive information about drought causes, drought factors and vulnerability assessment provides publication from Nagarajan, 2010. Additional materials containing information about current research drought vulnerability are for example from Hagenlocher et al., 2019 or from Ionescu et al., 2009 in context of climate change.

The first notable vulnerability assessment was conducted by Wilhelmi and D. A. Wilhite (2002) by subjective weighting individual factors over Nebraska. A similar approach used by Jain et al. (2015) in the region of Ken river in India, Shahid and Behrawan (2008) in Bangladesh. Ekrami et al. (2016), Hoque, Pradhan, and Ahmed (2020) improve exactness of assessment by employing analytical hierarchy process (AHP) method. Fuzzy logic methods employed D. Zhang et al. (2011) in Liaoning province in China, Dayal et al. (2018) in Queensland in Australia and Hoque, Pradhan, Ahmed, and Sohel (2021) in northern New South Wales in Australia. A new technique - machine learning models such as Support Vector Machine, random forest or boosted regression trees utilised Rahmati, Falah, et al. (2020) for vulnerability assessment of part of Queensland, Australia. A similar study, only with an artificial neural network, was developed a year later (Rahmati, Panahi, et al., 2020). The last two studies are important sources as they assess vul-

nerability to drought with machine learning.

2.2 MACHINE LEARNING

Machine learning as a discipline encompasses various algorithms, methods and approaches. In recent decades, as a consequence of the rapid growth of computational performance, the field of machine learning experienced a renewal after the so-called AI winter. Interest in machine learning and artificial intelligence surged in recent years following the availability of big data from internet users. Therefore, machine learning became a popular tool for data analysts. Demands for theoretical and practical is reflected by a huge amount of publications. For example, the publication Pattern Recognition and Machine Learning from Bishop (2006), which belong to the most cited in the field, provides main information from a theoretical perspective. A valuable theoretical base also provides older publication from Mitchell (1997). Other examples of well written publication with a more contemporary approach are from Alpaydin (2020) and Hastie et al. (2009). Machine learning evolution from a less mathematical, more exemplary perspective offers Russell and Norvig (2002) in their more than 2000 pages long book. Several authors tried to provide more specialized theoretical knowledge such as Kuhn, Johnson, et al. (2013) publication devoted to statistical prediction or Berk (2020), who wrote about machine learning from a regression perspective. A lot of publication offers more applied information. James et al. (2013) deliver mathematically lightly put content with code examples in R targeted at beginners.

As was stressed, machine learning comprises plenty of methods and approaches, including algorithms. The popular ones are the artificial neural network, support vector machine, linear and logistic regression and last but not least, algorithms based on a decision tree. Because decision tree is simple, intuitive and easy to interpret, their variants are known from ancient times. One of the first notable use decision tree-like structures is Porphyrian tree, from Greek philosopher Porphyry (3. century CE). Hierarchical tree represent scale of being, from "supreme genus" to "species" and "individuals" (Lima, 2011). Earliest modern utilisation of a decision tree, according to de Ville (2013) trace back to Belson (1956). The term Classification and Regression Tree was coined in the publication by Breiman et al. (1984), in which authors describe the methodology used to construct decision trees. The first true tree-like machine learning algorithm was developed by Quinlan, called "Interactive Dichotomizer" (ID3) (J. Ross Quinlan, 1986), subsequently refined to C4.5 algorithm (J Ross Quinlan, 1992). The next improvement of decision tree-based algorithms was achieved by incorporating bootstrap sampling, developed by Efron (1992). Bootstrap can be described as random sampling with replacement, and in the context of decision tree leads to lower bias and variance. This approach led to the design of Adaptive boosting (Freund, Schapire, et al., 1996) or short Adaboost. The algorithm creates many shallow trees (called stump) and iteratively train and assign

weight to them. Ho (1995), and later Breiman (2001) further refined bootstrapped decision tree by incorporating randomization of input features, an improved algorithm is known as RF.

The excellent performance of RF attracted the attention of users and researchers alike, which resulted in a high number of publications discussing properties, variants and extensions. Genuer et al. (2008) contribute with general methodological insight along with advice for feature selection. Feature importance caught attention of several authors; Grömping (2009) compares RF features importance with linear regression, Strobl, Boulesteix, Zeileis, et al. (2007) researched bias in feature importance results, Strobl, Boulesteix, Kneib, et al. (2008) and Archer and Kimes (2008) inspect performance with high dimensional dataset. Probst, Wright, et al. (2019) review pieces of information concerning hyper tuning RF.

Increasingly available machine learning algorithms found application in geography. Fotheringham, Charlton, et al. (1996) published adjustment to the linear regression to be better suited for spatial data, known as Geographic Weighted Regression (GWR). The concept is further broadened in their publication (Fotheringham, Brunsdon, et al., 2003). Fotheringham, Crespo, et al. (2015) later extended this concept to the time dimension. Georganos et al. (2019) applied the same principle to RF, creating GRF. Another approach to incorporate spatial aspect to machine learning are geographical covariates in RF (Hengl et al., 2018), (Meyer et al., 2019). Because GRF is relatively new concept, only a handful of studies utilised it. Santos et al. (2019) used GRF and dasymetric mapping to evaluate forest change drivers in Ecuador. Hokstad and Tiganj (2020) compared performance of GRF to regular RF and Kriging in spatial modelling of gas deposit.

2.3 DROUGHT ASSESSMENT IN CZECHIA

Fur the purpose of validation vulnerability assessment, it is important to describe relevant publications and projects, which monitor and evaluate drought events in Czechia. Czech Hydro-Meteorological Institute provides warnings on possible drought events using Standardized Precipitation Index (SPI) and Standardised Precipitation-Evapotranspiration Index (SPEI) for meteorological and available soil water for soil drought. Intersucho (Miroslav Trnka, Hlavinka, et al., 2014) is a multidisciplinary project established in 2012, which provides a short-time and long-time prediction of drought severity and monitors the current condition of soil and vegetation. Water balance model SoilClim is used for estimation evapotranspiration and soil moisture. According to Hlavinka et al. (2011) model has been tested and performs well within the area of Czechia.

Publication from Brázdil and Miroslav Trnka (2015) is a comprehensive source of information about drought in Czechia. The authors describe means of monitoring drought with indices, historical drought events and the impact of

drought on various sectors of the economy, mainly agriculture, forestry and water management. The last part of the publication is dedicated to future trends, including a description of vulnerable regions and estimated drought vulnerability for Czechia in form of a map. The vulnerability assessment was conducted with datasets provided with project intersucho and project EnviSec. Project EnviSec arises from the cooperation of several research institutes, financed by the ministry of interior, with aims to developed integrated methods for monitoring and evaluation of global changes that impact to environmental security of Czechia. The project began in 2012 and finished in 2015. The study conducted by (Pártl et al., 2017), one of the outcomes of the project, assess integrated risk as a function of hazard and vulnerability.

Several other studies assess vulnerability to drought. For example, Trnka et al. (2009) identified a drought-prone region with new developed combined drought index. Miroslav Trnka, Semerádová, et al. (2016) focus on assessing the vulnerability of agricultural areas. As an indicator of drought median number of days with a saturation of topsoil less than 30 % is used. Vulnerability of forest to drought in the context of climate change discusses Hlásny et al. (2014). The authors estimate future climate data from historical records and regional climate models. Results show that forests within Czechia are relatively resilient to climate change-induced droughts.

2.4 SUMMARY

The chapter describes the current state of research in the field of drought vulnerability assessment using machine learning models, specifically the RF model. The first part focus on different understanding of drought hazard, vulnerability assessment frameworks and distinct assessment methods used. Case studies (drought indicators and assessments) are mostly located in arid and semi-arid regions such as Australia, India or Iran. On the other hand, the number of studies, which would be localised in Central Europe or similar climate is very small. The contribution of such studies from different climatic environments to the thesis is smaller.

Second part focus on machine learning models and specific models, which deals explicitly with spatial information. This field is relatively new, as the GRF concept has been developed recently. Therefore, only a small number of studies exist, which might be used for the comparison of results. The third part deals with existing monitoring and vulnerability assessments in Czechia. The problem is similar. Despite many studies deals with drought hazards, a small number of them assess environmental vulnerability to agricultural drought. Existing studies present vulnerability in insufficient scale and without distinction to land cover, which further hinders comparison and validation.

In conclusion, the spatial applicability of machine learning models is not thoroughly explored. Only a small number of studies utilise spatially explicit models such as GRF. In addition, unique property of RF - feature im-

portance assessment was not explored in this context. Similarly, use of machine learning models in vulnerability assessment is not wide-spread. The aim of this thesis is to provide insight into usability of spatially explicit RF model in vulnerability studies.

3 THEORETICAL BACKGROUND

This chapter aims to describe theoretical concepts, which are required for developing a new methodology for assessment of agricultural drought with a spatially sensitive RF algorithm. The first part is dedicated to drought. It is important to describe the environment and factors, which influence the severity of drought. The next part comprises theoretical information about machine learning principles. Next, RF and GRF are described. Lastly, sampling methods are described.

3.1 DROUGHT

Compared to droughts, earthquakes, tropical storms or floods belong to more prominent and visible hazards. Subsequent devastation, which occurs quickly is easy to capture and quantify. On the other hand, "drought is a creeping phenomenon that accumulates over a period of time across a vast area, and the effect lingers for years even after the end of drought" (Sivakumar, 2011). Drought is often a trigger of more serious effects, for example, drought can cause crop failure, which results in famine and violence. Drought ranks first among natural disasters sorted by several criteria such as severity, length, area extent or loss of life (Bryant, 2005). According to Blaikie et al. (2014) impacts of drought contribute 86.9 % of all deaths caused by natural hazards between the years 1900 - 1999. Despite successful mitigation of negative effects of drought attribute mainly to humanitarian aid and accelerated economic development, 1.5 billion people were affected by drought in 1998 - 2017 (Wallemacq, 2018). Because of huge economic and human losses, there is a constant effort to develop methods to detect, monitor, predict and mitigate drought hazards.

Traditionally, drought is classified into four categories. Meteorological drought refers to the shortage of precipitation over an area and period. Hydrological drought relates to an insufficient supply of surface and subsurface water. Agricultural drought is linked to meteorological and hydrological drought and refers to lack of soil moisture and subsequent crop failure. A plants demand for water is dependent on prevailing meteorological conditions, biological characteristics of the specific plant, its stage of growth, and the physical and biological properties of the soil (D. A. Wilhite and Glantz, 1985). Thus agricultural drought severity may vary from plant to plant. Lastly, socio-economical drought expresses the failure of meeting market demand. Mishra and Singh (2010) suggest adding groundwater drought, which refers to low levels of groundwater resources.

It is important to identify drought hazard accurately. Imprecise identification leads to inaccuracies, which are propagated to all subsequent steps and will hinder model building process. To achieve this goal, three properties need to be obtained; initiation and termination, duration and severity of the drought (Yevjevich, 1967), (Mohan and Rangacharya, 1991). A. D. Wilhite (2011) adds spatial extent, which is needed to distinguish regional droughts. The magnitude of the drought is calculated as the ratio of severity and duration (Dracup et al., 1980b). The longer the duration, the smaller the magnitude of drought. Dracup et al. (1980a) lists several steps:

1. The nature of water deficit: This point refers to a particular definition of analysed drought. In the case of agricultural drought, it is a deficit of available soil moisture and subsequent crop failure.

2. Identification of variable: This step was added by Mohan and Rangacharya (1991) and is closely linked to the first point. This step includes the selection of indicators, which is used to quantify drought. Various indices were developed to measure drought and the selection of indicators will be discussed further below.

3. Identification of averaging period: The selection of time step (for example hours, days, months) over which analysed data are aggregated. A shorter averaging increment will result in a larger number of drought events and a longer increment in a smaller number. Secondly, a shorter time step tends to result in a bigger serial correlation than a larger time increment.

4. The truncation level: Component which serves as a divider between a time series of into drought period and normal period. Truncation value can be set arbitrary, but mostly a statistical variable is used. Several truncations or threshold values might be used for different categories of drought.

3.1.1 Drought indices

According to Mishra and Singh (2010), drought indicator is a prime variable for assessing the effect of drought and defining different drought parameters, which include intensity, duration, severity and spatial extent. The selection of appropriate indicator is essential as it will be the dependent variable, which will be predicted by model. Firstly, it needs to distinguish between drought index and indicator. Indicators are variables or parameters used to quantify drought hazard. A typical indicator is the amount of precipitation or soil moisture for a selected time scale. Indices are usually calculated numerical representations of drought and are computed from indicators (Svoboda, Fuchs, et al., 2016). Besides providing information about drought events, indicators and indices help compare quantitative drought impacts over variable scales of geography and time and facilitating the communication of drought conditions among various interested entities (Zargar et al., 2011). There are more than a hundred drought indicators and indices, yet no one is generally accepted as perfect. This subsection describes

drought taxonomy, most important indices and other issues connected to drought indices.

Same as drought, indicators are classified into meteorological, hydrological and agricultural. Socio-economic and agricultural drought will not take place without one of the former. The indicator of socio-economic drought would be monetary (Keyantash and Dracup, 2002), which is difficult to assess. Meteorological indices use precipitation and temperature as input. Hydrological indicators are derived from stream characteristics, lastly, agricultural indicators are concerned mainly with soil moisture and vegetation conditions. Niemeyer et al. (2008) adds the category of comprehensive indices which combine variables from more fields, for example, precipitation and soil moisture. Indicators can be classified through the acquisition of variables needed to construct indices. Indicators constructed from in situ observation rely on a network of meteorological stations. Because of a long history of meteorological records, these indices may be used in historical research. On the other hand, in the region with sparsely located meteorological stations records might not be sufficient. The second option is to utilise remote sensing. The obvious advantage compared to in situ is spatial resolution, which is uniform across borders. Research scope of thesis predetermines the selection of agricultural indicator, however other categories are often use in agricultural drought research and shows decent performance. Some of the most prominent are described further as they are widely available and well-known, which eases reproducibility.

One of the oldest and most prominent indexes is the Palmer Drought Severity Index (PDSI) (Palmer, 1965). Index use water supply and water demand calculation, which requires precipitation, temperature data and available water content of the soil to approximate soil moisture. Soil is divided into two layers, the upper layer, which is 25 mm deep and the lower layer. Firstly, the top layer is saturated, then lower and lastly run-off will occur. Potential evaporation (demand side) is calculated using a model developed by Thornthwaite (1948). Drought will occur when all moisture in the soil is evaporated. Despite being popular and widely used, there are several concerns and objections. According to Heim Jr (2002), index needs to be calibrated if a comparison between different regions or months is needed. This issue resolves Z-index, created by the same author, which does not require data from previous months. The index is not as sensitive as PDSI to calibration and is more suitable to short term assessment of drought (Karl, 1986). Alley (1984) emphasizes a number of arbitrary assumptions that were made during the development of the PDSI. Firstly, the model developed by Thornthwaite is one of many available models, other models could be easily used. The second problem is an arbitrary designation of drought severity classes. Also, PDSI was developed for a semi-arid region of the United States and it is inaccurate for mountainous regions. Despite that index has come under scrutiny in recent studies, its variations provide accurate means of drought identification, thus making it a suitable indicator in drought assessment.

SPI developed by McKee et al. (1993). The index is calculated from long-term records of precipitation. The records are fitted to the probability function and then transformed using the Gaussian function, which results in the mean of SPI being zero and variance being one. According to Guttman (1999), who compared several distribution models, Pearson type III and Generalized additive model achieved the best performance, especially for wet events. Time interval is flexible, usually computed with 1, 3, 6 or 12 months intervals. Despite being a meteorological index, SPI is often used for agricultural drought. Szalai, Szinell, and Zoboki (2000) investigate relations between soil moisture, PDSI and SPI in which 2-months SPI achieved the best correlation with soil moisture. Zargar et al. (2011) summarise the advantages and disadvantages of the index. SPI is simple to obtain because relies on precipitation records only. It is adaptable, maybe use for various times scales and various types of drought. In addition, it can be used to compare different climates and monitor the wet period as well. On the other hand, SPI is loosely connected to ground conditions and additional variables are required. Lastly, long precipitation records are essential for accurate assessment. One of the weak points of the SPI tried to solve (Vicente-Serrano et al., 2010), who introduced the SPEI. The index incorporates potential evaporation and SPI. However, potential evaporation requires several variables, which might be hard to obtain, for example, surface temperature, air humidity or water vapour pressure. SPI would not be the first choice for the indicator as it lacks a link to the link to others features, however, it is easily obtained and widely known. SPEI is linked to ground conditions, on the other hand, it is not as easily obtained.

Combined Drought Indicator CDI belongs to the category of combined or comprehensive indices. Was developed by Sepulcre-Canto et al. (2012) to monitor agricultural drought over Europe. CDI combines SPI, anomalies of soil moisture and the Fraction of Absorbed Photosynthetically Active Radiation (fAPAR). Classification scheme includes five categories “watch”, “warning”, “alert”, “partial recovery” and “full recovery” after specific condition is achieved. For example, for the first category three months, SPI must bust be lower than -1 or two months SPI lower than -2. Recently, the index was update by Cammalleri et al. (2021) to include new category “temporary recovery”. Despite being well suited for Europe and being an agricultural indicator, categorical values of CDI are not suited for a regression problem. Transforming indicator to continuous values would increase inaccuracy, which makes indicator somewhat less convenient.

Soil moisture, especially within the root system of the plants, is an accurate indicator of agricultural drought. There have been several attempts to develop a model, which simulates a water balance in soils layers. For example, Huang et al. (1996) created a model which uses monthly mean temperature and precipitation. However, with the development of remote sensing, assessment of soil moisture from synthetic aperture radar emerged. Soil moisture detection utilises an increase in the backscattering of microwave radiation. Soil saturated with water has a higher dielectric constant than dry soil, however during temperatures below freezing point, water content appears

similarly to dry soil. Several physical and empirical models to extract information about soil moisture were developed. Most notable, multi-temporal method developed by Wagner, Lemoine, et al. (1999) utilises archives of microwave imagery. For every pixel is found a record with the lowest value, which equates to minimal soil moisture and base roughness of the terrain. This base value is then subtracted from other images resulting in soil moisture. Information about soil moisture might be represented by the Soil Water Index (SWI), which provides relative information about the water saturation of the soil. Compare to other mentioned indicators SWI is not as widely used and studied, thus making it less comparable and known. This is not necessarily a negative attribute, but compared to other indicators its applicability to drought assessment is not as much examined.

Besides the aforementioned indices, several others need to point out. In a category in situ indices belongs Crop moisture index (Palmer, 1968). Index is constructed from mean weekly temperature values, total weekly precipitation and index value for the previous week. Opposite to PDSI, which maps long term drought, index is better suited to evaluate short term soil moisture shortage. Surface water supply index developed by Shafer and Dezman (1982). Index is calculated from the monthly probability of precipitation, reservoir storage, streamflow and snowpack. The index is used to detect anomalies in the surface water supply. However, index is regional index computed primarily for river basins in just the western United states (Heim Jr, 2002) thus its transferability to other climatic zones is questionable. From a remote sensing perspective, most utilised is the Normalized Difference Vegetation Index, which is calculated from red and near-infrared bands. Most satellites designed for terrestrial observation contains these bands, which make index accessible. Tucker et al. (1991) demonstrated the usability of index for agricultural detection in the Sahel region in Africa. Enhanced Vegetation Index (Huete et al., 1994) builds on concept Normalized Difference Vegetation Index. New index minimize effect atmospheric and soil background effects (Sivakumar, 2011).

There have been many attempts to find the best drought indicator. Computed soil moisture followed by soil moisture anomaly index and Z index was found best by several criteria such as robustness, transparency or extendibility (Keyantash and Dracup, 2002). Performance of drought indices is region-specific due to the variability in meteorological variables and streamflow characteristics which are used for deriving indices (Mishra and Singh, 2010). Numerous studies compare the most prominent indices - PDSI and SPI. According to Szalai and Szinell (2000), SPI is more suitable and flexible. Raible et al. (2017) tested several indices in Europe from 2000 years of records. SPI was found to be most efficient in west Europe, whereas for southern and eastern Europe Standardized Precipitation Latent Heat Evapotranspiration Index (SPLEI), which incorporates temperature and evaporation, was found to be better. PDSI, which involve water supply and demand suits better northern Europe. Similar results conclude Bachmair et al. (2018); SPI is a better indicator for drought in western Europe and SPLEI is more effective in eastern and southern Europe. In conclusion, there is no single

indicator that outperforms others. Effectiveness depends on the location and availability of meteorological and hydrological records.

3.1.2 Drought risk assessment

Firstly, it is needed to distinguish between predicting drought, monitoring drought and assessment of drought vulnerability. Drought modelling (forecasting) refers to process of predicting the onset, duration and severity of drought time period in advance. These properties are forecast by various methods, for example regression analysis, time series models or probability models. Input typically consist of hydro meteorological variables and climate indices such as sea surface temperature. Drought monitoring involves continuous observing of drought indicators and evaluating its severity. Vulnerability assessment on the other hand, is not concerned with exact period in future, rather evaluates hazard risk for selected extent in general. Vulnerability assessment as one of the task of thesis is described in detail.

It has been demonstrated that drought hazard causes huge economic loses and even famine. Thus, it is important to assess hazard risk and identify vulnerable regions and populations. Firstly, it is needed to describe and select a framework within which several concepts will be defined. For example, the basic idea of "Pressure and release" model (Blaikie et al., 2014) is that a disaster is a result of two opposing forces - hazard and vulnerability of the environment. The vulnerability has three components; root causes, dynamic pressures and unsafe conditions. The framework of MOVE model (Joern Birkmann et al., 2013) is much more complex, authors further distinguish vulnerability to exposure, susceptibility and fragility, and lack of resilience. These factors interact with hazard resulting in a degree of risk. In Hazard and risk model (Office for Disaster Risk Reduction, 2004), widely used across scientific community, risk is seen as the results of two components; hazard and vulnerability. The last framework provides clear and simple explanation of hazard, which is well suited for thesis aims. Within the framework standpoint other terms are defined.

According to Office for Disaster Risk Reduction (2004), risk can be defined as "the probability of harmful consequences, or expected losses (deaths, injuries, property, livelihoods, economic activity disrupted or environment damaged) resulting from interactions between natural or human-induced hazards and vulnerable conditions". Drought hazard has been defined earlier. In the hazard and risk model drought is characterizes by its severity and probability. Probability is a relative chance of hazard occurring in time. The severity of drought is quantified by drought indices. Probability aspect is omitted in analysis as there is no option to model relation between relative chances and predefined time period.

The concept of vulnerability is often blurred and vaguely defined. Wilhelmi and D. A. Wilhite (2002) summarise several authors who agree that vulnera-

bility describes the degree of susceptibility of society to a hazard. Lavell et al. (2012) define vulnerability as "propensity or predisposition to be adversely affected by hazard and is a result of diverse historical, social, economic, political, cultural, institutional, natural resource and environmental conditions and processes". Vulnerability is often further classified to better describe specific aspects of the affected place. Joern Birkmann et al. (2013) distinguished dimension of vulnerability; social, economic, physical, cultural, environmental and institutional. The environmental dimension, for example, describes the propensity for ecological and biophysical systems to be damaged. In more practical meaning, the environmental dimension of vulnerability can be assessed in terms of a crop's areas ability to resist the effects of drought (Han et al., 2016).

Vulnerability varies across the space and economic background of those affected. Vásquez-León et al. (2003) found vast differences in socio-economical vulnerability and response to drought on both sides of the US-Mexico border. In developing nations drought vulnerability constitutes a threat to livelihoods and the ability to maintain productive and stable economies. In developed countries, drought poses economical risks to individual and public enterprises (Bakker and Downing, 2000). Vulnerability is linked to infrastructure and socio-economical conditions, thus the poor suffer more from hazards than the rich (Wilhelmi and D. A. Wilhite, 2002). The environmental vulnerability may differ on the other side of the border as well. For example, it is known that different farming practises and field sizes on the opposite sides of the Czech Austrian border, affect soil erosion (Čermáková et al., 2014), therefore soil ability to cope with precipitation deficit.

Vulnerability has also a time dimension (Wilhelmi and D. A. Wilhite, 2002). Vulnerability changes over time as a result of technological, economical and demographic progress. Often, vulnerability decrease as a result of policies aimed to improve the situation after a previous disaster. However, inadequate response to hazard events may increase the vulnerability of the environment in time. Another important aspect of vulnerability assessment is its scale. Before the assessment, it is needed to define spatial and temporal scale, which will suit research aims. Unfortunately, spatial scale is often driven by data availability (Fekete et al., 2010).

According to Omar Dario Cardona et al. (2012) exposure refers to the inventory of elements in which hazard event may occur. No hazard risk exists if economic resources or population were not located in the affected region. It is necessary to be exposed, while vulnerable. In the case of assessing agricultural drought risk exposure may refer to crop area or the presence of vegetation. As the thesis deals with agricultural drought, primary focus is on soil used for crop production.

Vulnerability assessment can be described as a process of identifying, quantifying and scoring the vulnerabilities in a system (De Stefano et al., 2015). Most studies concerning mapping drought vulnerability were issued in the last two decades. According to Wilhelmi and D. A. Wilhite (2002) two main

reasons led to this increase; firstly, the recognition of the importance of vulnerability in hazard assessment and secondly, the growing availability of GIS technology, which allows integration of spatial data from various sources. For accurate assessment is needed to specify "the vulnerable entity, the stimulus to which it is vulnerable and the preference criteria to evaluate the outcome of the interaction between the entity and the stimulus" (Ionescu et al., 2009). In the case of environmental vulnerability assessment, the vulnerable entity is vegetation, the stimulus is the variety of factors, for example, amount of precipitation and the criteria is selected indicator.

Various techniques for mapping drought vulnerabilities were developed. First studies, incorporated several factors which influence drought vulnerability, for example, climate factors, soil properties, land use or irrigation proximity, to create an agricultural drought vulnerability. Each factor was assigned weight, which value was based on informed assumptions on the relative contribution to drought vulnerability. The apparent problem is a subjective assumption of weight for each factor. The weight would differ for each scientist and such assumptions will be the source of error and inaccuracy. An improvement over subjective assumptions provides utilisation of AHP technique, in which a questionnaire is answered by experts in agriculture. Different approach is to use the natural break method (Zeng et al., 2019). This method classifies factor into a defined number of classes by natural break classification, which seeks to minimize the average deviation from the class meanwhile maximizing the deviation from the other classes.

Machine learning algorithms can be applied to vulnerability assessment. Contrary to others method, no weight to factors is assigned in the process. A prediction (classification or regression) model is created; drought factors are independent variables and drought index (or other vulnerability indices) is dependent. A drought vulnerability map is created from a trained model. Opposite to previous methods, the impact of individual factor cannot be assessed. The exemption is a RF algorithm, which provides the relative importance of independent variables. The main advantage of the utilization of machine learning algorithms is no need to subjective assign weights. The efficiency of the model, and therefore the assessment is quantified by selected metric, for example, recall or RMSE. On the other hand, it is not possible to create hazard, exposure and vulnerability assessment separately.

3.1.3 Drought vulnerability factors

Agricultural drought is linked to many factors, which need to be considered in vulnerability assessment. In these subsections, various properties of the environment will be described. Also, meteorological variables typically belonging to the hazard category are included. This is because of the nature of the drought index or indicator, which values are obtained in time and place and influenced by meteorological characteristic. Thus, these properties need to be included to compensate for temporal variation. An abundance of fac-

tors was used in vulnerability assessment, for example, terrain characteristic, soil properties or land use. The listed factors were described because of their well-founded link to the agricultural drought vulnerability in Central Europe.

Terrain characteristic refers to quantitative descriptions of the physical features of the land. Topography alters various climatic condition. It is the most important factor controlling soil water redistribution, organic matter, nutrients, soil textural composition and other properties, which affect plants well being (Dinaburga et al., 2010). *Altitude* translates to lower temperatures and different conditions, which sustain different processes than in lower altitude regions. Vegetation in mountainous regions subscribes to different patterns of climatic conditions and developed specific adaptation. In higher altitude snow cap protects plants from frost and its absence may harm the mountainous ecosystem and make them more vulnerable to drought in spring and summer. Compare to highland plants, lowland vegetation is more resilient to winter drought (Rosbakh et al., 2017).

Slope of area affects the run-off, recharge and movement of surface water. Flat terrain areas have relatively high infiltration rate, on the other hand, the areas with steeper slopes have low infiltration rate and higher run-off (Shekhar and Pandey, 2015). Steeper slopes tend to be more susceptible to erosion and soil degradation. If a drought event occurs, vegetation in steeper slopes is more vulnerable than vegetation in less steep areas due to deficit of soil moisture Hawthorne and Miniat, 2018.

Another topographic factor is *aspect*, which refers to the orientation of slope. The aspect of slope can influence local climate because of the length of the exposure to sun rays. West and south-facing slopes will be warmer than east and north-facing slopes, therefore have lower soil moisture and higher evaporation rate (Magesh et al., 2011). The importance of aspect is somewhat diminished, when vegetation is present (Oorthuis et al., 2021). More vulnerable to drought are south-west facing slopes, especially if vegetation cover is missing.

Topographical Wetness Index TWI developed by Beven and Kirkby (1979) describes proclivity of place to accumulate water based on topographic information. It is calculated as follows Mattivi et al., 2019;

$$TWI = \ln\left(\frac{SCA}{\tan\phi}\right)$$

where ϕ is slope angle and SCA is specific catchment area. SCA is computed by different algorithms, which are classified into single flow direction and multiple directions, depending on how water flow is distributed between the grid. The most notable include D8 (O'Callaghan and Mark, 1984), in which water flows in a grid is distributed to one adjacent cell through the steepest gradient and MD∞ (Seibert and McGlynn, 2007), which splits grid cells into triangles and then redistributed water flow proportionally to the slope gradient. According to the authors, this methodology improves dispersion on

planar and concave hillslopes. MFD-md (Qin et al., 2007) algorithms transfer water flow to all neighbourhood cell based on the linear function of a gradient. The study by Raduła et al. (2018) demonstrated better performance of MFD-md algorithms in comprehensive comparison of several algorithms. TWI is widely used to assess information about the spatial distribution of wetness condition due to the requirements of the only terrain model. On the other hand, index is static and relies on the assumption that slope is an adequate indicator for the effective downslope gradient. This is often not true on flat terrain (Grabs et al., 2009).

Meteorological factors. The connection between agricultural drought and meteorological patterns is clear. *Precipitation* is the only source of moisture for the environment with exception of irrigation, which is available for fraction of cultivated areas. Snow precipitation is important as its cover and protects soil (Dinaburga et al., 2010), provides soil moisture in early vegetation circle, which facilitates root growth (Cairns et al., 2011) and in higher altitudes is the source of water supply for river basin located downstream. *Temperature* influence the rate of transpiration, higher temperatures increase the transpiration rate. A region with higher temperatures is, therefore, more prone to drought. However, precipitation deficit impacts are greater than high temperatures in general (Yang et al., 2020).

Soil properties are important factors influencing the environment ability to cope with drought. Soil acts as a substrate for plants roots, provide them with water and nutrients. Soil characteristic influence these function to various degree. Soil properties refer to the physical and chemical attributes of the soil. Not all properties are significant, only a handful of them will be taken into account. *Soil texture* refers to the composition of solid particles, which soil is composed of. Sandy soil particles are smaller than 2 mm and bigger than 0.05 mm. Silt particles are smaller than sand but larger than 0.002 mm. Tinier than silt is clay. Soil water available for plants is stored in space between soil particles. Sands soil has the least space for water, which tends to evaporate faster. Clay soils have the most total spore space, however, water is held too tightly for plants to access (Sullivan, 2000). Silt soil provides a balance between water capacity and accessibility.

Soil bulk ratio convey the ratio of the dry mass to the total volume occupied in the soil. It is also an indicator of soil porosity. Soil with high bulk density is susceptible to surface run-off and erosion because water is restricted from moving through the soil. On the other soils with lower bulk ratio are more prone to vertical leaching of nutrients (Easton, Bock, et al., 2016). Soil bulk ratio can be altered by physical management processes, for example, tillage or by tunnelling activities of worms.

One of the most important factors is the amount of *organic matter* in soil. The mechanisation of agriculture, use of pesticides and use of intensification of agriculture reduce organic matter and its beneficial impact on soil quality. However, fertilizers and tillage can not fully substitute the function of the organic part of the soil, which reduces the resilience of soil and plants to

environmental hazards. According to Bot and Benites (2005) organic content increase water infiltration and water holding capacities, increasing diversity and activity of soil organism, provides nutrient availability. Organic matter that covers the soil surface protects from raindrops impacts, thus increasing rainwater infiltration, reducing run-off and erosion. Organic matter increases to greater activity of earthworms and other insects, which improves other soil properties. The pore space in the soil is enlarged with increased organic matter, making the soil more capable of holding water during heavy rains (Tirado and Cotter, 2010). Lastly, soil cover reduces water evaporation and shell soil from the negative heating effects of the sun.

Plants need appropriate *soil depth*. Otherwise, plants roots are too shallow and the plant is more vulnerable to be carried away by wind or excessive water flow. Shallow soils have lower available water. According to Scherer et al. (2017) if soil depth is less than 90 cm available water to plants is decreased. Above mentioned soil parameters and their relation to water retention can be summarized in *Available Water Capacity (AWC)* characteristic. AWC refers to the amount of water in the soil that can be removed by plants. It is estimated by the difference in soil water content between field capacity and permanent wilting point (Cassel and Nielsen, 1986). Field capacity refers to the level of soil moisture left after drainage of the gravitational water. The wilting point is defined as soil moisture low enough to be non-extractable for plants. Most crops are permanently damaged.

Land cover is intertwined with water demand and coping abilities of the environment to drought hazard. Land use describes how society uses land, land cover refers to physical features of the land. In case of vulnerability to drought, scientific community classify several types - mainly agricultural field, grassland, forest, barren lands, urban areas and water bodies (Jain et al., 2015), (Thomas et al., 2016), (Hoque, Pradhan, Ahmed, and Sohel, 2021). In agriculture cultivation methods has impacts on other factors, mainly soils. Tillage with heavy machinery, use of fertilizers and pesticides disrupts natural processes, which regenerates soil, reduce organic matter and make the soil more suspecting to erosion. Ecological agriculture mitigates many of mentioned negative effects (Tirado and Cotter, 2010). For example, conventional agriculture tends to cultivate monoculture, which is more prone to hazards. Ecological agriculture promotes biodiversity resulting in resilience to drought (Di Falco and Chavas, 2008).

Forests, on the other hand, are most resilient to drought (Peng et al., 2019). Forests produce organic matter in higher proportions than other ecosystems, which increases the water capacity of soils. Forests also transpire more water. Plants in woodlands ecosystem are more resilient than plant in grasslands, because of larger degradation of soil than in forested areas (Zhao et al., 2008). Grasslands are often used as pastures. Grazing of livestock has been associated with decreased soil infiltration rate, porosity and hydraulic conductivity, which leads to enhanced run-off rates (Weatherhead and Howden, 2009). Urban areas are typical for low infiltration rates and high run-off. Closeness to water bodies is another factor. Plants in close proximity are less

vulnerable because of abundance of surface and subsurface water. However, the benefit of water abundance is limited to close proximity to water source.

3.2 MACHINE LEARNING

Term machine learning is often interchangeable with terms pattern recognition and statistical learning. The main goal of statistical learning theory "is to provide a framework for studying the problem of inference, that is of gaining knowledge, making predictions, making decisions or constructing models from a set of data" (Bousquet et al., 2003). Definition of machine learning is very broad and often is described by examples. Mitchell (1997) states; "A computer program is said to learn from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E". Simply put machine learning can be defined as computational methods using experience to improve performance or to make accurate predictions (Mohri et al., 2018). Machine learning might be seen as a extension of statistical learning to computer environment. Alpaydin (2020) notes that role of computer science is twofold: First, in training we need efficient algorithms to solve the optimization problem, as well as to store and process the massive amount of data. Secondly, once a models learned, its representation and algorithmic solution for inference needs to be efficient as well.

Statistical models have two main tasks; prediction and explanation. Predictive modeling aims to create a statistical model capable of predicting new values from known observation. Explanatory or descriptive modeling intents to reveal relation between independent and dependent variables. By revealing relations prior hypothesis can be tested. Predictive model would be build and tuned differently than explanatory model as Shmueli et al. (2010) stressed: "In explanatory modeling the focus is on minimizing bias to obtain the most accurate representation of the underlying theory. In contrast, predictive modeling seeks to minimize the combination of bias and estimation variance, occasionally sacrificing theoretical accuracy for improved empirical precision." Depending on goal different models might be appropriate. Linear models allow for relatively simple and interpretable inference, but may not yield as accurate predictions. In contrast some of the highly non-linear models provide more accurate predictions (James et al., 2013).

The traditional tasks of machine learning include classification and regression. Classification concerns with predicting categories, regression predicts values. Other tasks include ranking (predicting order according to selected criterion), dimensionality reduction, and clustering. Machine learning can be divided into several categories. Supervised learning is the most common scenario where labels are known for each observation in the training phase. On the other hand, the unsupervised learning algorithm receives unlabelled observation. In between are semi-supervised learning, transductive inference, on-line learning, reinforcement learning, or active learning which receive labelled observation to some degree at the beginning or during the

process.

Ultimate goal is build a model which will have high accuracy, in other words low testing error. This can be achieved by sufficient training of the model, which will result in simultaneously low variance and low bias. This balance is also called bias variance trade-off. "Bias represents the extent to which the average prediction over all data sets differs from the desired regression function and variance measures the extent to which the solutions for individual data sets vary around their average" (Extent to which the function is sensitive to the data set.) (Bishop, 2006).

Among the most used supervised algorithms are kernel methods (most known is support vector machines), based on decision trees (random forest or boosted regression trees) or artificial neural networks.

3.2.1 Decision trees

Decision trees based models make predictions by dividing prediction space into several subregions and have tree like hierarchical structure. Decision tree consist of decision node and leaf node with beginning at root node. Decision trees can be used for classification or regression problem, given the scope of thesis, only regression would be described. The building of a decision trees follow two steps. Firstly, the predictor space is divided into two subspaces. Secondly, for every observation which belong to subspace a prediction is made - mean of known samples. The problem is to find value dividing predictor space most efficiently. The threshold value is calculated so that the overall sums of squares error are minimized (Kuhn, Johnson, et al., 2013).

$$\text{SSE} = \sum_{i \in S_1} (y_i - \bar{y_1})^2 + \sum_{i \in S_2} (y_i - \bar{y_1})^2$$

where $\bar{y_1}$ and $\bar{y_2}$ are the averages of the training set outcomes within groups S_1 and S_2. However, sums of squares error not calculated only for one split at the time. Computing the best split for all predictor subspaces would be infeasible. This is top-down, greedy approach. Top-down because starts at the top of the tree and greedy because trees is divided at each split. This process is also known as recursive binary splitting.

Decision trees have many advantages. Decision trees are easy interpret (Kuhn, Johnson, et al., 2013), especially when are shallow. Trees are also relatively fast to compute. Another advantage is handling of missing values. When tree splits, only observation with selected predictor are used. After first split, substitute variable is created which mimic first predictor and split. On the other hand single decision trees have low predictive power and have high variance which results in overfitting. By nature of its construction, tree models split predictor space into rectangular regions, which might be not suitable for every feature. Often a small change in the data can result in a very different series of splits. The major reason for this instability is the hierarchical nature of the process: the effect of an error in the top split is

propagated down to all of the splits below it (Hastie et al., 2009).

High variance of decision trees can be reduced by process called bagging, short for bootstrap aggregation. Bagging belongs to ensemble techniques which incorporates many models to improve prediction efficiency. First step is to prepare bootstrapped samples. Bootstrap is random sampling with replacement, which means some observation will be represented more and some wont. On average 63.2 % of observations are included (Efron, 1983). Bagging is averaging results of several decision trees and can be formulated as;

$$\hat{f}_{bag}(x) = \frac{1}{B}\sum_{b=1}^{B} \hat{f}^{*b}(x)$$

where B is the number of bootstrapped samples. One third of observation, which were not sampled with bootstrap (Out-of-Bag (OOB) samples) can be utilized as testing samples. Bagging increases accuracy and reduces variance. On the contrary the readability of models is worsened.

3.2.2 Random Forest

As was stated, bagging can reduce variance by averaging trees. However, by nature of its building process, bagged trees are strongly correlated. Single trees are similar, especially at the top, because most important predictors are used first. This is also know as tree correlation and limits possible reduction of variance. RF reduce trees correlation by considering only a portion of predictors in every split.in regression by default only one third of variables are considered (in classification is $\sqrt{n}$). Algorithm is described below (Hastie et al., 2009);

1: **for** b = 1 to B **do**
2: Draw a bootstrap sample Z of size N from the training data.
3: **repeat**
4: Select m variables at random from the p variables.
5: Pick the best variable/split-point among the m.
6: Split the node into two daughter nodes.
7: **until** $n_{size} = n_{min}$
8: **end for**
9: Output the ensemble of trees $\{T_b\}_1^B$
10: To make a prediction at a new point x: $\hat{f}_{rf}^B(x) = \frac{1}{B}\sum_{b=1}^{B} T_b(x)$

The size of variance can be described by equation (Hastie et al., 2009);

$$Var\hat{f}_{rf} = \rho(x)\sigma^2(x)$$

where $\rho(x)$ is the sampling correlation between any pair of pairs of trees. $\sigma^2(x)$ is the sampling variance of any single randomly drawn tree. Variance is dependent on sampling variability of bootstrapped data, number of variables in splitting pool and their mutual correlation. In addition to mentioned

factors, low values of minimum node size tends to overfit the model (Segal, 2004). Bias remain same as bias of any single tree used in the ensemble.

One of the advantage of tree based models is ability to return variable importance. Breiman (2002) describe several methods or metrics to quantify variable importance;

- *Random permutation*. For every split in every tree, split values are randomly permuted. Then is calculated new error. Difference between of permuted and original error is variable importance. The larger the error, the more important the variable.
- *Gini impurity*. At every split for every variable decrease in the gini impurity is summed. Variable with highest decrease in giny impurity is the most important. For regression, MSE is used.

Gini impurity metric lacks accuracy when calculating importance for variables with many categorical values (Strobl, Boulesteix, Zeileis, et al., 2007), therefore random permutation is preferred method. The main problem of random permutation methods overestimates importance of correlated predictors (Strobl, Boulesteix, Kneib, et al., 2008). This is especially true in cases with many predictors and small number of observation. In such cases, prediction of variable importance tends to be unstable (Genuer et al., 2008). Kuhn, Johnson, et al. (2013) illustrate an example; Suppose critical predictor had an importance of X. If another predictor is just as critical but is almost perfectly correlated as the first, the importance of these two predictors will be roughly X/2.

Genuer et al. (2008) state, that variable importance assessment is concluded for two main reason;

- To find important variables highly related to the response variable for interpretation purpose,
- to find a small number of variables sufficient to a good prediction of the response variable.

Dıýaz-Uriarte and De Andres (2006) suggests a method based on recursive eliminations of predictors. Firstly, RF variable importance is calculated. Secondly, certain number (authors used 20 %) of predictors is eliminated. Then the process is repeated. Different approach proposed Genuer et al. (2008). Procedure consist of two step. Compute the random forest variable importance and eliminate m least important variables. Remaining variables order in decreasing order of importance. For interpretation purpose build nested random forest model which include n variables. Then select variables, which lead to smallest error. For prediction build ascending sequence of RF models by invoking and testing the variables stepwise. The variables of the last model are selected.

The RF algorithm achieves one of the highest predictive accuracy compared to other algorithms for broad field of tasks in behavioural, social, and biomedical sciences (Berk, 2020). One of the advantage of RF is great performance

for high dimensional data, when amount of predictors is higher than amount of observation. This is especially useful in field microbiology and genetics (see Dıýaz-Uriarte and De Andres (2006)). RF rarely overfit; deep trees may cause slight overfiting, that can be easily resolved (Hastie et al., 2009). Another reason to choose RF is great computational performance, which is native to all tree based algorithm. Compared to bagging, "RF is more computationally efficient on a tree-by-tree basis since the tree building process only needs to evaluate a fraction of the original predictors at each split" (Kuhn, Johnson, et al., 2013). RF can be running simultaneously on more machines and results can be aggregated afterwards (Liaw, Wiener, et al., 2002).

3.2.3 Spatial extension

Machine learning algorithms mentioned so far might be described as aspatial. This may be a problem for a dataset containing observations, which are localized in space. The same feature or same value of an attribute has a different or opposite influence on the dependent variable. For example prediction of residential units in Prague; the age of houses might have a different impact depending on location. Old houses close to the historical center will have a higher price because were built as a spacious residence for wealthy elites. On the other hand, old houses in the outskirts of Prague will tend to have a lower price because were built as residences for workers in agriculture. These two opposite influences on price in different parts of Prague may negate or reverse the final effect. Reversal of results when observations are analyzed separately and together is called Simpson's paradox Simpson, 1951.

The variety of influence on different places is called spatial non-stationarity (Fotheringham, Charlton, et al., 1996). Fotheringham, Brunsdon, et al. (2003) lists tree reasons which cause spatial non-stationarity. Firstly, observations in different places are not the same. There is sampling variation. Secondly, some relationships are intrinsically different across space, which is especially true for social processes. And lastly, there is an option that one or more important variables are missing from the model. Inaccuracy which is caused by reasons behind spatial non-stationarity might be reduced by using spatially sensitive statistical models.

The problem of aspatial statistical models is well known and many authors contributed to the solution by creating new spatial models or extending established statistical models. Examples are spatial expansion method (Casetti and Jones III, 1991) spatial regression models, namely kriging (Krige, 1966), which is a geostatistical method. One of the more prominent methods is GWR developed by Fotheringham, Brunsdon, et al. (2003). GWR works the same as regular regression but each observation is weighted by its distance to the regression point. GWR is written as;

$$\hat{y}_i = \beta_0 (u_i, v_i) + \sum_n \beta_n (u_i, v_i) x_{in} + \epsilon_i$$

where (u_i, v_i) denotes the coordinates of the ith point in space.

The space in which the points are included in regression is called a kernel. The size of the kernel or the maximal distance is bandwidth. Kernel size can either fixed or adaptive. The distance of fixed kernels remains the same for each regression points, which is adequate when observations are in a regular grid, but insufficient when observations are localized irregularly. This is often the case. Adaptive kernel defines the size of bandwidth for each regression point by a predefined amount of closest points. This is useful when regression points are located in the periphery, where tends to be fewer observations.

It is important to calibrate the appropriate bandwidth. The bigger bandwidth or more observations included lower the standard error and variance. Contrary, bigger bandwidth will introduce bias. There are more options to create an adaptive kernel with varying bandwidth. Firstly, choose n nearest neighbours and apply the desired function, for example, exponential. The second option is to rank observations by their distance; the closest point weights 1, the second weights 2 etc. Lastly, the sum of weights will be a predefined constant for every regression point.

Another approach is to include spatial covariates. Spatial information is then treated same as other attributes. Hengl et al. (2018) lists some of the most used;

1. *Geographical coordinates.* Easting and northing.
2. *Euclidean distances.* Distance to reference points or area. For example, distance to the center and edges of the study area or distance to the closest ocean.
3. *Euclidean distances to sampling locations.* Distances from observed locations.
4. *Downslope distances.* Distances within a watershed.
5. *Resistance distances or weighted buffer distances.*

The most widely used and simplest are geographical coordinates. They are not ideal for most algorithms, because the prediction ignores spatial pattern appearing at the sample point (Ahn et al., 2020). However, algorithms based on decision trees such as RF, are naturally good at handling coordinates in geographic datasets, because latitude and longitude might be chosen during splitting process, thus leaf node is within the area defined by its ancestral nodes (Deng et al., 2020), even if tends to create linear boundaries (Hengl et al., 2018).

Adding additional distance covariates (distance to corner or center) may increase accuracy as in the study from Behrens et al. (2018) in which coordinates achieved less predictive power than distances to corners and center of the map. Together, coordinates and distances obtained the best results. Calculating distances to each observations might solve problems of coordination

as covariates. With few observations performance of such a model tend to be low, contrary with more observations computational requirements increase steeply (Ahn et al., 2020). Spatial covariates are a necessity in the cross-validation process. In case of RF, coordinates may contribute to overfitting of model and predictor should be chosen according to its importance (Meyer et al., 2019).

The concept of GWR applied to the realm of random forest was introduced by Georganos et al. (2019). For each regression point, a local RF model is computed, which includes a subset of nearest observations. Contrary to GWR, a global model is also computed and the result is a weighted average of the global and local model. Similarly to GRF adaptive and fixed kernel might be used. The main benefit is that GRF can retain low bias spatial non-stationarity and at the same time capture a low variance global model. Bandwidth for local model and weight should be treated as a hyperparameter.

Unfortunately, compare to RF, computational complexity of GRF is higher. RF is in general fast algorithm with a computational complexity expressed by following equation (Hassine et al., 2019);

$$O\left(n \log\left(n\right)\right)$$

where n is the number of training samples. Complexity is dependent on other parameters such as number of trees, or number of features. Here, for simplicity are these parameters omitted. On the other, hand for GRF is complexity much higher because of computation of local models and can be expressed by equation;

$$O\left(n \log\left(n\right) + n\left(bn \log\left(bn\right)\right)\right) \rightarrow O\left(n^2 \log\left(n\right)\right)$$

where b is ratio of samples in bandwidth. Lower bandwidth reduce computational complexity. For example, bandwidth of one tenth of study area encompasses only one hundredth of all samples. GRF for big data might be impractical or with a regular desktop PC impossible.

3.2.4 Random forest tuning

Process of building machine learning model involves setting up number of parameters, which affects stability and performance. In this subsections, procedure of tuning RF algorithms will be evaluated including description of parameters and their influence on stability, accuracy and variable importance score and tuning strategies.

RF contains various parameters which can be tweak and tuned. Here, most influential will be described;

- Number of randomly selected features (max_features): Parameter is defined as a size of pool from which are randomly drawn candidates when growing a tree. Typical value for regression is $nfeatures/3$.

Lower values of parameter result in more stable trees, but lower accuracy. Lower value of parameter induce greater randomness to splitting process, which will makes trees more diverse and potentially more informative. However, a too strong randomization produces trees that do not suit problem enough - high bias. On the other hand, not enough diverse trees will overfit the data - high variance (Bernard et al., 2009). Parameter should be set high in case, that dataset contains only small number of influential features or all features have same information value. Computational complexity decrease with lower parameter value.

- Number of trees (n_estimators): Amount of trees is not a typical tuning parameter. According to Probst and Boulesteix (2017) tuning is unnecessary during classification, in case of regression more trees result in lower error with diminishing returns. Oshiro et al. (2012) do not advise larger values trees as it only increase computational cost, which grow linearly. Number of trees possible depends on other parameters, which create less correlated, more diverse trees may require higher number of trees which will offset their negative effect.

- Sample size (max_samples): Sample size define number of sample to be selected from full dataset for training each tree. Lower sample sizes result in less correlated trees and better prediction accuracy and lower stability. It is similar to max_features parameter. According to case study of Martıýnez-Muñoz and Suárez (2010) parameter is data dependent and can be tuned. Lower sample size also lower computational cost, which is always advantageous.

- Node size (min_samples_split): Parameter determines sample size in a terminal node. Lower values will allow deeper trees lowering bias and increasing variance. Parameter can be tuned and rising node size can increase accuracy especially in low-dimensional and large sample size case (Lin and Jeon, 2006). Limiting tree depth by node may significantly improve computational cost, which decrease exponentially with increasing node size (Probst, Wright, et al., 2019).

- Tree depth, leaf node size, impurity decrease (max_depth, min_samples_leaf, min_impurity_decrease): Similarly to node size parameters controls tree depth by different means and should be treated as such.

- Splitting rule (criterion): During splitting process rule, which minimise either Root Mean Squared Error (RMSE) (weighted variance) or Mean Absolute Error (MAE) is used.

Traditional approach to tuning machine learning algorithms is to use k-fold cross-validation. This strategy randomly divides training data into k parts of equal size. The model is trained on $k-1$ parts and then tested on the last one. Special case of cross-validation is Leave-One-Out Cross-Validation (LOO CV) when only one sample is used for validation and the remaining samples are used for training. Obviously, this strategy is usable only with very small number of samples. K-fold strategy needs to be executed only k

times, typically three, five or ten. Additional advantage is higher accuracy. LOO CV values achieve very low values of bias, on the other hand very high values of variance because model is trained on lot of almost identical observation, resulting in correlated results (James et al., 2013). Compare to other algorithms RF can utilise OOB samples feature to estimate error during training. Each tree is built on a bootstrapped sample, approximately one third of OOB samples can be used to test error. Advantage over k-fold cross-validation is time. Model does not need to be trained reapeatably k but only once. OOB method compare to cross-validation underestimates error, however the difference is very small and both methods are effective in hyperparameter tuning (Ljumović and Klar, 2015).

Several methods for identification the best parameters has been proposed. The simplest is to manually identify parameters. This process is only applicable in models with small number of parameters. Manual selection is fast, provides insight into parameters influence and has no computational barriers. However, results of manual search are hard to reproduce (Bergstra and Bengio, 2012). Computationally most demanding option is grid search. Method is evaluating every possible combination of parameters, which might be computationally infeasible. Search can be easily parallelized (Petro Liashchynskyi and Pavlo Liashchynskyi, 2019) and usually achieves best possible results. Third options is random search, which in contrast with grid search evaluate randomly chosen values of parameters. Number of selection is input parameter of grid and can be adjusted by availability of resources. Lower computational demand is main advantage of this approach. In case of accuracy, random search can reach grid search if features, which contributes to performance change are known (Bergstra and Bengio, 2012). Besides manual, grid and random search other less conventional methods for parameter tuning such as genetic algorithms (Loussaief and Abdelkrim, 2018). This methods imitates darwinian selection by creating artificial chromosomes, which compete between each other and subsequently evolve. Surviving elements represent optimal values. Discussed search methods consider only one metric for selecting optimal parameter - accuracy. Besides accuracy, stability and computational complexity can be introduced as performance metric for parameter tuning (C. B. Liu et al., 2017).

Even without tuning RF algorithm performs well with default settings. Compare to other algorithms such as SVM, tuning contribution to performance is very small (Probst, Boulesteix, and Bischl, 2019). Van Rijn and Hutter (2018) found across several testing dataset, that parameters leaf node size and max features contributed most to change in performance. Even if leaf node size, tree depth and node size controls similar characteristic of tree, leaf node size is much more advantageous then latter parameters.

RF tuning is exhausting and time consuming process. This is especially true in case of GRF tuning. One of the solution is to use sampled dataset to tune hyperparameters. However, results from sampled dataset might not fit complete dataset. DeCastro-Garcıýa et al. (2019) proved that, third of all combination between parameters is statistically different in sampled dataset.

Although, size of the dataset does not impact accuracy in critical way.

3.2.5 Sampling

It is often the case, that researchers are limited by time or resources and cannot access complete dataset or research full population. For example, most visible cases are opinion polls, which receive information from small reference sample. The process of selecting reference sample is know as sampling. In general, sampling methods can be classified into non-probability and probability sampling. In non-probability sampling researcher sample data based on his subjective judgement and it is often used in qualitative studies. In probability it is known the chance of sample being selected into sampled dataset. In this sections, tree methods for probability sampling will be discussed - random sampling, stratified sampling and Latin Hypercube Sampling (LHS).

- *Random sampling* can be described as "sampling design in which n distinct units are selected from dataset in such a way that every possible combination of n units is equally likely to be the sample selected." (Thompson, 2012). Obvious advantage of random sampling is ease of assembling sampled data. Random sampling is also representative of original dataset if all observation are available to sampling. Only luck can influence the sampled dataset, which results in sampling error (Sharma, 2017).
- *Stratified sampling* divides datasets into N subsets called stratas. Selection of samples from subset is independent. Compare to random sampling, information about dataset are needed to divide data into stratas. True stratified sampling divides dataset along all dimension, which is impractical for higher dimensional dataset. Stratified sampling can be divided in disproportionate and proportionate sampling. In former, the number of samples selected from each subset is not proportional to their share in total population. In latter, selected samples are proportional.
- *LHS* developed by McKay et al. (1979) operates by dividing N-dimensional space intro stratas. Compare to true stratified sampling LHS might be viewed on a opposite spectrum possible stratification methods. LHS stratifies each dimensions individually and true stratified sampling stratifies space simultaneously (Shields and J. Zhang, 2016). LHS performs much better in high dimensional space.

Generalization of dataset from a random or stratified sampling required appropriate size of sample to avoid high bias and sampling error. At the same time sample has to be small enough to keep processing time as low as possible. Several guides focus to random sampling, almost no LHS. In general, LHS achieve better performance with smaller sample size (Matala, 2008). Swidzinski and Chang (2000) in their study achieved same variability with

200 samples selected by LHS and with 1000 samples selected by random sampling.

3.2.6 Performance metrics

Performance metrics are essential element in building, tuning and evaluating machine learning model. Metric quantify prediction results, and simplifying them to one, easily readable value. Many various metric exists and would be inappropriate to describe every one them. Thus, only the most popular will be discussed - RMSE and MAE.

One of the most used metric is RMSE. Metric is the square root of mean of the average of squared errors. RMSE is disproportionately impacted more by larger values of error, thus penalising outliers. The formula is expressed as;

$$\mathrm{RMSE} = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}}$$

Similar metric, Mean Squared Error (MSE) is often used. Compare to RMSE is less readable, because of calculated values outside the range of prediction values. RMSE and MSE is often criticised for its counterintuitiveness and ambiguity. Squared errors can not be meaningfully compared to other squared errors, because of unknown variability of the errors (Willmott et al., 2009). However, Chai and Draxler (2014) argues, that if errors distribution is normal rather than uniform, RMSE is better metric than non-squared metrics.

Most popular non-squared error is MAE, which can be calculated using formula below;

$$\mathrm{MAE} = \frac{1}{n} \sum_{i=0}^{n-1} |y_i - \hat{y}_i|$$

MAE is simple mean of errors, which does not penalise outliers. RMSE and MAE have informative value, thus it is advisable to use both. Another popular metric is Mean Absolute Percentage Error (MAPE) calculates as below;

$$\mathrm{MAPE} = \frac{100}{n} \sum_{i=1}^{n} \left| \frac{\hat{y}_i - y_i}{\hat{y}_i} \right|$$

MAPE has major drawbacks, for instance can not be used if actual value is zero. Also, it is known that MAPE overstates errors (Tayman and Swanson, 1999).

3.3 SUMMARY

The chapter contains theoretical information for two distinct topics - drought and machine learning. The first part is dedicated to drought definition and identification. The precise identification is important as it is the phenomenon that is modelled. It is required that drought is present to the desired extent. Next, drought is quantified by the drought indicator. Several indicators were described and their positive and negative aspect was discussed. As was stressed there is not a universal best indicator as climatic conditions in which was indicator created differs across the world. Another important aspect is the availability of data. More complex, experimental or recently developed indicators are often not obtainable. Based on those conditions indicators derived from soil moisture are most suitable. Soil moisture influences the plant's well-being and its deficit cause crop failure. Compare to SPI, soil moisture is impacted by other ground conditions, for example, soil properties, from which vulnerability can be assessed. In addition, it is available within the Copernicus project.

In the next subsection, the frameworks within which would be vulnerability understood are described. The most appropriate framework is the Hazard and Risk model from Office for Disaster Risk Reduction (2004). However, the framework does not distinguish environmental vulnerability, which had to be defined. Most vulnerability assessments are carried out by subjective weighting of factors, for example, method AHP. Machine learning techniques eliminate this negative aspect. The thesis methodology further developed the machine learning approach. The next subsection describes several categories of factors that affect vulnerability. As RF algorithms can handle noisy data (features without importance) the more available feature the better. Limitation remains in the availability of quality data.

The second part is focused on machine learning. The introduction contains a definition of machine learning and a description of the two main goals of the statistical model. Both prediction and explanation purposes are utilised in the thesis. The main tool, RF is described as an evolution of the basic decision tree model. Decision trees models tend to overfit because trees grow until predictor space is divided by all features. RF solves this by incorporating bagging and random selection from a subset of features. RF variable importances can be assessed by two methods; random permutations and Gini impurity. The latter is used more frequently, which makes a comparison to other studies easier and is more accessible from a software standpoint. If categorical features are not used, Gini impurity is the preferred method.

There are two possible techniques to make the model able to perform better with spatial data. The easier option is to include spatial covariates into features. RF is known to be good with handling coordinates, which makes them better options than other possible geographical covariates. The second option is to give each observation weight based on their distance to predicted locations. This is the of GWR or GRF. These techniques are not exclusive, both should be applied in methodology. GRF is known to be computation-

ally demanding and some concessions have to be made. A lower number of observations, better hardware or more effective coding (utilisation of all core of CPU).

Tuning of the model is important as it increases the accuracy of prediction. RF and GRF have several parameters which can be tuned. The most important parameter is the number of randomly selected features and node size, which can be controlled by other parameters such as leaf node size. The number of trees is not a traditional parameter because with increasing value the error only decreases. However, a higher number of trees increase the runtime. The GRF parameters bandwidth and local weight are not explored in literature as much as parameters of RF. There are two possible options when choosing a tuning strategy; either choose all possible parameters and use random search or manually select the most important parameters and tune them with grid search. The latter option is better as many parameters are connected or unimportant.

Three sampling methods were described - Random sampling, stratified sampling and LHS. The last one tends to be the best option as it sample observations evenly. However, all methods should be evaluated empirically and the best option used to sample data to test and train set. Lastly, three evaluation metrics are described; RMSE, MAE and MAPE.

4 METHODOLOGY

This chapter describes the methodology used to address the research question stated in the first chapter. The proposed methodology is based on theoretical concepts reviewed in previous chapters. Three stages are identified; pre-processing of data, parameter tuning and evaluation of algorithms, and lastly vulnerability assessment. The workflow of methodology is depicted in the scheme below.

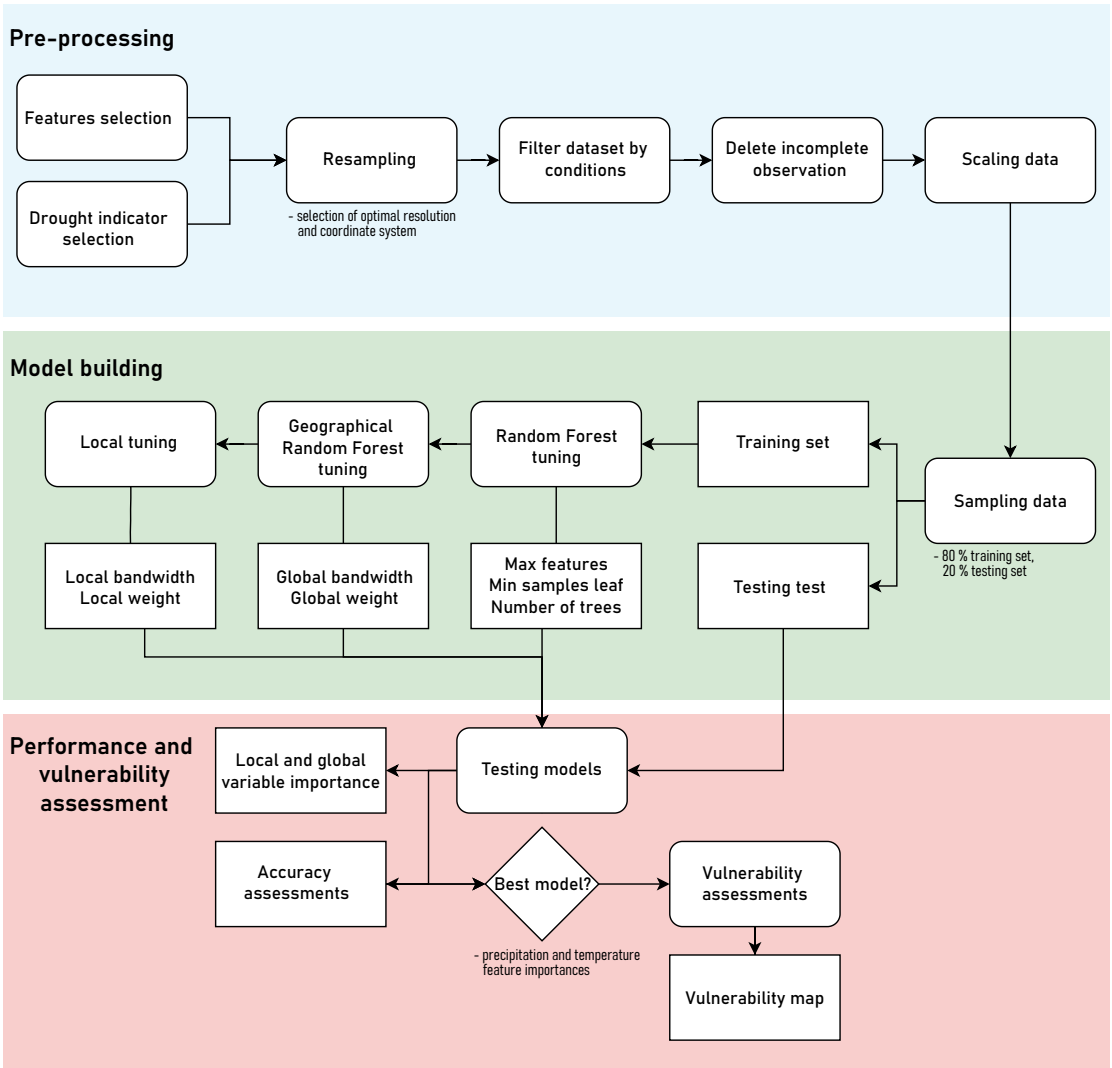


Figure 1: Methodology workflow.

4.1 PRE-PROCESSING

4.1.1 Data selection and preparation

The first step is to select a drought indicator, which can sufficiently quantify drought. Various indicators and their advantages and disadvantages has been described in previous chapter (Section 3.1.1). Drought indexes or indicators are published in a time dimension most often hours or days. Therefore, the period needs to be selected and aggregated in the next step. For example, Rahmati, Panahi, et al. (2020) use a relative difference of period mean to annual mean of soil moisture, selected period to correspond to drought period. It is important, that the drought index has a wide range of values, to make the model more versatile. Therefore, a month with drought partially present should be chosen. However, no study dedicated to drought period selection exists.

SWI was chosen as drought indicator for several reason; SWI can reasonably represent soil moisture condition in soil layers, which is good proxy for crop condition. It is available in sufficient spatial and temporal resolution and soil moisture based indicators has been successfully used in similar studies, for example Rahmati, Panahi, et al. (2020) and Rahmati, Panahi, et al. (2020). As a representative time period august of 2018 was chosen. According to Masante and Vogt (2018), Central Europe was affected by drought. CDI identified medium and high risk for parts of Czechia, Poland and Germany. SPI and fAPAR indicate severe drought in Elbe valley.

The timespan of an indicator is linked to the timespan of meteorological features. Features influence on indicator values is time-dependent. Therefore, the timespan of features should be similar to the indicator, but not identical. The selection of time depends on the type of indicator. Vegetation indexes are affected by precipitation two to three months in advance (Méndez-Barroso et al., 2009). Soil moisture is most affected in the period of one day to two weeks in advance. After two weeks precipitation influence weakens (Lauzon et al., 2004). However, coherence exists between earlier periods of precipitation and deeper soil depth.

The selection of features is based on proved links to drought severity described in the theoretical part (Section 3.1.3). The selection of continuous data is emphasised, because of their suitability in analysis. Discrete data needs to be converted to continuous data, which cause inaccuracies or translate to several features, which inflates feature space. The selection process is also influenced by the availability of data. Selected features are listed in table below.

Table 1: Selected features and their category.

Category	Feature	Category	Feature
Terrain	Elevation	Spatial	X-coordination
	Slope		Y-coordination
	Aspect	Land Cover	Built-up area
	TWI		Forest
Soil	Organic matter		Agricultural area
	AWC		Grassland
	Bulk density	Meteo	Precipitation
	Coarse particles		Temperature
	Clay content	Other	Water bodies proximity
	Sand content		

First category contains elevation, slope, aspect and TWI. Elevation is plain value from Digital Elevation Model, slope and aspect were calculated in degrees. Values of aspect are circular (value 0° and 360° are the same), thus two values are needed to describe feature. This would make feature less interpretable and highly correlated in variable importance assessment. This is less desirable than a small loss of accuracy, single value feature was selected. TWI was calculated using MFD-md method.

Category soil contains organic matter content, AWC, bulk density, coarse particles and sand content. Soil category content choice is affected by availability of suitable data sources. Clay and sand content is expressed as share of total soil matter. Silt content was not included as it is negatively correlated with sum of two other shares and would not be contribute to model fitting as much. Similarly to previous, coarse fragments values are share of total soil mass. Bulk density, AWC and organic matter is provided as mass over volume.

As spatial covariates plain coordinates were selected, because of their simplicity and readability. Water bodies proximity express distance to closest water body. Places with zero value are bordering water.

Land cover distinguish several classes; built-up area, forest, agricultural area and grassland. Features built-up area and forest are used later in filtering inappropriate regions. Features represent share of given land cover. From wide variety of land covers were disqualified wetlands because of their small representation in study area and water bodies, which can not be assessed by drought indicators.

From meteorological category temperature and precipitation is selected. Period is selected in considering timespan and type of drought indicator. From soil moisture derived SWI has timespan of one month. Optimal period is timespan of indicator plus two weeks. Same period is selected for both meteorological features - 16.07.2018 to 31.08.2018. Temperature is average along time dimension while precipitation is summed.

Spatial resolution and coordinate system differ across datasets, thus uniform parameters need to be chosen. As most viable coordinate reference system ETRS89-LAEA[1] was selected for several reasons. The system is suitable for mapping with regions within Europe, mapping units are in meters, which is readable and lastly, many available datasets for Europe are in ETRS89-LAEA. Spatial resolution is determined concerning several aspects. High spatial resolution increases the number of samples, which raises runtime. On the other hand, if the spatial resolution is low, properties of places are too generalized and spatial patterns diminish. Secondly, the spatial resolution should be chosen to be close to the resolution of available datasets to prevent distortions during interpolation. In order to balance both condition resolution of 1 km selected.

In case that datasets are in different resolutions, they are interpolated to desired resolution by bilinear interpolation. Similarly, missing values for independent variables are interpolated by bilinear interpolation. Compare to the nearest neighbour method, bilinear interpolation takes into account other near values. Other methods, which are superior in accuracy such as kriging are too complex and not as available as former methods. Lastly, missing observations for the dependent variable are dropped. Missing values are mostly localized in metropolitan areas and mountains.

Cleaned data are filtered by two following conditions;

1. Forest should not cover more than 20 % of the pixel area.
2. Built-up are should not cover more than 20 % of the pixel area.

These conditions prevent the analysis of forest and metropolitan areas, which are not suffering from agricultural drought in the same way as agricultural areas or grasslands. The inclusion of such areas would distort the model, which would prioritise land cover classification. Variable importance would be skewed towards land cover classes if land cover features are included.

Lastly, all features are scaled, to provide more accurate results. Variables are measured in different scales, which might not contribute evenly to the model fitting and create a bias. Therefore, it is useful to transform features into the same range. Compare to standardization scaling does not change the distribution of data. Transformation is expressed by the following formula;

$$x_{sc} = \frac{x - min(x)}{max(x) - min(x)}$$

4.1.2 Data exploration

An explanatory analysis is the process of investigation of data structure using mainly visualization methods. The primary purpose of visualization is to make data and the corresponding phenomena perceptible to the mind

1 European Terrestrial Reference System 1989 on Lambert Azimuthal Equal Area projection. EPSG: 3035

or imagination (N. Andrienko and G. Andrienko, 2006). According to Telea (2014) visualization provides insight about data and studied phenomena and helps answer concrete questions about a given problem. Firstly, insight may provide information, which is important to the ensuing analysis. For example, data visualization shows unnatural patterns in one of the features, that indicates damaged or fake data. In the second case, data visualization may provide answers to specific questions, which results from the assessment. Data can be visualized and explored in a variety of ways.

The first part of data exploration investigates the correlation between all variables and correlation between dependent and independent variables in detail. A high correlation between features may result in their lower variable importance. A high correlation between drought indicators and features may indicate linear relation. In the second part relation between individual features and indicators is explored and described for purpose of detecting possible functions.

4.1.3 Sampling

As will be describes in next section dataset is split into two parts - train and test set. The splitting process is achieved by using one of three sampling methods - random, stratified and LHS. The performance of each methods is rated by difference of sampled set to original dataset. As evaluation metrics were used differences of the mean and standard deviation between original and sampled dataset. The smaller the difference is, the closer the sampled dataset is to the original. It is assumed that a similar dataset will select parameters close to optimal values.

4.2 MODEL BUILDING

4.2.1 Random Forest

To achieve the best possible accuracy it is necessary to tune parameters. For regular RF, three parameters were tuned - max features, min samples leaf and number of trees. No additional parameters were considered because of their minuscule benefits to performance or redundancy. Parameter min samples leaf encompass several others parameters such as tree depth, leaf node size or min samples split. The number of trees parameter was tuned independently, as it is not a classical parameter. It is needed to identify a break value, the number of trees below makes the model unstable and values above increase runtime with diminishing returns.

To tune parameters, the Grid-search method was utilised, which is viable because of the low dimensionality of parameters (only two). Dataset was divided into two parts; training and testing set in 80/20 ratio. The best

sampling method is used. Model is tuned on training set evaluating OOB samples. RMSE and MAE metrics are calculated to find optimal parameters.

4.2.2 Geographical Random Forest

After tuning regular RF, the next step is to train GRF. Optimal parameters from regular RF are employed in GRF. Tuning parameter for each local model would be computationally infeasible. It is assumed, that optimal parameters for global model will suit local models as well. Among variants, GRF with an adaptive kernel was chosen. Data points are not in a regular grid and a fixed kernel would results in an uneven number of samples for each local model. On the other hand, for models created in isolated places without close neighbours the average distance to samples will be higher than in more populated regions. Bandwidth is considered as a tuning parameter, the second parameter local weight can be tuned without repetitive training and testing of the model.

The first step is to calculate the distances of each point to each other point, which will be saved in the matrix of size (n, n). For larger sizes of samples, more space needs to be allocated. For example, for a sample size of 70 000, approximately 36 GB of memory space in float64 data type. Therefore, data were converted to int16 format, which requires less memory. Secondly, the weight of b (according to bandwidth) closest samples are set to 1, other samples are assigned 0.

Next, the algorithm loop over locations and for each creates and trains an RF model. Samples with zero weight are not considered in the local model. After local models are trained, one global model is created. Lastly, prediction from both local and global will be weighted by values from 0.1 to 1 and summed. From summed prediction, RMSE and MAE are calculated. Best weight and bandwidth is selected from the lowest value of RMSE.

Compare to RF, the tuning process of GRF is different. It is not possible to utilise OOB for evaluation of accuracy, because observation for a location to predict is missing. Similarly, k-fold Cross-validation for a small subset would be pointless. However, in some sense tuning, GRF can be seen as LOO CV because for every observation, the model is trained (only on a portion of data) where this observation is missing. The tuning process consists of repeating training and predicting for each sample for several bandwidths. Not for all observations are predicted values. Observations outside the bounding box are used only for training.

4.2.3 Geographical Random Forest with local tuning

Lastly, a new variant of spatial RF was developed. LT GRF further expand the idea of Georganos et al. (2019). In the GRF case, bandwidth and local

weight are universal for all samples. However, in LT GRF for each sample is found best possible bandwidth and local weight.

Tuning local bandwidth does not require more runtime than GRF. Similarly to GRF, the model is repeated for each value of bandwidth. Local weight is estimated from prediction after training time, therefore does not requires training at all.

4.3 PERFORMANCE AND VULNERABILITY ASSESSMENT

4.3.1 Accuracy assessment

After finding optimal parameters, each model is tested on the testing dataset. Four variants of RF algorithms were evaluated. RF with spatial coordinates model (RF_XY model) is trained on a training set. After training, the model is fit to testing data and new values are predicted. These values are compared to real values and RMSE and MAE are calculated from differences. Relative values of metrics are calculated for better interpretability. Secondly, RF without spatial covariates (RF model) is tested. Benefit of including spatial covariates is assessed as difference in error. GRF is tested with the same parameters as regular RF and with bandwidth and global weight find during the tuning process. Local parameters for GRF were found only for the training set and must be interpolated for the test set. Inverse distance interpolation was used. It is unknown if the type of interpolation significantly affects the performance of the model. Lastly, the LT GRF model with a global weight and local bandwidths is tested. Testing of separate LT GRF is executed to found if the model can be improved significantly with local weighting.

4.3.2 Variable importance

The process of extracting variable importance from the model is described in the previous chapter (Section 3.2.2). From two possible options Gini Impurity was chosen for the following reasons; compare to random permutations, the chosen method does not overestimate the importance of correlated predictors and is widely used and available. Random permutations options might be especially unstable with low bandwidth parameters in GRF.

Variable importance is computed from the RF model once for all observations. In GRF and LT GRF variable importance is computed from all local models, which is n times more. To make these results readable, values are aggregated and compared to the global model. Values are visualized for each location on a map.

4.3.3 Vulnerability assessment

Lastly, a vulnerability map is produced from the best geographical model (GRF or LT GRF) in terms of prediction power. Meteorological features importance (temperature and precipitation) are summed. This value would be relative indicator a drought hazard. It is assumed, that high importance of temporal features is indicative of low resilience to extreme events such as precipitation deficit or heat waves. However, this approach does not take into account actual precipitation and temperature condition, which varies across space. If meteorological features are important and at the same time precipitation is adequate, region is not vulnerable to drought.

The uneven distribution of meteorological variables is compensated by weighting the feature importance by relative precipitation and temperature. Relative value v_{rel} for each location is calculated by;

$$v_{rel} = \frac{v_m - v_p}{v_m} \times 100$$

where v_i is long-term (2012 - 2017) mean value for selected period for all study area and v_p is actual value for each locations. Relative values are scaled, averaged (weighted mean based on importances) and multiplied by importance.

Due to filtering of dataset by land cover condition, data grid contains gaps. Intermittent grid is difficult to interpret, therefore a new hexagon grid was created with radius of 3 km. Values within each cell are averaged.

4.4 SUMMARY

The chapter describes individual steps from selecting and pre-processing data to model building and vulnerability assessment. In the first stage, features are selected based on theoretical links to drought hazard. Selection of drought indicators was more difficult as only a handful of research studies focus on machine learning drought assessment. Similarly, delimitation of the time period was not studied at all and therefore might be a source of inaccuracy. In the next step, all datasets are standardized and filtered by conditions. Land cover features remained preserved despite their diminished information value after filtering. If their contribution to model fitting remains significant, even a small share of given land cover can alter drought severity.

In the second stage, three models are tuned. Firstly, datasets are divided into test and train dataset. Parameter number of trees, maximum features and minimum samples leaf are tuned for regular RF and bandwidth and local weight for GRF.

Tuned models are tested on a testing set and their accuracy is evaluated by RMSE and MAE metrics. Secondly, variable importance output is inter-

preted and compared to global importance. Lastly, a vulnerability assessment from a best geographical model is produced. Meteorological features importances from model are summed and visualized.

5 DATA, TOOLS AND STUDY AREA

This chapter contains a description of datasets sources, study area and tools used in the thesis. The chapter aims to provide an evaluation of available data sources in terms of their accuracy, spatial and temporal resolution and other qualities. Choice of datasets is important as it was one of the prerequisites to feature selection. The next subsection determines the study area. Lastly, tools used in the thesis were listed.

5.1 DATASETS

Drought indicator - SWI is available in sufficient spatial resolution within Copernicus Global Land Service. SCATSAR-SWI is computed from data fuse of products Sentinel-1 SSM and ASCAT SSM/SWI, which assess soil moisture. The first product provides high spatial resolution (1/112°), second provides high temporal (1 day). Both datasets are resampled to an independent spatial grid. The correlation layer is created by calculating Pearson correlation coefficient. Next, the weighting layer is formed from signal to noise values, which are interpreted as quality scores. The score is provided in a separate layer for each day and location. Values of two datasets are matched using the cumulative distribution function. An important parameter in the final calculation is variable T, which determines weights for observations in time. Higher T values will results in higher weights for older observations (Bauer-Marschallinger and Pfeil, 2021). Dataset is provided with various T values from 1 to 100. Values of 20 is selected correlated best with subsoil conditions (10 - 20 cm below surface) (Paulik et al., 2014) and have uniform quality score across the study area. T value was selected in consideration of other datasets, mainly soil properties.

Accuracy assessment of the product is done by comparison to in situ measurements. There is no dedicated soil moisture network for the study region, closest is WegenerNet in Austria and TERENO in Germany. However, RMSE for all stations is consistent with a value of approximately 0.04, which can be considered as accurate (Bauer-Marschallinger and Pfeil, 2021). Dataset is provided with a daily temporal resolution from the year 2015 and a spatial resolution of 1 km. It is also available with a coarser spatial resolution of 12.5 km with temporal resolution from the year 2007. Data are in WGS 84[1] coordinate system which results in higher spatial resolution in the study region (approximately 650 m). However, not all values are available. Especially in winter and spring snow cover prevents soil moisture assessment resulting in absent information. Network Common Data Form (netCDF) format was

1 World Geodetic System 1984. EPSG: 4326

used to distribute data.

European Digital Elevation Model (EU-DEM) was designated as the input dataset for elevation features and derived terrain characteristics because of availability and good accuracy. EU-DEM was created as a data fuse between SRTM[2] and Aster GDEM by weighted approach and it is administered by European Environment Agency under the framework of the Copernicus programme. The first version of EU-DEM was available in 2009. Vertical accuracy south of 60 °N is assessed to have RMSE of 2.23 metres with a mean error of -0.56 m. Should be noted that the model becomes less accurate with increasing slope and density of tree cover (Tøttrup and Sørensen, 2014). According to citeeuropean2021eudem, in the current version (EU-DEM 1.1) several improvements were achieved, for example, removal of artefacts or systematic correction of geo-positioning issues. EU-DEM v1.1 is available at 25 m spatial resolution and in 32bits GeoTIFF format. Model is provided in ETRS89-LAEA coordinate system.

All soil properties except Soil organic matter were acquired from Topsoil Physical Properties for the Europe dataset (TPPE), which is based on Land Use and Cover Area frame Statistical Survey (LUCAS) dataset. LUCAS is the largest harmonized soil dataset in Europe overseed by the Statistical Office of the European Union, which consisted of in situ measurements from more than 22 000 locations (Orgiazzi et al., 2018). The first survey was conducted in 2009. TPPE dataset consists of interpolated values from LUCAS dataset using Multivariate Adaptive Regression Splines with a normalized error between 4 % and 10 % (Ballabio et al., 2016). Besides LUCAS soil samples other environmental covariates were included in the model, for example, normalized difference vegetation index and enhanced vegetation index, CLC, climate data and soil data from European Soil Database. TPPE is delivered in a high spatial resolution of 500 m.

Another dataset derived from LUCAS is Soil Organic Matter (SOM) fractions. Cotrufo et al., 2019 utilized more than 9 400 points, to interpolate point data to a grid with 1 km spatial resolution using RF algorithm. Organic matter is divided by size into particulate and mineral-associated organic matter (less than 53 µm). Datasets are delivered in GeoTiff data format and ETRS89-LAEA coordinate system. Both datasets, TPPE and SOM are distributed by European Soil Data Centre (Panagos et al., 2012). Other datasets were considered, for example, European Soil Database v2[3] in the raster version offers more soil attributes, on the other hand, datasets are in categorical values.

Land cover information was obtained from CLC. Dataset provides a raster or vector representation of land cover classified into 44 classes. The current version was produced in the years 2017 - 2018, published in 2020 and it is the fifth iteration of the product. CLC is created by manual and semi-automatic classification of satellite data, mostly imagery from Sentinel-2. Evaluation of dataset was done by blind interpretation of classes on samples generated

2 Shuttle Radar Topography Mission
3 https://esdac.jrc.ec.europa.eu/content/european-soil-database-v2-raster-library-1kmx1km

by stratified sampling to independent imagery, for example, Bing maps or Google maps. According to product manual (Büttner et al., 2021) overall accuracy reached 93.2 % for blind analysis. CLC is delivered in raster GeoTIFF with a spatial resolution of 100 m or vector with a minimum mapping unit of 25 ha in ETRS89-LAEA coordinate system.

Features were created by aggregating existing CLC classes. Classes were selected based on general knowledge of land cover. New, generalized features and original classes are listed in the table below;

Table 2: Original and aggregated classes for land cover.

Aggregated class	**Former classes**
Built-up area	Continuous urban fabric, Discontinuous urban fabric, Industrial or commercial units, Road and rail networks and associated land, Port areas, Airports, Mineral extraction sites, Dump sites, Construction sites
Agricultural area	Non-irrigated arable land, Vineyards, Fruit trees and berry plantations, Annual crops associated with permanent crops, Complex cultivation patterns
Grassland	Pastures, Natural grasslands, Moors and heathland
Forest	Broad-leaved forest, Coniferous forest, Mixed forest, Transitional woodland-shrub

CLC was compared to Pan-European High-Resolution Layers, which contains classes similar to aggregated classes. However, the agricultural area is not included. The importance of agriculture to drought was described earlier (Section 3.1.3), therefore CLC is preferred.

Meteorological data are acquired from E-OBS dataset maintained by European Climate Assessment & Dataset project. E-OBS is interpolated from point data gathered from the national meteorological station across Europe. According to project website (Project Team ECA&D, 2021) Czechia has above average density of stations (770 km² for precipitation and 913 km² for temperature per station). Point data are interpolated with an extended linear model (generalized additive model) with an RMSE of 3.63 mm for precipitation and 1.15 °C for temperature (Cornes et al., 2018). The dataset is available in netCDF format with a 12.5 km resolution. Data were downloaded in version 23.1e, which was available in March 2021.

5.2 STUDY AREA

State boundaries of Czechia encompass the study area. Due to the nature of GRF, locations outside the country which are close to borders are part of the train set. Extent reaches approximately 50 km beyond border north to Poland, west to Germany, south to Austria and east to Slovakia. Total area

of Czechia is 78,871 km². The total number of observations with a spatial resolution of 1 km is 78 975. Condition satisfied (forest and built-up area share) 25 011 observation or 31.67 % of the country. The area and dominant land cover is shown below (Figure 2).

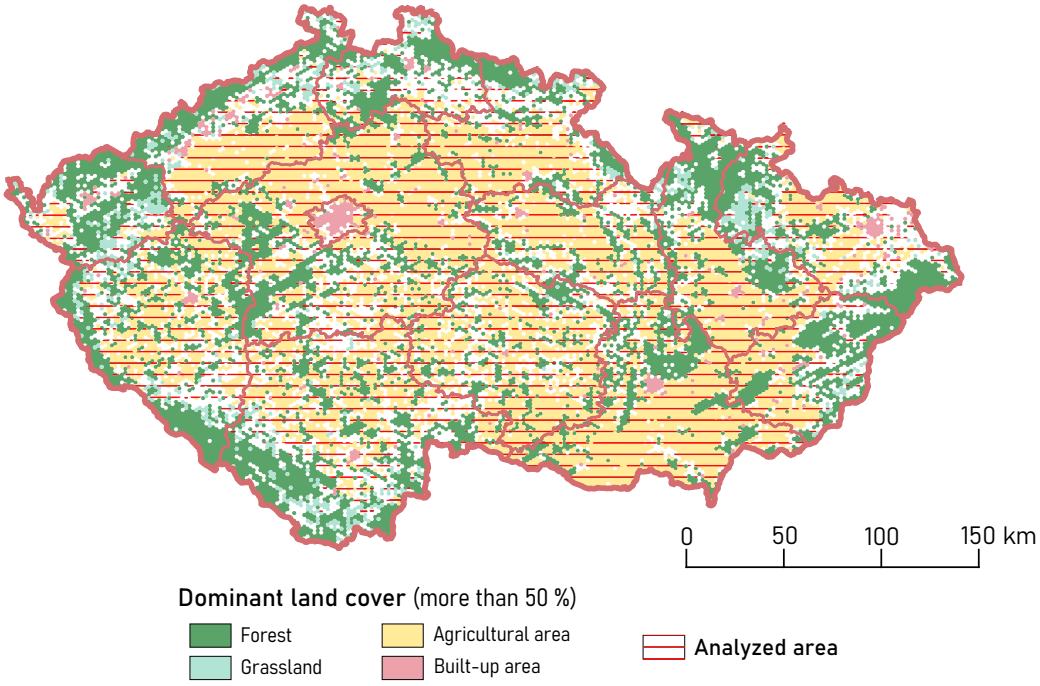


Figure 2: Dominant land cover and filtered study area.

5.3 TOOLS

Python programming language was the main tool for processing, visualization, analysis and model building. Numerous python libraries were utilized to expand python functionality namely;

- Xarray, Pandas and NumPy to make possible import and pre-processing of datasets in effective manner.
- SciPy for interpolation of data.
- Scikit-learn provides tools for machine learning model building.
- Matplotlib and Seaborn are libraries which facilities data visualization.

QGIS was used for cartographic visualization, visual exploration of spatial data and together with SAGA GIS to create terrain analysis which would be impractical in python.

5.4 SUMMARY

The chapter describes selected data, designated study area and used tools. The drought indicator dataset is in satisfactory spatial and temporal resolution. Most reviewed and selected datasets are created and maintained by institutions or projects subordinated to European Commission. All datasets are available for free. As the study area is within one Czechia, national data sources were considered. Unfortunately, at the time of gathering the data national institution have not been providing desired datasets for free. All used datasets are available for Europe, which makes thesis results more transmittable. Similarly to datasets, all tools are freely available.

6 RESULT AND DISCUSSION

The aim of the chapter is to present the results of experiments in an appropriate form, describe and discuss them. The chapter is structured similarly to methodology. Firstly, visualization of original data is presented. The focus is on the exploration of a possible relation between dependent and independent variables. Next, the choice of sampling method is described. Secondly, results of the tuning process for regular RF, GRF and LT GRF are presented. In the last section, the final results are presented. Accuracy metrics are listed for each tested model and their performance is discussed. Variable importance output for the regular and geographical model is visualized and compared. Finally, vulnerability assessment is presented and compared to other assessments.

6.1 PRE-PROCESSING

6.1.1 Exploration

SWI, the dependent variable has an approximately normal distribution. The minimal value is 10.89 and the maximal is 78.74, range of values is 67.85. If taken into account dataset without outliers (1 - 99 %) range shrinks to 47.39, with a minimum value of 17.71 and maximal value of 65.11. The range of values with a higher percentile (5 - 95 %) further shrinks to 36.14 with a maximum of 58.22 and a minimum of 22.07. Determine the range of values for the majority is important for the final assessment of accuracy, outliers might skew the perception of accuracy for the tested model.

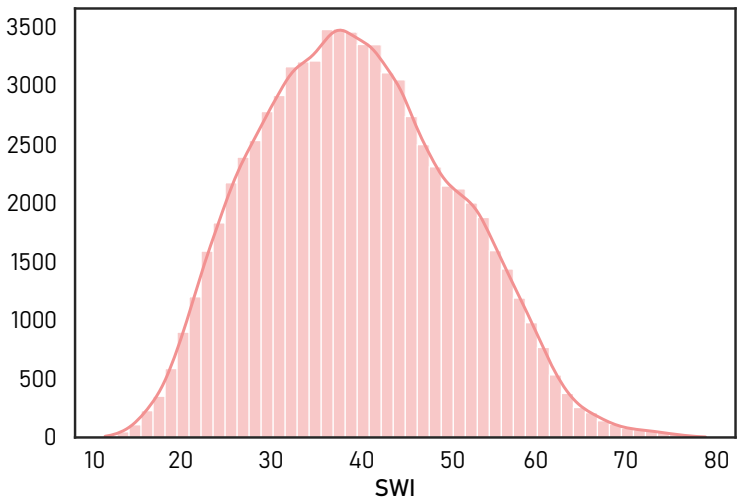


Figure 3: Histogram for SWI.

The correlation between SWI and features is not high in general. The highest positive correlation is x-coordinate (0.5), which is relatively small. The highest negative correlation is with Y-coordinate (0.4), which is even smaller. The highest positive and negative correlation is shown in the table. A small north-western gradient can be identified, SWI increase from west to east and from north to south.

Table 3: Highest positive and negative correlation between features and indicator.

Positive		**Negative**	
Feature	Value	Feature	Value
X-coordinate	0.504	Y-coordinate	-0.419
Precipitation	0.294	Coarse fragments	-0.224
Temperature	0.188	Bulk density	-0.079
Forest	0.089	Agriculture area	-0.073
Clay content	0.066	Water proximity	-0.042

High correlation (> 0.7 or < -0.7) exists between several features. TWI is highly correlated with slope (-0.8), AWC with clay content (0.93). High correlation is expected because of the known relation between them; the slope is input in the calculation of TWI and sandy soil are well to known to have lower AWC. Next, a high correlation between elevation and temperature (-0.72), and elevation and bulk density (0.73).

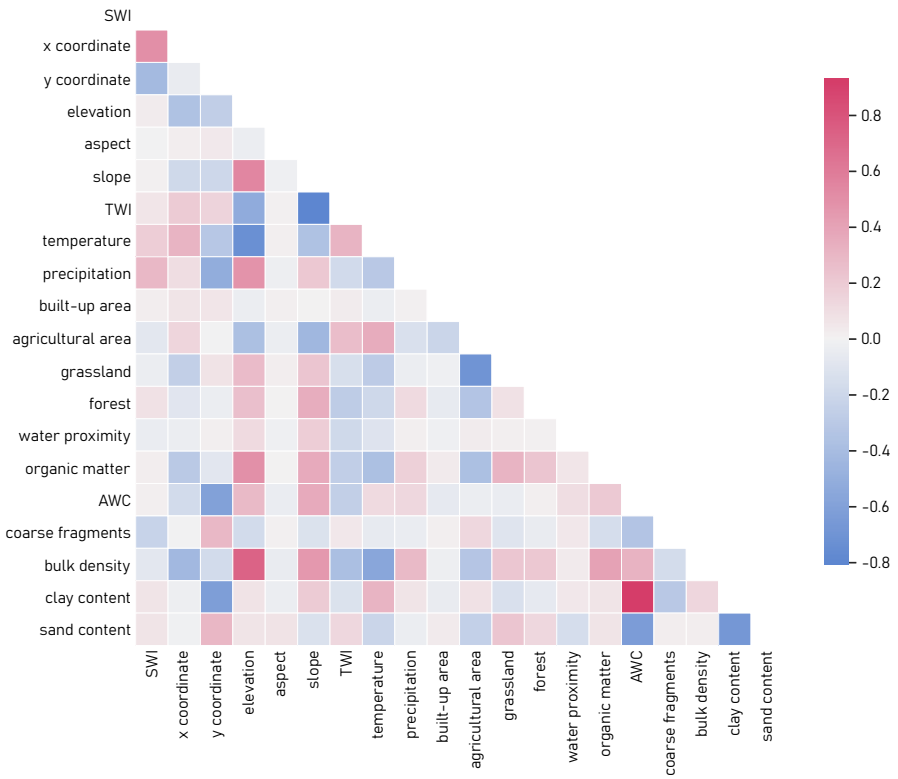


Figure 4: Correlation matrix.

6.1.2 Sampling

Sampling was repeated numerous times with different number of drawn samples. Mean difference and mean standard deviation of all features from original dataset were visualized in plot (Figure 5). Values of stratified and random sampling are oscillating around zero with maximal values of 12 and 30 for standard deviation, which is 0.12 % and 0.3 %. Values of LHS are much more higher. At 2000 samples difference of mean is more than 250, with rising sampling size value decline to 150. Similar trend reports standard deviation. Achieved values of LHS are too high to be suitable. According to theory, LHS should be most accurate in sampling from original dataset. The most probable explanation is that, the instance could not high share of sampled dataset size to original dataset size in high dimensional space. In sampling from existing dataset LHS effectively becomes true stratified sampling, which does not work in high dimensional dataset. Stratified sampling was selected as sampling method.

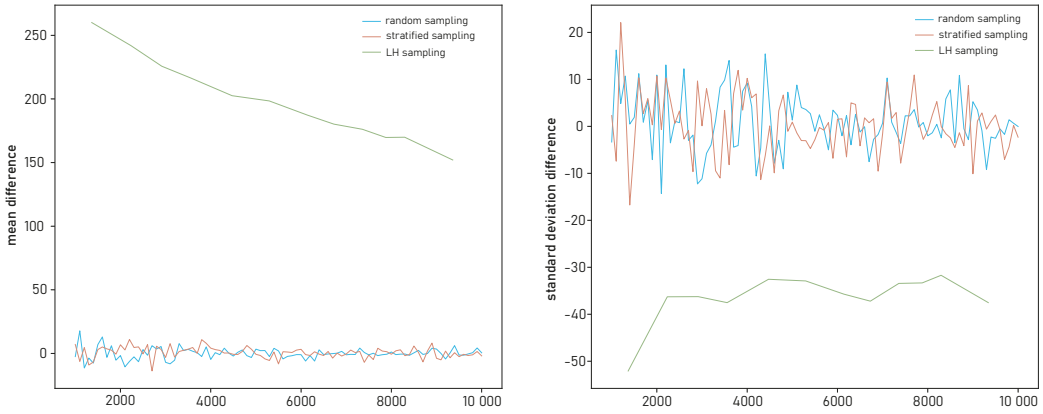


Figure 5: Mean and standard deviation difference for sampling methods.

6.2 MODEL BUILDING

The first model to be tuned is regular RF using the OOB method. Tuned (max features and min samples leaf) parameters and tested values are in the table below. Optimal values are highlighted. The difference in accuracy in the tuned and not tuned model is not substantial. The performance for parameter recommended in literature (max features = $n/3 = 6$) RMSE is 3.69, for optimal (max features = 16) is 3.51. The decrease is 0.18, which is 0.26 % or 0.38 % for range without outliers.

Table 4: Tuned parameters for RF

parameter	values
max_features	4 (sqrt), 6 (n/3), 8, 10, 12, 14, **16**
min_samples_leaf	**5**, 10, 20, 30, 40, 50, 60, 70

The decrease in error due to a higher number of trees is even more marginal. RMSE steeply decrease in interval 100 to 300 trees, from 3.545 to 3.515, then decrease with diminishing returns. A value of 300 was chosen as a breakpoint value.

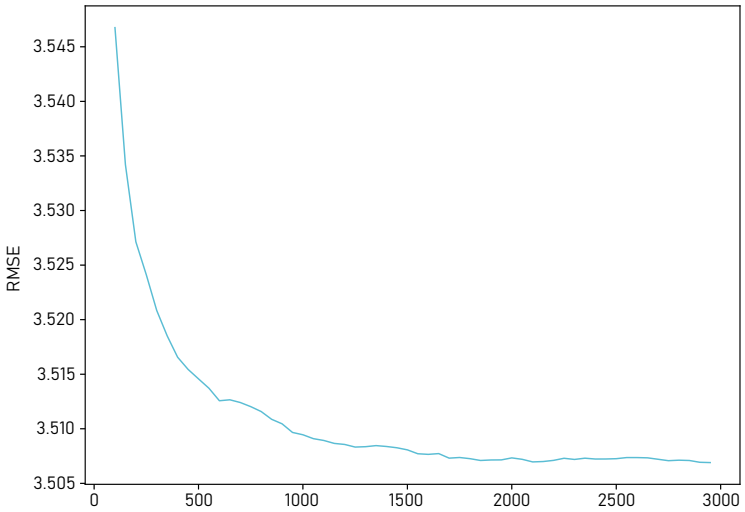


Figure 6: Number of trees and RMSE.

As the second step in the model building process, GRF parameters bandwidth and local weight were tuned. Following values for parameter bandwidth (number of closest samples) were tested; 25, 50, 75, 100, 150, 200, 250, 500, 750 and 1000. For local weight values from 0 to 1 with 0.1 increments were tested. Model is tested with the adaptive kernel, locations average distance to observations varies. The relation between bandwidth and absolute distance is displayed in the boxplot (Figure 7). The minimal average distance for bandwidth 25 is 1992 m for locations in observation rich areas, the maximal average distance is 17 407 m in very sparsely populated (by observations) regions. The average distance is 3206 m. The distance increase by the square root of bandwidth.

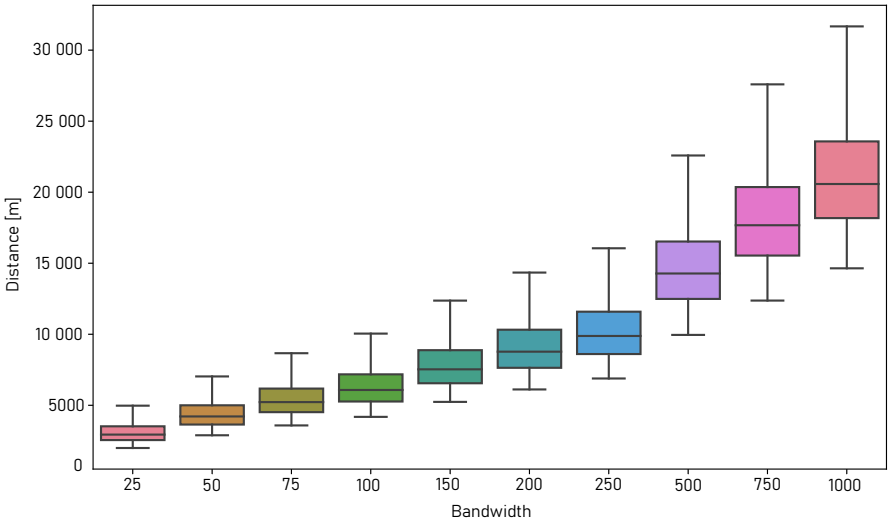


Figure 7: Distance and bandwidth values.

The error values for parameters are displayed in a line plot (Figure 8). Errors values are higher than values of tuned regular RF. This discrepancy resulted from different tuning methods, regular RF was tuned with OOB samples. RMSE of RF in the case of tuning GRF is 3.74 (MAE = 2.82), compare to RMSE of 3.51 of regular RF. OOB method tends to underestimate error.

The optimal bandwidth is 25 observations, which is also the smallest one. Bandwidth achieved RMSE of 3.58 or 2.67 MAE (for full weight for local models). The decrease is small. Even smaller is the contribution of global to local weight. Optimal weight is 0.7 for local models (70 % of local model values, 30 % of global). Combined weight decrease RMSE from 3.58 to 3.56 and MAE from 2.67 to 2.662. With increasing bandwidth accuracy decreases and converges to the regular RF error values. The biggest jump is within the first 100 observations, then accuracy decreases slower.

The low value of bandwidth can be partly explained by the nature of phenomena and the way the dependent variable is created. The indicator was created by upsampling low-resolution products to a higher resolution. As most of the observations remain in grid, the surrounding of each observation contains very similar samples. They are similar no only in the value of the indicator, but also in features value. For example, sand content, elevation or precipitation values tend to change gradually and not abruptly over space. The smallest possible bandwidth value is a result of continuous values of independent and dependent variables.

Several studies concerning GRF address the tuning process. Georganos et al. (2019) tuned parameter bandwidth in range 100 to 1000 observations (adaptive kernel). Low values of bandwidth display high error, with increasing bandwidth error, decreases and later increases. Optimal bandwidth is within the range of 300 to 500 observations. Three local weights were tested, 0.75 in favour of the global model was found to be most accurate. Santos et al. (2019) tested a similar range of values and found a value of 400 to be optimal. A

smaller bandwidth size of 25 found Hokstad and Tiganj (2020), which tested range from 25 to 100. The optimal local weight was found to be 0.5.

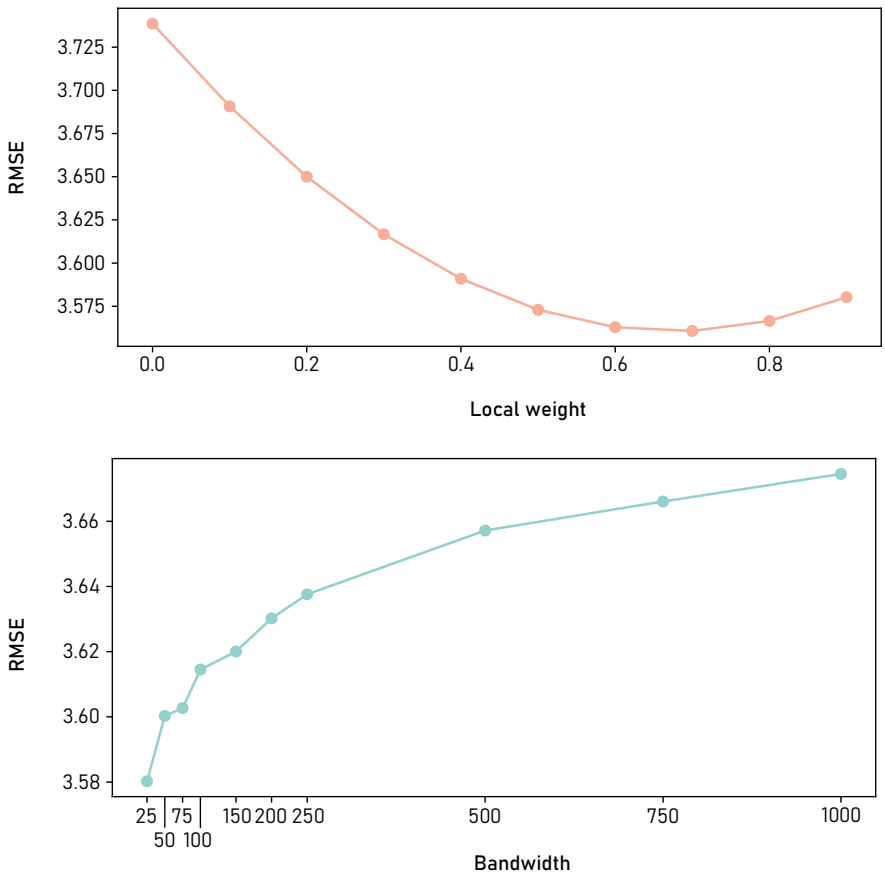


Figure 8: RMSE and parameters of GRF.

During tuning of LT GRF is for each training observation found the best bandwidth and local weight. The count of each value of the parameter is displayed in a histogram (Figure 9). The most abundant value for bandwidth is 25 constituting 33.4 % of all values. Second place belongs to value 50 with 12.9 % share. Other values constitute a smaller portion than 10 %. GRF assign to all locations one universal value, however, as can be seen, it is not optimal for the vast majority of locations. In the case of local weight, the situation is more uneven. The most abundant local weight value is 1 (only local model is employed) which constitutes 68.3 % of all values. The second most populous is value 0 (only global model is employed) with 10% contribution to all values. Other values are represented less, the count decreases with lower local weight. However, the best value achieved by GRF tuning is 0.7. This value is not optimal for more than 97 % of all locations. Therefore it is assumed that LT GRF can improve accuracy. The RMSE for the training set is 2.75 and 1.74 for MAE, which is a significant decrease in error. However, the applicability of LT GRF is dependent on the existence of spatial patterns in parameter values.

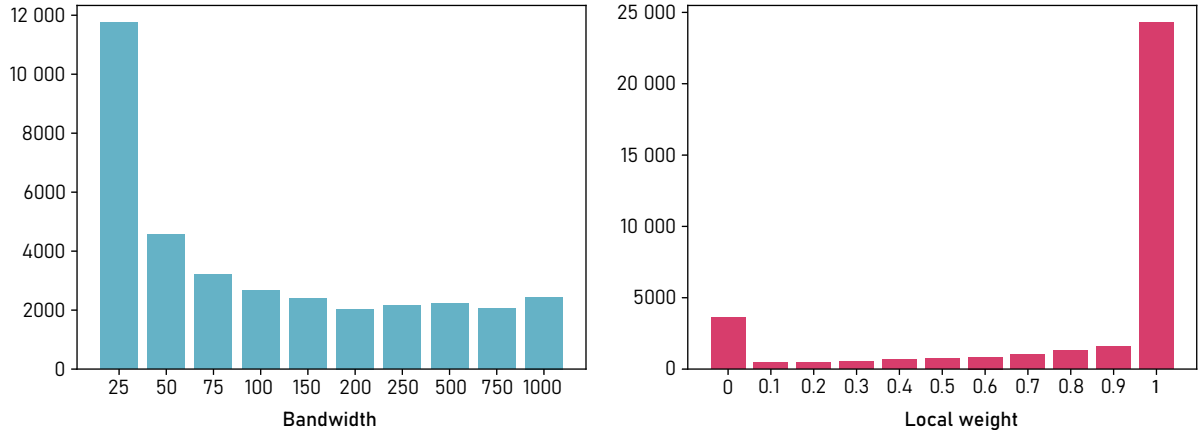


Figure 9: Histogram for bandwidth and local weight for LT GRF.

Spatial patterns of parameters were examined and displayed in a map (figure 9). No clusters of similar values were detected by visual exploration. Correlation between geographical coordinates or other geographical features were not detected. Values appear to be localized randomly. Visual exploration conclusion is further supported by variogram (Figure 11). The sill (correlated range) does not exists as first lag value (2) is similar to all others.

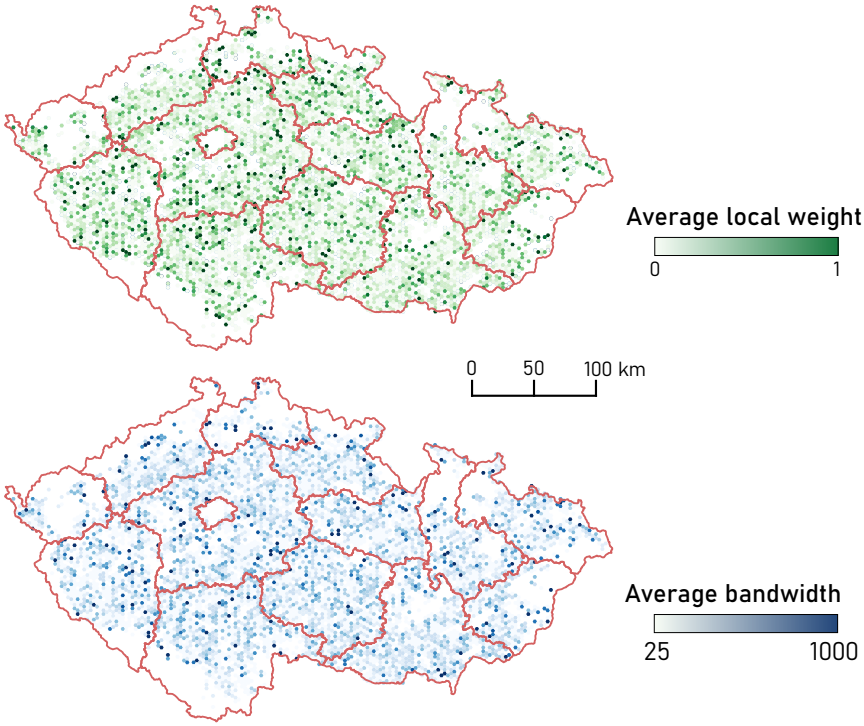


Figure 10: Spatial patterns of LT GRF parameters.

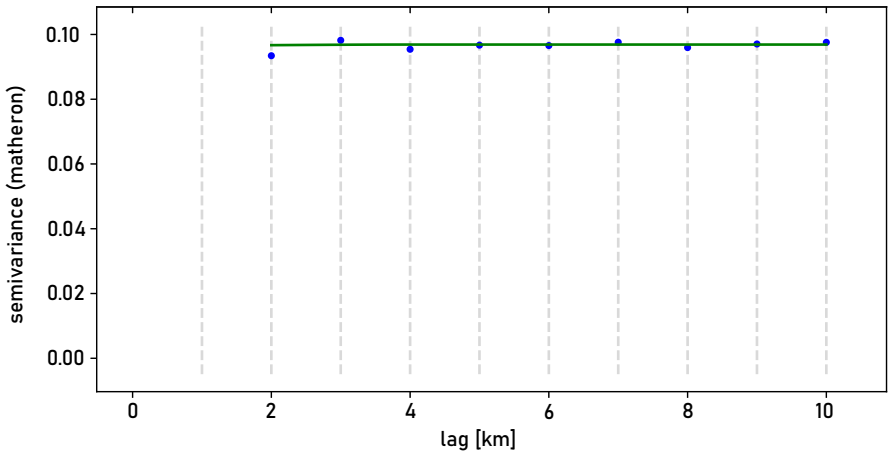


Figure 11: Variogram for local weight.

6.3 PERFORMANCE AND VULNERABILITY ASSESSMENT

6.3.1 Accuracy assessment

Each tuned model was trained and tested. The train set comprises 60 817 samples and a test set of 8862. The ratio is not strict 80/20, because observations outside the study area bounding box were deleted. RF without spatial covariates achieved RMSE 5.04 (MAE of 3.8), RF with spatial covariates achieved RMSE of 3.748 (MAE of 2.82), GRF of 3.6 (MAE of 2.69) and LT GRF of 3.57 (MAE of 2.65). In relative values, RF has an accuracy of 94.47 %, GRF 94.69 % and LT GRF 94.74 %. With consideration of range without outliers (1 - 99 %) accuracy is 92.09 %, 92.4 % and 92.46 %. Values are listed in the table below.

Table 5: Accuracy metrics for each tested model expressed in relative and absolute values.

	RMSE		MAE	
	abs	rel [%]	abs	rel [%]
RF model	5.0421	92.569	3.79986	94.399
RF_XY model	3.74811	94.476	2.82005	95.844
GRF model	3.60067	94.693	2.68956	96.036
LT GRF model	**3.57132**	**94.736**	**2.65191**	**96.092**

The best model in terms of accuracy is LT GRF. Compare to RF with spatial covariates is more accurate by 0.177 RMSE and to GRF by 0.029 RMSE. New method decrease error by 4.7 %. The decrease is very small. The minuscule difference in error between models can be explained in several ways. Firstly, regular RF achieves very good results. Accuracy of more than 92 % is very high and room for improvement is limited. It is possible that model accuracy can not be significantly improved any further. Secondly, spatial non-linearity is explained well by spatial covariates, which are input features in the global

model. In other words, the global model has not left any space for local models to improve. Difference between the RF model and RF_XY model of 1.29 RMSE or decrease in 25.663 % support this claim. Lastly, between the RF_XY model and observation does not exist any significant spatial non-stationarity, which could local models capture better than the global model. However, this hypothesis is rejected based on the difference in accuracy between the RF_XY model and RF model and the examination of variable importance in the next sections.

Surprisingly low decrease in RMSE is between GRF and LT GRF. Based on tuning LT GRF, in which RMSE decrease to 2.75, is decreased from 3.6 to 3.57 very small. RMSE was almost the same for nearest neighbour and linear interpolation for location in the test set. The unconvincing result proves the conclusion of visual examination for bandwidth. There is no or very little spatial correlation between bandwidth and local weight and a decrease in error. Values are localized randomly as a residue of random error.

GRF creates local models on a subset of original datasets. This process can be reinterpreted as a huge number of created decision trees with a very small number of observations. A similar situation can be recreated with regular RF with parameter maximum samples set to a value of best bandwidth (25). However, experiments show that such a model is very inaccurate (RMSE of 7.91) and this hypothesis can be declined.

Performance of GRF and LT GRF was compared to performance in other studies. In the study by Georganos et al. (2019) GRF with spatial covariates achieved RMSE of 0.606, the global model achieved RMSE of 0.65. The error decreased by 6.76 %. Master thesis by Hokstad and Tiganj (2020) compared RF with spatial covariates to GRF. RF achieved RMSE of 17 944 and GRF 16 705, 6.9% decrease in error. In both studies, a decrease in error between RF with spatial coordinates and without them is more significant.

Improvement of GRF or LT GRF over regular RF is small and computational runtime is much higher. For large datasets (more than 100 000 samples) desktop PC is not sufficient and less accessible and a more expensive solution needs to be employed. Therefore, GRF might not be advantageous to use. For example, if funds are limited and maximum accuracy is not imperative, regular RF with spatial covariates is sufficient. However, in case every decrease in error is transformed to higher profit and computational power is not limited, GRF is a better option.

6.3.2 Feature importance assessment

Feature importances for RF model and GRF model are displayed in barplots (Figure 12) and (Figure 13).

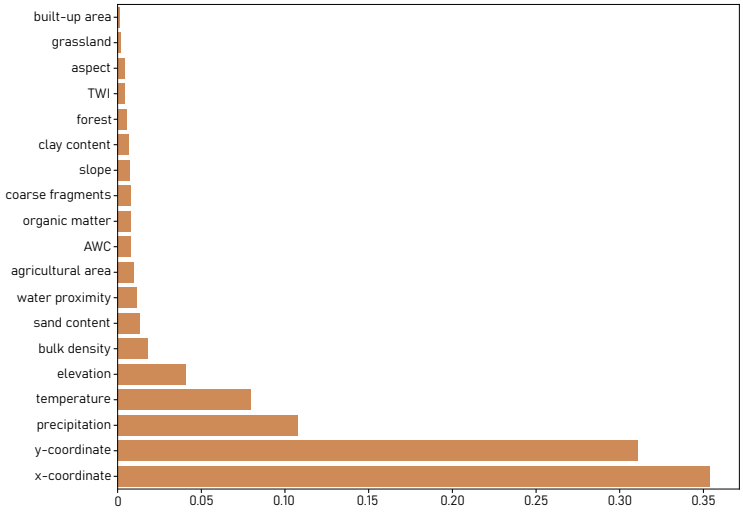


Figure 12: Feature importances for RF.

The most important features for the RF are spatial covariates - x-coordinate and y-coordinate. Together with accounts for almost two-third (35.3 % and 31 %). Third and fourth place occupy meteorological features. Precipitation accounts for 10.7 % and temperature for 7.97 %. Elevation accounts for 4 % and other features are responsible for less than 2 % of importance. The huge importance of spatial covariates points out to the spatial continuity of predicted variable. Simply put, new values are mostly predicted from close observations. The importance of meteorological features was expected. Precipitation and temperature affect soil water to a large degree. However, the insignificance of soil properties is unexpected. Similarly to SWI, soil characteristics change gradually in space, thus their change is captured by spatial covariates. Secondly, many of them are correlated, therefore their impact is underestimated.

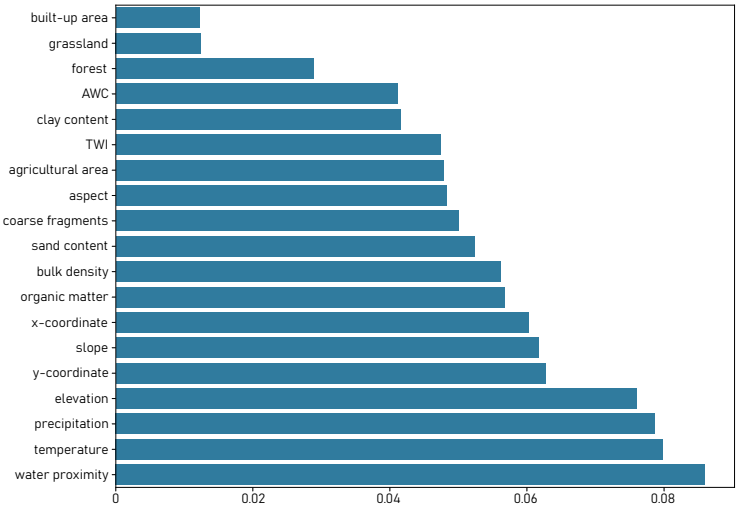


Figure 13: Feature importances for GRF.

The feature importances for the local model are much more balanced. The most important feature is water proximity with 8.5 %, followed by temperature 7.9 % and precipitation 7.8 %. Other features are responsible for 4 % to 7 % importance. Land cover classes (forest, grassland and built-up area) are least important. Spatial covariates are much less significant compare to the global model. Local kernel substitutes coordinates, which became meaningless in a small subset of data. Soil properties constitute 30 %, which is much more compared to the global model. Land cover classes occupy the last position in both models. This can be interpreted as an absence of effect on SWI. However, land cover classes often have a value of zero - class is absent, which diminished its information value.

Most important feature by mean value - water proximity was examined more closely (Figure 14). Importance varies over space. Maximum values reaches 60 %, on the other hand in many location importance is less than 5 %. Several small cluster can be identified, for example cluster of high values north of Kolín or cluster of low values south-west of Prague. The importance is not dependent on value of water proximity. Correlation coefficient is only 0.036 which suggest no correlation at all. This is also suggested by visual exploration. In conclusion, the impact of vicinity of water bodies need to be investigated individually especially in places with high importance.

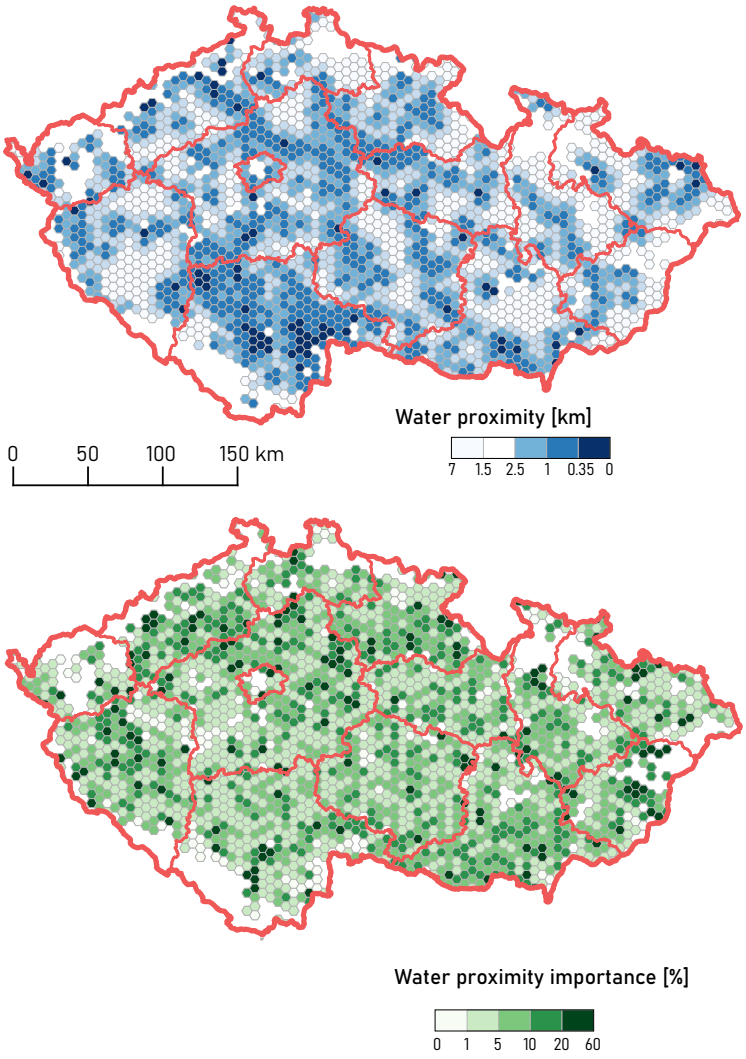


Figure 14: Water proximity importance.

6.3.3 Vulnerability assessment

Lastly, a map of vulnerability assessment was created (Appendix 1) from meteorological features importances of most accurate model (LT GRF) Locations of high (or low) vulnerability degrees are not grouped in a huge cluster rather are intertwined with regions with low or medium vulnerability. Despite that, several vulnerable regions can be identified. Two big clusters can be found in the South Moravia region - area in proximity of village Mikulov and north-east of Hodonín. Higher vulnerability clusters are found in the vicinity of cities Pardubice and Chrudim, near the city of Turnov and in the Žatec region. Patches of vulnerable regions are located in the foothills of the Šumava mountains.

Comparing vulnerability to the others studies is difficult because of different scales, cartographic visualization used and filtered land cover applied in thesis. Brázdil and Miroslav Trnka (2015) identified as highly vulnerable regions Southern Moravia, Krušné hory region, area south-west of city of

Kadaň, which is forested military area. Most highlighted regions are located in forested regions, which were not included in thesis. Miroslav Trnka, Semerádová, et al. (2016) recognise as vulnerable areas Southern Moravia and Krušné hory foothills. In general Southern Moravia region, Krušné hory foothills and surrounding of the city of Kolín. Contrary, to regions listed above thesis identified Šumava foothills as prone to drought.

The degree of vulnerability is not correlated with any of included features. The highest correlation (positive or negative) is between elevation (-0.18). The correlation is so small, that none of the features can be identified as having any effect on vulnerability. Unfortunately, the assessment of drought influences can not be investigated. Compare to other traditional approaches, assessment from feature importances is not subjected to manual set up of weights. The method is dependent on the selection of drought factors, especially meteorological features. Accuracy can be quantified by the accuracy of the model. On the other hand, the validation of assessment is limited due to the absence of similar studies or environmental vulnerability assessment of Czechia.

7 CONCLUSION AND FUTURE DEVELOPMENT

7.1 CONCLUSION

The thesis deals with building three regression models - RF, GRF and LT GRF. Models are used to predict values of the drought indicator from environmental data. Feature importance is extracted from each model and compared. Lastly, from meteorological features importances is created a vulnerability map.

The first task comprises of constructing and evaluating machine learning models trained and tested on spatial data. It was assumed, that spatially sensitive models will perform better. The RF model achieved RMSE of 5.04, RF model with spatial covariates (coordinates) achieved RMSE of 3.74, GRF of 3.6, and lastly LT GRF attained RMSE of 3.57. The biggest decrease results from including spatial covariates. This option is from a computational standpoint very easy; the new spatial features will not increase runtime and the approach does not require additional programming in comparison to GRF. The decrease in error to GRF is smaller (0.15 RMSE or 4.7 %). The decrease is small but comparable to other studies, which used GRF. However, the decrease from GRF to LT GRF model is minuscule (0.03 decrease in RMSE). The LT GRF model is not benefiting from the tuning of bandwidth and local weight parameters. It was found that there is no spatial continuity between values of both parameters, implicitly by visual exploration and huge difference between values of RMSE from training and testing set, and explicitly by variogram. LT GRF in this particular case does not decrease significantly error. On the other hand, GRF might be useful in some cases, for example, if the computational power is unlimited and accuracy requirements very high.

The second task was comprised of evaluating and comparing local variable importance to the global one. The most important features in the global model are spatial covariates. X and Y coordinate were responsible for more than 66 % of importance, followed by meteorological features (precipitation and temperature) with less than 19 %. The variable importance is divided between features unevenly. The first four features account for more than 85 % of importance. The local models are very different. The spatial covariates are not as important, which is the result of the small size of the subset. The importances are divided much more evenly. The most important feature is water proximity, followed by meteorological features and elevation. The soil properties have a much higher share of importance than in the global model. The water proximity feature was explored in closer detail, however, no correlation between the importance and water proximity values or any significant clusters were found, mainly because the importance is relatively low. The local importances can serve to assess local causes of drought, which would

not be possible from a global model.

The last task included developing of vulnerability assessment. As more vulnerable areas were identified southern Moravia, vicinity of cities Pardubice and Chrudim or are north of the city of Kolín. The more vulnerable areas are intertwined with relatively less vulnerable areas. This suggests inconsistency in the results. The validation process is hindered by the scarcity of suitable vulnerability assessments of Czechia regions. These assessments have insufficient scale or are not aimed at the agricultural land cover. On the other hand, from a subjective standpoint, the results of the assessment look promising.

In conclusion, the spatial version of RF - GRF provides several advantages over aspatial algorithm. Besides a small increase in accuracy, the GRF provides variable importance, which is localized for each location. This feature can be utilised in the development of vulnerability assessments.

7.2 FUTURE DEVELOPMENT

As a continuation of this thesis, several directions can be explored:

- Including higher number of features, especially meteorological one, which could improve performance of GRF and vulnerability assessment from feature importance values.
- Including higher number of training samples, from more than one drought episodes. This would decrease variance and improve accuracy for feature importances.
- Extend the GRF concept to time dimension similarly to extension of GWR (Fotheringham, Crespo, et al., 2015). Time dimension allows creation of a more dynamic model, which would perform better in context of climate change.

BIBLIOGRAPHY

Adnan, Shahzada, Kalim Ullah, Li Shuanglin, Shouting Gao, Azmat Hayat Khan, and Rashed Mahmood

2018 "Comparison of various drought indices to monitor drought status in Pakistan", *Climate Dynamics*, 51, 5, pp. 1885-1899.

Ahn, Seongin, Dong-Woo Ryu, and Sangho Lee

2020 "A Machine Learning-Based Approach for Spatial Estimation Using the Spatial Features of Coordinate Information", *ISPRS International Journal of Geo-Information*, 9, 10, p. 587.

Alley, William M

1984 "The Palmer drought severity index: limitations and assumptions", *Journal of Applied Meteorology and Climatology*, 23, 7, pp. 1100-1109.

Alpaydin, Ethem

2020 *Introduction to machine learning*, MIT press.

Andrienko, Natalia and Gennady Andrienko

2006 *Exploratory analysis of spatial and temporal data: a systematic approach*, Springer Science & Business Media.

Archer, Kellie J and Ryan V Kimes

2008 "Empirical characterization of random forest variable importance measures", *Computational statistics & data analysis*, 52, 4, pp. 2249-2260.

Bachmair, S, Maliko Tanguy, Jamie Hannaford, and K Stahl

2018 "How well do meteorological indicators represent agricultural and forest drought across Europe?", *Environmental Research Letters*, 13, 3, p. 034042.

Bakker, Karen and T. Downing

2000 "Drought discourse and vulnerability", *Drought: A global assessment*, 2 (Jan. 2000), pp. 213-230.

Ballabio, Cristiano, Panos Panagos, and Luca Monatanarella

2016 "Mapping topsoil physical properties at European scale using the LUCAS database", *Geoderma*, 261, pp. 110-123.

Bauer-Marschallinger, Bernhard, Christoph Paulik, Simon Hochstöger, Thomas Mistelbauer, Sara Modanesi, Luca Ciabatta, Christian Massari, Luca Brocca, and Wolfgang Wagner

2018 "Soil moisture from fusion of scatterometer and SAR: Closing the scale gap with temporal filtering", *Remote Sensing*, 10, 7, p. 1030.

Bauer-Marschallinger, Bernhard and Isabella Pfeil

2021 *Soil Water Index - Copernicus Global Land Operations "Vegetation and Energy" Water*, tech. rep., Technical University of Vienna.

Behrens, Thorsten, Karsten Schmidt, Raphael A Viscarra Rossel, Philipp Gries, Thomas Scholten, and Robert A MacMillan

2018 "Spatial modelling with Euclidean distance fields and machine learning", *European journal of soil science*, 69, 5, pp. 757-770.

Belson, William A

1956 "A technique for studying the effects of a television broadcast", *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 5, 3, pp. 195-202.

Bergstra, James and Yoshua Bengio

2012 "Random search for hyper-parameter optimization." *Journal of machine learning research*, 13, 2.

Berk, Richard A

2020 *Statistical Learning from a Regression Perspective*, Springer.

Bernard, Simon, Laurent Heutte, and Sébastien Adam

2009 "Influence of hyperparameters on random forest accuracy", in *International workshop on multiple classifier systems*, Springer, pp. 171-180.

Beven, Keith J and Michael J Kirkby

1979 "A physically based, variable contributing area model of basin hydrology/Un modèle à base physique de zone d'appel variable de l'hydrologie du bassin versant", *Hydrological Sciences Journal*, 24, 1, pp. 43-69.

Birkmann, Joern, Omar D Cardona, Martha L Carreño, Alex H Barbat, Mark Pelling, Simon Schneiderbauer, Stefan Kienberger, Margreth Keiler, David Alexander, Peter Zeil, et al.

2013 "Framing vulnerability, risk and societal responses: the MOVE framework", *Natural hazards*, 67, 2, pp. 193-211.

Bishop, Christopher M

2006 *Pattern recognition and machine learning*, springer.

Blaikie, Piers, Terry Cannon, Ian Davis, and Ben Wisner

2014 *At risk: natural hazards, people's vulnerability and disasters*, Routledge.

Bot, Alexandra and José Benites

2005 *The importance of soil organic matter: Key to drought-resistant soil and sustained food production*, 80, Food & Agriculture Org.

Bousquet, Olivier, Stéphane Boucheron, and Gábor Lugosi

2003 "Introduction to statistical learning theory", in *Summer School on Machine Learning*, Springer, pp. 169-207.

Brázdil, Rudolf and Miroslav Trnka

2015 *Historie počasí a podnebí v Českých zemích. minulost, současnost, budoucnost*, Centrum výzkumu globální změny Akademie věd České republiky, Brno, ISBN: 978-80-87902-11-0.

Breiman, Leo

1996 "Bagging predictors", *Machine learning*, 24, 2, pp. 123-140.

2001 "Random forests", *Machine learning*, 45, 1, pp. 5-32.

2002 "Manual on setting up, using, and understanding random forests v3. 1", *Statistics Department University of California Berkeley, CA, USA*, 1, p. 58.

Breiman, Leo, Jerome Friedman, Charles J Stone, and Richard A Olshen

1984 *Classification and regression trees*, CRC press.

Bryant, Edward

2005 *Natural Hazards*, Cambridge University Press.

Büttner, György, Barbara Kosztra, Gergely Maucha, Róbert Pataki, Stefan Kleeschulte, Gerard Hazeu, Marian Vittek, Christoph Schröder, and Andreas Littkopf

2021 *CORINE Land Cover - Copernicus Land Monitoring Service*, tech. rep., European Environment Agency (EEA).

Cairns, JE, SM Impa, JC O'Toole, SVK Jagadish, and AH Price

2011 "Influence of the soil physical environment on rice (Oryza sativa L.) response to drought stress and its implications for drought research", *Field Crops Research*, 121, 3, pp. 303-310.

Cammalleri, Carmelo, Carolina Arias-Muñoz, Paulo Barbosa, Alfred de Jager, Diego Magni, Dario Masante, Marco Mazzeschi, Niall McCormick, Gustavo Naumann, Jonathan Spinoni, et al.

2021 "A revision of the Combined Drought Indicator (CDI) used in the European Drought Observatory (EDO)", *Natural Hazards and Earth System Sciences*, 21, 2, pp. 481-495.

Cardona, Omar Dario, Maarten K Van Aalst, Jörn Birkmann, Maureen Fordham, Glenn Mc Gregor, Perez Rosa, Roger S Pulwarty, E Lisa F Schipper, Bach Tan Sinh, Henri Décamps, et al.

2012 "Determinants of risk: exposure and vulnerability", in *Managing the risks of extreme events and disasters to advance climate change adaptation: special report of the intergovernmental panel on climate change*, Cambridge University Press, pp. 65-108.

Casetti, Emilio and John Paul Jones III

1991 *Applications of the expansion method*, Routledge.

Cassel, DK and DR Nielsen

1986 "Field capacity and available water capacity", *Methods of soil analysis: Part 1 Physical and mineralogical methods*, 5, pp. 901-926.

Čermáková, Olga, Miloslav Janeček, Andrea Jindrová, Jan Kořıýnek, et al.

2014 "The impact of farming and land ownership on soil erosion", *Acta Universitatis Agriculturae et Silviculturae Mendelianae Brunensis*, 62, 5, pp. 883-890.

Chai, Tianfeng and Roland R Draxler

2014 "Root mean square error (RMSE) or mean absolute error (MAE)?–Arguments against avoiding RMSE in the literature", *Geoscientific model development*, 7, 3, pp. 1247-1250.

Cornes, Richard C, Gerard van der Schrier, Else JM van den Besselaar, and Philip D Jones

2018 "An ensemble version of the E-OBS temperature and precipitation data sets", *Journal of Geophysical Research: Atmospheres*, 123, 17, pp. 9391-9409.

Cotrufo, M Francesca, Maria Giovanna Ranalli, Michelle L Haddix, Johan Six, and Emanuele Lugato

2019 "Soil carbon storage informed by particulate and mineral-associated organic matter", *Nature Geoscience*, 12, 12, pp. 989-994.

Dayal, Kavina S, Ravinesh C Deo, and Armando A Apan

2018 "Spatio-temporal drought risk mapping approach and its application in the drought-prone region of south-east Queensland, Australia", *Natural Hazards*, 93, 2, pp. 823-847.

De Stefano, Lucia, Itziar Tánago, Mario Ballesteros Olza, Julia Urquijo Reguera, Veit Blauhut, James Stagge, and Kerstin Stahl

2015 *Methodological approach considering different factors influencing vulnerability - pan-European scale*, tech. rep., Universidad Complutense de Madrid (UCM), Albert-Ludwigs-Universität Freiburg, Germany (ALU-FR), and Universitetet i Oslo, Norway (UiO).

DeCastro-Garcıýa, Noemıý, Ángel Luis Muñoz Castañeda, David Escudero Garcıýa, and Miguel V Carriegos

2019 "Effect of the Sampling of a Dataset in the Hyperparameter Optimization Phase over the Efficiency of a Machine Learning Algorithm", *Complexity*, 2019.

Deng, Liangdong, Malek Adjouadi, and Naphtali Rishe

2020 "Inverse Distance Weighted Random Forests: Modeling Unevenly Distributed Non-Stationary Geographic Data", in *2020 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, IEEE, pp. 41-46.

De Ville, Barry

2013 "Decision trees", *Wiley Interdisciplinary Reviews: Computational Statistics*, 5, 6, pp. 448-455.

Di Falco, Salvatore and Jean-Paul Chavas

2008 "Rainfall shocks, resilience, and the effects of crop biodiversity on agroecosystem productivity", *Land Economics*, 84, 1, pp. 83-96.

Dinaburga, Gundega, Dainis Lapins, Janis Kopmanis, et al.

2010 "Differences of soil agrochemical properties in connection with altitude in winter wheat", in *Proceedings of Engineering for Rural Development Conference*, pp. 27-28.

Dıýaz-Uriarte, Ramón and Sara Alvarez De Andres

2006 "Gene selection and classification of microarray data using random forest", *BMC bioinformatics*, 7, 1, p. 3.

Dracup, John A, Kil Seong Lee, and Edwin G Paulson Jr

1980a "On the definition of droughts", *Water resources research*, 16, 2, pp. 297-302.

1980b "On the statistical characteristics of drought events", *Water resources research*, 16, 2, pp. 289-296.

Easton, Zachary M, Emily Bock, et al.

2016 "Soil and soil water relationships".

Efron, Bradley

1983 "Estimating the error rate of a prediction rule: improvement on cross-validation", *Journal of the American statistical association*, 78, 382, pp. 316-331.

1992 "Bootstrap methods: another look at the jackknife", in *Breakthroughs in statistics*, Springer, pp. 569-593.

Ekrami, Mohammad, Ahmad Fatehi Marj, Jalal Barkhordari, and Kazem Dashtakian

2016 "Drought vulnerability mapping using AHP method in arid and semiarid areas: a case study for Taft Township, Yazd Province, Iran", *Environmental Earth Sciences*, 75, 12, pp. 1-13.

Esposito, Floriana, Donato Malerba, Giovanni Semeraro, and J Kay

1997 "A comparative analysis of methods for pruning decision trees", *IEEE transactions on pattern analysis and machine intelligence*, 19, 5, pp. 476-491.

European Enviroment Agency

2021 *EU-DEM*, https://land.copernicus.eu/imagery-in-situ/eu-dem.

Fekete, Alexander, Marion Damm, and Jörn Birkmann

2010 "Scales as a challenge for vulnerability assessment", *Natural Hazards*, 55, 3, pp. 729-747.

Field, Christopher B, Vicente Barros, Thomas F Stocker, and Qin Dahe
2012 *Managing the risks of extreme events and disasters to advance climate change adaptation: special report of the intergovernmental panel on climate change*, Cambridge University Press.

Fotheringham, A Stewart, Chris Brunsdon, and Martin Charlton
2003 *Geographically weighted regression: the analysis of spatially varying relationships*, John Wiley & Sons.

Fotheringham, A Stewart, Martin Charlton, and Chris Brunsdon
1996 "The geography of parameter space: an investigation of spatial non-stationarity", *International Journal of Geographical Information Systems*, 10, 5, pp. 605-627.

Fotheringham, A Stewart, Ricardo Crespo, and Jing Yao
2015 "Geographical and temporal weighted regression (GTWR)", *Geographical Analysis*, 47, 4, pp. 431-452.

Freund, Yoav, Robert E Schapire, et al.
1996 "Experiments with a new boosting algorithm", in *icml*, Citeseer, vol. 96, pp. 148-156.

Füssel, Hans-Martin
2007 "Vulnerability: A generally applicable conceptual framework for climate change research", *Global environmental change*, 17, 2, pp. 155-167.

Genuer, Robin, Jean-Michel Poggi, and Christine Tuleau
2008 "Random Forests: some methodological insights", *arXiv preprint arXiv:0811.3619*.

Georganos, Stefanos, Tais Grippa, Assane Niang Gadiaga, Catherine Linard, Moritz Lennert, Sabine Vanhuysse, Nicholus Mboga, Eléonore Wolff, and Stamatis Kalogirou
2019 "Geographical random forests: a spatial extension of the random forest algorithm to address spatial heterogeneity in remote sensing and population modelling", *Geocarto International*, pp. 1-16.

Grabs, Thomas, Jan Seibert, K Bishop, and H Laudon
2009 "Modeling spatial patterns of saturated areas: A comparison of the topographic wetness index and a dynamic distributed model", *Journal of Hydrology*, 373, 1-2, pp. 15-23.

Grömping, Ulrike
2009 "Variable importance assessment in regression: linear regression versus random forest", *The American Statistician*, 63, 4, pp. 308-319.

Guttman, Nathaniel B
1999 "Accepting the standardized precipitation index: a calculation algorithm 1", *JAWRA Journal of the American Water Resources Association*, 35, 2, pp. 311-322.

Hagenlocher, Michael, Isabel Meza, Carl C Anderson, Annika Min, Fabrice G Renaud, Yvonne Walz, Stefan Siebert, and Zita Sebesvari
2019 "Drought vulnerability and risk assessments: state of the art, persistent gaps, and research agenda", *Environmental Research Letters*, 14, 8, p. 083002.

Han, Lanying, Qiang Zhang, Pengli Ma, Jianying Jia, and Jinsong Wang
2016 "The spatial distribution characteristics of a comprehensive drought risk index in southwestern China and underlying causes", *Theoretical and Applied Climatology*, 124, 3-4, pp. 517-528.

Hassine, Kawther, Aiman Erbad, and Ridha Hamila
2019 "Important complexity reduction of random forest in multi-classification problem", in *2019 15th International Wireless Communications & Mobile Computing Conference (IWCMC)*, IEEE, pp. 226-231.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman
2009 *The elements of statistical learning: data mining, inference, and prediction*, Springer Science & Business Media.

Hawthorne, Sandra and Chelcy Ford Miniat
2018 "Topography may mitigate drought effects on vegetation along a hillslope gradient", *Ecohydrology*, 11, 1, e1825.

Heim Jr, Richard R
2002 "A review of twentieth-century drought indices used in the United States", *Bulletin of the American Meteorological Society*, 83, 8, pp. 1149-1166.

Hengl, Tomislav, Madlene Nussbaum, Marvin N Wright, Gerard BM Heuvelink, and Benedikt Gräler
2018 "Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables", *PeerJ*, 6, e5518.

Hlásny, Tomáš, Csaba Mátyás, Rupert Seidl, Ladislav Kulla, and Katarıýna Merganicová
2014 "Climate change increases the drought risk in Central European forests: what are the otions for adaptation?", *LESNICKY CASOPIS-FORESTRY JOURNAL*, 60, pp. 5-18.

Hlavinka, Petr, Miroslav Trnka, Jan Balek, Daniela Semerádová, Michael Hayes, Mark Svoboda, Josef Eitzinger, Martin Možní, Milan Fischer, Eric Hunt, et al.
2011 "Development and evaluation of the SoilClim model for water balance and soil climate estimates", *Agricultural Water Management*, 98, 8, pp. 1249-1261.

Ho, Tin Kam
1995 "Random decision forests", in *Proceedings of 3rd international conference on document analysis and recognition*, IEEE, vol. 1, pp. 278-282.

Hokstad, Vegard and Dzenana Tiganj

2020 *Spatial Modelling of Unconventional Wells in the Niobrara Shale Play*, MA thesis, Norwegian school of economics.

Hoque, Muhammad Al-Amin, Biswajeet Pradhan, and Naser Ahmed

2020 "Assessing drought vulnerability using geospatial techniques in northwestern part of Bangladesh", *Science of The Total Environment*, 705, p. 135957.

Hoque, Muhammad Al-Amin, Biswajeet Pradhan, Naser Ahmed, and Md Shawkat Islam Sohel

2021 "Agricultural drought risk assessment of Northern New South Wales, Australia using geospatial techniques", *Science of the Total Environment*, 756, p. 143600.

Huang, Jin, Huug M van den Dool, and Konstantine P Georgarakos

1996 "Analysis of model-calculated soil moisture over the United States (1931–1993) and applications to long-range temperature forecasts", *Journal of Climate*, 9, 6, pp. 1350-1362.

Huete, A, C Justice, and H Liu

1994 "Development of vegetation and soil indices for MODIS-EOS", *Remote Sensing of environment*, 49, 3, pp. 224-234.

Ionescu, Cezar, Richard JT Klein, Jochen Hinkel, KS Kavi Kumar, and Rupert Klein

2009 "Towards a formal framework of vulnerability to climate change", *Environmental Modeling & Assessment*, 14, 1, pp. 1-16.

Jain, Vinit K, RP Pandey, and Manoj K Jain

2015 "Spatio-temporal assessment of vulnerability to drought", *Natural Hazards*, 76, 1, pp. 443-469.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani

2013 *An introduction to statistical learning*, Springer, vol. 112.

Karl, Thomas R

1986 "The sensitivity of the Palmer Drought Severity Index and Palmer's Z-index to their calibration coefficients including potential evapotranspiration", *Journal of Climate and Applied Meteorology*, pp. 77-86.

Keyantash, John and John A Dracup

2002 "The quantification of drought: an evaluation of drought indices", *Bulletin of the American Meteorological Society*, 83, 8, pp. 1167-1180.

Krige, DG

1966 "Two-dimensional weighted moving average trend surfaces for ore-evaluation", *Journal of the South African Institute of Mining and Metallurgy*, 66, pp. 13-38.

Kuhn, Max, Kjell Johnson, et al.

2013 *Applied predictive modeling*, Springer, vol. 26.

Lauzon, Nicolas, François Anctil, and Juan Petrinovic

2004 "Characterization of soil moisture conditions at temporal scales from a few days to annual", *Hydrological Processes*, 18, 17, pp. 3235-3254.

Lavell, Allan, Michael Oppenheimer, Cherif Diop, Jeremy Hess, Robert Lempert, Jianping Li, Robert Muir-Wood, Soojeong Myeong, Susanne Moser, Kuniyoshi Takeuchi, et al.

2012 "Climate change: new dimensions in disaster risk, exposure, vulnerability, and resilience", in *Managing the risks of extreme events and disasters to advance climate change adaptation: Special report of the Intergovernmental Panel on Climate Change*, Cambridge University Press, pp. 25-64.

Liashchynskyi, Petro and Pavlo Liashchynskyi

2019 "Grid search, random search, genetic algorithm: a big comparison for NAS", *arXiv preprint arXiv:1912.06059*.

Liaw, Andy, Matthew Wiener, et al.

2002 "Classification and regression by randomForest", *R news*, 2, 3, pp. 18-22.

Lim, Tjen-Sien, Wei-Yin Loh, and Yu-Shan Shih

1998 "An empirical comparison of decision trees and other classification methods".

Lima, Manuel

2011 *Visual Complexity: Mapping Patterns of Information*, Princeton Architectural Press, ISBN: 1568989369.

Lin, Yi and Yongho Jeon

2006 "Random forests and adaptive nearest neighbors", *Journal of the American Statistical Association*, 101, 474, pp. 578-590.

Liu, CH Bryan, Benjamin Paul Chamberlain, Duncan A Little, and Ângelo Cardoso

2017 "Generalising random forest parameter optimisation to include stability and cost", in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, Springer, pp. 102-113.

Liu, Xianfeng, Xiufang Zhu, Yaozhong Pan, Shuangshuang Li, Yanxu Liu, and Yuqi Ma

2016 "Agricultural drought monitoring: Progress, challenges, and prospects", *Journal of Geographical Sciences*, 26, 6, pp. 750-767.

Ljumović, Milica and Michael Klar

2015 "Estimating expected error rates of random forest classifiers: A comparison of cross-validation and bootstrap", in *2015 4th Mediterranean Conference on Embedded Computing (MECO)*, IEEE, pp. 212-215.

Lloyd-Hughes, Benjamin

2014 "The impracticality of a universal drought definition", *Theoretical and Applied Climatology*, 117, 3, pp. 607-611.

Loh, Wei-Yin and Yu-Shan Shih

1997 "Split selection methods for classification trees", *Statistica sinica*, pp. 815-840.

Loussaief, Sehla and Afef Abdelkrim

2018 "Convolutional neural network hyper-parameters optimization based on genetic algorithms", *International Journal of Advanced Computer Science and Applications*, 9, 10, pp. 252-266.

Magesh, Nochyil S, Nainarpandian Chandrasekar, and John Prince Soundranayagam

2011 "Morphometric evaluation of Papanasam and Manimuthar watersheds, parts of Western Ghats, Tirunelveli district, Tamil Nadu, India: a GIS approach", *Environmental Earth Sciences*, 64, 2, pp. 373-381.

Martıýnez-Muñoz, Gonzalo and Alberto Suárez

2010 "Out-of-bag estimation of the optimal sample size in bagging", *Pattern Recognition*, 43, 1, pp. 143-152.

Masante, Dario and Jürgen Vogt

2018 *Drought in Central-Northern Europe – August 2018*, tech. rep., JRC European Drought Observatory (EDO).

Matala, Anna

2008 "Sample size requirement for Monte Carlo simulations using Latin hypercube sampling", *Helsinki University of Technology, Department of Engineering Physics and Mathematics, Systems Analysis Laboratory*.

Mattivi, Pietro, Francesca Franci, Alessandro Lambertini, and Gabriele Bitelli

2019 "TWI computation: a comparison of different open source GISs", *Open Geospatial Data, Software and Standards*, 4, 1, pp. 1-12.

McKay, Michael D, Richard J Beckman, and William J Conover

1979 "A comparison of three methods for selecting values of input variables in the analysis of output from a computer code", *Technometrics*, 21, 2, pp. 239-245.

McKee, Thomas B, Nolan J Doesken, John Kleist, et al.

1993 "The relationship of drought frequency and duration to time scales", in *Proceedings of the 8th Conference on Applied Climatology*, 22, Boston, vol. 17, pp. 179-183.

Méndez-Barroso, Luis A, Enrique R Vivoni, Christopher J Watts, and Julio C Rodrıýguez

2009 "Seasonal and interannual relations between precipitation, surface soil moisture and vegetation dynamics in the North American monsoon region", *Journal of hydrology*, 377, 1-2, pp. 59-70.

Meyer, Hanna, Christoph Reudenbach, Stephan Wöllauer, and Thomas Nauss

2019 "Importance of spatial predictor variable selection in machine learning applications–Moving from data reproduction to spatial prediction", *Ecological Modelling*, 411, p. 108815.

Mingers, John

1989a "An empirical comparison of pruning methods for decision tree induction", *Machine learning*, 4, 2, pp. 227-243.

1989b "An empirical comparison of selection measures for decision-tree induction", *Machine learning*, 3, 4, pp. 319-342.

Mishra, Ashok K and Vijay P Singh

2010 "A review of drought concepts", *Journal of hydrology*, 391, 1-2, pp. 202-216.

Mitchell, Tom M

1997 *Machine Learning*, McGraw-Hill.

Mohan, S and NCV Rangacharya

1991 "A modified method for drought identification", *Hydrological Sciences Journal*, 36, 1, pp. 11-21.

Mohri, Mehryar, Afshin Rostamizadeh, and Ameet Talwalkar

2018 *Foundations of machine learning*, MIT press.

Moore, Samuel A, Daniel M D'addario, James Kurinskas, and Gary M Weiss

2009 "Are decision trees always greener on the open (source) side of the fence?", *weather*, 1, p. 40.

Mukherjee, Sourav, Ashok Mishra, and Kevin E Trenberth

2018 "Climate change and drought: a perspective on drought indices", *Current Climate Change Reports*, 4, 2, pp. 145-163.

Nagarajan, Ramanathan

2010 *Drought assessment*, Springer Science & Business Media.

Niemeyer, Stefan et al.

2008 "New drought indices", *Options Méditerranéennes. Série A: Séminaires Méditerranéens*, 80, pp. 267-274.

O'Callaghan, John F and David M Mark

1984 "The extraction of drainage networks from digital elevation data", *Computer vision, graphics, and image processing*, 28, 3, pp. 323-344.

Office for Disaster Risk Reduction

2004 *Living with risk: a global review of disaster reduction initiatives.* United Nations.

Oorthuis, Raül, Jean Vaunat, Marcel Hürlimann, Antonio Lloret, José Moya, Càrol Puig-Polo, and Alessandro Fraccica

2021 "Slope Orientation and Vegetation Effects on Soil Thermo-Hydraulic Behavior. An Experimental Study", *Sustainability*, 13, 1, p. 14.

Orgiazzi, Alberto, Cristiano Ballabio, Panagiotis Panagos, Arwyn Jones, and Oihane Fernández-Ugalde

2018 "LUCAS Soil, the largest expandable soil dataset for Europe: a review", *European Journal of Soil Science*, 69, 1, pp. 140-153.

Oshiro, Thais Mayumi, Pedro Santoro Perez, and José Augusto Baranauskas

2012 "How many trees in a random forest?", in *International workshop on machine learning and data mining in pattern recognition*, Springer, pp. 154-168.

Palmer, Wayne C

1965 *Meteorological drought*, US Department of Commerce, Weather Bureau, vol. 30.

1968 "Keeping track of crop moisture conditions, nationwide: the new crop moisture index".

Panagos, Panos, Marc Van Liedekerke, Arwyn Jones, and Luca Montanarella

2012 "European Soil Data Centre: Response to European policy support and public data requirements", *Land use policy*, 29, 2, pp. 329-338.

Park, Haekyung, Kyungmin Kim, et al.

2019 "Prediction of severe drought area based on random forest: using satellite image and topography data", *Water*, 11, 4, p. 705.

Pártl, Adam, David Vačkář, Blanka Loučková, and Eliška Krkoška Lorencová

2017 "A spatial analysis of integrated risk: vulnerability of ecosystem services provisioning to different hazards in the Czech Republic", *Natural Hazards*, 89, 3, pp. 1185-1204.

Paulik, Christoph, Wouter Dorigo, Wolgang Wagner, and Richard Kidd

2014 "Validation of the ASCAT Soil Water Index using in situ data from the International Soil Moisture Network", *International journal of applied earth observation and geoinformation*, 30, pp. 1-8.

Peng, Yu, Qinghui Wang, Huiting Wang, Yongyi Lin, Jingyi Song, Tiantian Cui, and Min Fan

2019 "Does landscape pattern influence the intensity of drought and flood?", *Ecological Indicators*, 103, pp. 173-181.

Probst, Philipp and Anne-Laure Boulesteix

2017 "To tune or not to tune the number of trees in random forest." *J. Mach. Learn. Res.*, 18, 1, pp. 6673-6690.

Probst, Philipp, Anne-Laure Boulesteix, and Bernd Bischl

2019 "Tunability: importance of hyperparameters of machine learning algorithms", *The Journal of Machine Learning Research*, 20, 1, pp. 1934-1965.

Probst, Philipp, Marvin N Wright, and Anne-Laure Boulesteix

2019 "Hyperparameters and tuning strategies for random forest", *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 9, 3, e1301.

Project Team ECA&D

2021 *Countries - European Climate Assessment & Dataset project*, https://www.ecad.eu/countries/.

Qin, Chengzhi, A-X Zhu, Tao Pei, Baoluo Li, Chenghu Zhou, and Lin Yang

2007 "An adaptive approach to selecting a flow-partition exponent for a multiple-flow-direction algorithm", *International Journal of Geographical Information Science*, 21, 4, pp. 443-458.

Quinlan, J Ross

1992 "C4. 5: programs for machine learning".

1986 "Induction of decision trees", *Machine learning*, 1, 1, pp. 81-106.

1987 "Simplifying decision trees", *International journal of man-machine studies*, 27, 3, pp. 221-234.

Raduła, Małgorzata W, Tomasz H Szymura, and Magdalena Szymura

2018 "Topographic wetness index explains soil moisture better than bioindication with Ellenberg's indicator values", *Ecological Indicators*, 85, pp. 172-179.

Rahmati, Omid, Fatemeh Falah, Kavina Shaanu Dayal, Ravinesh C Deo, Farnoush Mohammadi, Trent Biggs, Davoud Davoudi Moghaddam, Seyed Amir Naghibi, and Dieu Tien Bui

2020 "Machine learning approaches for spatial modeling of agricultural droughts in the south-east region of Queensland Australia", *Science of The Total Environment*, 699, p. 134230.

Rahmati, Omid, Mahdi Panahi, Zahra Kalantari, Elinaz Soltani, Fatemeh Falah, Kavina S Dayal, Farnoush Mohammadi, Ravinesh C Deo, John Tiefenbacher, and Dieu Tien Bui

2020 "Capability and robustness of novel hybridized models used for drought hazard modeling in southeast Queensland, Australia", *Science of The Total Environment*, 718, p. 134656.

Raible, Christoph C, Oliver Bärenbold, and Juan Jose Gomez-Navarro

2017 "Drought indices revisited–improving and testing of drought indices in a simulation of the last two millennia for Europe", *Tellus A: Dynamic Meteorology and Oceanography*, 69, 1, p. 1287492.

Rosbakh, Sergey, Annette Leingärtner, Bernhard Hoiss, Jochen Krauss, Ingolf Steffan-Dewenter, and Peter Poschlod

2017 "Contrasting effects of extreme drought and snowmelt patterns on mountain plants along an elevation gradient", *Frontiers in plant science*, 8, p. 1478.

Russell, Stuart and Peter Norvig

2002 *Artificial intelligence: a modern approach*, Pearson.

Santos, Fabián, Valerie Graw, and Santiago Bonilla

2019 "A geographically weighted random forest approach for evaluate forest change drivers in the Northern Ecuadorian Amazon", *PloS one*, 14, 12, e0226224.

Scherer, Thomas F, Larry Cihacek, and David Franzen

2017 "Soil, water and plant characteristics important to irrigation".

Segal, Mark R

2004 "Machine learning benchmarks and random forest regression".

Seibert, Jan and Brian L McGlynn

2007 "A new triangular multiple flow direction algorithm for computing upslope areas from gridded digital elevation models", *Water resources research*, 43, 4.

Sepulcre-Canto, G, SMAF Horion, A Singleton, H Carrao, and J Vogt

2012 "Development of a Combined Drought Indicator to detect agricultural drought in Europe", *Natural Hazards and Earth System Sciences*, 12, 11, pp. 3519-3531.

Shafer, BA and LE Dezman

1982 "Development of surface water supply index (SWSI) to assess the severity of drought condition in snowpack runoff areas", in Proceeding of the Western Snow Conference.

Shahid, Shamsuddin and Houshang Behrawan

2008 "Drought risk assessment in the western part of Bangladesh", *Natural hazards*, 46, 3, pp. 391-413.

Sharma, Gaganpreet

2017 "Pros and cons of different sampling techniques", *International journal of applied research*, 3, 7, pp. 749-752.

Sheffield, Justin and Eric F Wood

2012 *Drought: past problems and future scenarios*, Routledge.

Shekhar, Shashank and Arvind Chandra Pandey

2015 "Delineation of groundwater potential zone in hard rock terrain of India using remote sensing, geographical information system (GIS) and analytic hierarchy process (AHP) techniques", *Geocarto International*, 30, 4, pp. 402-421.

Shields, Michael D and Jiaxin Zhang

2016 "The generalization of Latin hypercube sampling", *Reliability Engineering & System Safety*, 148, pp. 96-108.

Shmueli, Galit et al.

2010 "To explain or to predict?", *Statistical science*, 25, 3, pp. 289-310.

Simpson, Edward H

1951 "The interpretation of interaction in contingency tables", *Journal of the Royal Statistical Society: Series B (Methodological)*, 13, 2, pp. 238-241.

Sivakumar, Mannava VK

2011 "Agricultural drought—WMO perspectives", in *Agricultural drought indices proceedings of an expert meeting*, pp. 24-39.

Strobl, Carolin, Anne-Laure Boulesteix, Thomas Kneib, Thomas Augustin, and Achim Zeileis

2008 "Conditional variable importance for random forests", *BMC bioinformatics*, 9, 1, p. 307.

Strobl, Carolin, Anne-Laure Boulesteix, Achim Zeileis, and Torsten Hothorn

2007 "Bias in random forest variable importance measures: Illustrations, sources and a solution", *BMC bioinformatics*, 8, 1, p. 25.

Sullivan, Preston

2000 *Drought resistant soils*, tech. rep., Appriopriate Technology Transfer for Rural Areas (ATTRA).

Svoboda, Mark, Brian Fuchs, et al.

2016 *Handbook of drought indicators and indices*, World Meteorological Organization (WMO).

Swidzinski, JF and K Chang

2000 "A novel nonlinear statistical modeling technique for microwave devices", in *2000 IEEE MTT-S International Microwave Symposium Digest (Cat. No. 00CH37017)*, IEEE, vol. 2, pp. 887-890.

Szalai, S and CS Szinell

2000 "Comparison of two drought indices for drought monitoring in Hungary—a case study", in *Drought and drought mitigation in Europe*, Springer, pp. 161-166.

Szalai, S, CS Szinell, and J Zoboki
2000 "Drought monitoring in Hungary", *Early warning systems for drought preparedness and drought management*, 57, pp. 182-199.

Tayman, Jeff and David A Swanson
1999 "On the validity of MAPE as a measure of population forecast accuracy", *Population Research and Policy Review*, 18, 4, pp. 299-322.

Telea, Alexandru C
2014 *Data visualization: principles and practice*, CRC Press.

Teweldebirhan Tsige, Dawit, Venkatesh Uddameri, Farhang Forghanparast, Elma Annette Hernandez, and Stephen Ekwaro-Osire
2019 "Comparison of meteorological-and agriculture-related drought indicators across ethiopia", *Water*, 11, 11, p. 2218.

Thomas, T, RK Jaiswal, Ravi Galkate, PC Nayak, and NC Ghosh
2016 "Drought indicators-based integrated assessment of drought vulnerability: a case study of Bundelkhand droughts in central India", *Natural Hazards*, 81, 3, pp. 1627-1652.

Thompson, S.K.
2012 *Sampling*, Wiley Series in Probability and Statistics, Wiley, ISBN: 9780471540458.

Thornthwaite, Charles Warren
1948 "An approach toward a rational classification of climate", *Geographical review*, 38, 1, pp. 55-94.

Tian, Liyan, Shanshui Yuan, and Steven M Quiring
2018 "Evaluation of six indices for monitoring agricultural drought in the south-central United States", *Agricultural and forest meteorology*, 249, pp. 107-119.

Tirado, Reyes and Janet Cotter
2010 "Ecological farming: Drought-resistant agriculture", *Exeter, UK: Greenpeace Research Laboratories*.

Tøttrup, Christian and Kamp Mikael Sørensen
2014 *EU-DEM Statistical Validation*, tech. rep., European Environment Agency.

Trnka, M, M Dubrovský, M Svoboda, D Semerádová, M Hayes, Z Žalud, and D Wilhite
2009 "Developing a regional drought climatology for the Czech Republic", *International Journal of Climatology: A Journal of the Royal Meteorological Society*, 29, 6, pp. 863-883.

Trnka, Miroslav, Petr Hlavinka, Daniela Semerádová, Jan Balek, Martin Možní, P stěpánek, P Zahraíček, Michael Hayes, Josef Eitzinger, and Zdeněk Žalud
2014 "Drought monitor for the Czech Republic-www. intersucho. cz", *Mendel and Bioclimatology*, pp. 630-638.

Trnka, Miroslav, Daniela Semerádová, Ivan Novotní, Miroslav Dumbrovskí, Karel Drbal, Franti ek Pavlıý k, Jan Vopravil, Adam Vizina, Jan Balek, Petr Hlavinka, et al.

2016 "Assessing the combined hazards of drought, soil erosion and local flooding on agricultural land: a Czech case study", *Climate Research*, 70, 2-3, pp. 231-249.

Tucker, CJ, WW Newcomb, SO Los, and SD Prince

1991 "Mean and inter-year variation of growing-season normalized difference vegetation index for the Sahel 1981-1989", *International Journal of Remote Sensing*, 12, 6, pp. 1133-1135.

Van Rijn, Jan N and Frank Hutter

2018 "Hyperparameter importance across datasets", in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 2367-2376.

Vásquez-León, Marcela, Colin Thor West, and Timothy J Finan

2003 "A comparative assessment of climate vulnerability: agriculture and ranching on both sides of the US–Mexico border", *Global Environmental Change*, 13, 3, pp. 159-173.

Vicente-Serrano, Sergio M, Santiago Begueriýa, and Juan I López-Moreno

2010 "A multiscalar drought index sensitive to global warming: the standardized precipitation evapotranspiration index", *Journal of climate*, 23, 7, pp. 1696-1718.

Wagner, Wolfgang, Guido Lemoine, and Helmut Rott

1999 "A method for estimating soil moisture from ERS scatterometer and soil data", *Remote sensing of environment*, 70, 2, pp. 191-207.

Wagner, Wolfgang, Klaus Scipal, Carsten Pathe, Dieter Gerten, Wolfgang Lucht, and Bruno Rudolf

2003 "Evaluation of the agreement between the first global remotely sensed soil moisture data with model and precipitation data", *Journal of Geophysical Research: Atmospheres*, 108, D19.

Wallemacq, Pascaline

2018 *Economic Losses, Poverty & Disasters 1998–2017*, tech. rep., Centre for Research on the Epidemiology of Disasters (CRED).

Weatherhead, EK and NJK Howden

2009 "The relationship between land use and surface water resources in the UK", *Land use policy*, 26, S243-S250.

Wilhelmi, Olga V and Donald A Wilhite

2002 "Assessing vulnerability to agricultural drought: a Nebraska case study", *Natural Hazards*, 25, 1, pp. 37-58.

Wilhite, A. Donald

2011 "Quantification of Agricultural Drought for Effective Drought Mitigation and Preparedness: Key Issues and Challenges", in *Agricultural Drought Indices: Proceedings of an Expert Meeting*, pp. 13-23.

Wilhite A, Donald

2000 "Drought as a Natural Hazard: Concepts and Definitions", in *Drought: A Global Assessment*, Routledge, pp. 3-18.

Wilhite, Donald A and Michael H Glantz

1985 "Understanding: the drought phenomenon: the role of definitions", *Water international*, 10, 3, pp. 111-120.

Willmott, Cort J, Kenji Matsuura, and Scott M Robeson

2009 "Ambiguities inherent in sums-of-squares-based error statistics", *Atmospheric Environment*, 43, 3, pp. 749-752.

Yang, Mingxia, Yuling Mou, Yanrong Meng, Shan Liu, Changhui Peng, and Xiaolu Zhou

2020 "Modeling the effects of precipitation and temperature patterns on agricultural drought in China from 1949 to 2015", *Science of the Total Environment*, 711, p. 135139.

Yevjevich, Vujica M

1967 *Objective approach to definitions and investigations of continental hydrologic droughts, An*, PhD thesis, Colorado State University. Libraries.

Zargar, Amin, Rehan Sadiq, Bahman Naser, and Faisal I Khan

2011 "A review of drought indices", *Environmental Reviews*, 19, NA, pp. 333-349.

Zeng, Zhaoqi, Wenxiang Wu, Zhaolei Li, Yang Zhou, Yahui Guo, and Han Huang

2019 "Agricultural drought risk assessment in Southwest China", *Water*, 11, 5, p. 1064.

Zhang, Dan, Guoli Wang, and Huicheng Zhou

2011 "Assessment on agricultural drought risk based on variable fuzzy sets model", *Chinese Geographical Science*, 21, 2, p. 167.

Zhao, Zhongqiu, Yunlong Cai, Meichen Fu, and Zhongke Bai

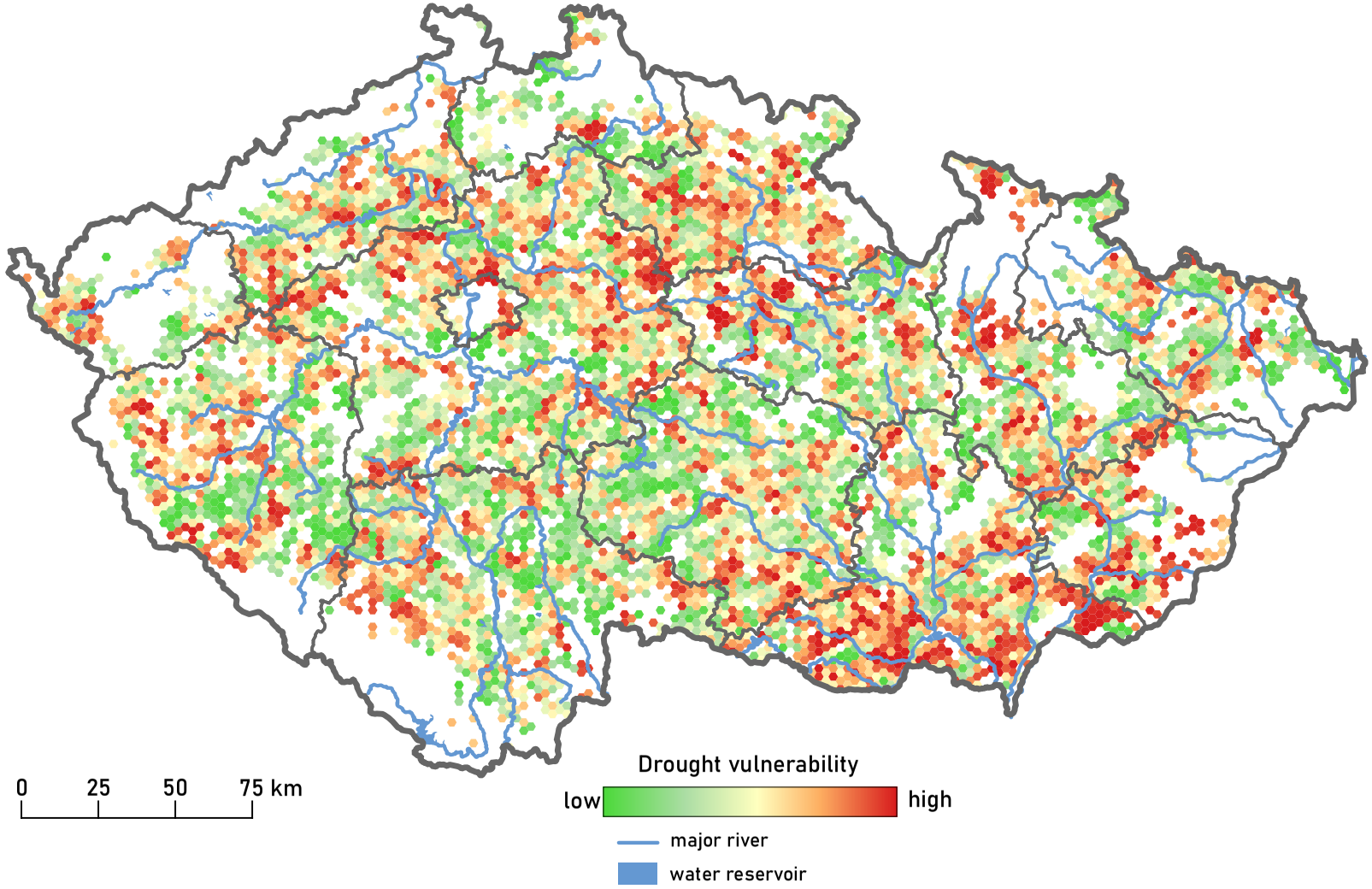
2008 "Response of the soils of different land use types to drought: eco-physiological characteristics of plants grown on the soils by pot experiment", *ecological engineering*, 34, 3, pp. 215-222.

APPENDIX

Appendix 1: Environmental vulnerability to drought hazard.

Environmental vulnerability to drought hazard

Weighted by mean precipitation and temperature



Daniel Bicák, 2021
datasources: ArcČR® 2
coordinate system: ETRS89-LAEA

Appendix 1: Environmental vulnerability to drought hazard.