

Palacký University Olomouc  
Faculty of Science  
Department of Geoinformatics



Study program: **P1301 Geography**  
Specialization: **Geoinformatics and Cartography**

# **Advanced methods for landslide assessment using GIS**

PhD Thesis

Candidate: MSc Miloš MARJANOVIĆ

Supervisor: Prof. RNDr. Vít VOŽENÍLEK, CSc.

Olomouc 2013.

*To my family*

## **Author's Statement**

I declare that this PhD thesis of P1301 Geography study program has been completed independently, under the supervision of Prof. RNDr. Vít Voženílek, CSc. All the materials and resources are cited with regard to the scientific ethics, copyrights and laws protecting intellectual property. All provided and created digital data will not be publicly disposed without the consent of the Department for Geoinformatics, Faculty of Science, Palacký University Olomouc.

In Olomouc, 15. April, 2013.

---

## Acknowledgement

Firstly, I would like to credit the people who have directly supported my research and the completion of the thesis in a timely manner. I would like to thank my supervisor, examiner and research associate, Prof. Dr Vít Voženílek, for his continual support and guidance, and my research associates Prof. Dr Branislav Bajat, and Assoc. Prof. Dr Miloš Kovačević, for their cooperation and help throughout the research. I would further like to thank the people who have helped my research indirectly, with their advices, discussions and tutoring through the landslide assessment and other related topics: Assoc. Prof. Dr Biljana Abolmasov and her team, Dr Fausto Guzzetti, and his IRPI team, Prof. Dr Leonardo Cascini and his LARAM team, all my colleagues from the Department of Geoinformatics at Palacký University Olomouc. Special acknowledgement goes to those kind people by whose courtesy I have been able to access the data from different study areas: Prof. Dr Rada Pavlović and Prof. Dr Branislav Trivić, Prof. Dr Snježana Mihalić, and once again Prof. Dr Vít Voženílek. I owe another acknowledgement to Prof. Dr Vít Voženílek, and especially to my sister Dr Milena Marjanović for peer-reviewing the text and helping me to reach a more correct and more comprehensive expression of the thesis. For the same reason, only regarding the text written in Czech I have to thank MSc Lukáš Marek, MSc Marea Grinvald and my girlfriend MSc Bojana Božanić. Finally, I would like to thank my closest ones for their unconditional understanding, support and encouragement throughout my PhD studies – Mirjana, Dragiša, Milena and Bojana, thank you!



# Table of Content

<b>1 Introduction</b> .....	1
<b>2 Objectives</b> .....	4
<b>3 Theoretical Background</b> .....	5
3.1 Landslide Phenomenology – Definitions and Scope .....	6
3.2 Susceptibility, Hazard and Risk – Definitions and Scope .....	10
3.3 Landslide Assessment Concepts, Principles and Problematic .....	11
3.3.1 Data Acquisition Issues in Landslide Assessment .....	12
3.3.2 Modeling Approach Issues in Landslide Assessment .....	15
3.3.3 GIS Issues in Landslide Assessment.....	17
3.3.4 Other Issues in Landslide Assessment.....	18
<b>4 Related Work</b> .....	21
<b>5 Methods and Procedures</b> .....	28
5.1.1 Attribute Selection Methods.....	28
5.1.2 Chi-Square .....	30
5.1.3 Information Gain (IG).....	30
5.2 Landslide Assessment Methods .....	31
5.2.1 Heuristic Approach .....	31
5.2.2 Statistical Approach.....	33
5.2.3 Machine Learning Approach.....	39
5.2.4 Deterministic Approach .....	49
5.3 Model Evaluation Methods .....	52
5.3.1 Kappa Statistics.....	53
5.3.2 Receiver Operating Characteristics (ROC) .....	53
5.4 Research Workflow .....	55
5.5 Data and Software Specifications for Different Methods .....	55
<b>6 Case Studies</b> .....	58
6.1 Fruška Gora Mountain (Serbia) .....	59
6.1.1 Setting.....	60
6.1.2 Dataset.....	61
6.1.3 Implementation, Results and Discussion .....	73
6.2 Starča Basin (Croatia) .....	92
6.2.1 Setting.....	92
6.2.2 Data .....	96
6.2.3 Implementation, Results and Discussion .....	99
6.3 Halenkovice Area (Czech Republic) .....	105

6.3.1 Setting.....	106
6.3.2 Data .....	108
6.3.3 Implementation, Results and Discussion .....	115
<b>7 Main Achievements.....</b>	<b>121</b>
<b>8 General Discussion.....</b>	<b>123</b>
<b>9 Conclusions.....</b>	<b>125</b>
9.1 Benefits and Drawbacks.....	125
9.2 Applicability .....	126
9.3 For Further Notice .....	126
<b>10 References.....</b>	<b>128</b>
<b>Complete List of Author's Publications</b>	
<b>List of Figures</b>	
<b>List of Tables</b>	
<b>Appendices</b>	
<b>Shrnutí</b>	

# 1 Introduction

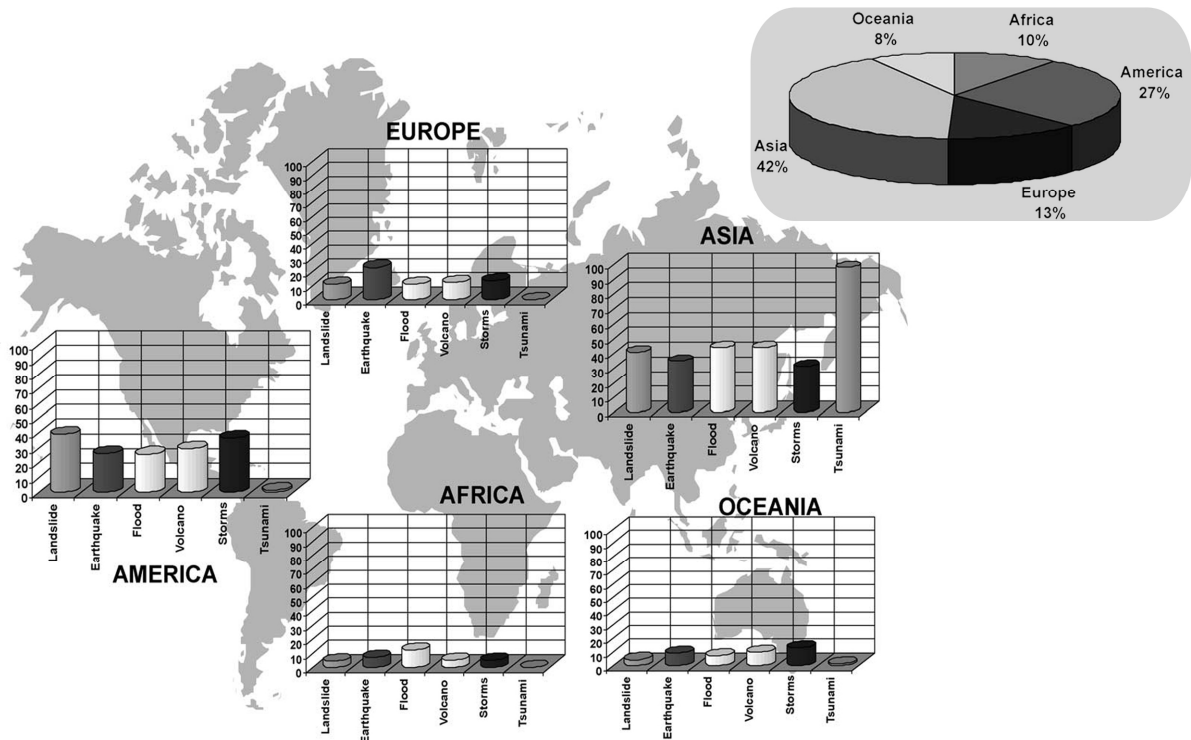
In the world of growing needs for land urbanization and exploitation of resources, yet with raising concerns on unstable climate conditions, the matters of environmental safety come under the spotlight. As the world population condenses by over-settling existing areas or settling new unpopulated areas, bigger and bigger volumes of geological environment become disturbed, encountered biomass suffers drastic reduction, while microclimatic effects amplify (Turner & Shuster 1996, Maskrey et al. 2009), which all allows various influences to embark and jeopardize the subjected population. In such circumstances, natural hazards stand among the most threatening influences to the safety of human lives and property. Different natural hazards unfold with different scenarios, causing various disasters, but more dangerously, they tend to combine and chain-up their effects. Thus, they all superimpose, making their separate influences and outcomes indistinct, which leaves their authenticity unclear (Bell 1999). For instance, an earthquake near the coastline can generate a tsunami wave, and they both can then trigger landslides, cause subsidence, storms, or provoke technological and other types of hazards, and in turn, completely devastate an area (like in a tragic episode that stroke Japan in March of 2011). In the aftermath of such a scenario it is difficult to separate effect of each hazard individually. Holistic (general) approach to the natural hazard and risk is therefore the only reasonable solution, but for now, it is an initiative to strive toward rather than a feasible praxis (Lee & Jones 2004). In order to have at least some idea on the overall hazard and risk distribution, a global and regional view on the natural hazard is yet necessary, and it is traditionally interpreted by plain statistical analysis of the historical records (Alcántara-Ayala 2002). It is apparent that analysis of longer time intervals brings inconsistency in the data acquisition (awareness and attention to the phenomena has changed over time, as well as standards and methodology of the data acquisition), so the figures and percentages are only portraying the principle trends. In such quantification, the most common natural hazards include earthquakes, volcanoes, floods, storms and landslides/mass movements (Bell 1999). The latter, with their 5% share (Fig. 1) in the total tolls<sup>1</sup> of natural disasters, are to be regarded hereinafter.

Indisputably, the landslide hazard awareness reached considerable quota in recent years. It involves not just the general population, but the academic circles and political officials, as well. Google's insight for search (covering 2004 – present time span) indicates considerable average ascent in interest for keywords such as landslide, debris flow, landslide hazard and susceptibility. The interest is particularly high right after the events that had caused considerable damage to society and that have been followed-up by media. Consequently, the interest is the highest in the affected areas (Fig. 2). These speculations are relative and disregard the influence of technological literacy and capability (availability of computer/internet configurations). The insight that one can track within the academic circles is by far more objective, and it is based on the publication activity records (Gokceoglu & Sezer 2009). Scientific and research teams have shown rising interest in the landslide-related topics since the late '80s. This abrupt and exponential (Gokceoglu & Sezer 2009) increase in activity has resulted in more than 150 scientific articles per year. Widening of the problematic, technological innovations and holistic approaches to the solution are promising prolific activity within research community in the future. Numbers of scholars per publication is also rising, presumably indicating the growth within the multidisciplinary teams that have been dealing with such matter. Records also show that researchers follow geographical distribution of the events (Nadim et al. 2006), meaning that they also mostly come from the areas affected by the landslides (Circum Pacific Region, mountainous regions in the Alps,

---

<sup>1</sup> According to some recent research results there has been nearly 90000 landslide casualties worldwide in the past decade (9000 per year), which is much more than 5% of global natural hazard toll. The research has been done by using Google News services (Petley 2012).

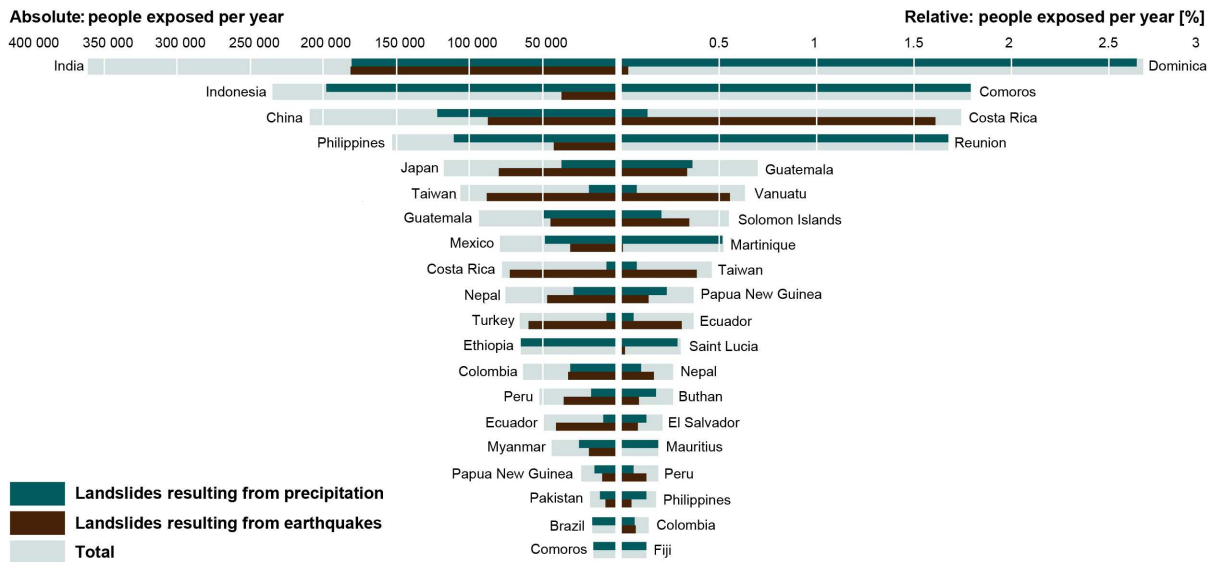
the Himalaya, and other volcanic and seismic areas worldwide). However, they show that the most of the researchers and the most significant researchers come from just a few countries: Italy, USA, Canada, UK, China, France, Japan and Spain (in the respective order) (Gokceoglu & Sezer 2009, Chacón et al. 2006, Petley 2012). The rest of the world is left with a few research teams and individuals to cope with the problematic. In such context, the problem is treated singlehandedly, and although it represents one of the most complex, versatile and the most wide-spread natural phenomena (Varnes 1984, Chacón et al. 2006, Gokceoglu & Sezer 2009, Guzzetti et al. 2012) and although the researchers' curiosity has an exponential response thus far, the number of studies, contributions and researchers has yet to grow. Only then the problematic of the landslides and the hazard they produce will be fully understood.



**Figure 1.** Global (per continent) distribution of different hazard types. The inset to the right shows total hazard distributions per continent (after Alcántara-Ayala 2002).

These global trends were thus the principal motifs for this research: inevitable need for broadening of the landslide hazard researchers pool, fact that landslides affect the society more frequently and more broadly than before, rising awareness to the problematic among the planner/decision-maker pools, availability of the advanced methodology and technology to remediate and monitor the landslides, opening funds for regional research projects. This study shows particular interest in the regional type of studies, due to their applicability on one hand, and scientific contribution on the other. It is further quite appreciable to work with such problematic in a Geographic Information System (GIS) environment, which allows various numerical, statistical or heuristic implementations to be conducted relatively easily. It was actually expected that the researchers will incline toward regional studies in nearer future, since such studies directly contribute to the landslide hazard mitigation (Gokceoglu & Sezer 2009, Brenning 2012), especially if they represent systematical comparisons of multiple modeling approaches and techniques (Brenning 2005, Yilmaz 2009). It was also expected to have various combinations of slope stability modeling, monitoring and landslide hazard modeling, and development of the Early-Warning Systems. These prognoses turned quite reasonable if one looks at the contribution lists and contents of the most recent, major landslide forums, congresses and symposia, organized by the leading communities in the field (International Association for Engineering Geology – IAEG, International Consortium on

Landslides – ICL, International Society of Soil Mechanics and Geotechnical Engineering – ISSMGE, International Society of Rock Mechanics – ISRM). Finally, the researchers at present also experiment with the incorporation of the landslide hazard into a holistic hazard assessment, and will continue to strive toward that final goal in the future, but this matter brings about complexity and compatibility issues and requires simultaneous development of all natural hazard branches in analogue or similar frameworks.



**Figure 2.** People exposure to landslides (app. 2.2 million per year). Note that small island countries have high relative exposure, while in the absolute exposure Asian countries – India, Indonesia, China and others dominate (after Maskrey et al. 2009).

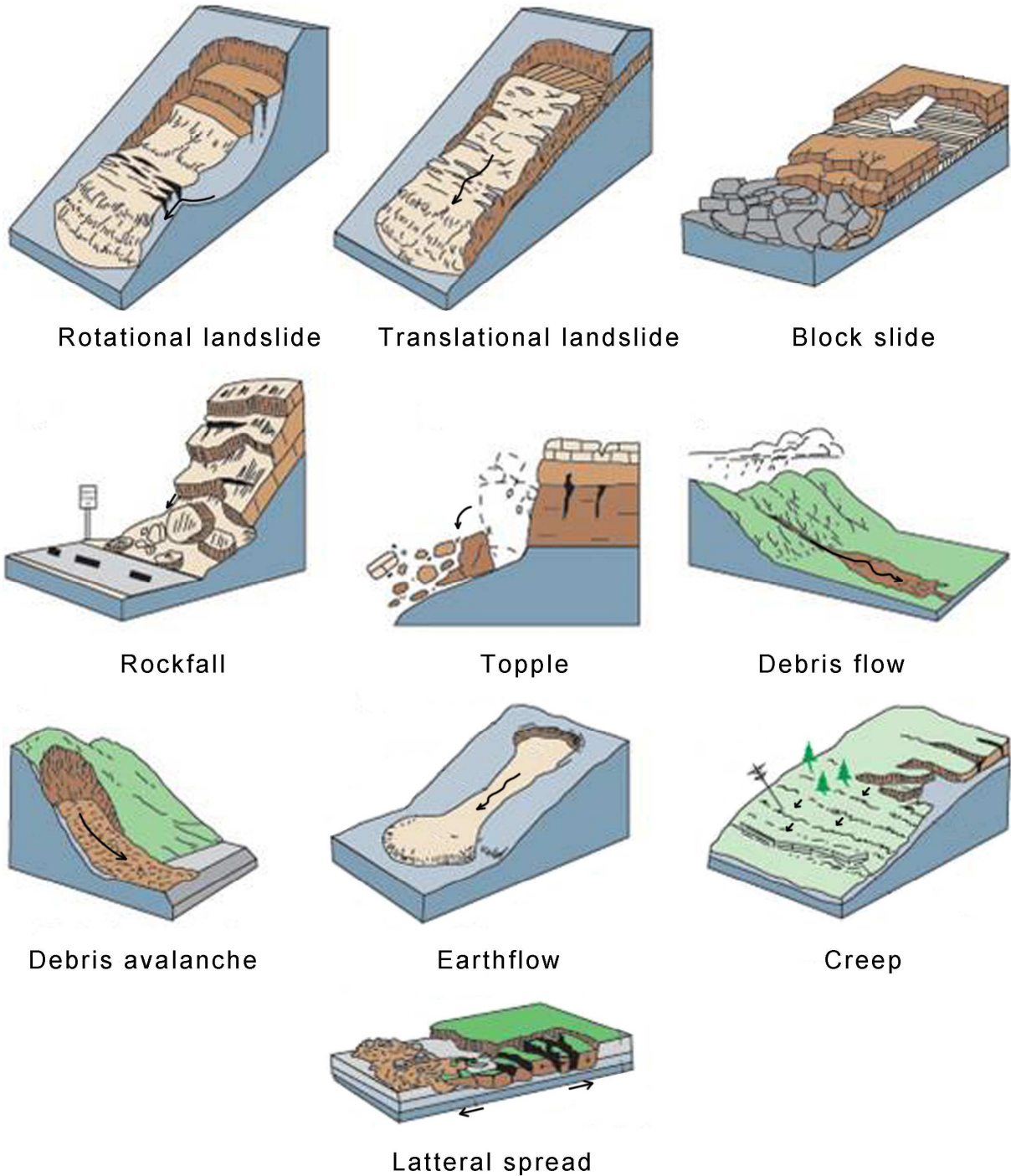
## **2 Objectives**

Resting on the abovementioned motifs, this research was shaped to meet the standardized requirements (Varnes 1984, WP/WLI 1995, Fell et al. 2008, Lynn & Bobrowsky 2008, Gerath et al. 2010, Brenning 2012) in terms of methodology of data acquisition and manipulation, choices of the advanced modeling approaches for landslide assessment, as well as the model evaluation techniques, and finally, the visualization choices, all via GIS. These objectives could be structured as follows:

1. Exploiting only low-cost data resources (available or open-source topographic, geological, satellite imagery and other repositories) and open source software packages.
2. Inspecting of the phenomena from different case-studies, including similar, but sufficiently different terrains (in order to compare the modeling results and test the capabilities of proposed methodological solutions).
3. Standardizing the data acquisition regarding the data type, scale, preprocessing procedures and so forth (in order to have fully comparable models from different case-studies) using GIS.
4. Implementing a variety of well-known modeling approaches, but also experimenting with the state-of-the-art techniques, advanced methods and unprecedented solutions for landslide assessment using GIS. Resulting models are to present transient relative values over the area, pinpointing landslide-endangered zones and safe zones (which shall be further elaborated).
5. Evaluating the results, i.e. the models performance in the most appropriate fashion, obtaining qualitative and quantitative descriptors of the models performance using GIS in combination with statistical tools.
6. Visualizing and publishing the results in the form of generic maps per each case-study using GIS, and web-GIS and estimating their applicability.

### 3 Theoretical Background

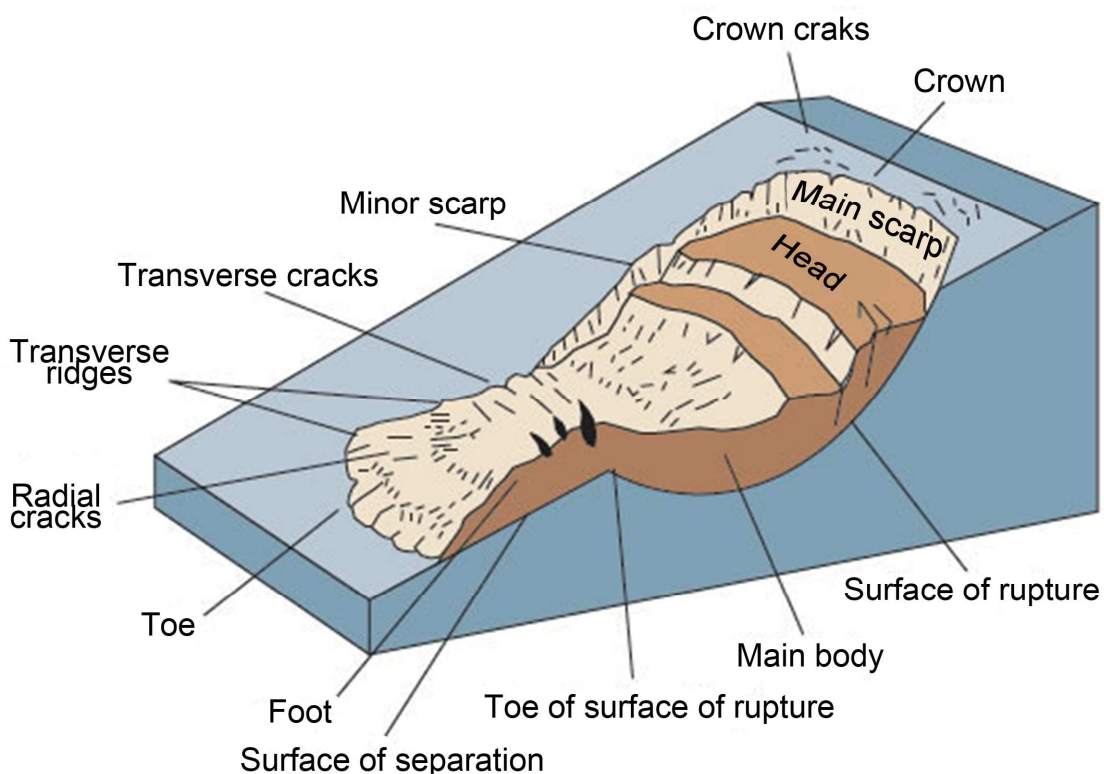
In order to present the problematic of this thesis systematically, it is first necessary to define and communicate the basic theoretical background behind the landslide phenomenology, comprehension of qualitative landslide assessment, impact of available technology which is in service of landslide assessment and the way in which GIS is enrolled in it.



**Figure 3.** A simplified illustrative landslide classification after Varnes (after Lynn & Bobrowsky 2008).

### 3.1 Landslide Phenomenology – Definitions and Scope

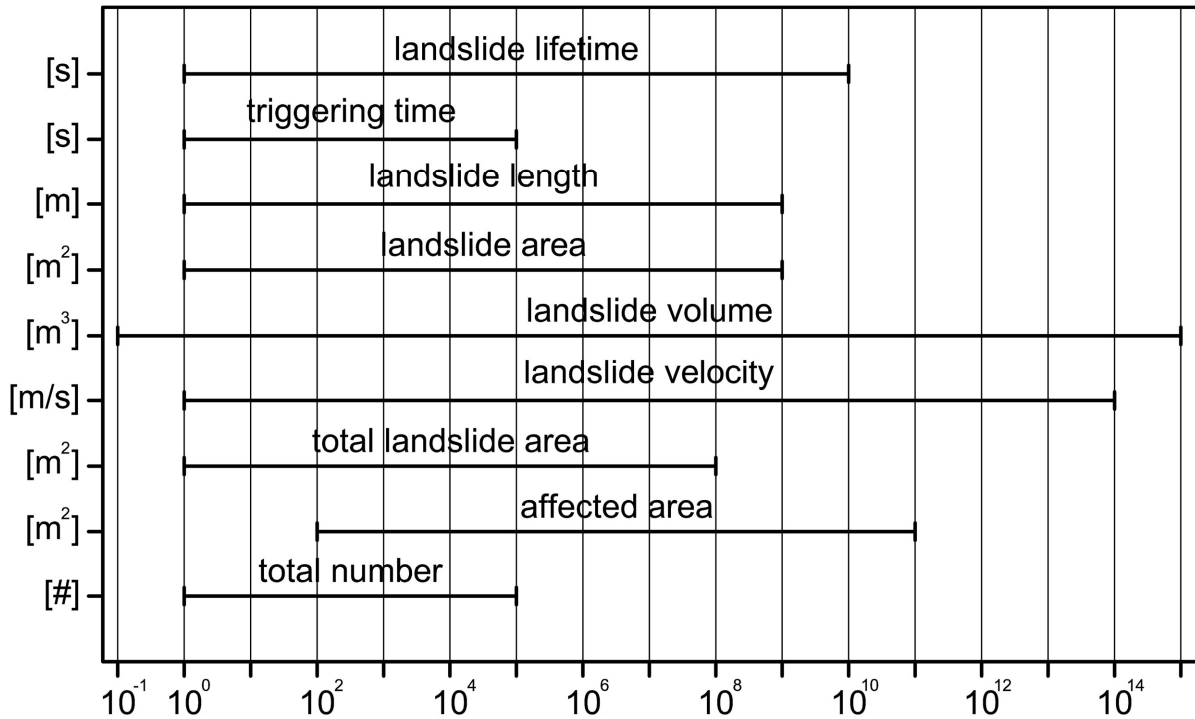
As in many cases of terminological disputes, landslide as a term has endured various interpretations, as the scientific disciplines that treat it changed and developed over time. It is also the matter of different conception of the term by different research-schools worldwide. In regard with the latter, landslide can express more specific or more general phenomenon. It seems logical that more general definitions broaden the phenomena, making it more complex to understand and offer solutions. Herein, such debates will not be of particular interest (although they are further influential to selection of the modeling techniques or modeling approaches for instance), since abundant information on that topic could be found elsewhere (Lee & Jones 2004, Chacón et al. 2006). Hereinafter, one of the broadest definitions and classifications endorsed by the leading communities and consortiums (Varnes 1984) is adopted (Fig. 3). The following paragraphs define the main terms and principles of the landslide phenomenology rather informally, in order to introduce the main problematic of this research.



**Figure 4.** Landslide elements (after Lynn & Bobrowsky 2008).

**Landslides** are downward movements of rock, debris or earth masses, usually developed along predefined planar discontinuities. These are called slip-surfaces (simple planar or higher order – complex surfaces), which propagate throughout the mass and clearly separate intact bedrock material from the moved material above. Other (morphological) elements of a landslide include crown and head, separated by a scarp; main body, channeled by flanks; foot, terminated by a toe; depletion zone capturing upper and accumulation zone capturing lower portions of a landslide (Fig. 4) Landslides can drastically vary in size and area, as well in some other measurements (Fig. 5). They can develop in natural or engineered/constructed slopes. Consequently, they have been studied in various scientific branches ranging from Geology and Geomorphology to Geological, Geotechnical and Civil Engineering, from a variety of aspects (Varnes 1984, Lee & Jones 2004, Bell 2007).





**Figure 5.** Spans of different measures of landslides (after Guzzetti et al. 2012). Note how broad and variable the orders of measures are, making landslides very diverse and very complex phenomena.

**Slope stability** rests on the equilibrium between the forces that act upon a slope. Displacement takes place when the resisting forces are succumbed by the driving forces, which in turn generates irreversible change to the slope. The former, resisting forces, are represented by shear strength and cohesion of the material, as well as friction along a slip surface, which all further depend on the nature and condition of the slope material (freshness – weathering degree; structural elements – presence of joints and fissures; heterogeneity – contrasts of water permeability or deformability; presence/absence of vegetation), as well as on the slope morphology/geometry (steepness, elevation, curvature etc.). Driving forces on the other hand, usually involve: increase of weight or shear stress (via water saturation, adding load and rearranging of the slope geometry), loss of support (via erosion and rearranging of the slope geometry) or dynamic influences. The features that influence driving and resisting forces and their balance are commonly called **Conditioning Factors**. In regional scales these are different geological, geomorphological and environmental properties of the ground. In other words, Conditioning Factors are providing the background of the landslide occurrence. Once the terms are reached, the process unfolds under the influence of different **Triggering Factors** or in their combination. The most typical Triggering Factors are: earthquake, volcanic eruption, intense rainfall, abrupt groundwater regime change, flood, rapid snowmelt, successive erosion and human intervention (Lee & Jones 2004, Bell 2007).

**Landslide activity** is another important parameter, which requires attention. Since landslides develop progressively and cyclically once they enter the process, it is important to estimate the state of their current activity in order to scale the future displacement rates. In particular, relative displacements are the highest during the initial activation, and decrease per each reactivation cycle, but the frequency of the events increases as a landslide progresses toward an active stage. Stage of the first failure is followed by the stage of reactivation, which are separated by suspended and dormant stages and this repeats per every cycle until the active stage is reached (Fig. 6).

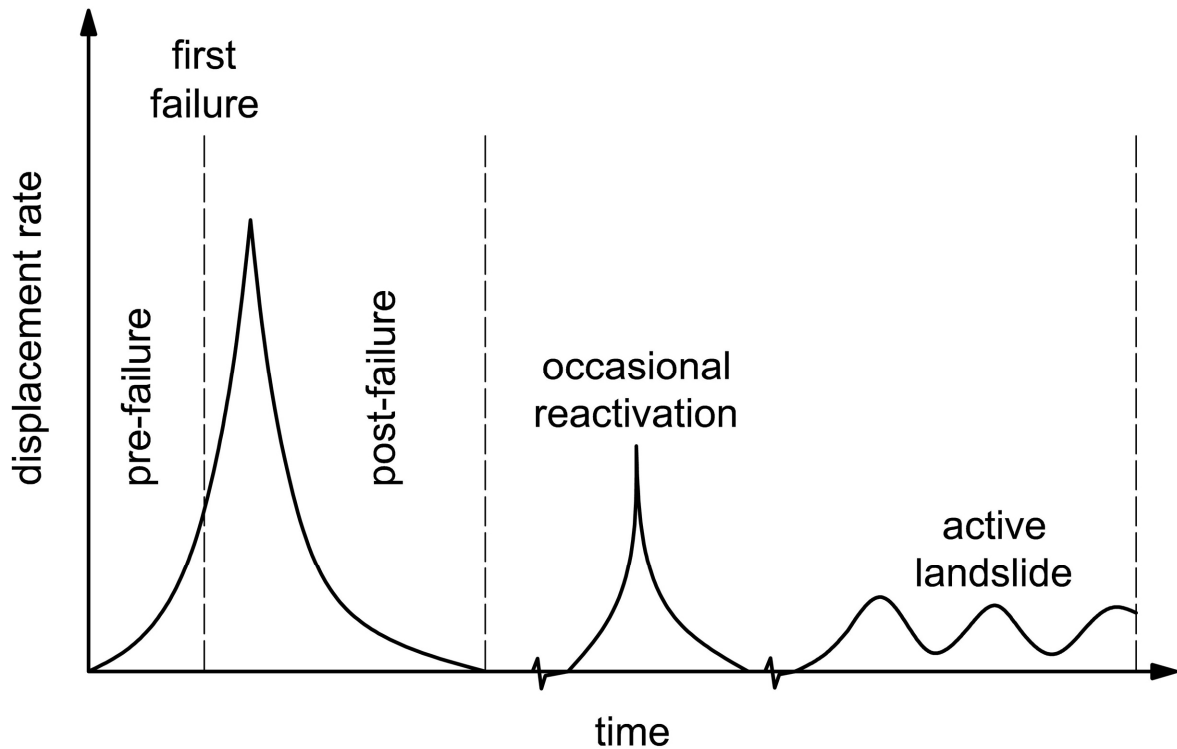
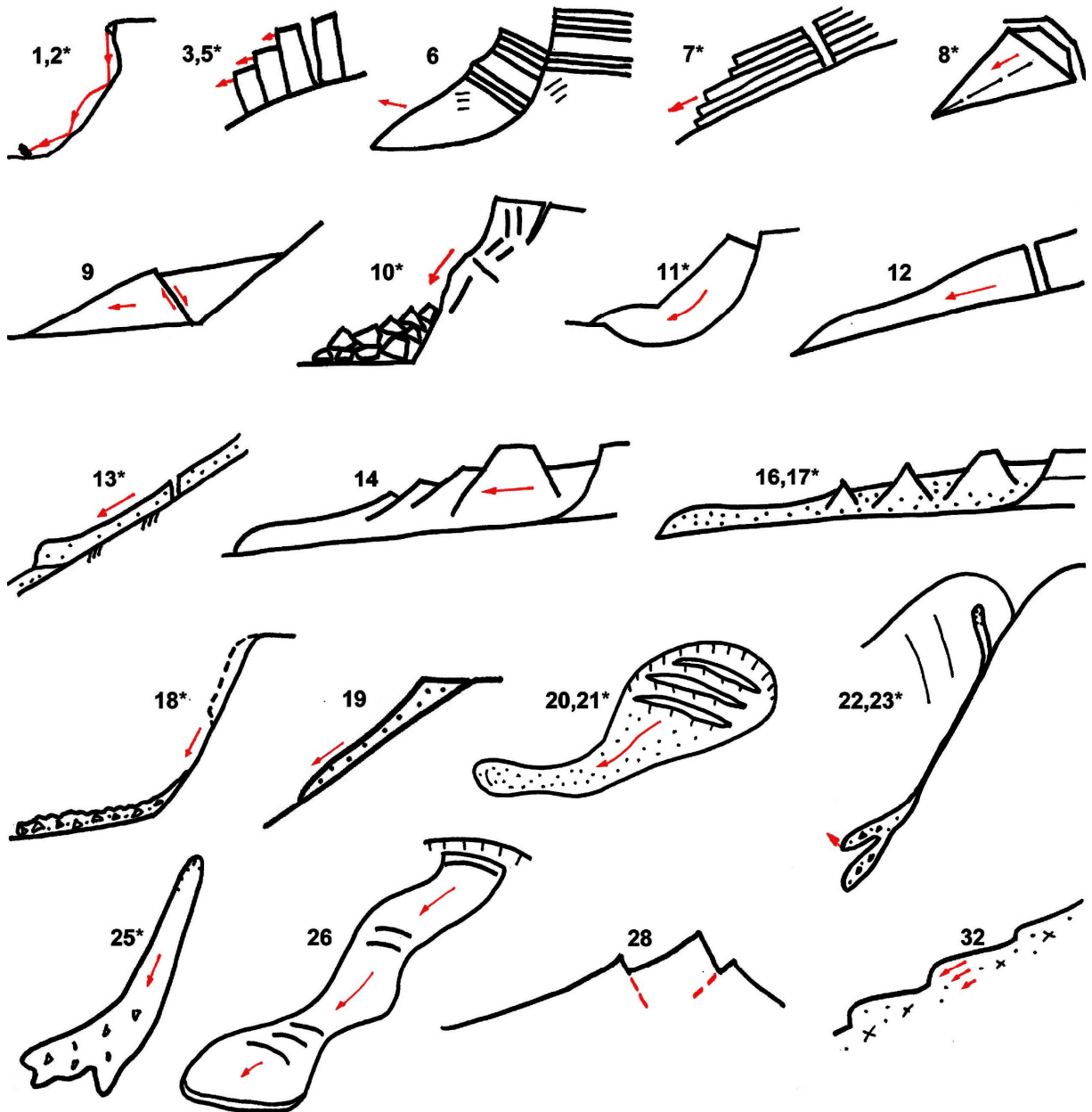


Figure 6. Landslide activity stages (after Leroueil et al. 1996).

Table 1. Updated Varnes landslide classification (after Hungr et al. 2012).

type of movement	rock	soil
fall	1 <sup>*</sup> rock fall	2 <sup>*</sup> boulder/debris/silt fall
topple	3 <sup>*</sup> rock block topple 4 rock flexural topple	5 <sup>*</sup> gravel/sand/silt topple
slide	6 rock rotational slide 7 <sup>*</sup> rock translational slide 8 <sup>*</sup> wedge slide 9 rock compound slide 10 <sup>*</sup> rock collapse	11 clay/silt rotational slide 12 clay/silt translational slide 13 <sup>*</sup> gravel/sand/debris slide 14 clay/silt compound slide
spread	15 rock slope spread	16 <sup>*</sup> sand/silt liquefaction spread 17 <sup>*</sup> sensitive clay spread
flow	18 <sup>*</sup> rock avalanche	19 sand/silt/debris dry flow 20 <sup>*</sup> sand/silt/debris flow slide 21 <sup>*</sup> sensitive clay flow slide 22 <sup>*</sup> debris flow 23 <sup>*</sup> mud flow 24 debris flood 25 <sup>*</sup> debris avalanche 26 earth flow 27 peat flow
slope deformation	28 mountain slope deformation 29 rock slope deformation	30 soil slope deformation 31 soil creep 32 solifluction



**Figure 7.** Schematic illustration of landslide types according to the updated Varnes landslide classification given in Table 1. (some of the examples are missing). Asterisks indicate high velocity movements. Scale of the illustrations varies (after Hungr et al. 2012).

**Displacement mechanism** can further define the landslide movement typology as follows: *fall, topple, slide, flow, spread and composite* (creeps are sometimes another separate category) (Fig. 3, 7). Together with the information on the type of the moved material it makes a basis for conventional classifications. Moreover, the **landslide velocity**, as another important landslide descriptor, is also governed by material type and movement mechanism, and can vary from extremely slow (mm per year in creep) through extremely rapid (m per second in debris flows). Displacement mechanisms have characteristic behavior which will not be described in detail herein, but can be found well elaborated in work of Varnes (1984), who basically founded the terminology that is later to be adopted by the international community (Turner & Shuster 1996).

Described features represent the base for development of the modified **Varnes classification** (Hungr et al. 2012) (Tab. 1, Fig. 7). Thus, every landslide could be classified

in accordance with this system by combining principally material and movement type, complemented with the estimation of the activity state and velocity. However, there are exceptions, which make this system more complex (this is the reason behind which the local classifications are occasionally preferred), and encourage its further refinement, since it suffers from a certain simplification and subjectivity, just as any other classification system (Guzzetti et al. 2012).

### 3.2 Susceptibility, Hazard and Risk – Definitions and Scope

Hitherto, terms hazard and risk might have been used in their broader, intuitive meaning, spoken as of something that poses a danger that come from a certain natural phenomenon. Following paragraphs are dedicated to elaborate and articulate their meaning in the analytical, quantitative framework, which is consistent with the internationally approved terminology of Geo-Engineering communities. From this chapter on, their articulation will be used only as such.

**Landslide susceptibility** ( $M$ ) stands for the spatial distribution and magnitude estimation of landslides which exist or may potentially occur over an area. It could also be treated as a pure spatial probability of landslide occurrence (Fell et al. 2008). Landslide magnitude can be expressed by means of total area, volume or velocity (if applicable) of a landslide. Although it is intuitive that more susceptible slopes will be affected more frequently than less susceptible ones, susceptibility remains explicitly in a spatial frame, with no temporal component. Terms landslide potential, sensitivity, relative hazard, total landslide density and likely frequency partly match term landslide susceptibility, but have not been used in suffice (Lee & Jones 2004, Chacón et al. 2006).

**Landslide hazard** ( $H$ ) stands for a probability of damaging landslide occurrence over an area within a given time period (temporal probability –  $p_t$ ). It could be regarded as a temporal extension of susceptibility. It is sometimes confused with susceptibility, but it is sufficient to notice its temporal dimension to make a distinction. Actually, susceptibility could be regarded as a special case of hazard that has a single temporal perspective instead of a time series (Einstein 1988, Lee & Jones 2004). In a broader sense, hazard is founded on the estimates of the landslide magnitude (area or volume) and probability of its recurrence (Eq. 1) (Fell 1994).

$$H = M \cdot p_t \quad (1)$$

**Element at Risk** ( $ER$ ) is any entity (any component of the terrain) which is potentially affected by a damaging phenomenon. It involves population, objects of personal property (real-estate and movables), engineering works and infrastructure, economic activities, public services and environmental valuables. While susceptibility and hazard analysis are not influenced by the choice of Element at Risk, Risk itself is, and could be separated in different categories according to the chosen element (Lee & Jones 2004, Fell et al. 2008).

**Vulnerability** ( $V(ER)$ ) is denoted as a degree of loss of an Element at Risk within the affected area. It could be also interpreted as a measure of exposure toward the hazard or potential to suffer damage. It can vary spatially, temporally and individually, hence according subtypes of vulnerability could be derived. For instance, hospitalized persons of a nursing home for the elderly which is directly facing a landslide would have greater total vulnerability than workers in a factory nearby, outside the landslide accumulation zone (temporal exposure of the workers is reduced to only several working hours per day, spatial vulnerability is reduced because the factory is not directly facing a landslide, workers are more agile and vital, so their individual capability to survive damaging event is greater, thus the individual vulnerability is smaller) (Fell 1994, Lee & Jones 2004, Fell et al. 2008).

$$R = H \cdot V(ER) \quad (2)$$

**Risk** ( $R$ ) is formulated (Eq. 2) as a measure of landslide occurrence probability and severity of its effects. Risk comprehension is somewhat difficult, due to the linguistic flexibility of the term itself (and intuitive similarities with terms susceptibility, hazard, vulnerability) and the fact that it resides in the future. It turns difficult to concept, especially to decision-makers who need to act upon risk estimations in advance, before disasters strike. As aforementioned, several risk categories can be segregated, given a different Element at Risk, i.e. societal risk, individual risk, group risk, but also categories derived particularly for decision-making, such as acceptable and tolerable risk (Fell 1994, Lee & Jones 2004).

It is pertinent to accept this internationally approved terminology for quantitative risk assessment in order to avoid any misunderstandings. In this thesis, as well as in the author's preceding work (Marjanović 2009, Marjanović et al. 2009, Marjanović 2010a, Marjanović 2010b, Marjanović & Čaha 2011, Marjanović et al. 2011a, Marjanović et al. 2011b, Marjanović et al. 2011c, Marjanović 2012, Marjanović 2013), all terminology is in accordance with the appropriate conventions (WP/WLI 1995, Fell et al. 2008, Lynn & Bobrowsky 2008, Gerath et al. 2010, Varnes 1984). Due to the nature of the subsumed research work, this thesis will mostly concentrate on susceptibility assessment, while the hazard and risk will be only speculated by their feasibility for further extensions of the research.

### 3.3 Landslide Assessment Concepts, Principles and Problematic

Herein, assessment is treated in a more specific sense than its original meaning instructs. It stands for a systematic gathering of the available information, processing/modeling with that information and forming a judgment about it in a transient workflow (Lee & Jones 2004). Landslide assessment workflow unfolds through phases of initiation of research (where the objectives, level of detail, scale, assessment type and study area are defined), acquisition (of data and background information), analysis and modeling (of landslide susceptibility/hazard/risk), evaluation, recommending/advising and reporting/publishing/visualizing (Gerath et al. 2010). In all mentioned stages, aspects of this problematic differ from one case to another, depending on the choices in the assessment approach. It involves not only the choice of the principal modeling approach, but also choices of other sub-stages, primarily regarding data acquisition and analysis. Hence, the aforementioned stages need a short insight and discussion hereafter, but first, it is necessary to articulate the basic principles on which the landslide assessment is founded.

The idea of landslide investigations and landslide assessment revolves around several principles and assumptions (Chacón et al. 2006, Guzzetti et al. 2012):

- Slope failures do not occur randomly or by chance, but as a result of interplay of different conditions, governed by different physical processes and laws.
- Landslides leave more-or-less distinct footprints (upon activation or after reasonable period of inactivity) that could be mapped in the field or remotely.
- Same types of landslide movement may result in similar landslide footprints.
- Principle of historical recurrence of landslides implies that the landslides are likely to reoccur on the same location, once activated in the past.
- Principle of uniformitarianism (past and present are keys for the future) implies that the slope failures are more likely to occur under those conditions that have led to instability in the past or at present at other, environmentally similar locations.
- Knowledge on landslides of some area can be generalized and expanded to other areas where similar conditions apply.
- Implicitly, conditions that are not taken into account in the model do not change systematically in time or space (time/space invariant).

It is crucial to understand the limitations and conditions under which all these assumptions apply, and to single-out special cases and exceptions, to reach a common (standardized) level of resulting products: Landslide Inventory maps, landslide susceptibility maps, landslide hazard maps and eventually, landslide risk maps. These postulates are approved by conventions (Varnes 1984, WP/WLI 1995, Fell et al. 2008, Lynn & Bobrowsky 2008, Gerath et al. 2010), as well as the concepts and methodology which are further to be described. This thesis, as well as the author's previous researches (Marjanović 2009, Marjanović et al. 2009, Marjanović 2010a, Marjanović 2010b, Marjanović & Čaha 2011, Marjanović et al. 2011a, Marjanović et al. 2011b, Marjanović et al. 2011c, Marjanović 2012, Marjanović 2013) stand in accordance with the latter.

### 3.3.1 Data Acquisition Issues in Landslide Assessment

Data acquisition in the landslide assessment framework is usually classified in respect to the type of investigation, i.e. its methodology and technology (Lee & Jones 2004). One can easily separate among mining of the historical records, field mapping techniques, instrumental monitoring techniques and Remote Sensing techniques. Furthermore, one can speculate between old, conventional and new methods for data acquisition (Guzzetti et al. 2012). These are all affected by the initial case study definition, i.e. required scale, level of detail, landslide size, mechanism type, configuration of the terrain, availability of the repositories, and they all bring about specific problematic, precision/accuracy issues, certainty issues and so forth.

Common or conventional methods have been established for a long time and have been proven in practice, but yet suffer from specific limitations.

**Investigation of the historical records** is one of the necessary stages of any landslide-related endeavor. It includes not only familiarizing with the facts on the landslides over an area, but also facts on geology, geomorphology, climatology, seismicity, Land Use, history of disasters and so forth. It is also presumed that one needs to get familiar with features of the wider surroundings of a chosen area, in order to have a better perspective on regional and local conditions in action. Principal investigation of historical records includes analyses of historical topographic and geological repositories, where applicable. Surprisingly, newspaper and diary reports on disastrous events can also be very resourceful, especially for hazard analysis. They can contribute to the existing databases, but must be treated with caution and criticism in order to avoid misconceptions, and where applicable, to be confirmed by other plausible resources. Tracking in such, merely unsystematic context is much easier nowadays, in the era of digital information and global networking. In particular, there are a number of websites, web-services and blogs dedicated practically only to the "landslide journalism", filtering-out all other, undesired journalistic contents ([www.geoprac.net](http://www.geoprac.net), [www.geohazards.usgs.gov](http://www.geohazards.usgs.gov), [www.landslideblog.org](http://www.landslideblog.org), and the most recent <http://landsliderisk.wordpress.com>) and there are also web tools such as RSS feeds and Google Alert and Google News for setting alarms for certain information in digital newspaper repositories.

**Field mapping**, albeit geological, geomorphological or engineering-geological (with special aspect on the slope processes) is confined by the practitioner's observational field of view, perspective of view, which might be obscured by the urban or vegetation cover, or by more recent geomorphological entities (in the case of larger or older landslides). Interpretation of larger landslide sites is therefore rather difficult. The interpretational subjectivity is also present throughout the map design (estimation of the landslide shapes and spread and their compilation at different scales), which leaves final result somewhat uncertain. This is usually reduced to some extent by augmented borehole testing (core mapping, specimen sampling, groundwater level checking), laboratory testing, *in-situ* testing, geophysical probing and so forth, but it additionally affects the research budget.

**Table 2.** Contemporary RS systems (after Guzzetti et al. 2012).

satellite	bands		resolution [m]	stereo mode	revisiting time	
	spectra	#			nadir	off-nadir
Landsat 7	pan	1	15	-	16	-
	r, g, b	3	30			
	nir, swir, mwir	3	30			
	tir	1	60			
Terra (ASTER)	gy, or	2	15	al	16	5
	nir, swir	6	30			
	tir	5	90			
SPOT 5	pan	1	5	al/ac	26	5
	gy, or	2	10			
	nir	1				
	swir	1				
IRS	pan	1	5.8	-	24	5
	gy, or	2				
	nir	1	23			
	swir	1	70			
ALOS	pan	1	2.5	al	46	2
RESOURCESAT 1	gy, or	2	5.6	-	5	
	nir	1	5.6			
CARTOSAT 1	pan	1	2.5	al	125	5
FORMOSAT 2	pan	1	2	-	1	1
	r, g, b	3	8			
	nir	1				
EROS A1	pan	1	1.8	al/ac	7	2.5
IKONOS 2	pan	1	1	al	3	1.5
	r, g, b	3				
	nir	1	4			
QuickBird 2	pan	1	0.6	al	3.5	1
	r, g, b	3	2.4			
	nir	1				
WorldView 1	p	1	0.5	al	5.4	1
GeoEye 1-2	pan	1	0.4	al	8.3	2.8
	r, g, b	3	1.6			
	nir	1				

R=red, G=green, B=blue, OR=orange-red, GY=green-yellow, nir=near-infrared, swir=short-wave-infrared, mwir=mid-wave-infrared, tir=thermal-infrared, al=along-track, ac=across track

**Visual interpretation of aerial photographs** by using stereoscopic techniques and equipment is also a well-known conventional method, which overcomes synoptic issues, allowing the practitioner to observe much wider areas with better perspective. It is also affected by presence of vegetation or infrastructural and urban objects (especially for shallow landslides and debris flows). In contrast to field mapping, the analysis is relatively easily combined and compiled across different scales (with some georeferencing difficulties due to the spherical geometry of acquisition which needs further orthogonal re-projection), and seems independent of field conditions (apart from the final evaluation stage, which requires certain amount of field work and therefore depends on field conditions). As for the subjectivity, it is even more pronounced than in the field mapping techniques due to the individual visual perception capabilities. However, some standard criteria for landslide

recognition do exist<sup>2</sup>, thus uncertainty prevails only to some extent. Important benefit is possibility of analyzing different time series and scales (some countries have multiple records in different scales, from several periodic surveys, usually restricted in repositories for military purposes or other purposes in different national institutions).

New methods for data acquisition primarily involve novel Remote Sensing (RS), field (*in-situ*) and surveying instrumentation technologies (Savvaidis 2003), complemented by according software and hardware development in order to support their full capability.

**Contemporary RS** came along with the advent of the high-tech satellite technologies and several Earth Observation programs. New opportunities have been introduced by widening perceptual capabilities with new sensors, focused at different parts of wide electro-magnetic spectrum (Tab. 2). Since unprecedented parts of the spectrum became available, new spectral features have been exploited. From the most recent perspective, the latter involve global coverage by multi-channeled sensors, i.e. multispectral and hyperspectral sensors for visible, but also infra-red and thermal spectral domains, as well as microwave sensors, with unparalleled spectral, temporal and spatial resolution. At airborne and terrestrial level, microwave and laser techniques appeared and brought unprecedented precision. These involved Light Detection and Ranging (LiDAR) and Side Aperture Radar (SAR) techniques, particularly interferometric (InSAR), differential interferometric (DInSAR) and polarimetric techniques (PolSAR), as well as Small Baseline (SB) and Permanent Scatter (PS) techniques (Ferretti et al. 2007). These approaches (LiDAR and InSAR in particular) have promoted surface-based monitoring, since systematic, high-resolution, on-demand surveys became possible. This facilitated production of high resolution Digital Elevation Models (DEM), allowing near-real-time tracking of surface deformation, by imaging at desired temporal frequency (temporal resolution). Geophysical satellite/airborne systems also fall in this group, providing even more details on the geological and physical conditions of the terrain (by means of different gravimeters, accelerometers, magnetometers, gamma-spectrometers and so forth). Further upgrades in RS technology can be expected principally due to the increase of the spatial and spectral resolution, conditionally temporal too, since many satellite programs today tend to satisfy the principle of data continuation (e.g. Landsat series) so that their data could be considered compatible with the missions that they have substituted. Benefits of using RS techniques in landslide assessment are multiple, including, but not limiting to: synoptic view, georeferenced data, lower expense of research, encouraged raster modeling approach, possibility of quantitative modeling method implementation (pixel and object-based classifications implementation, pixel and object-based classifications through combination of advanced statistics and Machine Learning with GIS) and therefore reduced subjectivity in design, possibility for urgent response and Early-Warning Systems for disastrous landslide events, even enabling on-screen visual 2-3D analysis, via special hardware/software configurations (Guzzetti et al. 2012). Special attention in the most recent technology is drawn by the unmanned vehicles and micro-vehicles, which are capable of producing high-resolution imagery at extremely low cost. Limitations on the other hand, are mostly technical: unavailability of specific sensor at the site (particularly, pricey and rare airborne/terrestrial LiDAR and SAR data), relatively short operational history of RS programs (only several decades, through which the data are not entirely consistent in terms of resolution and other technical features), and therefore limited applicability for temporal (hazard) framework. It is probably the most advisable to combine novel RS techniques and conventional aerial photography in order to achieve the optimal acquisition.

**Field (*in-situ*) instrumentation**, often referred to as geotechnical instrumentation, has also undergone some technological improvement, primarily toward near-real-time and

---

<sup>2</sup> Criteria for geomorphological landslide signature exist and usually include: shape, size, tone, color, texture, pattern of shadows, pattern of objects, overall topography and setting. It is assumed that occurrence of landslides causes characteristic optical properties of mentioned elements.



real-time data acquisition and distribution, based on the advent of wireless and internet networking technology. *In-situ* measurements of displacements, carried out by standard inclinometers, tiltmeters, extensometers or electro-optical systems (such as Time Domain Reflectometry – TDR) with the highest precision (in mm), provides valuable information for the assessment of landslide activity. Moreover, it involved measuring of the physical parameters of the triggering event (rainfall amount/intensity/duration, earthquake magnitude, water table level/pore-water pressure etc. via pluviometers, seismometers and piezometers, respectively), working toward modeling of the trigger-landslide relation, which in combination with real-time data distribution eventually allowed development of Early-Warning Systems, crucial for the suppression and mitigation of the landslide risk. The major drawback is the equipment cost, together with the installation and maintenance requirements, and localized information, rarely transferable from one study area to another.

**Surveying** equipment has experienced improvements from several aspects including, faster acquisition time with sufficient precision, more precise optical-laser systems for distance measuring (which reimbursed field mapping precision), but mostly via Global Navigation Satellite System (GNSS) and synergy of Photogrammetric and high-resolution optical imaging (terrestrial, airborne and satellite). GNSS receivers tend to gain higher and higher precision, allowing very precise systematic surveys of chosen critical points of the terrain. In turn, the information on total 3D displacement of these points and possibility to model landslides at different scales are provided. Near-real-time, i.e. robotic total station surveys are rather experimental for now, but seem to meet the precision and efficiency requirements, and could be easily assembled alongside with the standard *in-situ* instrumentation. Less proficient mobile GNSS receivers are being regularly used as complementary equipment for field mapping, reimbursing for the subjectivity in the mapping process. They also complement the commercial digital cameras which enable better precision in photo-documentation of the landslide events. The only drawback of GNSS technology in such framework is its dependence on the terrain physiographic condition (configuration and setting, screening by vegetation cover and urban objects). On the other hand, reception of the satellite signal tends to increase with appearance of new missions (existing GPS and GLONAS systems are soon to be joined by GALILEO mission, which will bring more satellites in the GNSS constellation and accordingly, better reception on the ground). Photogrammetric survey advancements are simply related to the imaging (spatial) resolution of aerial and terrestrial sensors – higher the resolution, higher the precision. It also necessarily follows the advent of the associated software/hardware configurations. Nowadays, it is possible to use Photogrammetric technique to produce high precision DEMs, paralleling the quality of LiDAR-based or InSAR-based DEMs. Its principal limitation is the engagement of the practitioner, making it time-consuming and resource-intensive (Savvaidis 2003, Guzzetti et al. 2012).

It is probably the most advisable to combine as many of the acquisition techniques as possible and never to rely entirely on a single one. Those older, conventional methods, especially aerial photography interpretation, are not to be neglected among acquisition techniques, and should be cherished in the landslide assessment practice (Guzzetti et al. 2012). Novel techniques, which are developing toward automatic (semi-supervised) landslide mapping, will hardly reach sufficient levels of certainty, since they face different, non-compensable limitations unlike visual, expert-driven interpretation.

### 3.3.2 Modeling Approach Issues in Landslide Assessment

Once the data are chosen and structured, they need to be fed to a proper modeling method, where particular choices strongly influence the quality and type of the outcomes. Model's predictability is a feature which can be adopted as a criterion for distinction between two separate cases: temporally predictive and temporally non-predictive (spatial analysis) models (Brenning 2012), although in practical situations this distinction remains deficient, due to the

tendency of obtaining more interpretable models (transition from predictive to non-predictive), thus mixing of the two.

Predictive models are based on non-linear supervised classification problem upon spatial or temporal reference (Brenning 2005). In particular, predictive models can relate spatial conditions of an area with its past landslide occurrence, in turn localizing endangered zones in adjacent areas. Alternatively they can relate several generations of past occurrences within the same area and predict the future events. They both require that general principles and assumptions apply (see the postulates in Chapter 3.3) and also require certain structure and type of the data. It is for instance indispensable that analyzed areas contain thematic variables (Conditioning Factors), including geological, geomorphological or even geotechnical parameters on one hand, and reliable Landslide Inventory or multi-temporal inventory on the other. Even though the resulting model provides numerical, i.e. quantitative measure (usually probability of spatial/temporal occurrence), relative scoring is yet preferred due to the great deal of assumptions which trouble the quantitative way of expressing the landslide susceptibility/risk/hazard<sup>3</sup>. It is further advisable to treat a non-linear problem with non-linear techniques, even though the current practice has shown different, but this is probably due to the lack of cases with properly applied non-linear techniques. Systematic comparative studies (Brenning 2005, Yilmaz 2009) hence give preference to linear or moderately non-linear techniques on behalf of both, their performance and their simplicity (in respect of time consumption and processing intensity). Implicitly, some advanced techniques, such as Machine Learning-based ones, turned less efficient than regression methods, discriminant analysis or even general additive modeling. This fact however, should not discourage experimenting with advanced techniques, on the contrary.

In non-predictive approach the objective is realized through the spatial analyses of different thematic variables (Conditioning Factors), and chiefly involves determination of their total contribution to the landslide susceptibility/hazard/risk, by exploring the statistical relation between the factor and landslide occurrence (landslide presence/absence), but also the relation among themselves. They call for a simpler, i.e. less time/computation-demanding techniques, in order to be testable by statistical hypotheses. In turn, non-predictive approach comes up with quantified values of individual impact of each factor. Although these methods decrease the uncertainty by surpassing some of the assumptions that are commonly made in the predictive modeling, they tend to subject uncertainties through the data preparation, due to arbitrary/empirical rearrangement of the raw data (slicing/ranging continual data into intervals, transforming the data, quantifying non-numerical data and so forth). The most appreciated techniques involve multivariate statistical tools (e.g. different types of regression techniques), which fully explore statistical possibilities (to relate the Landslide Inventory with thematic variables, but also thematic variables among themselves), and different linear and non-linear tools (such as odds indexing, entropy scores, conditional probability weighting, and other general additive modeling techniques). Important benefit of this approach is its quantitative nature, which is relatively easily communicable to non-landslide experts, planners and decision-makers (Brenning 2012).

One can alternatively discuss the modeling choice and brief the problematic which it brings, by accommodating a more conventional perspective. The most usual classification of methodological approaches sorts them into (i) heuristic/empirical or expert-driven, (ii) statistical and (iii) deterministic/physical. In respect to the preceding passages, only statistical methods qualify as predictive approach, but could also be enlisted among non-predictive, while the remaining two only qualify as non-predictive approaches. In brief, (i) use thematic data (variables such as geological, geomorphological, Land Use, infrastructure and so forth) and suffer from uncertainty related to the subjectivity of the practitioner in both, data

---

<sup>3</sup> Some assumptions are taken into account but some of the uncertainties usually remain unconsidered, and it is therefore disputable to measure susceptibility/hazard/risk in absolute quantitative scale.

preparation and modeling itself. Statistical modeling (ii) can suffer from uncertainty due to the data preparation, but the tendency of using advanced techniques, such as Machine Learning algorithms, might be helpful due to their capability of canceling-out these sources of uncertainty. Deterministic models (iii), regard only the simplest mechanisms (in regional scales) and introduce numerous assumptions into the modeling (Montgomery & Dietrich 1994), thus their uncertainty is relatively high. In conclusion, it seems that statistical methods, especially the more advanced (predictive) ones turn out to be the most promising and least limited for the exploration, but they do not necessarily grant the optimal solution (Bonham-Carter 1994).

### 3.3.3 GIS Issues in Landslide Assessment

Past few decades of landslide assessment had been witnessing paramount improvements in computer science and technology, eventually resulting in GIS. As GIS gained more attention in all spatially-featured disciplines, a prompt scientific evolution took place due to both, new possibilities for better data manipulation and more advanced modeling opportunities (Carrara & Pike 2008). This particularly applies to regional landslide studies, where geological, geomorphological and Land Use variables are preferred, unlike sight-specific studies where geotechnical parameters are required and sampled through a series of instrumental measurements and laboratory tests.

A firm relation between a landslide occurrence and conditions which host it is accomplished through morphological features of the terrain surface. Thus, development of DEM through GIS and RS virtually led to the morphometric revolution by introducing new, unparalleled tools for surface features extraction and creation of novel thematic spatial layers, unachievable through conventional – analogue practice. Other geo-environmental features became available in digital format, and also came about as thematic layers in a GIS environment. Features within such layers, presented by point/line/polygon vectors or gridded in the case of raster formats, became analyzable, synthesizable, decomposable, combinable, scalable, in other words, fully spatially operable (Bonham-Carter 1994). However, with such great capabilities came even greater difficulty of suiting the data for a specific research. Using different data resources, types and scales brings about the data quality and compatibility issues, which are topics on their own, but will be regarded in approaching chapters and case studies (see Chapter 6.).

Numerous modules onboard GIS platforms made majority of spatial modeling techniques available, including even complex, time/software/hardware-demanding techniques. These in particular include regression methods and Machine Learning algorithms. It is important to mention that raster format has made a major breakthrough for implementation of these advanced methods.

Once properly prepared for the desired modeling concept<sup>4</sup>, raster formats allow bulk spatial information to become easily reproducible and directly accessible to different module demands, including filtering, sampling and calculating, while keeping the spatial reference consistent. Such flexibility put very demanding modules into play. To this end, strong efforts are made to aggregate more and more complex algorithms through GIS environment.

Although the most critical GIS aspects in landslide assessment are presented above, it should be mentioned that GIS turned revolutionary for a number of other, more general innovations, valid for any spatial context, starting with the data structure. It allowed allocating each data input within a thematic layer, as well as its frequent editing/updating, networking and storing. It also allowed attributing practically unlimited amount of features to a single data

---

<sup>4</sup> Spatial/spectral/temporal resolution of the grid cell needs to be justified by the nature of a phenomenon, e.g. Landsat images are sufficient to monitor landslides bigger than 30 m in diameter, which leave visual imprints within the Landsat band spectra (VIS, NIR, SWIR, TIR), and do not change significantly over 16 days (repeat cycle of Landsat series).

input within each thematic layer, which had met even very demanding modeling requirements. The same applies to any thematic layer which is appended, while spatial/geometrical consistency among the layers remains persistent. The later is particularly important for working with time-series in hazard/risk assessment scenarios. This consistency is provided by georeferencing, i.e. attributing a particular geographical coordinate to each piece of information, making it always properly positioned in space, and therefore interoperable with other data and ready to subject to any desired spatial operation. GIS also became indispensable in terms of visualization of both, the input data and the resulting models. Once locally dense spatial information became much easier to visualize in multidimensional 2/3/4-D displays. In combination with web-GIS systems, global (GoogleEarth, BingMaps) or local GIS portals for different (usually administrative) purposes, landslide information became much better disseminated and visualized (Guzzetti et al. 2012).

The most recent trends imply merging of Remote Sensing, (Geo)statistical and GIS platforms especially among the open-source communities (e.g. SagaGIS, GrassGIS, ILWIS, R, SEXTANTE, and many other platforms). This is beneficial for extending analytical power of a practitioner or assembling a cross-disciplinary team of practitioners to work under the same framework and achieve better communication, better interoperability and eventually better results. The downside of this increasing “user-friendliness” of GIS platforms is recognized in lack of criticism, particularly by neglecting the input data quality issues and focusing rather on the complexity of the data manipulation and model implementation (Carrara & Pike 2008). Data manipulation or modeling technique, however sophisticated, can never compensate for inadequate quality, scale or theme of the inputs, due to intrinsic error that is continually replicated within. On the other hand, data availability and open-source policies are one of the most significant issues in research budget design, and lack of affordable data could lead to the decreasing of assessment quality, but this is rather financial than scientific issue to discuss.

### **3.3.4 Other Issues in Landslide Assessment**

Although a couple of preceding passages revealed the most critical concerns in landslide assessment framework, there is still suffice of other issues in landslide problematic, ranging from scientific, practical, technical, to social speculations.

Uncertainty is definitely one of the major issues for plausibility of resulting landslide models. It can originate from the data, from the modeling procedure choice and from the environment (real-world conditions). The former two were partly discussed before (see Chapter 3.3.1 and 3.3.2), but some specific details are to be emphasized:

Fuzziness and randomness are omnipresent in measured/imaged spatial data and contribute to the total uncertainty. Fuzziness is contributing by appending local imprecision, while randomness contributes by preventing regularity in patterns of distribution of data values. Both are especially pronounced in noised, biased or skewed data. In the case of non-predictive modeling, these sources of uncertainty can be treated and taken into account through the probability theory.

Incompleteness is nourished by oversimplification in the modeling stage. Landslide assessment has many assumptions (see the postulates in Chapter 3.3) and therefore it is highly affected with simplification (especially in deterministic modeling). One delicate issue within is the exclusion of conditionally unimportant data. There is a debate (van Westen et al. 2006, Carrara & Pike 2008) on whether any data shall be excluded, even if biased. On the other hand, some data are not excluded on purpose, but due to the lack of resources for the corresponding phenomenon (e.g. involving parameters of the trigger, geotechnical parameters, weathering parameters, soil thickness or groundwater parameters might be critical for yielding a reliable model, but having them sampled at regional scales is inapt due to the intolerable costs and strong distributional variability) or it is simply unforeseen as a pertinent factor by mistake or insufficient knowledge.

Environmental or real-world uncertainty is highly unpredictable and includes consequential actions of administration, public or individuals (conscious or subconscious deeds that can deliberately drive even high quality predictions off the course), and which are impossible to account for. This applies in both directions, actions in the future, which are impossible to specify, and actions from the past (especially from a long time ago), which have remained unknown. Due to this dimension of uncertainty, the further away one is from the present moment higher is the difficulty to predict landsliding (Lee & Jones 2004).

Quality check of data and results is another important issue, which troubles the production of reliable landslide susceptibility/hazard/risk maps. Data quality check is a necessary step, particularly due to the rising resourcefulness in contemporary researches (data from different scales, different spatial reference, different geometries, different precision and level of detail), but it is this plentitude of resources that limits possibility to standardize and objectify the quality check. Two basic quality check requirements should be met: appropriate strategy for model performance evaluation and actual valorization of the model (van Westen et al. 2006, Carrara & Pike 2008, Brenning 2012). It is hence recommendable to propose spatially-driven sampling strategy (which applies only for predictive models), for distinguishing training-testing-valorization sets within a model by a meaningful physical (spatial) splitting. For non-predictive models, the evaluation is intrinsic, governed by the statistical confidence intervals attached to the particular technique. Actual valorization of the model implies its confirmation through time, and particularly concerns hazard/risk models due to their temporal dimension.

Most of the times practitioners are more concerned with their modeling choice, while attempting to design the best model to suite the universal circumstances, which is most likely futile, assuming that the best model is the most complex and robust one (Carrara & Pike 2008, van Westen et al. 2006). Instead, the above mentioned result and data quality check might be an apt response to the particular problem posed before them. In some cases the Occam's razor directly applies, so that the simplest solution – the simplest modeling method can provide optimal solution. It would provide the optimal balance between the quality and complexity of the model (Lee & Jones 2004, Brenning 2012).

Another peculiarity due to the resourcefulness of the data comes along with the rising popularity of RS products in landslide assessment (aerial photographs, multi/hyperspectral satellite imagery, LiDAR and SAR products, i.e. DEMs, Principal Components, different indices such as ratios, vegetation indices, change detection indices and so forth). Usage of raw products is the easiest but irresponsible solution, since each one of them contains intrinsic noise, which foremost requires determination of its type, quantity and propagation. Subsequently, noise filtering is managed through image preprocessing, i.e. pansharpening, orthorectification, coregistration and radiometric correction, in this respective order (Mondini et al. 2011). Working with initial noise is qualified as a systematic error and will affect, perhaps even sabotage the model. Working with time series in the case of hazard assessment further perplexes the problematic by requiring pre-event and post-event noise filtering, as well as technical consistency in acquisition (images need to be taken in very similar conditions, such as view angle, altitude, mode of the sensor and also in compatible meteorological conditions).

Temporally dependent assessment (hazard/risk) also suffers from the possible misdating of a landslide vs. trigger(s) records and it is therefore usually reduced to separation of pre/post-event observations. The latter introduces substantial temporal tolerance, during which pertinent changes among Conditioning Factors could have taken place (with the exception of geological setting and to some extent, geomorphological features, but with the emphasis on Land Use). Postulate of temporal/spatial invariant (see the postulates in Chapter 3.3) is therefore merely plausible (Brenning 2012).

One important research hint infers distinction of the landslide assessment according to the landslide typology, because different landslide mechanisms, e.g. debris flows and

deep-seated slides, will exhibit completely different behaviors, but more importantly, will be induced under different circumstances in respect to the Conditioning Factors and Triggering Factors. For this reason alone, it is advised to assess only one landslide type at a time, and eventually combine these separate assessments later on (van Westen et al. 2006). It is further advisable to concentrate on characterization of the depletion areas, since they host the slope failure, while the accumulation zone is only the collateral result of the posterior downward movement, and its characterization does not lead to the real cause of the failure. Such division is rather difficult in regional studies, due to the size of the occurrences and spatial continuity of source and accumulation, but yet feasible in some cases (Guzzetti et al. 1999).

Difficulties also arise from purely technical causes, such as the lack of independent, long-lasting, institutionalized landslide agencies on national level, which would focus on all the aspects of landslide problematic, including their assessment and provide the research continuity. At present, individual projects at universities or institutes are treating this problem, but only during the project lifetime. At best, there are cases where multi-scaled and nationwide researches are involved, but most commonly landslide assessment is disconnected into separate case-studies, and focused on very specific project objectives, rather than revealing of the fundamental breakthroughs in landslide knowledge (van Westen et al. 2006, Carrara & Pike 2008).

## 4 Related Work

It is rather difficult, if even feasible to systematically and chronologically summarize all published landslide assessment references. Some earlier attempts to systematize the landslide literature could be found (Aleotti & Chowdhury 1999, Chacón et al. 2006), but even these were focused on particular topics. Contemporary volumes of literature would be probably much more challenging to review, thus eventual systematization of such kind would probably end-up in even stronger secession of topics, e.g. review of contributions in susceptibility/hazard mapping domain only, or in risk mapping only, only in GIS-based modeling or in input data evaluation or evaluation of the model. The following passages are therefore intended for reviewing only those critical contributions in landslide susceptibility/hazard mapping, which radicalized landslide assessment practice and pioneered advanced methods, particularly Machine Learning-based ones. Hereupon, the work which turned out to be the most inspiring for research design and the topic of this thesis are to be scrutinized in greater detail.

Early works in the GIS-based landslide susceptibility assessment came along as the GIS software/hardware components became more available to practitioners, i.e. as the related field of computer science has emerged in 1970's (Brabb et al. 1972). Pioneering attempts involved simple solutions including heuristic and simple statistical non-predictive models, but these rapidly changed toward the implementation of more sophisticated mathematical and statistical models. Availability of digital formats of previously analog input data, such as topographic, geological and geomorphological maps, subordinately, Land Use, pedological, seismic maps and so forth, propelled implementation of statistical methods via GIS. One of the most productive groups, gathered around IRPI institute in Italy, with Guzzetti as the most cited author in entire landslide assessment literature (Gokceoglu & Sezer 2009), have contributed in domain of GIS-based statistical assessment approach since their early opus (Guzzetti et al. 1999, Guzzetti et al. 2000), but also kept perfecting their practice till present (Rossi et al. 2010, Brunetti et al. 2010, Guzzetti et al. 2012). Copious as it is, their work is difficult to present in detail, but several points are to be discussed hereafter. Their early practice regarded the basic capabilities of GIS-based landslide analyses, and more importantly, the choice and the manner of partitioning of the slope units, prior to the analyses (Carrara et al. 1991). These are being developed on the basis of geomorphological and watershed analyses and have been further improved (Carrara et al. 1999, Guzzetti et al. 1999). Another important issue addressed in this early stage regarded the importance of Landslide Inventory (its certainty and quality), thereto speculating capabilities of producing multi-temporal inventories for hazard assessments (Malamud et al. 2004). Such thematic was recently revisited, and resulted in a very systematic perspective on inventory problematic (Guzzetti et al. 2012). Their practice sublimed in a series of articles and case studies where optimal techniques (according to their findings) have been practiced (Rossi et al. 2010). Numerous case studies from various aspects have been subjected to their practice, but the focus was on central Italy (Umbria region), where all of their breakthroughs were first experimented. Meanwhile, they have contributed toward the development of national landslide information system and Early-Warning System for rainfall-triggered landslides, which became the principal preoccupation of the group henceforward (Guzzetti et al. 2007, Guzzetti et al. 2008). Finally, some contributions regarding the state-of-the-art technology attract attention, principally involving various InSAR and integrated GNSS techniques (Guzzetti et al. 2009, Santangelo et al. 2010). In conclusion, the most relevant findings that relate to the topic of this thesis regard the sampling strategy, which suggests 1:2 size ratio between the training and testing samples (Rossi et al. 2010). Despite of introducing the bias in the testing sample, such ratio ensures the robustness of the model, which has been acknowledged by the thesis author in his preceding work (Marjanović et al. 2011a,b). Another good example of the influence of this group is featured through the selection and manipulation of the input data (Guzzetti et al. 2006, Rossi et al. 2010). In particular, the

group has committed very exhaustive work on defining relevant data inputs, ranging from essential geological, geomorphological and environmental layers<sup>5</sup>, through their conceptual derivatives<sup>6</sup> and statistical/computational derivative variables<sup>7</sup>. Due to the different approach in appreciation of the base unit by the thesis author (the grid cell approach over slope unit approach has been preferred in this work), the input data strategies presented by the group could not be directly adopted in the thesis case studies, but they have been proven indispensable for any true hazard assessment and turned inspiring for the future work of the thesis author. The group has thoroughly implemented various landslide susceptibility modeling techniques by means of predictive modeling, and their experience is duly noted. Another very original finding on behalf of this group is the combination of the forecasts, i.e. the optimization of the multiple models by means of the ultimate supervised classification, which turns to be a very inspiring proposition for model post-processing, particularly when multiple modeling methods are in use (Rossi et al. 2010). Regardless to the introduced bias by domination of specific individual models in the final classification, and occasionally better performance of individual models over the post-processed one, higher certainty is achieved in post-processed models (because of the lower standard deviation rate) than in individual models (Rossi et al. 2010). Finally, the group suggested extensive qualitative/quantitative performance evaluation by always using several statistical descriptors ( $\kappa$ -index, ROC, bootstrapped error) and emphasizing the role of False Negative error in model's predictability, which is also adopted by the thesis author (Marjanović 2013).

Another prominent group, gathered around Joint Technical Committee on Landslides and Engineered Slopes (JTC-1), has been experimenting over numerous case studies worldwide, and has included different methodological aspects and has processed voluminous problematic in order to optimize susceptibility/hazard/risk assessment. Their long experience has sublimed in form of the guidelines for further researchers (Fell et al. 2008). Their range of interest goes from basic assessment, i.e. landslide zoning to advanced monitoring techniques, mitigation measures and management toward better Land Use planning. Yet, the accent was always on rainfall-induced landslides, involving rainfall in combination with groundwater interplay as a triggering mechanism for all slow moving (Cascini et al. 2010a) and flow-like landslides (Fell et al. 2007, Cascini et al. 2010b, Cascini et al. 2011), as well as for rockfalls (Corominas et al. 2005, Mavrouli et al. 2009), wherein the group has exhibited a significant amount of modeling. Particular attention has been directed and significant contribution for future researchers has been introduced in their work on the problematic of scale (Cascini 2008). Possibilities for differently-scaled landslide zoning were scrutinized, and involved the problems of different data sources and different modeling techniques at different scales. Their guidelines have since been adopted by the most of the researchers, including the thesis author (Marjanović et al. 2011a,b). Another authentic contribution that could be linked to the group is regarding the analysis of the landslide frequency, i.e. mostly rockfall frequency via dendrochronology (Corominas & Moya 2010), as a significant step in transition from susceptibility toward hazard assessment. The group in general, rather prefers deterministic approach in landslide modeling and concentrates on shallow and flow-like landslides. The latter involves a very complex geological environment, represented by unsaturated soils<sup>8</sup>, which entails further complexity in the laboratory experimenting stage (special conditions, longer experiments, special equipment). In compensation,

---

<sup>5</sup> Variables like lithological units, slope units – hydrogeological units, Land Use and so forth.

<sup>6</sup> Geological domains layer in particular, which are derived by combining geological maps and aerial photograph interpretation and relate the bedding attitude to the slope inclination.

<sup>7</sup> Various morphometric sub-variables, different buffers, means and standard deviation and other statistical descriptors of original variables.

<sup>8</sup> Acknowledging that the soil exists through all its phases: solid, liquid and air, instead of approximating to the solid and liquid only.



acquired/measured parameters tend to be more realistic, and therefore improve the reliability/certainty of their models (Sorbino et al. 2010).

Deterministic approach in landslide assessment has been pioneered relatively early (Montgomery & Dietrich 1994) and there have been several succeeding developments involved. They all gradually perplexed the model and introduced more variables, by decreasing the number of approximations, but their reach in applicability has been disputed, i.e. limited to a very specific, homogeneous ambient and conditions, very rarely present in actual terrains. Recent attempts (Cotecchia et al. 2010) have been led by the experts in clay mechanics and slope stability, and proposed a very interesting, but complex solution, which combines regional variables with site-specific variables in a GIS database (specially designed to enclose both regional and site-specific data). The assessment is conducted by using different level of analysis (ranging from 1<sup>st</sup> level – preliminary heuristic analysis, 2<sup>nd</sup> level – limit equilibrium modeling to 3<sup>rd</sup> level – advanced numerical modeling by Finite Elements) in two stages: acquisition (identification of relevant pre-Conditioning Factors at regional scale – geological, geomorphological, environmental, but also at site-specific scale – geomechanical and hydraulic properties of the geological units) and selection (which entails defining of representativeness of particular sites). Throughout the first stage, different geo-hydro-mechanical set-ups are determined across the entire area, as well as different landslide mechanisms, to be subsequently generalized as representative for that area of interest in the second stage of the research. Once detected representative, they are to become principal concern of the 2<sup>nd</sup>–3<sup>rd</sup> level analysis, through which their general trends and overall characteristics are then either confirmed and extended to a wider area (with similar setups), either denied (Cotecchia et al. 2010). Such approach has not been exploited in its full extent, and many additional case studies, with many different geological configurations are necessary for the future affirmation of the approach, or its definite abandoning. In conclusion, the deterministic approach, remains an open issue, and remains a promising foreground for future investigators, which is why it has been partly integrated in this thesis.

Further fruitful ground for landslide assessment research turned out to be a combination of pure statistical and heuristic models. There have been several authors who have revisited the heuristic methods since the very first attempts at the very beginning of the GIS-based landslide assessment practice, as they combined/matched them together with statistical methods, claiming that heuristic touch introduces the necessary non-linearity into otherwise very ordered statistical models, i.e. they mirror the uncertainty imbedded in each thematic layer, and therefore yield better accuracy in resulting susceptibility or hazard maps (Ercanoglu et al. 2008). Komac is to be singled-out as the author who has gone the furthest distance in combining of the two, heuristic and statistic approaches. His PhD project and associated articles are to witness his efforts in refining their relation, even though his expertise extends beyond susceptibility/hazard/risk assessment scope, and relates to engineering geological mapping, development of national (Slovenian) geological information system, and some other fundamental geological applications, applied Remote Sensing and environmental applications (Komac 2003, Komac 2005). In his work landslide susceptibility is approached by several aspects, regarding acquisition and building of the Landslide Inventory, data preprocessing and feature selection, and finally, heuristic analysis, paralleled and compared by statistical analysis. As for the acquisition, the advanced image fusion<sup>9</sup> techniques are proposed for apparently, quite accurate landslide detection and successful

---

<sup>9</sup> Medium resolution satellite images are herein combined with the high-resolution orthorectified aerial photographs. Principal Components of satellite images are extracted and PCA fusion (replacing the first PC by high-resolution image) was performed. Subsequently, reverse PCA transforms components back to the original satellite bands and an improved band stack is obtained. Unsupervised classification over such stack, by RGB clustering method then takes-over, and finally gets transformed into CIE L\*a\*b\* color model, producing numerous color composites. Generated classes from composites are then aggregated to only those related to landslides.

semi-automatic generation of the Landslide Inventory (Komac 2005) under given circumstances (conveniently the vegetation and urban cover were not abundant in his study areas). The author was also very dedicated to the input data pre-processing, by implementing laborious univariate statistics to rank each spatial data layer and determine its significance to the landslide occurrence. Even though it is not always supported due to their possible influence in the multivariate statistics stage (van Westen et al. 2006) the author turned to exclusion of insignificant variables to reduce computational cost. Further, multivariate statistics are typically involved in his practice, but what is exceptional is the parallel involvement of multivariate statistics in heuristic model, such as Analytical Hierarchy Process (AHP). More precisely, AHP matrix indices, obtained through the multivariate stage, are challenged against purely heuristically<sup>10</sup> chosen indices in the AHP engine (in a GIS environment). According to the standard deviation, the maps coming from this model yield a higher accuracy and higher certainty, as indicated above. However, the author also turns to the calculation of hazard, but only by estimating the spatial probability, i.e. by reclassifying the generated susceptibility map, which is inconsistent with the hazard definition. Such choice is even more unusual if one learns that the repositories used in the research have relatively consistent temporal dimension (dated displacements), thus providing elements for multi-temporal approach via landsliding frequency, instead of speculating probability via standard deviation. The author further tends to evaluate the societal landslide risk over the area by overlaying population distribution and road infrastructure (as GIS layers) to the hazard maps, and qualitatively describes the proportions of endangered and risk-free areas (Komac 2006). Having in mind the aforementioned shortcomings of the hazard maps, these risk qualifications are also to be regarded with some reserve, but the author's authenticity and expertise in the susceptibility part is indisputable. As a paradigm in this thesis research the heuristic approach has been integrated and its applicability has been speculated, as well as in the author's previous work (Marjanović 2009, Marjanović et al. 2009, Marjanović 2010b, Marjanović & Caha 2011, Marjanović et al. 2011a, Marjanović 2013).

Weather using ordinary Fuzzy Sets, or fuzzy measures, or even combining fuzzy with other statistical or classification approaches (Dempster-Shafer, *k*-Means, Neural Networks) the ultimate advantage of fuzzy approach is recognized in providing a substantial possibility for standardization of the analysis, since Fuzzy Logic procedure tends to be repeatable, adjustable and reliable (Jiang & Estman 2000). When it comes to the landslide assessment analysis in particular, a number of researchers have applied fuzzy approach to handle the embedded non-linearity. The Himalayan terrains were addressed in many investigations with Fuzzy Set Theory background, starting from standard Fuzzy Set approach (Chamapitray et al. 2006, Kanungo et al. 2009, Srivastava et al. 2010), through combinations of ANN-fuzzy (Kanungo et al. 2006) and risk-oriented fuzzy approach (Kanungo et al. 2008). Most of these studies agreed that plausible susceptibility models could be obtained by cautious application of fuzzy operators, with preference toward Cosine Amplitude method for obtaining memberships. Very similar conclusions with analogue methodology have been drawn in Iranian case studies (Tangestani 2004), and in Turkey (Ercanoglu & Gokceoglu 2006), China (Wang et al. 2009), and so forth. The latter is also interesting in respect of harmonizing expert-based and fuzzy-driven solutions, inferring that one does not exclude another, but supports it. Regmi with his research team (Regmi et al. 2010) has conducted one of the most consistent investigations, where many different fuzzy configurations were put to the test. Detailed elaboration of the choice of fuzzy operator type, optimal fitting of gamma operator as a method of preference, and some suggestions on handling multi-type landslide cases, can be found in that research. In addition, most of the researchers encourage the usage of the fuzzy method in other, similar (mountainous regions with flows and falls as dominant landslide types) or entirely different ambient, worldwide. The thesis author has been

---

<sup>10</sup> This means "not arbitrarily", but by interviewing relevant experts through standardized questionnaires, as proposed by AHP methodology (Saaty 1980).

practicing such methods under the influence of findings of aforementioned researchers (Marjanović & Čaha 2011).

Finally, the "advanced" methods used in landslide assessment are to be addressed hereafter. These are including the same methods termed "advanced" throughout this thesis and used in the thesis case studies, i.e. Multivariate Statistics and Machine Learning.

New solutions for non-linear classification problem were recognized in Machine Learning techniques such as Logistic Regression, Decision Trees, Artificial Neural Networks (ANN) and Support Vector Machines (SVM).

Logistic Regression has a longer tradition in natural hazard assessment, and landslide susceptibility is not an exception. It has been proven successful in numerous case studies, but lately it is being broadly challenged by other Machine Learning approaches. Logistic Regression has usually been involved in comparative case studies, but there are several contributions dedicated to Logistic Regression in greater detail (Falaschi et al. 2009, Bai et al. 2010). Their findings are confidently promoting the method as very reliable and very convenient in the landslide assessment framework. In extension, a very interesting approach has been proposed in a Southern Norway case study (Erener & Düzgün 2010), in which Geographically Weighted Regression variants have been utilized together with Global regression models (Logistic Regression and Spatial Regression). They have revealed that Geographical Weighting, i.e. incorporating spatial correlation structure in regression, aids global regression models and enhances their predicting performance.

Decision Trees are often denoted as classification data mining algorithms that reveal complex relation between the elements (instances) in data structure. The advantage is that those algorithms are not true black-box models like ANN or SVM algorithms (Hwang et al. 2009). Instead, obtained hierarchical relations are observable in the most of the cases. There were a few attempts to utilize Decision Tree algorithms within the landslide susceptibility framework. In the case study from the South Korea (Hwang et al. 2009) very extensive work has been applied to the national database of engineered slopes. All the abovementioned authors yet admit that the task turned distasteful, due to the imperfections and more importantly, the size of the database, but they were persistent in their goal to rule-out the most important attributes. Another coupled case study from Japan (Saito et al. 2009) encountered similar problem that was handled by automatic and manual filtering of the database for incomplete or unwanted content. The ranking of the database attributes was required prior to the implementation of the decision tree algorithm. The algorithm then examined how chosen attributes were related to the Landslide Inventory. Their potential is more related to the Expert Systems design since the most of the Decision Tree techniques give an insight into the particular conditions that are potentially correlated with landslide occurrences (decomposing the tree to a congregation of rules gives an insight into the attribute-landslide relationship). Both studies came to a similar conclusion that some important relations could be ruled-out (even though expected, as in the case of the seepage and precipitation relations to the landslide occurrence) but majority of relations remained too complex for interpretation. Both of the studies have resulted in a proper landslide susceptibility replication based on modeled rules with some 70% of accuracy. One comparative study (Brenning 2005), which will be addressed in greater detail later on, asserts that Decision Tree method copes with overoptimistic assessments due to the overfit of the input data (even if the data are pre-processed and filtered), so the study preferred SVM and Logistic Regression models as alternatives to the Decision Tree.

One of the most popular and most broadly used Machine Learning technique in the landslide assessment field is multi-layered feed-forward (neurons are processed from one layer to another) ANN with back-propagation learning algorithm (Lee et al. 2007). Among the numerous case studies some pioneering works as well as the comparative studies are here to be mentioned. Initially, the problem of multi-dimensionality and non-linearity of input data were solved by prediction of system's behavior with ANN algorithms, rather than by its

complex and never complete mathematical or statistical model. This has been first experimented in the case studies in South Korea (Lee et al. 2004) where ANN procedure, trained over likelihood ratio result obtained fairly precise susceptibility models. Therein, overfitting had been addressed as a serious drawback and usage of independent testing area (not included in the training stage) was suggested as a precaution measure. As in the most of the other studies (Aleotti & Chowdhury 1999, Ermini et al. 2005, Kanungo et al. 2006, Caniani et al. 2008, Nefeslioglu et al. 2008), the utmost advantages of the method included: no need for particular data distribution, mixing of ordinal and nominal data and the generalization power of the algorithm. The drawbacks were recognized in GIS integration issues, time-consuming data preparation (data normalization), sometimes very demanding fitting of the parameters of the Neural Network and associated optimization problems of the back-propagation learning algorithm, and durable evaluation period. In comparison to the other methods (Logistic Regression, cluster analysis, fuzzy approach etc.) in mentioned studies, ANN was characterized as significant and perspective technique in the landslide susceptibility and hazard evaluation.

The practice of SVM in geo-spatial modeling has quite recent history. Pioneering the application in landslide susceptibility (Yao & Dai 2006, Yao et al. 2008) compared single-class vs. two-class (binary) SVM in the Hong Kong area. The authors demonstrated how the latter provided better conditions for algorithm training and testing, since it is clearly favorable to know both, where landslides exist and where they do not. Naturally, this brings about whole another dimension to the problem, since geotechnical engineering practice turns more reliable in determining where landslides occur than where they are being absent. Another study (Yuan & Zhang 2006) regards only one aspect of the landslide phenomena, i.e. the debris flows, by comparing SVM and fuzzy approach. Since it outperformed fuzzy method in the testing mode, SVM method was considered appropriate and more convenient for this kind of assessment in the area of interest (Yunnan Province, China). In the framework of geotechnical engineering, but from another - deterministic modeling aspect, SVM was used to overcome the calculation difficulties of implicit expressions of *Safety Index* (and derived *Factor of Safety*) (Zhao et al. 2008). SVM was used to predict *Factor of Safety* in several scenarios, and proved effective for the slope reliability analysis. This aspect of SVM is certainly more interesting in site-specific scale, where deterministic slope stability models prevail over statistical or probabilistic methods. SVMs were also proven suitable for large scale geological studies involving 3D modeling of geological bodies from the drill core data samples (Smirnoff et al. 2008). Nevertheless, similar philosophy could equally hold true in the case of geotechnical 3D models, for advanced thematic interpolation, be it particular geological stratification, groundwater table or stress and strain distribution. Thus, as long as there are methods to measure and monitor parameters of subsurface conditions, SVM seems to be capable in retrieving proper interpretation after optimal training over measured data. The use of SVM in geotechnical engineering for the seismic liquefaction phenomena assessment (Goh & Goh 2007) presented another aspect of large scale geotechnical hazard, solved in similar fashion, but the authors were not satisfied with revealing only the potential of liquefaction, not knowing the internal relations of the input seismic parameters that had driven it, and they intend to particularize that problem in the future. Piling-up the individual case studies usually turns problematic, since the scholars point out to the specific merits or shortcomings (Carrara & Pike 2008). It is rather comparative researches that are illustrating true value of the method. Few contributions have been made in this sense. The first to mention concerned a case study from the Ecuadorian Andes (Brenning 2005) by employing Logistic Regression, Decision Trees and SVM. The author emphasized the necessity of thorough input data preparation, and pointed to the overoptimistic accuracy of the Machine Learning techniques, yet turning less efficient than Logistic Regression model. Finally, more recent comparative research appeared (Yilmaz 2009), giving a very complete perspective on the landslide assessment methodology. Various modeling methods have been considered and compared, including ANN and SVM Machine Learning. The study shows, that several methods turned very precise and efficient. However, it also underlines the GIS compatibility

issue as a serious drawback. Those last two studies gain gravity for their practical contribution, as the first ones to compare various approaches and to analyze their suitability. Subsequently, a host of researchers, including the thesis author, have been attracted by these findings, and encouraged to experiment with SVM on their own, albeit in typical landslide susceptibility framework (Yao et al. 2008, Marjanović et al. 2011a) or in more specific cases, involving particular landslide mechanisms (Xu et al. 2012a,b). Relevance Vector Machines (Tipping 2001) is another quite similar classifier and regression module, which surpasses the SVM drawbacks by introducing probability of classification, similarly as Logistic Regression does. Even though there have been attempts to apply RVM in classification and regression scenarios (Samui et al. 2011), the method seems to be unexploited, yet very promising for some future attempts.

## 5 Methods and Procedures

This chapter is structured in several subsections which depict implementation of different methodologies at different stages of the research, i.e. feature selection methods, landslide assessment methods and model evaluation methods. In addition, some other specifications of the methodology are to be presented in the workflow. Details on the chosen software solutions are also discussed.

Instead of a general style that could be found elsewhere, in various textbooks and articles, this chapter is explaining these different methods in the light of the GIS landslide assessment, using respective examples and descriptions, which brings the topic closer and with better comprehension.

### 5.1 Attribute Selection Methods

*This branch of methods is featuring the Objective 3 (see Chapter 2).*

Attribute Selection, also referred to as Feature Selection and Variable Selection is a preprocessing<sup>11</sup> tool that has been found to be useful in spatial modeling, particularly in classification tasks, where sophisticated classifiers come into play, but it also contributes to any type of spatial calculations (Varmuza & Filzmoser 2009, Witten et al. 2011). In the landslide assessment framework, as regarded in this thesis so far, the attributes are called Conditioning Factors, represented by various thematic inputs (i.e. thematic layers in a GIS environment), including geological, morphometric, hydrological and environmental parameters, as well as synthetic parameters derived by discretization, reclassification or performing statistical operations over original inputs. In turn, there could be so many attributes that further application of a particular landslide assessment method could be compromised. Attribute Selection usually leads to the reduction of the number of input variables, i.e. the reduction of the dimensionality of the input dataset (Varmuza & Filzmoser 2009).

There are a number of arguments to support this end, and they mostly underline the reduction of time and computational efforts as primary benefits of the Attribute Selection (Varmuza & Filzmoser 2009). What is even more pronounced is the case of the classifiers (i.e. implementation of the Machine Learning or regression tasks in landslide assessment), where another important benefit comes through better control of overfit, that can appear due to the sheer abundance of input data. On the other hand, there are counter-arguments which claim that in any multivariate classification framework it is not particularly meaningful to exclude variables because of their internal relations. A variable might not be relevant for the landslide occurrence directly, but by affecting the other variables (van Westen et al. 2006). The ultimate true for Attribute Selection is thus in trial-and-error, as with any relatively novel and unexplored method. Sometimes it provides a better ground for further modeling, but sometimes it can compromise the results. Therefore, in this thesis the Attribute Selection has been performed in the preparatory stages of each case study, but only to justify the choice and to better portray the importance of different attributes, without exclusion of the particular ones (Tab. 3). The exception is the last case study, where Attribute Selection had been more engaged.

---

<sup>11</sup> Other preprocessing techniques have been implemented throughout the research but they have been regarded rather common and therefore excluded from the description. They involved different basic manipulations on input data, such as normalization of monotonous and dichotomous data, quantification of nominal data, integerization, transformation, scaling, and so forth.

There are numerous Attribute Selection schemes elaborated thus far, starting from univariate and bivariate (filtering), to more complex ones, such as Principal Components or even those which are implementing learning tasks (wrapping) prior to the classification itself. The latter is an example usually related with the Decision Trees and successive linear classifiers (such as SVM). Decision Trees are actually based on the internal Attribute Selection (which takes place during the initial process of populating the tree) through which some of the attributes might never satisfy the criteria and enter the tree. If it turns out that the tree had gave good results, then only those attributes approved through the tree building procedure could have been fed to a chosen model (e.g. model which is very sensitive to the relevance of each attribute, like *k*-NN), leading to the improved performance of the latter model. Also, some linear models (linear classifiers), which are very appropriate for iterative, sequential adjustment of the internal modeling parameters, may have integrated Attribute Selection within the classification process, by starting with all of the available attributes and eliminating the ones with the lowest rank in every consecutive iteration (leave-one-out<sup>12</sup>). The optimal assembly of the attributes will appear in the model that shows the best performance (Mitchell 1997, Witten et al. 2011). However, it has been mentioned that the purpose of Attribute Selection in this thesis is more formal than that (the third case study is an exception – see Chapter 6.3.3), and serves mostly to ensure and justify the choice of attributes proposed in the literature.

In parlance of the latter, two statistical tests have been considered, each one with the emphasis on tracking the attributes with the weakest relation to the dependent variable (referent landslide classes). These are rather simple solutions and include bivariate descriptor – *Chi-Square* and entropy-based descriptor – *Information Gain (IG)*. One ranking scheme based on these two methods is shown in Table 3. below, as a part of the second case study within this thesis (see Chapter 6.2).

**Table 3.** Attribute ranking for Starča Basin case study

<b>Conditioning Factor</b>	<b>Chi-rank</b>	<b>Chi-Square</b>	<b>IG-rank</b>	<b>IG</b>
<i>lithology</i>	1	11225.111	1	0.06157
<i>channel network base elevations</i>	2	8546.6167	2	0.04034
<i>groundwater depth</i>	3	5311.0166	7	0.02680
<i>Stream Power Index (SPI)</i>	4	4626.1226	4	0.03038
<i>aspect</i>	5	4469.4704	5	0.02828
<i>altitude above channels</i>	6	4442.2272	3	0.03078
<i>Topographic Wetness Index(TWI)</i>	7	4223.8196	6	0.02789
<i>Land Cover</i>	8	3823.2823	10	0.02129
<i>downslope gradient</i>	9	3401.1757	8	0.02413
<i>LS factor</i>	7	3052.666	9	0.02241
<i>slope angle</i>	11	2749.7677	11	0.02100
<i>convergence index</i>	12	2441.8224	12	0.01723
<i>plan curvature</i>	13	1298.7882	14	0.00800
<i>distance from structures</i>	14	1255.0141	13	0.00938
<i>profile curvature</i>	15	948.26960	15	0.00605
<i>slope length</i>	16	940.12422	16	0.00600

<sup>12</sup> Leave-one-out technique considers iterative learning task. After each run the dimensionality (in this case the number of landslide Conditioning Factors) is reduced for one dimension (one factor) on the basis of model error or Feature Selection rank. Such reduced feature space then reenters the learning process until the optimal performance is achieved, i.e. in accordance with the Occam's razor, until a sufficient performance is reached with the minimal feature space size (Mitchell 1997).

### 5.1.1 Chi-Square

The *Chi-Square* is based on a cross-tabulation of dependant variable (Landslide Inventory in this case) with all of the independent variables (Conditioning Factor, i.e. geological, morphometric, hydrological, environmental or synthetic terrain attributes). *Chi-Square* statistic parameter  $\chi^2$  relates the apparent frequencies of the observed independent variable instances  $\varphi_o$  within the dependent variable classes (landslide classes), and expected frequencies of the observed independent variable  $\varphi_e$ , in the following fashion:

$$\chi^2 = \sum_{i=1}^l \sum_{j=1}^n \frac{(\varphi_{o,i} - \varphi_{e,i})^2}{\varphi_{e,i}}, \quad (3)$$

where  $l$  is the number of classes of a dependent variable, and  $n$  the number of the independent variable classes, i.e.  $l$  represents Landslide Inventory classes (landslide classes, such as *active*, *dormant*, *abandoned landslides* or simply, *landslide* and *non-landslide* class), while  $n$  disclose the classes of a particular Conditioning Factor, since  $\chi^2$  needs to pair every single factor with the dependent variable separately. A given Conditioning Factor disapproves the hypothesis of being statistically independent from the Landslide Inventory classes only if it fails to exceed the critical  $\chi^2$  threshold, defined by the level of confidence (in respect with the normal distribution) and degrees of freedom (defined by reduced product of  $l$  and  $n$ , i.e.  $(l-1)(n-1)$ ). In effect, this method reveals the relation of an attribute and the referent Landslide Inventory, but the ranking among multiple attributes is rather relative (unless the sets are subjectively normalized), mostly due to the measurement scale and unit dependence of  $\chi^2$  (Bonham-Carter 1994).

### 5.1.2 Information Gain (IG)

The second employed technique is *Information Gain*, defined as a reduction in entropy  $E(C)$  of a referent Landslide Inventory  $C$  (with  $l$  classes), due to the informational interference of a Conditioning Factor  $F$  (with  $n$  classes). Given the  $E(C)$  as a measure of homogeneity of  $C$ :

$$E(C) = - \sum_{i=1}^l \delta_i \log_2 \delta_i, \quad (4)$$

(where  $\delta_i$  is a proportion of the  $i^{\text{th}}$  class values within the entire set) and introducing  $m$  factor classes with values  $v_1, v_2, \dots, v_m$ , the *Information Gain*  $IG(F)$  partitions the entropy by a factor of weighted expected entropy  $E(F, v)$ .

$$IG(F) = E(C) - \sum_{v \in \{F_{1..m}\}} \frac{|C_v|}{|C|} E(C_v) \quad (5)$$

The latter comes as summed entropy of  $C_v$  subsets of  $C$ , matched with the factor's class value  $v$ , and weighted by the subset proportion to  $C$ .

This technique is integrated in the Decision Tree algorithm, and is going to be presented in greater detail in section 5.2.3.2. Here, it should be mentioned that unlike *Chi-Square* statistic, this parameter allows preliminary ranking, since it disregards measure scales and units of Conditioning Factors (Mitchell 1997).



## 5.2 Landslide Assessment Methods

*This branch of methods is featuring the Objective 4 (see Chapter 2).*

Landslide assessment methods used in this thesis involved numerous techniques, structured in several different approaches: heuristic, statistic, Machine Learning and deterministic. Herein, these are to be presented in detail, with particular focus on the Machine Learning approach.

### 5.2.1 Heuristic Approach

Heuristic approach implies the experience-based solution to the problem. Using heuristics is rather controversial issue in the landslide assessment (Barredo et al. 2000, Ercanoglu et al. 2008), but it is generally accepted that heuristics could and should be used for preliminary levels of research. It may be used for more detailed levels of research only when combined with more exact approaches. Even within the heuristic approach one can distinguish many different techniques, ranging from plain expert-opinion modeling through techniques that quantify the raw expert judgment to some extent and methods that include fuzzy/gray systems, pattern recognition etc.

#### 5.2.1.1 Plain Multi-Criteria Analysis

Multi-Criteria Analysis (MCA) has not been originally developed for the spatial modeling problems, but rather for decision support with multiple choices, when there are disagreements between different parties that offer those choices. However, it has been successfully applied in the landslide assessment with the analogically proposed problem.

In the most basic variant, the MCA comes down to the judgment of the importance of the multiple factors, which is the basic case in the landslide assessment framework. Herein, the choices are made on the arbitrarily assigned weights of importance of a Conditioning Factor,  $w_i$  (i.e. its level of influence on the landslide susceptibility), by a shear expert's judgment, or several experts' judgments<sup>13</sup>. Thus, each particular factor (such as lithological unit, slope angle, aspect, elevation, drainage buffer, Land Use unit and so forth) is being weighted (Eq. 6) and simple addition of the weighted factors delivers a MCA model in the GIS environment.

$$M_{MCA} = \sum_{i=1}^n w_i F_i = M_{AHP} \quad (6)$$

In the case of multiple experts' judgments, it is common that the weights vary significantly, so that averaged weights  $\bar{w}_i$  could come into play (Eq. 7).

$$M_{MCA} = \sum_{i=1}^n \bar{w}_i F_i \quad (7)$$

Even though the pool of experts might be chosen cautiously (so that the experts meet all the requirements, i.e. that they are very familiar with the approach and with the study area at hand), it is important to stress that the method faces a very high level of subjectivity, arising from entirely arbitrary judgments. Still, the method can be substantially refined, while still carrying that precious touch of personal experience in particular landslide hazard problem.

<sup>13</sup> These judgments are usually delivered by filling in the prepared questionnaires, where the importance (weight) of each factor (lithological unit, slope angle, aspect, elevation, drainage buffer, Land Use unit and so forth) is being scored on a custom predefined scale, usually 0–9, or 1-10.

### 5.2.1.2 Analytical Hierarchy Process

Analytical Hierarchy Process (AHP) was also developed for decision making based on economic and management grounds (Saaty 1980), but turned out to be a proper way of quantifying expert's judgment involved in the spatial analysis within GIS framework. In fact, it is now incorporated in a host of GIS platforms (commercial, like ArcGIS and open source, like SagaGIS for example) in the form of different modules, macros extensions or add-ins (Marinoni 2004). Another arising possibility is to combine the AHP quantitatively with some other technique (Komac 2005, 2006), e.g. bivariate statistics, Fuzzy Logic, cluster analysis, etc.

AHP is a convenient procedure for raster-based modeling in a multi-criteria hierarchical configuration (Geniest & Rivest 1994, Ercanoglu et al. 2008) and has thus been equally applied in landslide assessment, as in any spatial modeling framework (Komac 2006, Ercanoglu et al. 2008). It is important to mention that the AHP implementation in spatial analysis is usually restricted to the first level AHP, since true AHP implies  $k$ -fold structure, where levels from the 1<sup>st</sup> through the  $k-1$ <sup>th</sup> involve criteria analyses, while the last  $k$ <sup>th</sup> level involves selection of the alternatives<sup>14</sup>. Herein, the procedure is going to be explained in detail and illustrated in such context.

Prior to the obtaining of the true weights of the corresponding landslide factors (such as *lithology, slope angle, aspect, elevation, Land Use* and so forth), the procedure engages a gross estimation of each factor's importance score, established by the expert(s) judgment (through a personal advisement with the scholars and engineers or formally through the questionnaires). If  $n$  is the number of Conditioning Factors, then the total number of comparisons that an expert needs to establish is  $n(n-1)/2$ , which makes this procedure comfortable for no more than a dozen of factors.

The original technique (Saaty 1980) implies 9-leveled scoring scale<sup>15</sup>, but a different (arbitrary) range is also viable. The 9-leveled scoring system is then applied to a two-dimensional  $n \times n$  reciprocal matrix, also called the comparison matrix (Tab. 4), which is generated by pair-wising all of the factors across each other. Note that the scores are being transposed over the main diagonal of the matrix, so that the corresponding scores (1 through 9) turn reciprocal (1 through 1/9) symmetrically over the main diagonal.

To obtain the priority vector<sup>16</sup> as a vector of weights (Tab. 5, shaded column), the procedure further requires a normalization of the comparison matrix and averaging of scores from comparison matrix (Tab. 4) by their row sums (Geniest & Rivest 1994, Saaty 2003). Priority vector will represent the utter distribution of the weights  $w_i$  once the matrix turns consistent, i.e. when there is none or little contradiction in scoring. Since the vector is normalized, the weights sum should be 1 (100%). The procedure for shifting from inconsistent to near-consistent matrix is featured by versatile solutions, considered by

<sup>14</sup> In the landslide assessment, this would involve multi-leveled criteria analysis, where  $i$ <sup>th</sup>-level criteria set would be progressively singled-out for a Conditioning Factor that have turned the least important in the  $i-1$ <sup>th</sup> (previous level) criteria analysis. The alternatives would normally represent different landslide types. It could then be speculated which factors (lithological, morphometrical, environmental, hydrological etc.) would most strongly affect which landslide type (shallow landslide, debris flow, rockfall etc.). However, the knowledge and experience on the relations between a landslide type and Conditioning Factors is far from profound, which is why the spatial AHP analysis usually sticks with the 1-fold variant.

<sup>15</sup> Saaty's scale contains an array of {1/9, 1/8, 1/7, 1/6, 1/5, 1/4, 1/3, 1/2, 1, 2, 3, 4, 5, 6, 7, 8, 9}, where the following relations apply: 1 – equal importance (landslide-wise), 3 – weak dominance of the observed factor, 5 – strong dominance, 7 – demonstrated (witnessed) dominance, 9 – absolute dominance, 2,4,6,8 – transitive scores.

<sup>16</sup> Computationally, the priority vector is very similar to the principal Eigen vector of the matrix.

different authors (Geniest & Rivest 1994, Laininen & Hämäläinen 2003, Saaty 2003). However, their results prove to be just a fraction different from the simplified technique. Thus, it is understandable to control the matrix consistency on the simplest basis, i.e. by Saaty's consistency parameters *CI*, *RI* and *CR* (*Consistency Index*, *Random Index* and *Consistency Ratio*, respectively) using the following criterion:  $CR=(CI/RI)<0.1$ . By this manner, initial subjectivity of score distribution (Tab. 4) has been unbiased up to a certain level, leaving the refined scores depicting the final distribution of weights in the priority vector (Tab. 5, shaded columns). Finally, the priority vector or more appropriately, the normalized linear distribution of the weights can be defined as before (Eq. 6), where *F* corresponds to the Conditioning Factor, respective to their order of appearance in Table 4 ( $F_1 =$  lithology,  $F_2 =$  slope...  $F_7 =$  aspect), and  $w_i$  refers to the factor's weight, which reflects its overall importance in the landslide susceptibility model. The model is then directly calculated by multiplying and adding the appropriate variables in a GIS environment. The weights in the model are simply the multipliers of the thematic GIS layers, as they multiply each pixel (its *Digital Number – DN* value) of each raster layer and then sum all (multiplied) layers together, yielding a final raster model – the raw model of landslide susceptibility. It depicts spatial distribution of the susceptible zones (revealing low susceptibility by low, and high susceptibility by high overalls) in a custom scale. The custom scale is inappropriate so normalization procedure is used to arrange the scale in a more common fashion, e.g. in 0–1 span or 0–100%. It is further possible to choose more appropriate cut-offs and qualify intervals arbitrarily, e.g. Low, Moderate, and High susceptibility. However, the arbitrary (re)classification of such kind is entirely another issue, beyond the AHP scope, and depends on the particular case at hand.

**Table 4.** An example of AHP comparison matrix.

$F_i$	$F_1$	$F_2$	.	$F_n$
$F_1$	$a_{11}$	$a_{12}$	.	$a_{1n}$
$F_2$	$a_{21}$	$a_{22}$	.	$a_{2n}$
.	.	.	.	.
$F_n$	$a_{n1}$	$a_{n2}$	.	$a_{nn}$
$\Sigma$	$\Sigma a_{1n}$	$\Sigma a_{2n}$	.	$\Sigma a_{nn}$

**Table 5.** An example of AHP weights derivation.

$F_i$	$F_1$	$F_2$	.	$F_n$	$w_i$	%
$F_1$	$a'_{11}(=a_{11}/\Sigma a_{1n})$	$a'_{12}$	.	$a'_{1n}$	$w_1 (= \Sigma a'_{1i}/n)$	$100 \cdot w_1$
$F_2$	$a'_{21}$	$a'_{22}$	.	$a'_{2n}$	$w_2$	$100 \cdot w_2$
.	.	.	.	.	.	.
$F_n$	$a'_{n1}$	$a'_{n2}$	.	$a'_{nn}$	$w_n$	$100 \cdot w_n$
$\lambda_{max} = \dots; CI = \dots; RI = \dots; CR = \dots (CR < 0.1);$					$\Sigma = 1$	$\Sigma = 100$

## 5.2.2 Statistical Approach

Unlike expert-based approach above (which relies on experience and knowledge on the process), or deterministic approach described later on (which starts with theoretical conditions of the Limit Equilibrium on the landslide-affected slope), statistical approach starts with the available data (Conditioning Factors). It relates the values, distributions, aggregations and other data features with the consequence, in this case, a landslide occurrence. It extracts that relation and brings about a more objective prognostic dimension

to the model, although the prognosis is not temporal, only spatial, as explained in Chapter 3.3.2.

There is a host of various techniques that are normally used in this context, and choosing the optimal one is rather based on trial-and-error. In addition, they tend to provide similar modeling performance, according to the results of this research. One can distinguish between bivariate statistics and Multivariate techniques which are principally different in one single postulate, regarding the independency between the Conditioning Factor and landslide occurrence. While bivariate techniques observe one factor at a time and correlate it to the landslide occurrence, relying on its independence from the rest of the factors, Multivariate techniques correlate all the factors among themselves and factors with the landslide occurrence simultaneously (Bonham-Carter 1994).

### 5.2.2.1 Conditional Probability

In plain statistical terms, the primary goal of Conditional Probability is to increase the probability of predicting a variable, by having other variables at disposal to correlate against (Eq. 8).

$$p(C, F) = \frac{\#\{C \cap F\}}{\left\{ \frac{\#F}{T} \right\}} \quad (8)$$

$C$ : 1- $l^{\text{th}}$  landslide units in the area

$F$ : 1- $m^{\text{th}}$  class unit of 1- $n^{\text{th}}$  factor in the area

$T$ : total number of units (pixels) in the area

$p(C, F)$ : posterior (conditional) probability of landslide occurrence given the specific  $F$

In landslide assessment terms, the referent variable is landslide occurrence pattern and the variables that enter the correlation with the former are given as Conditioning Factors (lithological, morphometric, environmental etc.). Pattern of landslide occurrence has its own initial spatial probability, called prior probability. In bivariate context, the Conditioning Factors can be checked for correlation against the referent landslide pattern one at a time, yielding the posterior probability per each factor (spatial probability after the correlation). If the significant correlation is achieved, the posterior probability will be higher than prior spatial probability. Eventually, the posterior probabilities of all factors should be cumulated, giving the total increase of the probability, hence giving the model of spatial probability i.e. landslide susceptibility. There are a number of mechanisms for developing a Conditional Probability analysis, and they all involve different weighting measures, such as Likelihood Ratios, various Odds Ratios, Weights of Evidence, and so forth. Herein the latter shall be discussed in detail, while some of the ratios are also used and explained in the Fuzzy Logic section (see Chapter 5.2.2.2).

Weights of Evidence is a log-linear Bayes rule-based technique. It enables the prediction of a posterior probability of landslide occurrence by using its prior probability and enhancing it by weights, which are dependent on Conditioning Factors i.e. by generating correlative positive  $W^+$  and negative  $W^-$  weights of Conditioning Factors. Positive weights are differences between prior and posterior logits<sup>17</sup> of a Conditioning Factor's class, given the presence of landslide occurrences. Negative weights are also prior/posterior logits differences, given the absence of landslides, i.e. given the presence of non-landslides. The simplest case is to assume a binary class type of landslide occurrence (*landslide* and *non-landslide* classes), but multi-class types (*non-landslide*, *dormant*, *active*, *fossil*, *suspended* and other standard inventory classes) are also feasible, by simply extracting 1-1 binary cases

<sup>17</sup> Logit of a given probability  $p$  is defined as:  $\text{logit}(p) = \frac{1}{2} \cdot \ln\left(\frac{p}{1-p}\right)$ .

( $l$  being the number of landslide classes). What is important in this technique is that the weighting (extraction of the prior/posterior probabilities) in a GIS environment can be done by simple raster cross-tabulation (contingency table) of the Landslide Inventory on one side, and a classified Conditioning Factor on the other. From cross-tabulation matrix, the contingency portions of  $C \cap F$  (overlap of *landslide* class and the given factor's class),  $\neg C \cap F$  (overlap of *non-landslide* class and the given factor's class) and the other respective overlaps ( $C \cap \neg F$  – landslides out of class,  $\neg C \cap \neg F$  – non-landslides out of class) are easily extracted, and fed into equations (Eq. 9, 10).

$$W^+ = \ln \frac{\#\{C \cap F_i\}}{\#\{\neg C \cap F_i\}} \quad (9)$$

$$W^- = \ln \frac{\#\{C \cap \neg F_i\}}{\#\{\neg C \cap \neg F_i\}} \quad (10)$$

Thus, each class in each factor is provided with a pair of weights that describe the intensity and the character ( $\pm$ ) of its engagement with the landslide occurrence pattern. GIS environment eases the further operational effort toward the final spatial probability model. By assuming the conditional independence of all  $n$  Conditioning Factors (the prime postulate of bivariate statistics) the latter can be expressed as prior probability of landslides corrected by the sum of positive/negative weights of all  $m_i$  Conditioning Factor's classes, depending on the presence or absence of landslides, respectively (Eq. 11).

$$\text{logit}(p(C, F)) = \text{logit}(C) + \sum_{i=1}^n \sum_{j=1}^{m_i} W_{i,j}^{\pm} \quad (11)$$

Simple spatial calculation over a raster set of Conditioning Factors in respect to equation (Eq. 11) yields a final landslide susceptibility (raster) map (Bonham-Carter 1994).

It is also feasible to calculate the certainty of every class of every factor<sup>18</sup>, and mask all data instances that are too uncertain or missing.

Advantages of the technique certainly lies in objective, data-driven assessment, capability to operate with multiple inputs and multiple classes, ease of handling of missing or irrelevant data (by assigning zero or close-to-zero weights), to accompany the result with the certainty estimate and to make according exclusions where needed.

On the other hand, shortcomings are not numerous. These mostly regard the conditional independency assumption (bivariate postulate), which disables the interrelation among the factors, and the contribution which it might bring to the model. Furthermore, it does not work with nominal Conditioning Factors, but requires quantification and normalization prior to the analysis, while another promising, but time-consuming and computationally-intensive alternative is to segregate such nominal Conditioning Factors into  $m$  binary cases. Also, the numeric Conditioning Factors with ordinal and dichotomous scales cannot keep continual values, but have to be reclassified and ranged by arbitrary intervals, which can be optimized only through the trial-and-error variations of the number of classes ( $m$ ) and values of interval cut-offs. This turns very demanding with large number of such inputs, not to mention how it introduces considerable subjectivity in the procedure. Finally, the drawback is also high dependency on the proportion of independent variable's classes, in this case, the number of *landslide* vs. *non-landslide* instances, since very few landslides will not provide as reliable model as some other, more sophisticated techniques.

<sup>18</sup> Certainty could be extracted from standard deviation  $\sigma$  of posterior probabilities by approximate relations:  $\sigma(W^+) = \frac{1}{\#\{C \cap F\}} + \frac{1}{\#\{\neg C \cap F\}}$  and  $\sigma(W^-) = \frac{1}{\#\{C \cap \neg F\}} + \frac{1}{\#\{\neg C \cap \neg F\}}$ .

Relative certainty is then calculated as  $W^{\#}/\sigma(W^{\#})$ , which is usually called "Studentization".

### 5.2.2.2 Fuzzy Logic

Concepts of Fuzzy Logic have a long tradition in spatial analysis framework. The main purpose of Fuzzy Logic is to deal with vague information and with data that contain some kind of uncertainty (Zadeh 1965). When using the Fuzzy Set Theory or Fuzzy Logic in the landslide assessment, each class of each Conditioning Factor is given a value within 0–1 interval, indicating its membership to the landslide occurrence. This concept is very helpful for categorization of data and for decision making in landslide management framework, because unlike Boolean Logic it produces valid results with specific degree of truth. That helps finding not only the perfect match for a given criteria (in this case the criteria is to discern landslide susceptibility zones, i.e. to point-out to highly susceptible areas), but also showing how much each of the possibilities meet the given criteria. At some specific situations, when modeling physical geographical crisp sets, Boolean Logic fails to provide good results because of the nature of the phenomena at hand. In such cases, Fuzzy Set Theory and Fuzzy Logic provide solutions for dealing with imprecise and vague data, which would be hard or even impossible to process by any other means.

The most delicate part of the problem regards the fuzzy membership. Its value is determined by a membership function (Eq. 12), which stands for a function that maps all given elements to 0–1 interval of values:

$$\mu_C : F \rightarrow \langle 0,1 \rangle , \quad (12)$$

where  $\mu_C$  is a membership function (in this case of the *landslide* classes),  $F$  is a set of elements (in this case Conditioning Factors classes). Then for each instance value  $\mathbf{x}$  ( $\mathbf{x} \in F$ ),  $\mu(\mathbf{x})$  is a membership value of that instance (pixel of the grid) to the referent landslide set  $C$  (Zadeh 1965). There are a number of ways to express the membership, involving linear and non-linear functions. For purpose of this research, two common functions have been used for computing the membership values: *Frequency Ratio* and *Cosine Amplitude*.

*Frequency Ratio (FR)* gives proportion of the *landslide* instances in the specific class for each of the Conditioning Factors. It can be described as a ratio of relative frequency of *landslide* instances in the particular class to the relative frequency of all *landslide* instances in the area (Eq. 13):

$$FR = \frac{\#(C_j)/\#(F_i)}{\#(C)/\#(T)} , \quad (13)$$

where  $\#(C_j)$  is the number of *landslide* cells in the  $j^{\text{th}}$  class of a Conditioning Factor,  $\#(F_i)$  is the total number of instances of the  $i^{\text{th}}$  class of factor  $F$ ,  $\#(C)$  is the total number of all *landslide* instances and  $\#(T)$  is the total number of instances in the grid (total number of pixels). If the result is higher than 1 it shows higher density of *landslide* cells in the class than the dataset overall. Results lower than 1 point to the classes that have density of landslides lower than the dataset overall. To transform *FR* to membership values those outputs have to be normalized by dividing each *FR* by  $FR_{max}$  in the given group of classes. The membership values are then ordered into 0–1 interval (normalization). Higher this number is the higher is the influence of that particular class on the landslide occurrence.

Another method for determining the membership values of factor's classes to the set of landslide occurrence is *Cosine Amplitude (CA)* method (Eq. 14).

$$CA = \frac{\#(C_j)}{\sqrt{\#(F_i) \cdot \#(C)}} \quad (14)$$

In this case, the membership value is calculated as a ratio between the number of *landslide* instances in the class and the square root of the class size (in number of instances) product with the total number of *landslide* instances in the area. Unlike *FR* the output values do not have to be normalized because they already fall into 0–1 interval.

Once the fuzzy memberships are determined, the issue of the combinations between the sets (sets of Conditioning Factors) remains. The latter is solved by involving different fuzzy operators and choosing the optimal, at one or multiple levels. The best-known operators are AND and OR, but both of them suffer from the problem of extremes, i.e. one of the combined sets are preferred and their influence is predominant. In case of the operator AND the factor with the minimal membership classes is the one that shapes the output, while in the case of OR operator it is the one with the maximal membership values. The Fuzzy Algebraic Product, Fuzzy Algebraic Sum, Gamma Operation and Weighted Average are proposed as more objective solutions (Bonham-Carter 1994). Herein are given some basic relations that apply for these, but first it is necessary to acknowledge Equation 15. for better notation purposes used later on:

$$t = \sum_{i=1}^n m_i, \quad (15)$$

where  $m_i$  is the number of classes in  $i^{\text{th}}$  Conditioning Factor and  $n$  is the number of Conditioning Factors, which then makes  $t$  a total number of all membership functions to be combined. The above means that a membership function is calculated for every class of every Conditioning Factor at hand. In the Fuzzy Algebraic Product and Fuzzy Algebraic Sum the memberships are defined as:

$$\mu_{product} = \prod_{j=1}^t \mu_j(\mathbf{x}), \quad (16)$$

$$\mu_{sum} = 1 - \prod_{j=1}^t (1 - \mu_j(\mathbf{x})), \quad (17)$$

respectively, where  $\mu_j(\mathbf{x})$  is the membership function of the instance  $\mathbf{x}$  belonging to one of the  $m$  classes of the Conditioning Factor  $F_j$ . The Fuzzy Algebraic Product tends to produce output function lower or equal to the lowest function given, while the Fuzzy Algebraic Sum is complementary to the former, so it provides output function higher than all the inputs but never higher than 1. Regarding (Eq. 16–17) Gamma Operation can be defined by:

$$\mu_\gamma = (\mu_{sum})^\gamma \cdot (\mu_{product})^{1-\gamma}. \quad (18)$$

The exponent  $\gamma$ , which is a number from 0–1, allows optimization of the membership combination because it balances between  $\mu_{product}$  and  $\mu_{sum}$ . Setting  $\gamma$  to the extremes give either Fuzzy Algebraic Sum ( $\gamma=1$ ) or Fuzzy Algebraic Product ( $\gamma=0$ ). Weighted Average is defined as:

$$\mu_w = \frac{\sum_{j=1}^t w_j \cdot \mu_j(\mathbf{x})}{\sum_{j=1}^t w_j}, \quad (19)$$

where  $w_j$  is the weight of the  $j^{\text{th}}$  membership function, indicating its importance to the result. Weight system in this equation allows more interaction of the practitioner, because it allows emphasis of certain classes by choosing their weights arbitrarily.

### 5.2.2.3 Multivariate Regression Analysis

In contrast to the conventional statistics, based on the bivariate Bayes approach, stands a group of Multivariate techniques, which are found to be very convenient in spatial analysis, and much more elaborated, for that matter. The principal advantage revolves around the fact that multiple factors are examined simultaneously, and that their interplay is allowed. As mentioned before, independent variables that are poorly correlated with the dependant variable are not necessarily useless, because they could be mutually correlated with other

independents (which are for example very well correlated with the dependant), and therefore influence the final posterior probability of the dependant. Regression methods turned out to be among the most favorable ones (Brenning 2005), although the majority of Machine Learning techniques and discriminant analysis could be also classified as Multivariate. Hence, the struggle for supremacy in such context is tight, but it should be emphasized that the general preference on the most favorable technique cannot be made, since every case is specific in its own way. The only acceptable approach is to pair together several different techniques, which are initially suspected as effective, and then make a choice toward the optimal one.

Regression methods are numerous, and inconveniently, they all require some delicate presumptions that sometimes might not apply in reality (Süzen & Doyuran 2004). The crucial among these is the assumption on normal data distribution within all the variables, linear relationship between the independent variables (Conditioning Factors) and the preference of binary format of the dependent variable (i.e. Landslide Inventory is preferred to be structured by *landslide* vs. *non-landslide* classes, while multiclass Landslide Inventory poses a problem insurmountable for some of the regression methods, but for others, like ANN-regression and Logistic Regression it is not particularly difficult). Since Logistic Regression seems to be the one of the best elaborated methods in geo-spatial context (Brenning 2005), and requires the least presumptions to operate, the following passages are to communicate some further details on that technique.

Depending on the number of classes within the dependent variable, two separate cases of Logistic Regression could be distinguished, binomial and multinomial (Varmuza & Filzmoser 2009). Since the most usual cases in landslide assessment come from the analysis of binomial or binary dependants (Landslide Inventories), wherein *landslides* (=1) and *non-landslides* (=0) are the only classes, this simpler variant of Logistic Regression is going to be considered.

The principal difference of the Logistic Regression to the other regression methods is transformation of dependant into a logit variable (a natural logarithm of the occurrence odds of dependant variable, in this case the landslide occurrence odds). This implies fewer constrains regarding data distribution, linearity between independent variables, same type of data etc. In this way, the regression takes place on logit-transformed dependant variable, unlike other regression schemes (Fig. 8). Given  $n$  Conditioning Factors, the regression is performed as linear combination coordinates  $x_i$  of instances  $\mathbf{x}$  (as in any other multiple regression) via linked linear function  $z$  (Eq. 20), and posterior probability is finally modeled as a logistic function, which is basically a sigmoid function for *landslide* probability (Eq. 21), and inverse sigmoid function for *non-landslide* probability (Eq. 22).

$$z = b_0 + \sum_{i=1}^n b_i \cdot x_i \quad (20)$$

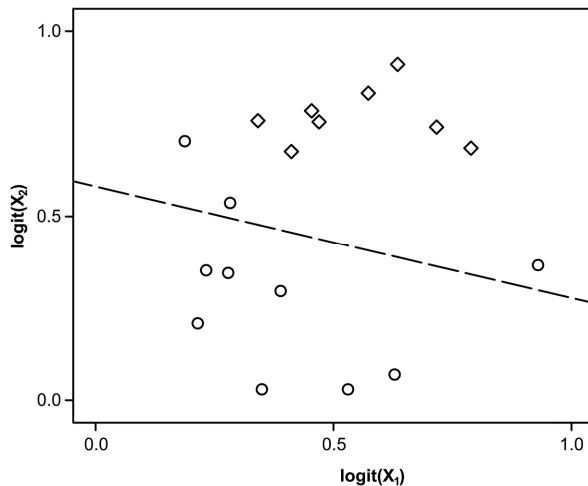
In the Equation 20.  $b_0$  stands for the interception of the fitted regression model and  $b_i$  for the regression coefficient (regression function slope) of the  $i^{\text{th}}$  independent variable  $x_i$ .

$$p_{\text{landslides}} = \frac{e^z}{1 + e^z} \quad (21)$$

$$p_{\text{non-landslides}} = \frac{1}{1 + e^z} = 1 - p_{\text{landslides}} \quad (22)$$

The calculated posterior probability of occurrence of both, *landslide* and *non-landslide*, falls into 0–1 range and sums-up to 1, since posterior probabilities of *landslide* and *non-landslide* instances are complement (Varmuza & Filzmoser 2009).





**Figure 8.** An example of Logistic Regression classification. Two random Conditioning Factors are given on the axes, *landslide* instances are represented by circles and *non-landslide* by squares, separation hyper plane (presented as dashed line) is a regression line with  $(b_0, b_1)$  parameters -0.23 and 0.6.

As indicated, the advantages are numerous:

- 0–1 (spatial) probability as a result (which is convenient for interpretation and comparison with other probabilistic methods),
- possible usage of mixed data types of independent variables (nominal and ordinal),
- less stringent requirements regarding the data: linearity, normal distribution, homogeneity of variance etc.

### 5.2.3 Machine Learning Approach

Machine Learning had to be singled-out as a separate approach for at least three reasons.

- The first one is that it represents an emerging field of computer science which studies computer algorithms that improve automatically through experience (Mitchell 1997, Kanevski et al. 2009, Witten et al. 2011). This learning concept is therefore different than any of the mentioned modeling approaches, and empowers some additional predictability in spatial domain (as exemplified in the case studies section, Chapter 6).
- The second is that it represents a mixture of so many different disciplines, which makes it difficult to subcategorize Machine Learning under a more general approach. It is defined as an interdisciplinary field, built on many different concepts, such as: probability and statistics, artificial intelligence, information theory, as well as philosophy, psychology, neurobiology and so forth (Mitchell 1997, Kanevski et al. 2009, Witten et al. 2011).
- The third is that it represents the essence of this thesis research, and therefore it deserves slightly higher hierarchical position than the other approaches. In fact, the advanced methods in the title of this thesis are mostly regarding the methods that will be presented hereinafter.

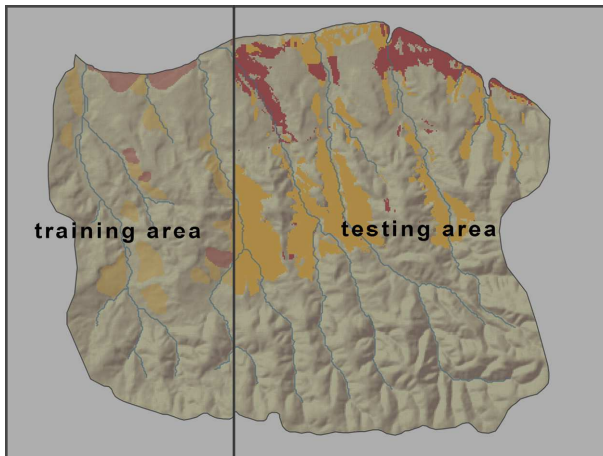
There is a host of different algorithms, which are all exploiting different capabilities for different learning tasks, such as clustering, classification and regression. Herein, these are going to be limited to the classification-related algorithms, among which some less prominent algorithms are preferred. For instance, the better-known algorithm and one of the first that had implemented artificial intelligence (based on neurobiological analogy), Artificial Neural Networks (ANN), turned out to be less comfortable for the implementation, with more parameters to model and with costly time consumption, than for example Decision Trees,

which are additionally providing the practitioner with the insight in hierarchical structure of the model, “lightening-up” the black-box learning concept.

### 5.2.3.1 (Supervised) Learning Problem Formulation

Before the particular techniques are scrutinized, it is necessary to pose a learning problem and make the problematic more comprehensive. This is to be illustrated for the landslide assessment framework, in order to make the later descriptions of different Machine Learning algorithms more appealing.

The main objective is to exploit the possibility of automating the process of landslide susceptibility mapping or landslide mapping, i.e. to make a plausible prediction of landslides spatial distribution by using Machine Learning techniques. The desired automated procedure assumes that after the initial acquisition of the necessary spatial data, an expert is presented with a (possibly small) representative region (training region). Such scenario assumes a supervised learning approach in which the expert performs mapping in the representative region. The algorithm subsequently uses that expert map for training, i.e. learning from instances of the expert map by linking his interpretation with a set of Conditioning Factors. Finally, after learning the mapping rule proposed by the expert, the algorithm extrapolates the rule in the rest of the area, and gives an automated prognosis of the spatial distribution of landslides.



**Figure 9.** An example of a manual training-testing area split (33-67% proportion).

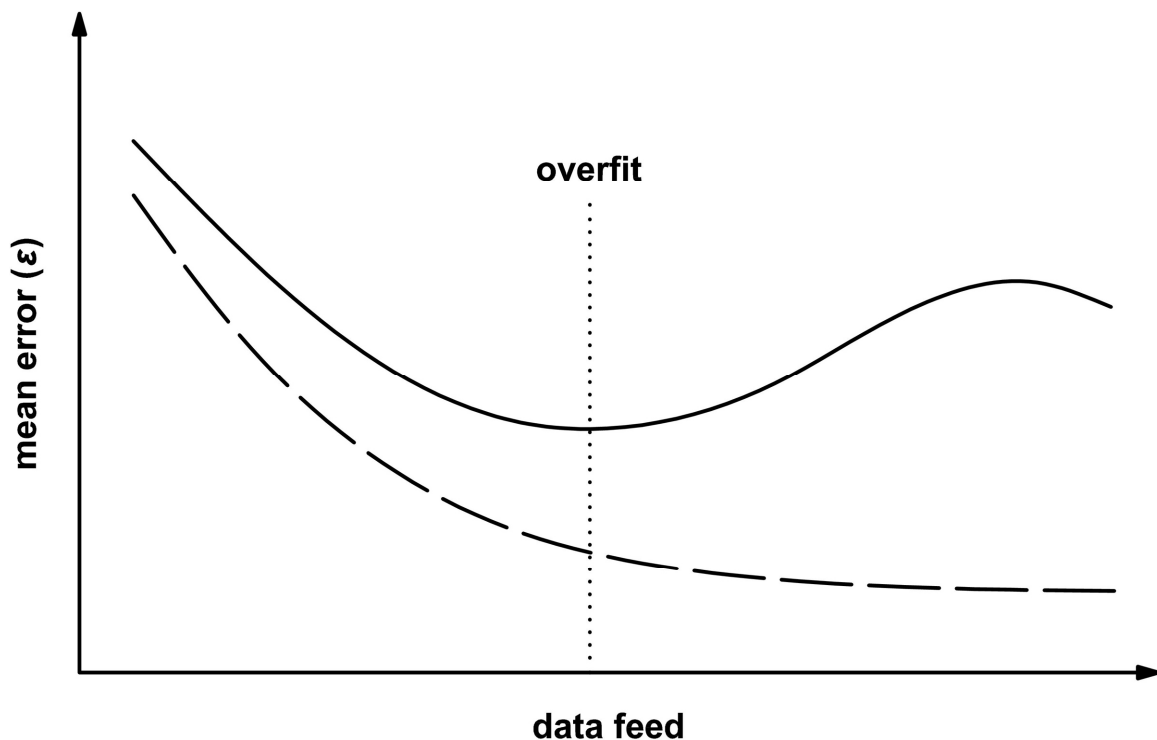
Firstly, it is necessary to assume that the input data are presented by 2D rasters of appropriate Conditioning Factors (geological, morphometric, environmental) and the referent Landslide Inventory map. The inventory is hypothetically necessary only for the training area (the area that has been assessed by an expert), but it is usually provided for the remaining area as a reference for evaluation of the model. The input rasters are organized in the way that each grid element (pixel) represents a data instance at a certain point of the area. Proposed approach leads to a classification task. The task is to place each pixel into an appropriate landslide category using the Conditioning Factor values associated with that pixel. The task applies only for the remaining area, usually called testing area (the area that has not been assessed by an expert). Selecting the size of the training area is very delicate, and requires particular strategies. An optimal approach is to build a sufficiently accurate model with a smaller number of training examples, thus leading to a reduced engagement of the expert. On the other hand, a practical value of a model in the landslide assessment framework lies in the model's prediction power, which implies more meaningful training sampling strategy. Therefore, it is desirable to have a training area that is physically separated from the testing area (Fig. 9). However, some of the algorithms (such as  $k$ -NN) require different sampling strategy, involving sparse and randomly sampled training instances.

The corresponding learning problem could be formulated as follows. Let  $P=\{\mathbf{x}|\mathbf{x}\in R^n\}$  be the set of all possible pixels extracted from the raster representation of a given area. Each pixel is represented as an  $n$ -dimensional real vector  $\mathbf{x}$ , where coordinate  $x_i$  represents the value of the  $i^{\text{th}}$  Conditioning Factor associated with the pixel  $\mathbf{x}$  (each pixel is represented by  $\mathbf{x}=\{x_1, x_2, \dots, x_n\}$ ). Further, let  $C=\{c_1, c_2, \dots, c_l\}$  be the set of  $l$  disjunctive, predefined landslide classes (a multinomial case). A function  $f_c:P\rightarrow C$  is called a classification if for each  $\mathbf{x}\in P$  it holds that  $f_c(\mathbf{x})=c_j$  whenever a pixel  $\mathbf{x}$  belongs to the landslide susceptibility class  $c_j$ . In practice, for a given terrain, one has a limited set of  $g$ -labeled examples  $(\mathbf{x}_q, c_j)$ ,  $\mathbf{x}_q\in R^n$ ,  $c_j\in C$ ;  $q=1, \dots, g$ ,  $j=1, \dots, l$  (where  $g$  is being a reasonably small amount of instances – training instances) belonging to  $P_q=P-P_d$  (where  $P_q$  is a training set and  $P_d$  testing set). The Machine Learning approach tries to find a function  $f_c'$  which is a good approximation of a real, unknown function  $f_c$  using only the examples from the training set  $P_q$  and a specific learning method.

\*\*\*

One common problem troubles all Machine Learning algorithms. It is primarily caused by the characteristics of the training set and the classifier's generalization power (Mitchell 1997) and it is called the overfit, also referred to as random error or noise. It is the problem of underperforming in the testing/validation set, while showing high performance in the training (Fig. 10).

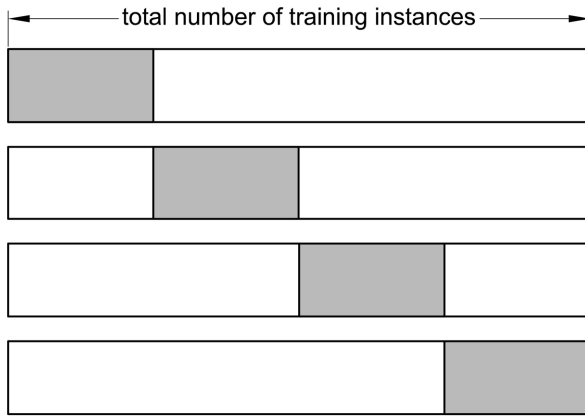
In other words, it is a paradox of reduced performance while having increased complexity of the model or bigger amount of data to build the model with. The algorithm is hence learning the noise as well, so its generalization is disputable (the learning becomes too specialized and the algorithm does not generalize well enough). The one is thus trading-off the model complexity for its fitness, i.e. the model's variance against its bias.



**Figure 10.** The overfit problem. Dashed line represents the training sample and bold line test sample. The functions are showing evident rise of the erroneous returns in testing mode despite the rise in data feed (amount of training/testing data or complexity of the model).

One way of dealing with the overfit is to optimize the generalization power of the algorithm. Another is to generate training and testing splits which will have balanced class

distributions, i.e. the sizes of all classes will remain proportional in both splits. The latter is not always feasible in spatial modeling, due to the usual abundance of one class and scarcity of another or several other classes (as in the case of landslide assessment, where *non-landslide* class is much bigger than the *landslide* class). This is especially pronounced if the adopted training/testing sampling strategy comes down to a physical, manual separation of the training and testing area as mentioned before (Fig. 9). A technique which partially prevents the overfit and involves specific optimization strategy during the training is the Cross-Validation (Mitchell 1997, Kanevski et al. 2009). It is probably the most efficient manner to deal with the overfit effect, and it is based on repetitive training and validation but only over the training split. It can be  $k$ -fold, where  $k$  stands for the number of partitions of the training split and therefore also represents the number of iterations. In the first run, one partition is taken for validation while  $k-1$  partitions are merged together for training. In every subsequent iteration, a different split takes the validation role, while the remaining  $k-1$  splits take the training role, until all  $k$  iterations are finished (Fig. 9). In turn, the procedure yields a result for one configuration/combination of the algorithm parameters. If one seeks the optimal parameter combination, giving the best generalization power to the algorithm, the Cross-Validation needs to be repeated for each parameter configuration. It is hence preferable that the algorithm does not have too many parameters to optimize.



**Figure 11.** A 4-fold Cross-Validation scheme. The rectangles schematize the training sample, wherein white parts represent CV training, and gray parts CV validation splits.

### 5.2.3.2 $k$ -Nearest Neighbor ( $k$ -NN)

This is amongst the simplest algorithms (Mitchell 1997), which classifies pixel instance  $\mathbf{x}$  containing  $x_i$  coordinates (containing an  $n$ -dimensional input space  $\mathbf{x}=\{x_1, x_2, \dots, x_n\} | \mathbf{x} \in R^n$ , where dimensions represent the values of the Conditioning Factors related to that particular pixel) by class values  $c_j$  of the  $k$  closest neighboring pixels  $\mathbf{x}_r$  surrounding  $\mathbf{x}$  ( $c_j$  is previously assigned in the training set by a practitioner as  $f_c(\mathbf{x}_r)$ ). The nearest neighbors are defined in terms of Euclidean distance  $d(\mathbf{x}, \mathbf{x}_r)$ , thus the classifier first calculates distances to  $k$  neighbors for each  $\mathbf{x}$  instance in the training set. Subsequently, a simple voting system assigns  $c_j$  class value (*landslide* class) to that particular pixel by class which predominates in neighboring instances (Eq. 23) or it alternatively assigns its mean value if the data are ordinal numeric (Eq. 24) (Fig. 12).

$$f'_c \leftarrow \arg \max \sum_{i=1}^k f(d(\mathbf{x}, \mathbf{x}_r), f_c(\mathbf{x}_r)); \quad \forall (\mathbf{x} \wedge \mathbf{x}_r) \in \text{nominal data type} \quad (23)$$

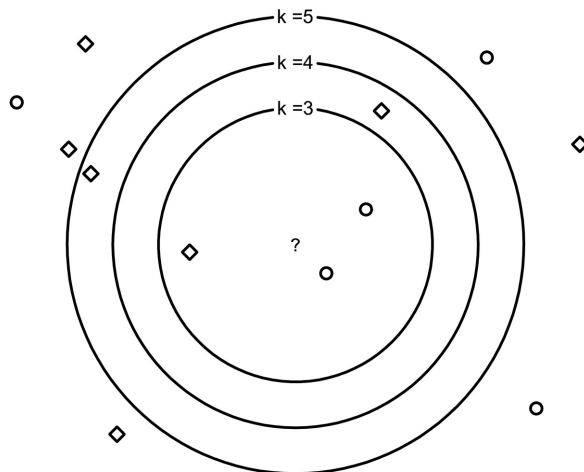
$$f'_c \leftarrow \arg \max \frac{\sum_{i=1}^k f_c(\mathbf{x}_r)}{k}; \quad \forall (\mathbf{x} \wedge \mathbf{x}_r) \in \text{ordinal data type} \quad (24)$$

Typically for  $k$ -NN there is no need for conventional training/testing procedures,  $f'_c$  is simply calculated for the remaining (testing) part of the dataset in the same way as in the

training mode (Varmuza & Filzmoser 2009). To avoid even votes, the number of neighbors is necessarily an odd number ( $k=1, 3, 5, 7\dots$ ). Since it is more probable that closer neighbors have a greater impact, it is further desirable to ponder each neighbor's proximity, thus upgrading to weighted  $k$ -NN (Mitchell 1997). It allows the algorithm becoming global<sup>19</sup> (Sheppard's method) but it requires sorting and weighting of distances per each pixel element (and each Conditioning Factor assigned to it) in the training set, resulting in a hardware-demanding and time-consuming procedure.

Having the classifier described, it becomes evident how biased it can become if all (relevant and irrelevant) Conditioning Factors are being fed together to the algorithm, because it will build (weighted or regular)  $k$ -NN relation per each, thus misleading the classification. In other words,  $k$ -NN is extremely sensitive to Conditioning Factor's relevance to the landslide occurrence, which is why a very strict Attribute Selection needs to be performed prior to the analysis. Alternatively, in the case of weighted  $k$ -NN, Euclidean distance axis could be stretched so that different Conditioning Factors would have different weights according to their relevance. Still, it does not solve the computational demands of this algorithm, especially when there are mixed data types, which call for a double procedure (due to the different distance calculations).

On the bright side, the algorithm is straightforward (the distances are the classification criteria, so there is no true black-box model behind it) and it can originate from a very sparse data, randomly sampled throughout the training set (which is sometimes convenient, but for the landslide assessment concept and prediction of the spatial landslide distribution it is of little relevance). It is also very convenient for experimenting, since only one parameter (the number of neighbors  $k$ ) needs to be optimized.



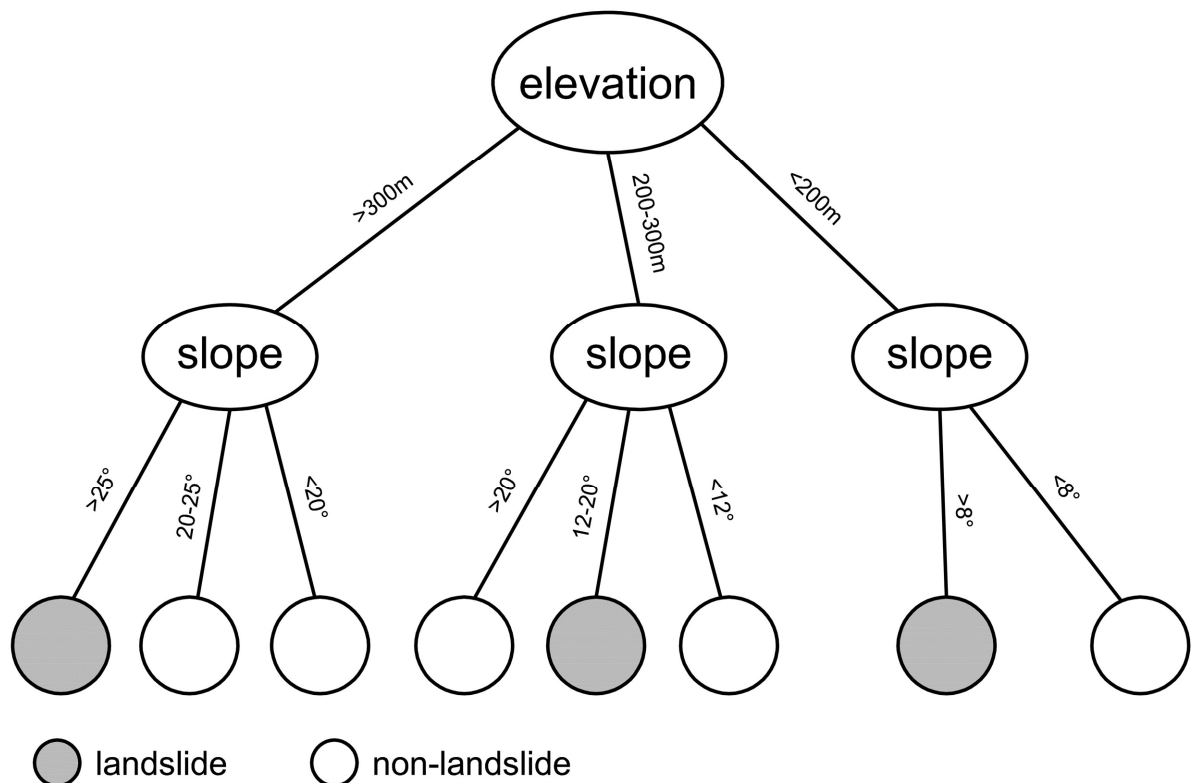
**Figure 12.**  $k$ -NN classification principle. Unclassified instance (?) is classified by the majority of neighbors into *landslide* (circle) or *non-landslide* (square) instance. Note that for  $k=3$  the instance is classified as *landslide*, for  $k=4$  it remains unclassified (votes are even 2:2) and for  $k=5$  the instance is classified as *non-landslide*.

### 5.2.3.3 Decision Tree C4.5

Decision Tree classifiers are the algorithms which resemble a tree structure from the root downward (Fig. 13), where every set of branches originate from the common node and extend to the further nodes and branches hierarchically. These eventually terminate in end-nodes, called leaves of a Decision Tree.

<sup>19</sup> A Machine Learning algorithm is *global* when it allows all training instances to participate in mapping function (in this case  $f_c$ ), otherwise it is *local*.

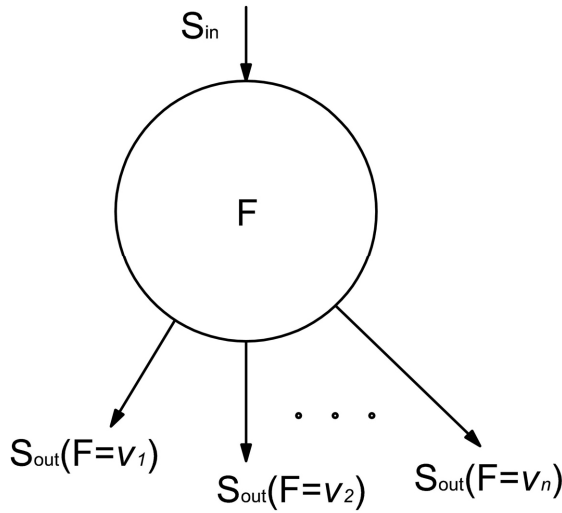
C4.5 is a well-known univariate decision tree classifier (Quinlan 1993). In this approach (and with respect to the landslide assessment framework), a pixel instance, described with a set of  $n$  Conditioning Factor values is classified by testing the value of one particular factor at each node, starting from the root of the tree. It then follows a certain path in the tree structure depending on the tests in previous nodes and finally reaches one of the leaf nodes labeled with a class  $c_j$  (*landslide/non-landslide* in binomial  $j=2$  case or accordingly to the designated landslide classes in multinomial  $j=l$  cases). Each path leading from the root to a certain leaf node (landslide class label) can be interpreted as a conjunction of tests involving Conditioning Factors. Since there could be more leaf nodes with the same class labels, one could interpret each class as a disjunction of conjunctions of constraints on the Conditioning Factors values of instances  $\mathbf{x}$  from the dataset ( $\mathbf{x}=\{x_1, x_2, \dots, x_n\}$ ). The interpretability of the derived model enables a domain practitioner to have better understanding of the problem and in many cases could be preferable over functional models such as SVM and ANN.



**Figure 13.** An example of a simple tree structure on landslide assessment example depending on two Conditioning Factors, *elevation* and *slope* (*elevation* is preferred over *slope* at the root node).

It is now left to briefly explain how the tree can be derived from the training data  $(\mathbf{x}_q, c_j)$ ,  $q=1, \dots, g$ , where  $c_j$  is one of  $l$  disjunctive classes ( $j=1, \dots, l$ ). C4.5 deals both with numerical and categorical attributes but for the sake of the simplicity it is assumed that all Conditioning Factors are categorical (nominal data scale). The tree construction process performs a greedy search in the space of all possible trees starting from the empty tree and adding new nodes in order to increase the classification accuracy on the training set. A new node (candidate Conditioning Factor test) is added below a particular branch if the instances following the branch are partitioned after the test in such way that the distinction between the classes becomes more evident. If the test on the Conditioning Factor  $F$  splits the instances in subsets in which all elements have the same landslide class labels, a perfect attribute choice is reached (those subsets become leaf nodes). On the other hand, if the instances are distributed so that in each subset there has been equal number of elements belonging to different landslide classes (leading to indecisiveness regarding that factor), then  $F$  would be

the worst attribute choice. Hence, the root node should test against the most informative Conditioning Factor concerning the whole training set. C4.5 uses *Gain Ratio (GR)* measure (Quinlan 1993) to choose between the available  $F$ s and is heavily dependent on the notion of *Entropy*. Figure 14. explains the calculation of *Gain Ratio*.



**Figure 14.** The illustration of the *Gain Ratio* on the decision tree node.

Let  $S_{in}$  be the set of  $N$  instances for which the preceding test in the parent node forwarded them to the current node. Let  $n_i$  be the number of instances from  $S_{in}$  that belong to class  $c_j$ ,  $j=1, \dots, l$ . *Entropy*  $E(S_{in})$  is defined as a measure of impurity (in respect to the class label) of the set  $S_{in}$  as:

$$E(S_{in}) = - \sum_{i=1}^l \frac{n_i}{N} \log_2 \frac{n_i}{N} . \quad (25)$$

If all instances belong to the same class, then entropy is equal to zero. On the other hand, if all classes are equally present, the entropy is maximal ( $\log_2 l$ ). In particular problem setting,  $F$  denotes the candidate attribute of an instance  $\mathbf{x}$ . Since it is assumed that  $F$  is categorical and can take  $m$  different values  $v_1, v_2, \dots, v_m$ , there are  $m$  branches leading from the current node. Each  $S_{out}(F=v_i)$  represents the set of instances for which  $F$  takes the value  $v_i$ . The informative capacity of  $F$  concerning the classification into  $l$  predefined landslide classes can be expressed by using the notion of *Information Gain (IG)*:

$$IG(S_{in}, F) = E(S_{in}) - \sum_{v \in \{v_1, \dots, v_m\}} \frac{|S_{out}(F=v)|}{N} E(S_{out}(F=v)) . \quad (26)$$

In Equation 26.  $|S_{out}(F=v)|$  represents the number of instances in the set  $S_{out}(F=v)$  and  $E(S_{out}(F=v))$  is the *Entropy* of that set calculated using Equation 25. Higher the  $IG$ , more informative the  $F$  for the classification in the current node, and vice-versa (Mitchell 1997).

The main disadvantage of the  $IG$  measure is that it favors the factors with many values (bigger  $m$ ) over those with fewer (smaller  $m$ ). This entails wide trees with many branches starting from the corresponding nodes. If the tree is complex and has a lot of leaf nodes, then it is expected that the model will overfit the data (it will learn the anomalies of the training data and its generalization capacity, i.e. classification accuracy on unseen instances will be decreased). In order to reduce the effect of overfitting C4.5 further normalizes  $IG$  by the *Entropy* calculated with respect to the factor's class values instead of landslide class labels (*Split Information – SI*) to obtain *Gain Ratio (GR)*:

$$SI(S_{in}, F) = - \sum_{v \in \{v_1, \dots, v_n\}} \frac{|S_{out}(F = v)|}{N} \log_2 \frac{|S_{out}(F = v)|}{N}, \quad (27)$$

$$GR(S_{in}, F) = \frac{IG(S_{in}, F)}{SI(S_{in}, F)}. \quad (28)$$

C4.5 uses *GR* to run a greedy search over all possible trees. If the factor is numerical (this is the case for the most of them) C4.5 detects the candidate thresholds that separate instances into different classes. Let pairs  $(F, c_j)$  be (50, 0), (60, 1), (70, 1), (80, 1), (90, 0) and (100, 0). C4.5 identifies two thresholds on the boundaries of different classes:  $F < 55$  and  $F < 85$ .  $F$  now becomes a binary attribute (true or false) and the same *GR* procedure is applied to select among the two thresholds, when considering the introduction of this attribute test into the growing tree.

Finally, C4.5 uses so-called post-pruning technique to reduce the size of the tree (complexity of the model). After growing, the tree that classifies all training examples as well as possible errors (overfitted model) is converted into a set of equivalent rules (e.g. *IF*  $F_1 = v_1$  *AND*  $F_2 < v_2$  *AND* ... *THEN*  $c_j$ ) per each leaf node (a path from the root to a leaf). It then prunes the rules by removing every condition that does not affect the estimated rule accuracy, and then sorts the pruned rules by their estimated accuracy. In the operational phase C4.5 uses sorted pruned rules for the classification of unseen instances (for testing set).

C4.5 calculates observed estimates for rules by using the training set as a whole (number of correctly classified instances/number of total instances per each leaf) and then calculating the standard deviation assuming binomial distribution. For a given confidence level, the lower bound estimate is taken as the measure of the rule accuracy. There are many variants of pruning techniques but all of them can be compared with adjusting parameter  $c$  in SVM algorithm (as will be shown later), since they both trade-off the training error versus the model complexity in order to increase the generalization power of the induced classification model (Quinlan 1993, Mitchell 1997, Varmuza & Filzmoser 2009).

### 5.2.3.4 Support Vector Machines (SVM)

After some time of digesting of the theoretical background and realizing all the advantages over conventional statistics or other Machine Learning systems, SVM-related improvements in learning theory were acknowledged in many different fields of classification and regression tasks, especially with complex and large datasets (Kanevski et al. 2009, Kecman 2005). Conveniently, spatial modeling in Geo-sciences commonly faces the problem of non-linearity and multi-dimensionality of input data, leaving ordinary statistic or probabilistic tools struggling while featuring-out a pattern or rule that could lead input data to an interpretative model. The potential of Machine Learning algorithms for solving such problems has only recently been exploited (as presented in Chapter 4). Herein, SVM algorithm utilization in classification task in respect to the landslide susceptibility assessment will be considered, but first, it is necessary to elaborate the learning mechanism and its formulation.

SVM Machine Learning was developed under the Vapnik-Chervonenkis generalization theory, with linear separating learning machines, extending to kernel-induced feature space, and with respect to the optimization theory. The fact that SVM has a convex optimization problem makes them quite unique and somewhat advantageous in comparison to the other supervised learning systems (Cristiani & Shawe-Taylor 2000). In fact, the learning problematic with SVM resembles classical statistic inference and ANN system (Fig. 15), yet with significant differences, concerning data distribution and optimization. SVM approach does not require normally-distributed datasets, and does not use pre-defined parameters (it is called non-parametric modeling due to the latter). Instead it tunes the parameters to match the needed learning capacity in relation to the data complexity. Conversely to ANNs and statistical or Fuzzy Systems, where systematical reducing of the



initial training error takes place until the estimated threshold is reached, SVM keep the training error fixed and reduce the confidence intervals instead (Cristiani & Shawe-Taylor 2000, Kecman 2005). Moreover, SVM method provides novel principle for error treatment in so-called, Structured Risk Minimization principle, which optimizes the algorithms performance on the basis of learning space reduction toward the most desirable learning capacity. In simple words, SVM procedure puts the learning capacity and desired accuracy in balance (Burges 1998, Cristiani & Shawe-Taylor 2000).

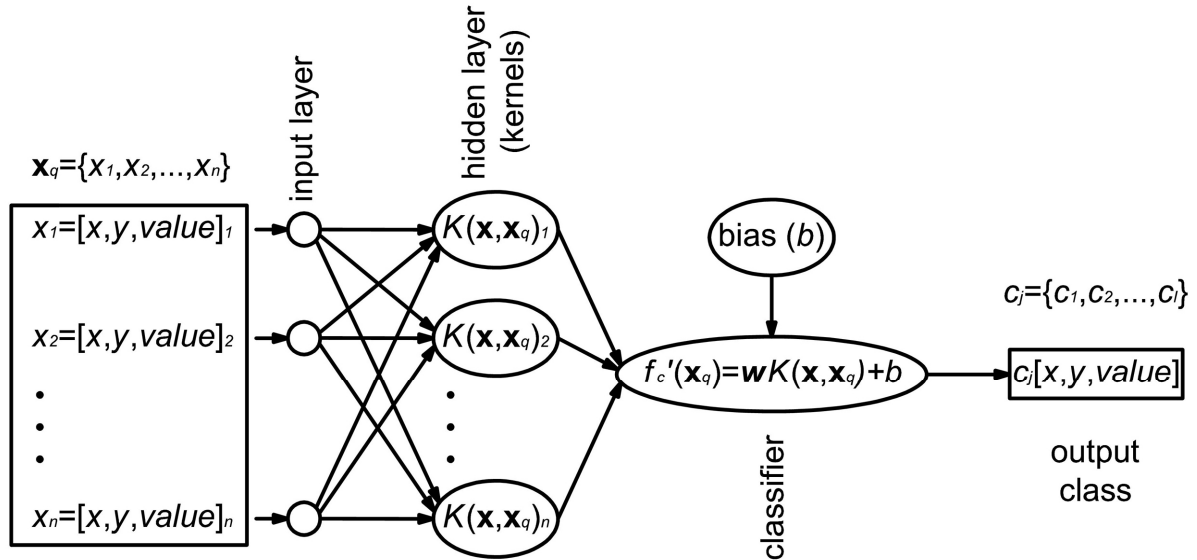


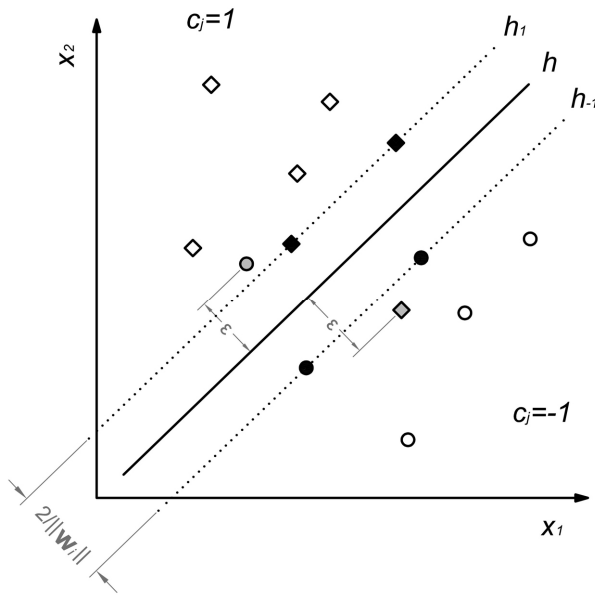
Figure 15. Architecture of SVM.

SVM network architecture (Fig. 15) is the same as in ANN RBF network design, in fact SVM are a sub-branch of ANN with a single hidden layer for kernel function operations: inputs are vector coordinates of a training pixel instance  $\mathbf{x}_q \in P_q$  (geological, morphometric and environmental Conditioning Factors) and *landslide* class, i.e.  $\mathbf{x}_q = \{x_1, x_2, \dots, x_n\}$ ,  $f_c(\mathbf{x}_q)$ , which are forwarded to the hidden layer nodes mapped by RBF kernel functions  $K(\mathbf{x}, \mathbf{x}_q)$ , and each is assigned a non-zero weight  $\mathbf{w}_i$  (in the case the node contains Support Vectors). Every node is solved for the weight vector  $\mathbf{w}_i$  and/or bias  $b$  by training on initial function  $f_c(\mathbf{x}_q)$  and relates to the function  $f'_c(\mathbf{x}_q) = c_j$  which maps the initial function into the new instances  $\mathbf{x}_p \in P_p$  (in the testing set).

Originally, SVM is a linear binary classifier (instances could be classified to only one of the two classes, e.g. *landslide* and *non-landslide*), but one can easily transform  $l$ -classes problem (multinomial landslide classes) into a sequence of  $l$  (one-versus-all) or  $l(l-1)/2$  (one-versus-one) binary classification tasks, where using different voting schemes lead to a final decision (Belousov et al. 2002). Given a binary training set  $(\mathbf{x}_q, c_j)$ ,  $\mathbf{x}_q \in R^n$ ,  $c_j \in \{-1, 1\}$ ,  $j = 1, \dots, l$ , the basic variant of the SVM algorithm attempts to generate a separating hyper-plane in the original space of  $n$  coordinates ( $x_i$  parameters in vector  $\mathbf{x}$ ) between two distinct classes (Fig. 16). During the training phase the algorithm seeks for a hyper-plane which best separates the samples of binary classes (classes 1 and -1). Let  $h_1: \mathbf{w}_1 \mathbf{x}_q + b = 1$  and  $h_{-1}: \mathbf{w}_{-1} \mathbf{x}_q + b = -1$ , ( $\mathbf{w}_i, \mathbf{x}_q \in R^n$ ,  $b \in R$ ) be possible hyper-planes such that majority of class 1 instances lie above  $h_1$  ( $\mathbf{w}_1 \mathbf{x}_q + b > 1$ ) and majority of class -1 fall below  $h_{-1}$  ( $\mathbf{w}_{-1} \mathbf{x}_q + b < -1$ ), whereas the elements belonging to  $h_1$ ,  $h_{-1}$  are defined as Support Vectors (Fig. 16). Finding another hyper-plane  $h: \mathbf{w} \mathbf{x}_q + b = 0$  as the best separating (lying in the middle of  $h_1$ ,  $h_{-1}$ ), assumes calculating  $\mathbf{w}_i$  and  $b$ , i.e. solving the nonlinear convex programming problem. The notion of the best separation can be formulated as finding the maximum margin  $MA$  between the two classes. Since  $MA = 2\|\mathbf{w}_i\|^{-1}$ , maximizing the margin leads to the constrained optimization problem (Eq. 29).

$$\min_{\mathbf{w}, b} \frac{1}{2} \|\mathbf{w}_i\|^2 + C \sum_i \varepsilon_i \quad (29)$$

$$\text{w.r.t: } 1 - \varepsilon_i - c_j(\mathbf{w}_j \cdot \mathbf{x}_q + b) \leq 0; -\varepsilon_i \leq 0; i = 1, 2, \dots, g; j = 1, 2, \dots, l \quad (30)$$



**Figure 16.** A general binary classification example, separating *landslide* (circles) from *non-landslide* instances (squares) in a simple 2D feature space (two Conditioning Factors/coordinates –  $x_1$  vs.  $x_2$  define the space). Shaded points represent instances that were misclassified. Solid (bolded) instances represent Support Vectors.

Despite of having some instances misclassified (Fig. 16) it is still possible to balance between the incorrectly classified instances and the width of the separating margin. In this context, the positive slack variables  $\varepsilon_i$  and the penalty parameter  $c$  are introduced. The slacks represent the distances of misclassified points to the initial hyper-plane, while parameter  $c$  models the penalty for misclassified training points, that trades-off the margin size for the number of erroneous classifications (bigger the  $c$  smaller the number of misclassifications and smaller the margin). The goal is to find a hyper-plane that minimizes the misclassification errors while maximizing the margin between the classes. This optimization problem is usually solved in its dual form (dual space of Lagrange multipliers):

$$\mathbf{w}^* = \sum_{q=1}^g \alpha_q c_q \mathbf{x}_q, \quad c \geq \alpha_q \geq 0, \quad (31)$$

where  $\mathbf{w}^*$  is a linear combination of training examples for an optimal hyper-plane. However, it can be shown that  $\mathbf{w}^*$  represents a linear combination of Support Vectors  $\mathbf{x}_q$  for which the corresponding  $\alpha_q$  Lagrangian multipliers are non-zero values. Support Vectors for which  $c > \alpha_q > 0$  condition holds, belong either to  $h_1$  or  $h_{-1}$ . Let  $\mathbf{x}_a$  and  $\mathbf{x}_b$  be two such Support Vectors ( $c > \alpha_a, \alpha_b > 0$ ) for which  $c_a = 1$  and  $c_b = -1$ . Now  $b$  could be calculated from  $b^* = 0.5\mathbf{w}^*(\mathbf{x}_a + \mathbf{x}_b)$ , so that classification (decision) function finally becomes:

$$f'_c(\mathbf{x}_q) = \text{sgn} \sum_{i=1}^g \alpha_i c_i (\mathbf{x} \cdot \mathbf{x}_q) + b^* \quad (32)$$

In order to cope with non-linearity even further, one can propose the mapping of instances to a so-called feature space of very high dimension:  $\Psi: R^n \rightarrow R^\omega$ ,  $n \ll \omega$ , i.e.  $\mathbf{x} \rightarrow \Psi(\mathbf{x})$ . The basic idea behind this mapping into a high-dimensional space is to transform the non-linear case into linear and then use the general algorithm, as already explained (Eqs. 29-32). In such space, dot-product from Equation 32. is being transformed into  $\Psi(\mathbf{x}) \cdot \Psi(\mathbf{x}_q)$ . A certain class of functions for which  $K(\mathbf{x}, \mathbf{x}_q) = \Psi(\mathbf{x}) \cdot \Psi(\mathbf{x}_q)$  holds, are called kernels (Cristiani & Shawe-Taylor 2000). They represent dot-products in some high dimensional dot-product

spaces (feature spaces), and yet could be easily computed back into the original space. Radial Basis Function (Eq. 33), also known as Gaussian kernel (Abe 2005), is one of such functions commonly implemented with SVM<sup>20</sup>.

$$K(\mathbf{x}, \mathbf{x}_q) = \exp\left(-\gamma\|\mathbf{x} - \mathbf{x}_q\|^2\right) \quad (33)$$

Now Equation 32. finally becomes:

$$f'_c(\mathbf{x}_q) = \text{sgn} \sum_{i=1}^g \alpha_i c_i K(\mathbf{x}, \mathbf{x}_q) + b^* = \text{sgn} \sum_{i=1}^g \alpha_i c_i \exp\left(-\gamma\|\mathbf{x} - \mathbf{x}_q\|^2\right) + b^* . \quad (34)$$

After removing all training data that are not Support Vectors and retraining the classifier by applying the function above, the same result would be obtained as in the case of classifying with all available training instances  $\mathbf{x}_q$  (Cristiani & Shawe-Taylor 2000). Thus, ones defined, Support Vectors could replace the entire training set containing all necessary information for the construction of the separating hyper-plane  $h$ .

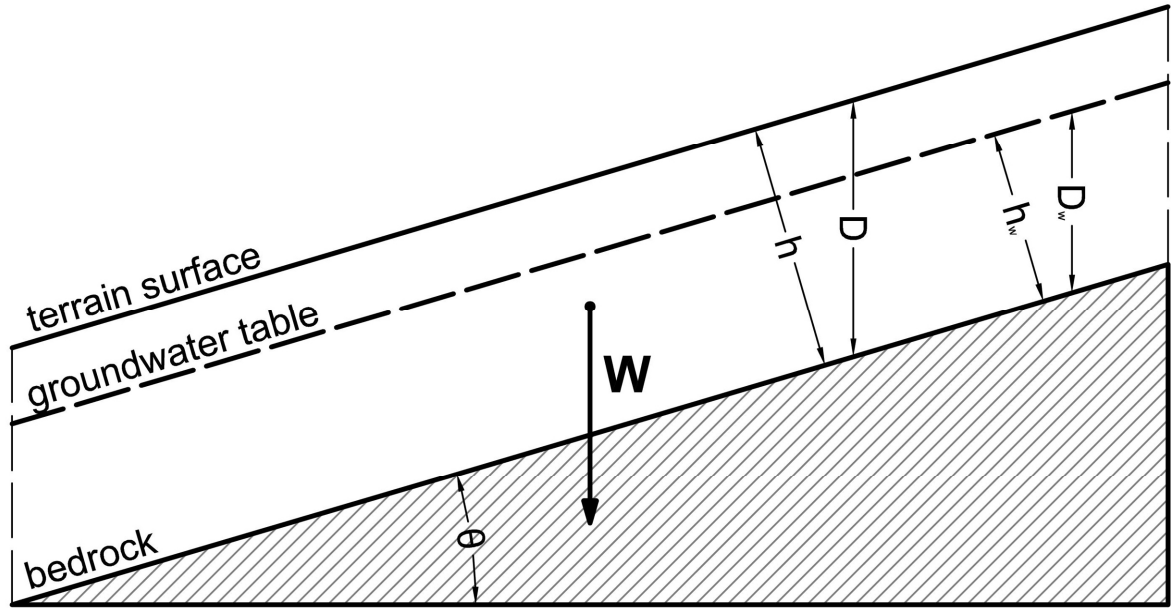
## 5.2.4 Deterministic Approach

This is an entirely different approach, which employs deterministic, i.e. physical nature of the landsliding phenomenon, and involves almost utterly different features to characterize masses susceptible to landsliding. Even though it could be said that the regional scale under GIS framework has remained in use, the deterministic modeling requires more localized data, which are acquired on the field and subsequently generalized for wider areas with presumably similar characteristics. Thus, instead of dealing with chiefly geological, morphometric and environmental themes of an area as in previous approaches, analysis now requires geotechnical features which are to determine mechanical properties of the masses throughout the area. This nears the problematic of susceptibility to the engineering geological framework, which dominates on local (site-specific) scales. The modeling itself is feasible due to the significant deal of approximation that has to be introduced in order to keep the regional character of the analysis and avoid the model complexity overload.

The GIS implementations of such kind are already well-known, and there will be some comments on particular extension packages for a variety of GIS platforms later on (see Chapter 5.5).

Deterministic analysis in regional GIS framework comes down to the simplest of the cases of slope stability analyses – the infinite slope model, and the simplest Triggering Factor involvement – shallow groundwater flow that is converging under stationary conditions. Otherwise, the slope stability analyses can be very detailed and require very high data sampling density, usually impossible to collect at regional scales. It can involve very elaborate modeling based on Finite Elements or Finite Differences, not to mention how complex the linked triggering model can become (profound expert knowledge on the triggering type would be required, involving possibly earthquake modeling, hydrological-meteorological modeling, hydrological-hydrogeological modeling, erosion modeling and so forth). However, the simplest forms of slope analysis follow the Limit Equilibrium Method (LEM), while the simplest groundwater flow model considers stationary flow under steady-state conditions. The general principle of any LEM type of method implies the balance of the driving and resisting forces acting upon a particular slip surface in the body of a slope. It involves relatively simple static composition of these forces and their ratio, represented by the *Factor of Safety* ( $F_s$ ) or *Stability Index* ( $S$ ). The latter is more common in the GIS practice (Montgomery & Dietrich 1994, Pack et al. 2001) and it will hence be discussed in the simplest, infinite slope scenario in the following passages.

<sup>20</sup> In all of the implementations of SVM in this thesis, presented in detail in Chapter 6, Gaussian kernel, containing only two parameters ( $c$  and  $\gamma$ ) for optimization has been used.



**Figure 17.** The infinite slope model.

Deterministic model at hand is coupled. It involves LEM-based *Stability Index* model and steady-state groundwater flow model. The amount of the approximations introduced by combining these two models is substantial and affects the model's applicability and limitations, as will be discussed later on.

LEM-based *Stability Index* initially assumes the infinite slope (Fig. 17), where the slip surface is parallel to the topographic slope. The consequence of the latter is enabling the GIS application in the deterministic concept, because it puts surface topography in control of the groundwater behavior, and it is needless to say how developed are the analysis of surface topography in GIS and how valuable data can be extracted using GIS. The method further assumes combined cohesion, and does not take into account reinforced and mitigated scenarios.

A simple Coulomb-Mohr's condition of failure along the slip surface (Eq. 35) is adopted, as well as the principle of effective stresses:

$$\tau = c_{soil} + \sigma'_{soil} \operatorname{tg} \varphi_{soil}, \quad (35)$$

where  $\tau$  is the shear strength of the soil along the slip surface,  $c$  is the combined cohesion (relative dimensionless cohesion),  $\sigma'_{soil}$  is the effective stress of the soil (total stress decreased by the pore pressure), and  $\varphi_{soil}$  is the friction angle of the soil. The *Factor of Safety*  $F_s$  is then built as a ratio between resisting and driving forces (Eq. 36).

$$F_s = \frac{\text{forces resisting the failure}}{\text{forces driving the failure}} = \frac{F_r}{F_d} \quad (36)$$

These forces could be composed by the geometry of the infinite slope (Fig. 17):

$$F_s = \frac{c_{soil} + \cos^2 \theta (\gamma_{soil} (D - D_w) + (\gamma_{soil} - \gamma_w) D_w) \operatorname{tg} \varphi_{soil}}{\gamma_{soil} \sin \theta \cos \theta} = \frac{c_{soil} + \cos \theta (1 - w_r r) \operatorname{tg} \varphi_{soil}}{\sin \theta}, \quad (37)$$

where  $\gamma_{soil/w}$  is the unit weight of soil/water,  $w_r = D_w/D = h_w/h$ , and  $r = \gamma_w/\gamma_{soil}$  is water to soil weight ratio or relative wetness, and combined (dimensionless) cohesion  $c_{soil} = (c_{root} + c'_{soil}) / (h \gamma_{soil})$ . *Stability Index* is defined as spatial and temporal probability of  $F_s$ , which will be discussed later-on.

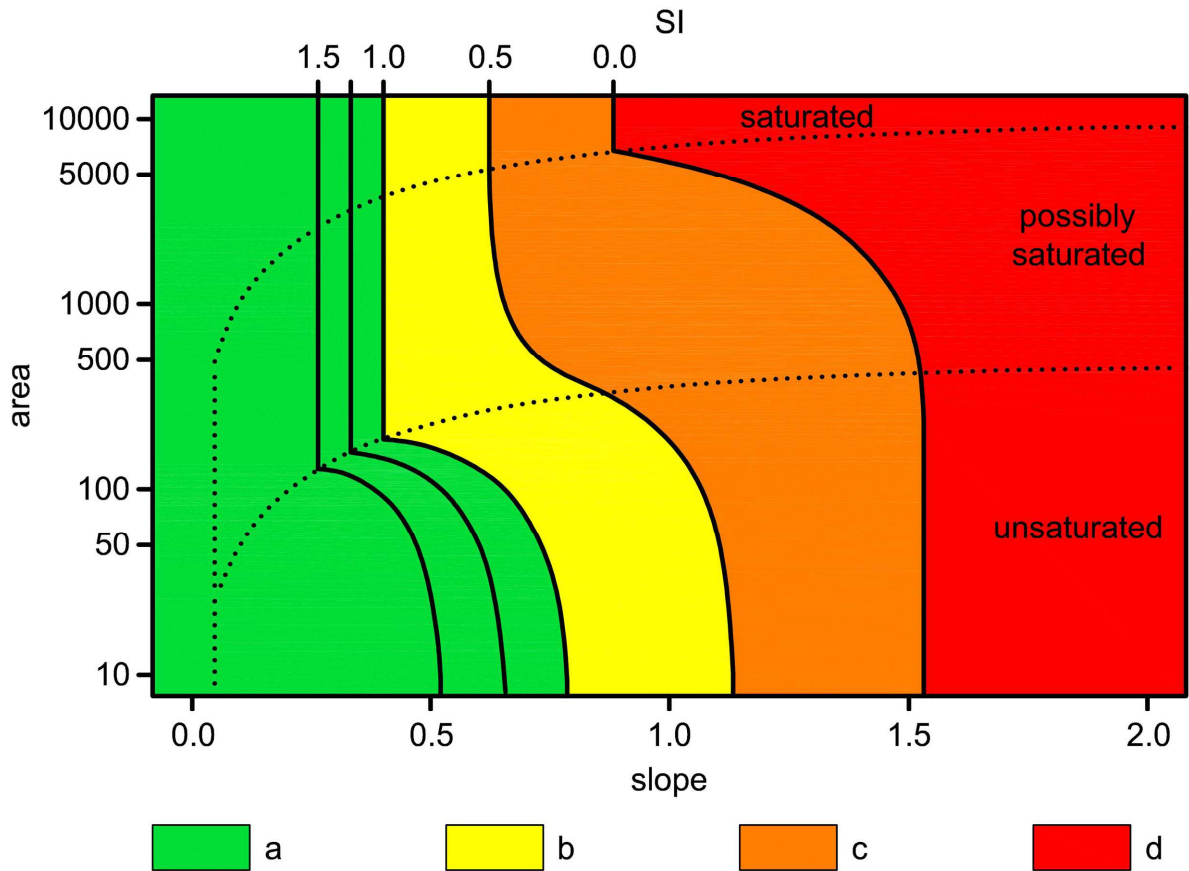
The second part of the deterministic model considers the trigger, which is hereby defined as convergent groundwater flow, collected solely from the rainfall. Assumptions infer the impervious bedrock, and porous soil mantle, so that the entire amount of precipitated water percolates through the soil with no lateral losses, and no superficial run-off (the worst case scenario). Furthermore, the subsurface flow parallels the topographic slope, and steady-state conditions apply to the flow, while lateral discharge stands in equilibrium with the steady-state recharge.

Given such assumptions, it is possible to define the flux at each point as  $T \sin \theta$ , where  $T$  stands for the ground water transmissivity [ $m^2/h$ ], i.e. hydraulic conductivity times soil thickness, and  $\theta$  stands for the slope angle. The assumptions further support formulation of depth-integrated lateral discharge  $q_{soil}$  [ $m^2/h$ ], by  $q_{soil} = Ra$ , where  $a$  is a specific catchment area that equals catchment area divided by the unit length of the slope ( $a = A/b$ ), and  $R$  is a steady-state recharge, i.e. its effective value which causes critical conditions (e.g. during the most extreme month or week of the year, precipitation-wise). The relative wetness  $w_r$  can be thus defined as a minimal recharge to flux ratio (Eq. 38).

$$w_r = \min\left(\frac{Ra}{T \sin \theta}, 1\right) \quad (38)$$

Now meteorological (precipitation) and hydrogeological factors are combined through  $w_r$ , and range from 0 to 1. Equation 37. now becomes:

$$F_s = \frac{c_{soil} + \cos \theta \left(1 - \min\left(\frac{Ra}{T \sin \theta}, 1\right) r\right) \tan \varphi_{soil}}{\sin \theta} \quad (39)$$



**Figure 18.** Stability Index in relation to stability classes in Area-Slope space (a=stable, b=lower than average potential of instability, c=higher than average potential of instability, d=unstable).

An important detail is here regarding the input data, since parameters  $a$  and  $\theta$  are extracted from the surface topography, while remaining  $c_{soil}$ ,  $r$ ,  $\phi_{soil}$ , and  $R/T$  are the data that need to be sampled or monitored. Thus, some uncertainty is introduced due to both, measurement errors and errors of generalization over wider areas. The solution is to turn to stochastic approach and to present a range of input variables, which presumably behave as normally or even better, uniformly distributed data. The data will hence have the worst-case-scenarios for the most unfavorable values, as well as the best-case-scenarios. If we assume  $r$  to be a constant ratio (or at least, constant for a specific part of the area, defined by the lithological units) the worst case would be when  $c_{soil}$  and  $\phi_{soil}$  are the lowest, and  $R/T$  are the highest. The *Stability Index SI* is then computed as  $\min(F_s)$  and all areas with such  $F_s$  greater than 1 are unconditionally stable (at least regarding the precipitation-groundwater trigger) and  $SI = \min(F_s)$ . If the  $\min(F_s)$  is less than 1 then a possibility (non-temporal, spatial probability) of a failure exists and  $SI = p(F_s > 1)$ . The other extreme is the opposite situation or the best case scenario, where if  $\max(F_s)$  is less than 1,  $SI = p(\max(F_s) > 1) = 0$ . The rest of the *SI* combinations are shown in Figure 18., where they are defined in the area-slope space ( $a$ ,  $\theta$ ), so that appropriate classes could be defined. In general, the lower the index is, the lesser the stability and vice-versa.

The limitations of the method are apparent. It requires only precipitation-induced landslide mechanisms, developed only in mantled soil, such as eluvial and delluvial Quaternary formations. It is typical for hummocky topography with well developed channeled hollows, where sliding mechanism needs to be translational over significantly less permeable bedrock. It further requires that certain points of the terrain have been previously defined as unstable by the expert, in order to calibrate the output class scale (Fig. 18). For these reasons the resulting classification is defined as the relative susceptibility/hazard because it partly contains the temporal dimension, expressed through the meteorological data.

### 5.3 Model Evaluation Methods

*This branch of methods is featuring the Objective 5 (see Chapter 2).*

The quality of classification emerging from any of the presented approaches, could be simply estimated as the relation between correctly and incorrectly classified *landslide* instances (accuracy), but the problem of proper evaluation becomes more complex (Frattini, et al. 2010), and requires more sophisticated solutions. These are commonly based on the confusion matrix or contingency table (Tab. 6), which introduces different types of classification hits and errors. *True Positives* are the instances where the model and the reference agree that landslide exists, *True Negatives* represent the instances where the model comply with the reference on *non-landslide* instances, while misclassifications are presented by *False Positives* (model claims landslides where they do not exist according to the reference) and *False Negatives* (vice-versa). Confusion matrix is obtained from the cross-tabulation of the model and the referent landslide map (in the case of the landslide assessment framework).

**Table 6.** Confusion matrix and appropriate error measures.

		Landslide Inventory		ROC space coordinates
		true	false	
model	positive	$tp$ (true positive)	$fp$ (false positive)	$tp_{rate} = tp / (tp + fn)$
	negative	$fn$ (false negative)	$tn$ (true negative)	$fp_{rate} = fp / (fp + tn)$

### 5.3.1 Kappa Statistics

Herein, a parameter called kappa index ( $\kappa$ -index) was proposed. It represents the measure of agreement between compared entities, rather than the measure of classification performance (Landis & Koch 1977). It turns quite convenient for comparison of the maps with the same classes (Bonham-Carter 1994), as it commonly is the case in Machine Learning-based classification experiments. The best way to compute  $\kappa$ -index is to derive it from a confusion matrix, an  $l \times l$  cross-tabulation table ( $l$  being the number of landslide classes) in which  $x_{ij}$  represents the number of pixels from the actual classes  $c_j$  that are paired with the  $c_i$  classes of the model.

$$\kappa = \frac{T \sum_i x_{ii} - \sum_i x_{i+} x_{+i}}{T^2 - \sum_i x_{i+} x_{+i}} \quad (40)$$

In Equation 40.  $T$  represents the total number of tested pixels, while  $x_{i+}$  and  $x_{+i}$  are the total numbers of observations in particular row and column of the confusion matrix, respectively. The idea of  $\kappa$ -index is to remove the effect of the random agreement between the two experts (here between a referent Landslide Inventory and a classifier). Obtained index ranges from -1 for the complete absence of agreement, to +1 for the absolute agreement, while zero value suggest that the agreement is random. Based on (Landis & Koch 1977, Fielding & Bell 1997)  $\kappa$ -index values falling in 0.61-0.81 range are categorized as substantial, and values higher than 0.81 are considered as nearly perfect.

In the multinomial case (more than two classes)  $\kappa$ -index can be performed per class, thus specifying which class are matching better or poorer. It is usually called conditional kappa  $\kappa_i$  and it follows a similar formulation as ordinary kappa (Eq. 41).

$$\kappa_i = \frac{T x_{ii} - x_{i+} x_{+i}}{T x_{i+} (1 - x_{+i})} \quad (41)$$

There are also some other formulations of  $\kappa$ -index, where the accent is on decomposing of the index into classes, or decomposing it to the histogram match and location match as in  $\kappa_{histo}$  and  $\kappa_{location}$  (multiplying these, the overall  $\kappa$ -index is obtained). This is an advanced idea in the domain of landslides (or any spatial-featured classification for that matter, such as Land Cover mapping for example), because it gives better insight on the actual matching of the model. In particular,  $\kappa_{location}$  introduces a factor which takes into account the class sizes and treats the errors with bigger tolerance, while  $\kappa_{histo}$  takes into account the distribution of values within a class<sup>21</sup>. Ultimately, there is a fuzzy measure  $\kappa_{fuzzy}$  which introduces the correction based on a distance decay function, which furthers the concept of higher tolerance for the particular error types. The classes can be even weighted differently (the distance decay membership function can be adjusted arbitrarily), so that the smaller classes get another decrease of misclassification penalty. It is essential that the practitioner is experienced with the model and the system of evaluation, so that the slack in tolerance that the model is being presented with is not abused (Hagen 2003).

### 5.3.2 Receiver Operating Characteristics (ROC)

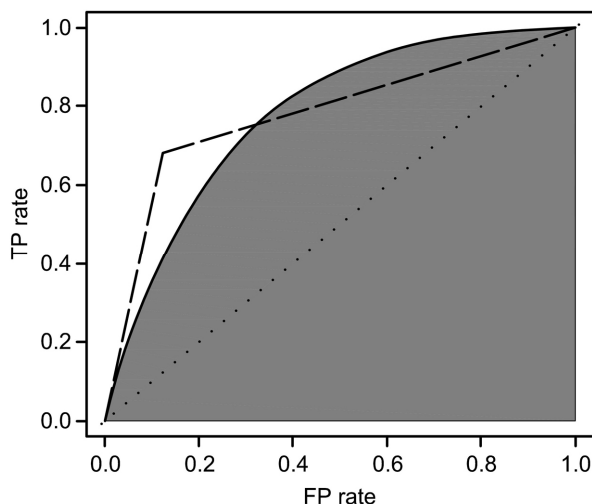
Receiver Operating Characteristics (ROC) represents an evaluation metric that depicts relative trade-offs between benefits and costs, i.e. *True Positive* rate ( $tp_{rate}$  or hit rate)

<sup>21</sup> For instance, if a model has correctly classified 10 landslide instances out of all 15 landslide instances existing in the area, it should be preferred over the model which mapped 10 out of 30 instances correctly. This would not be seen with ordinary  $\kappa$ -index, primarily because the size (hence the influence) of the landslide class is usually up to 10 times smaller than non-landslide class.

and *False Positive rate* ( $fp_{rate}$  or false alarm rate). These are the coordinates of a 2D plot defined as a ROC space (Fig. 19). The ROC curves, given their contingency table parameters (Tab. 6) at a given probability threshold intervals, are simply the performance functions in that space (Fawcett 2006). It is important to highlight that the term probability can refer to any 0–1 scoring scale. Accordingly, some models in this research represent spatial probabilities while others represent just relative scoring scales, but their comparison is hence plausible in ROC space. True probability would require a temporal dimension of the landslide distribution. Due to the usage of ratios of contingency table elements ( $tp/fp_{rate}$ ), indifferent to the actual class distribution, it is possible to avoid the compatibility issues between differently designed model scoring scales in ROC space. Other benefits that arise from within are that approach can handle the unbalanced classes (e.g. it is often that *non-landslide* instances predominate), and deal with cost-sensitive ( $fp_{rate}$ -sensitive) models. The latter is not necessarily beneficial for landslide mapping, where *False Negatives* are even less desirable than *False Positives* (favoring of safety – conservativeness).

The most common numeric parameter of evaluation in ROC space is *Area Under Curve* (*AUC*). The higher the *AUC* (within the 0–1 span) the better the performance and vice-versa. Additional characteristics of the ROC plot (e.g. random performance marker is the diagonal, conservative performance marker is lower left sector and liberal performance marker is the upper right sector of the plot) allow descriptive evaluation, useful when choosing among the models with a similar *AUC*. There are some other measures, which basically all emphasize some characteristics of the curve, but the *AUC* is conceptually the most comprehensive and the most exploited one.

Some particular details need to be specified for different types of models (Fig. 19). Apparently, ROC analysis requires class probabilities, while discrete classifiers models (such as Decision Trees and SVM) do not offer any. Instead, their task is to match the original classes that they were trained upon, thus giving no ranges based on probability. Only additionally discrete classifiers can be supplemented with the quasi-probability descriptor, by simply enacting probability iteratively, i.e. by using several model variants and averaging the final model. The model variants can be generated by varying the classifier parameters, or sampling splits. Otherwise, such discrete classifiers would not be represented by a curve in ROC space, but by a triangle over the main diagonal. In this way some features, especially qualitative features of the curve would remain unavailable for the analysis, thus slightly concealing all details of the modeling performance. Other, probabilistic models, usually called generative classifiers (such as Logistic Regression) do not suffer from such shortcoming (even if they have quasi-probabilistic scoring scales).

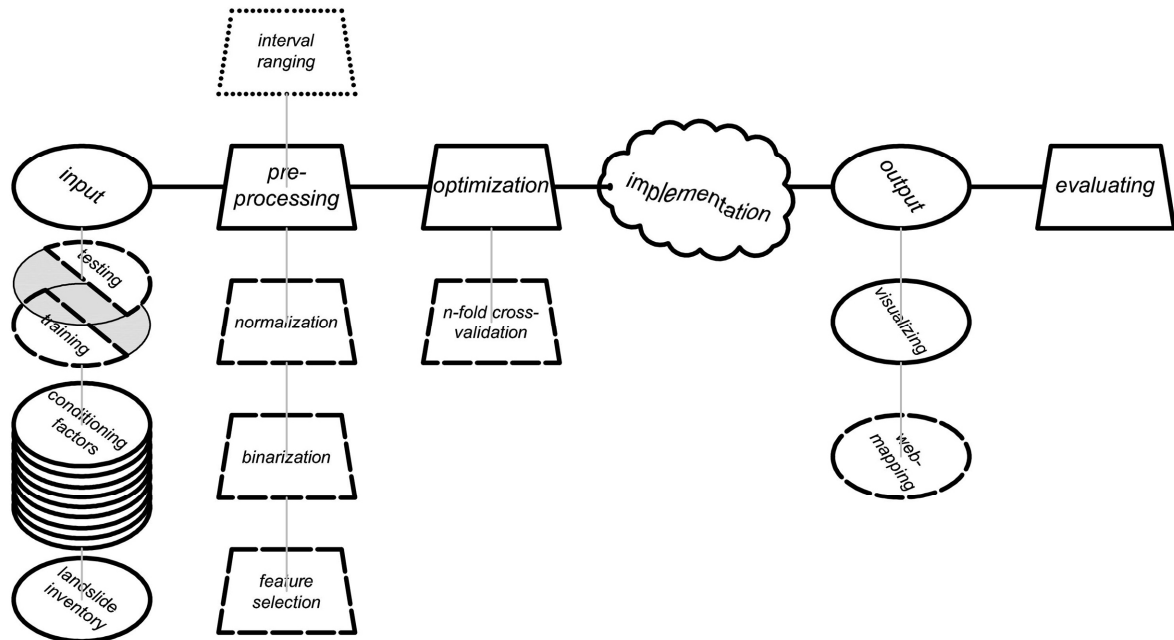


**Figure 19.** An example of a ROC curve for a generative/probabilistic classifier (solid line) and a discrete classifier (dashed line). Note that the *AUC* is approximately 0.8 for both classifiers (shaded area).



## 5.4 Research Workflow

The research workflow depended on a particular modeling method. It is possible to distinguish between slightly different courses of procedures, but the general scheme is uniform (Fig. 20).



**Figure 20.** A flowchart of the research. Dashed elements relate only to the advanced methods based on Machine Learning protocol, while dotted elements apply only to conventional methods.

## 5.5 Data and Software Specifications for Different Methods

The data for the spatial analysis in landslide framework are very typical, nearly standard (regarding the scale/resolution, unit area selection, data format), and involve different attributes, which correspond to different Conditioning Factors that potentially affect the landsliding process. The most of them are the miniature models of these factors, and will be presented in detail in respective case studies (see Chapter 6).

Herein, some general comments on the software are going to be given, since the most of the case studies used similar software combinations.

Primarily, the different GIS platforms have been used for generating the dataset and some of the models, where simple geo-computation was sufficient to deliver a model. ArcGIS Desktop 9 and 10 have been used in generation, geo-referencing, correcting, rescaling, evaluating, converting and actual modeling, via different modules, such as Spatial Analyst for example, but also some extension modules such as AHP. Simultaneously, some open-source alternatives have been implemented for generating additional attributes unavailable in the Spatial Analyst. These included SagaGIS (which has been used to generate compound DEM-based attributes, because the bigger variety than in the ArcGIS packages is available there) and MapWindow GIS (which served for placing a deterministic model SinMap 2). Some of the DEM-based attributes have been generated in order to compare the differences (due to the possible module differences) between ArcGIS-SagaGIS-MapWindowGIS. It turned out that the ArcGIS offered the most stable solutions, which did not have problems in data format conversions. For visualizations of all the data, the ArcGIS turned indispensable, as the most compatible and the most user-friendly platform, but for simple web-map outputs R package, plotGoogleMaps has been used (Kilibarda & Bajat 2012).

Some particular extensions have been utilized for AHP-driven models, since there is an open source extension for ArcGIS, while SinMap 2 refused to work properly in both versions of ArcGIS (which is why the MapWindow GIS has been implemented). Both of these could be done by combining MS Excel sheets, which turned convenient and stable in working with substantial amount of data, while the outcomes could be easily converted back to the .shp extension and introduced to GIS platforms. However, MS Excel (especially the older versions) suffers from limited data capacity, which in consequence can lead to tiling of the study area into smaller parts or downscaling of the dataset (Chapter 6). MS Excel has also been used as a stable platform for communication between the GIS platform and Machine Learning and Statistical software.

Some specific modeling (e.g. *Land Cover* attribute), which required Remote Sensing software platforms, have been conducted in IDRISI Taiga and ErdasImagine 10. These included modeling via supervised classification tasks and image ratioing, which are presented with very powerful modules in both applications.

Statistical modeling has been done in either MS Excel or R software. The latter prove to be much more powerful, giving ability to append specialized or custom packages. For the purpose of the research in this thesis R has been successfully implemented for generating ROC plots, otherwise very demanding for implementation without particular ROC packages in R. The MapComparisonKit, a standalone (ASCII communicable) software package has been used only for additional model performance measurements, such as kappa statistics, although ArcGIS in combination with MS Excel has also been proved sufficient for cross-tabulation-based measurements of the models performance (obtaining contingency tables).

Finally, Machine Learning has been implemented in R, via appropriate packages, but also in MachineLearningOffice, particularly GeoSVM standalone package, and most importantly, in Weka 3.7 software.

- R turned out to be very resourceful with numerous packages dealing with different types of classifiers, even kernel-based ones. The possibility of optimizing the classifier parameters (based on trial-and-error) is supported by providing all the necessary parameter adjustments and kernels selections. It is also at disposal of a very powerful engine for graphic representations of any kind (multi-dimensional as well). Significant drawbacks are the unfriendly, console-launched commands (with negligible amount of GUI elements), which require experience and routine in manipulation of the data, and not particularly good performance with large datasets (with millions of instances and multiple layers). It results in a very costly process, time-consumption-wise. For now, the only solution for the latter lies in multitasking, which itself requires additional packages and a cluster of similar machines (convenient for a classroom environment if available). R's recent merging with GIS platform (R GIS initiative, achieved under SEXTANTE platform) could prove useful for solving compatibility issues, which troubled the remaining two solutions (GeoSVM and Weka).
- GeoSVM turns to be rather small but powerful package. Its essential drawback is in handling of large datasets. It has limited options for experiment design and optimization of parameters, which comes down to only few combinations. The outputs (and inputs) with their .dat extension, suffer from incompatibility with the GIS platforms. It is specialized for SVM classifier, for which it is well equipped with all the necessary adjustment options (including selection of a variety of kernels), but it does not enable implementation of any other type of classifier desired in this research, such as Logistic Regression and Decision Trees.
- Finally, Weka seemed as the best choice, since it has been developed primarily for the Machine Learning tasks. It is stable, even with large datasets, provides several modes which in turn give the practitioner much better perspective and options for the experiment design. In the Experimenter mode, it provides possibility to optimize

parameters in as many combinations as desired. It also gives good perspective for visualizing the data relations and examining the dataset. Further, it directly involves different types of filtering and preprocessing of the data (including even the Attribute Selection), and also provides built-in performance evaluation metrics (overall and class-specific). It turned out to be the fastest among the tested software for the chosen pilot areas. The major drawback is the compatibility, which again has to be conducted in text or table editors (such as MS Excel), in order to receive the inputs and communicate the results to a GIS platform.

## 6 Case Studies

*This chapter is featuring the Objectives 1 & 2, and indirectly all the others (see Chapter 2).*

The proposed methodology has been implemented in three study areas, NW Slopes of Fruška Gora Mountain (Serbia), Starča Basin (Croatia) and Halenkovice Area (Czech Republic). These areas are rather different in terms of geological and other conditions of the ambient, but similar enough to be modeled by competitive techniques for landslide susceptibility assessment. The most prominent qualifications for such context could be listed as follows.

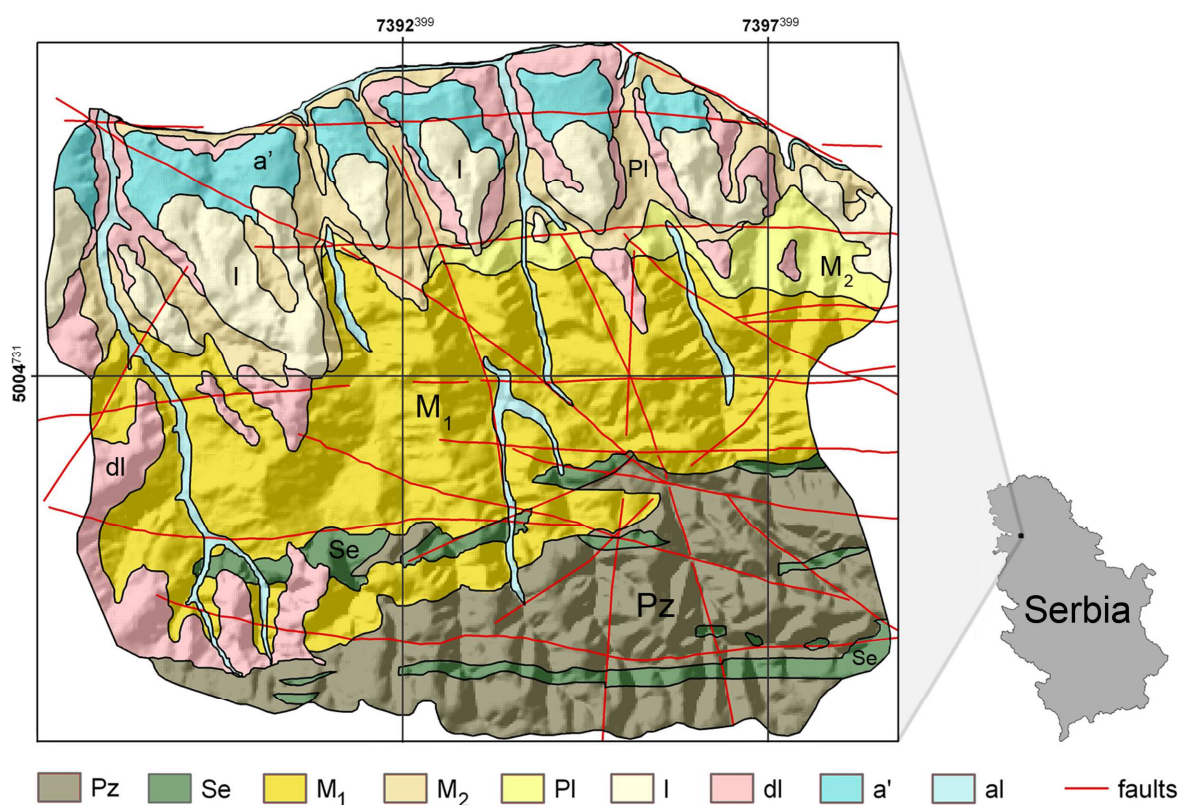
- All of the case studies are focused on a specific type of landslides and triggering mechanism. For instance, first case study (Fruška Gora Mountain, Serbia) is dominated by deep-seated slides hosted in Tertiary formations, wherein the landslides are primarily erosion-triggered, while in the second (Starča Basin, Croatia) they are predominantly shallow and triggered by the precipitation-groundwater dynamics. The Czech case study on the other hand, has also erosion as a major trigger, while the landslide typology is different and includes mostly translational slides hosted in shallow deteriorated mantle of the complex and stratified flysch formations.
- The scales of the input data are concurrent and range from 1:50000 to 1:5000 (or 10–30 m pixel size for raster input type), which is usually acknowledged as regional scale (Fell et al. 2008). This allowed full implementation of GIS and thereupon, typical themes of terrain attributes, such as those regarding ground surface morphology, geological features, subsurface, hydrological and environmental features, as well as some derived synthetic features.
- In all of the studies a reliable landslide reference (Landslide Inventory) was at disposal.
- Even though the temporal dimension has not been included, it is important that some of those time-dependent conditions, primarily climatic, are rather similar for all of the study areas, since they all belong to the continental climatic realm. It would not seem very meaningful to practice the same methodology if some of the studies were belonging to the entirely different ambient.

Nevertheless, some of the case studies have been evidently more elaborated than the others. This especially holds for the first pilot area on Fruška Gora Mountain (Serbia), which has been thoroughly experimented by the entire gamut of the modeling techniques, but there is another reason to it. It has not been addressed in similar context by previous investigators, which is going to be discussed later. Some of the methods have been abandoned in the following studies, and the focus has turned to specific details of particular techniques (the most successful techniques and techniques unprecedented for that study area). It practically came down to implementation of the Machine Learning techniques in the remaining two case studies.

## 6.1 Fruška Gora Mountain (Serbia)

The problematic of the landslide hazard in the context followed throughout this thesis was practically unattended in this area in the past. There has been a host of practical considerations, mainly small geotechnical projects and reports, tightly related to the landslide problematic for various purposes, mainly site-specific ones, for construction design, rarely studies at regional scales for urban and regional planning, or just some plenary researches targeted at different geological aspects, thus just barely scratching the surface of the landslide hazard and susceptibility problematic. Nevertheless, there was a national plan for the nation-wide engineering-geological mapping in 1:100000 scale, by the end of the 20<sup>th</sup> century. The sheets should have matched the existing geological map of the same scale, but the idea was not realized to date. Such situation with data availability affected the initial stage of this research, but the case study area turned resourceful after a compound study of a multidisciplinary team (the author being included) from the Faculty of Mining and Geology (University of Belgrade) in 2006. Segments of these data have been used for this research, as will be presented shortly hereafter.

It is important to outline that this pilot study area has been researched for four years and there have been different aspects of it elaborated and published for different occasions (Marjanović 2009, Marjanović et al. 2009, Marjanović 2010a, Marjanović 2010b, Marjanović & Čaha 2011, Marjanović et al. 2011a, Marjanović et al. 2011b, Marjanović et al. 2011c, Marjanović 2012, Marjanović 2013).



**Figure 21.** Geological setting of the study area (projection: Gauss Krüger – zone 7, Bessel 1841). Standard geological symbols apply: Pz=Paleozoic schists; Se=serpentinite; M<sub>1</sub>=Miocene limestone and marl; M<sub>2</sub>=Miocene sandstone, organic limestone and marl; Pl=Pliocene clay and marl; I=loess; dl=delluvium; a'=terrace sediments; al=alluvium.

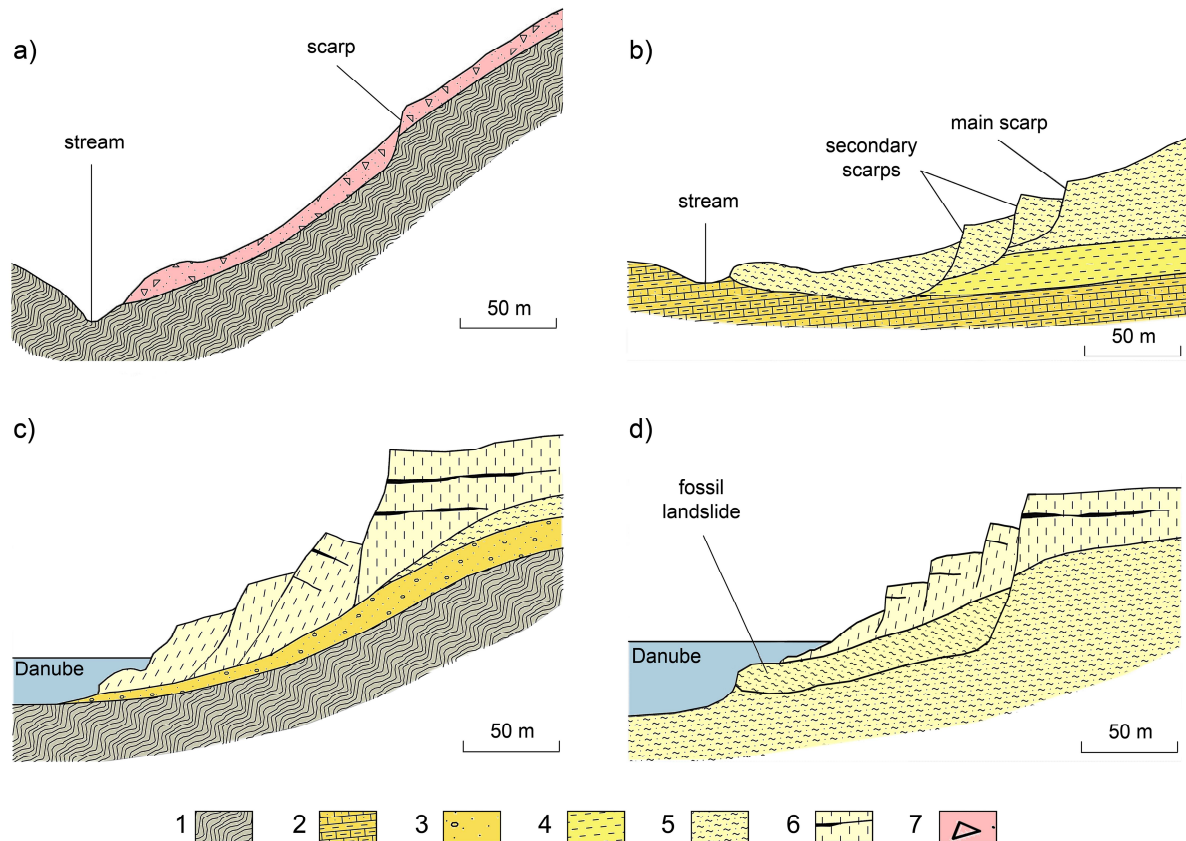
### 6.1.1 Setting

The study area encompasses NW slopes of the Fruška Gora Mountain, in the vicinity of Novi Sad, Serbia (Fig. 21). The site (N 45°09'20", E 19° 32'34" – N 45°12'25", E 19°37'46") spreads over approximately 100 km<sup>2</sup> of hilly landscape, but with interesting dynamics and an abundance of landslide occurrences.

From a wider point of view, stability is a very interesting issue along the river-banks of the Danube. Nearly the entire stretch of its right river-bank in Serbia has been very symptomatic for the landslides, due to the Danube's dynamics and related undercutting erosion of the river mainstream, which moves the banks southward. The Danube curves sharply in NW Serbia (near Novi Sad) and from meridian switches to parallel strike, following the major regional fault-lines such as Periadriatic Line. This causes characteristic dynamics along the right bank, leaving many landslide hotspots, surrounding Novi Sad, as well as other towns and cities Beška, Krčedin, Zemun, Belgrade, Vinča, Grocka, Smederevo etc. further downstream. Numerous morphological imprints of fresh and dormant landslides, especially large ones, are perceivable along the river-bank, where landslide scarps are dominating the upper slope crests toward the local ridges linked to the near-shore river islands, which all provides evidence for tracking landslide depletion and accumulation zones (Bogdanović & Bugarski 1984).

Geological setting of the entire mountain follows a zonal composition. It is caused by the complex E–W trended horst-anticline, shaping the core of the mountain. Traversing these base structures are the NW–SE faults of younger origin, some of which are neo-tectonically and seismically active. A typical succession (Pavlović, et al. 2005) starts with Paleozoic metamorphic rocks in the anticline base, encompassing the ground above 300 m and underlying all younger formations. Triassic basal sediments (conglomerate and sandstone gradually shifting toward limestone) imply localized subsidence in the relief at the time of the basin formation. The closing of that basin during the Jurassic–Cretaceous left typical oceanic crust evidence (ultramafic unit) as well as gulf limestone sequences. The Tertiary is chiefly represented by marine sediments, gaining more carbonate components as the basin turned more limnic during the late Neogene. The most widespread Quaternary unit is loess, which covers the lower landscape toward the Danube on the north.

In geotechnical practice it is believed that the superficial dynamics of the ground directly depends on the geological background, meaning that the rock behavior, under agencies of different processes yields diverse geodynamic outcomes (Janjić 1962). Geomorphological evidence supports these expectations for this relatively small study area (Fig. 21), where slope stability can be generalized into several scenarios (Fig. 22). The higher ground, chiefly composed of metamorphic rocks i.e. coated by the shallow mantle of metamorphic origin, is mostly shaped by the fluvial and slope processes, where shallow valleys and gullies are cut through the coated slopes all the way to the bedrock, forming a dense drainage pattern. Since they are not too steep, mostly vegetated and not severely tectonized, the dominating slope processes on the flanks of these valleys are screes, minor rockfalls and minor shallow landslides (Fig. 22a). The central plateau (Fig. 21) is formed of carbonate and clastic rocks and has insufficient thickness to develop karstification, so fluvial forms predominate. However, the slope processes are better developed in Miocene-Pliocene marlstone and clay, where deep-seated landslides (up to 10–20 m in depth) are being hosted (Fig. 22b), particularly on the slopes steeper than 20°. The morphology of the lowest ground that flattens northward is governed by the fluvial dynamics of the Danube. It is represented by sequences of river terraces, inundation plains and the alluvial fans of smaller tributaries. This is where the loess formations are facing the river in relatively steep cliffs. Despite of the general stability of loess, landslides on the cliff faces are quite common along the riverbank. The latter is caused by surges of groundwater that locally communicate through the bedrock (Fig. 22c) keeping the loess units in unfavorable conditions (saturation and suffosion). Moreover, loess overlays clay and marl units, that locally host the fossil landslides and seize the loess slabs above them (Fig. 22d).



**Figure 22.** Schematic models of instabilities present in the case study area: **a)** shallow landslide; **b)** rotational landslide in Neogene clay; **c)** collapse of loess slabs (caused by the groundwater dynamic relation with the river); **d)** relict landslide seizing loess slabs above. Legend shows 1=Paleozoic schists; 2=marl and limestone; 3=sand and gravel; 4=clay; 5=sandy clay; 6=loess; 7=delluvium.

It seems apparent that the instabilities in proposed study area are generally driven by geological, geo-morphometric, and environmental attributes (such as *lithology*, *elevation*, *slope angle*, *Land Cover* and so forth) which was further elaborated in this study. The trigger however, could be different, ranging from seismic activity along the nearby active fault zone, through groundwater and superficial water dynamics (which change the pore-pressure regime and according to the Terzaghi's principles, decreases the soil strength), to direct southward fluvial erosion of the Danube, which not only undercuts the banks but lowers the local erosion basis, causing all local streams in the watershed to intensify their vertical erosion. Heavy rainfall, which is typical for the late-summer periods (August–September) in the wider area (according to meteorological observatory in Novi Sad) also plays important role as a potential trigger. The problem of the trigger requires temporal analysis which was not directly involved in this research, due to the temporal data shortage, i.e. the shortage of the landslide activity records (precise dating and rate of displacement).

### 6.1.2 Dataset

Within the framework of The Vojvodina Province governmental project on Geological conditions of rational exploitation of the Fruška Gora Mountain area, completed in 2005, by the multidisciplinary expert team from the Faculty of Mining and Geology (University of Belgrade), a set of different thematic maps was generated for the entire Fruška Gora Mountain area, including Geological, Geomorphological, Photogeological, Pedological, Seismological and other maps. Thanks to the courtesy of the Faculty of Mining and Geology scholars some of those resources have been used in this research.



It has already been indicated that conventional techniques for this type of research involve the use of an input dataset with important terrain attributes, called Conditioning Factors, which are being selected according to their availability at selected regional scale, and their significance for the problem at question. Although a majority of researchers use various available data and measure their statistical dependence against landslide occurrences prior to their implementation, it is suggested that the input attributes are chosen more cautiously, due to their temporal and scale constrictions. In landslide susceptibility analysis it is considered that constant variables (regarding the duration of the landsliding process which is sometimes measured in decades) should be preferred, meaning that attributes such as *Land Use* and *Land Cover*, climate data, pore-water pressure data and other that have distinct annual or even diurnal variations, should be used with caution (Ohlemacher 2007). Furthermore, the caution is necessary for preparing synthetic data as well, especially interpolated data. For instance, a groundwater table map or soil mantle depth are hypothetically possible to extract from the borehole data, if the sampling density turned out to be good enough, so that some interpolation method (possibly advanced, e.g. based on non-ordinary kriging) could be used to interpolate data over the area. In any case, the interpolation implies switching from point-specific data to spatially distributed data, which inevitably introduces uncertainty error. It is hence disputable whether such data are contributing to the model or biasing it. Attribute Selection (see Chapter 5.1) is one way of dealing with such problem, but it is also arguable whether it should be used, especially in multivariate framework. In this study a *Chi-Square* and *Information Gain* ranking have been implemented, mostly though just to demonstrate and highlight which are the most dominant Conditioning Factors, and to prove their statistical independency to the Landslide Inventory, and not to exclude any low-ranked factors. It has actually been shown that excluding these factors have led to the poorer modeling performance, especially in Machine Learning case<sup>22</sup>.

The 2D raster thematic GIS layers, representing different Conditioning Factors and Landslide Inventory, have been acquired from different resources, entailing different levels of generalization and different scales. Subsequently, they have been compiled to an optimal 30 m cell resolution. Regarding the support problem, generalization to 30 m cells is plausible, because the most critical data had reasonably small scale (such as geology with 1:50000) and turned out to be too detailed for the research purpose. As a matter of fact, an additional generalization (by aggregating similar classes) took place over such inputs, thus compensating for the support problem (Dungan 2001). Assembling of the input dataset, has been prepared by combining ArcGIS and SagaGIS packages (Böhner, et al. 2006) stored in ESRI grid formats, and also in ascii format which was necessary for commuting with the Machine Learning software (in absence of fully integrated ML modules in present GIS platforms). The content of the input dataset was practically identical for all implemented methods, apart from deterministic model, which required different input variables (as will be specified later-on).

---

<sup>22</sup> SVM and Decision Tree algorithms for instance, already include some sort of Attribute Selection on their own, so no further exclusion by Attribute Selection is needed.



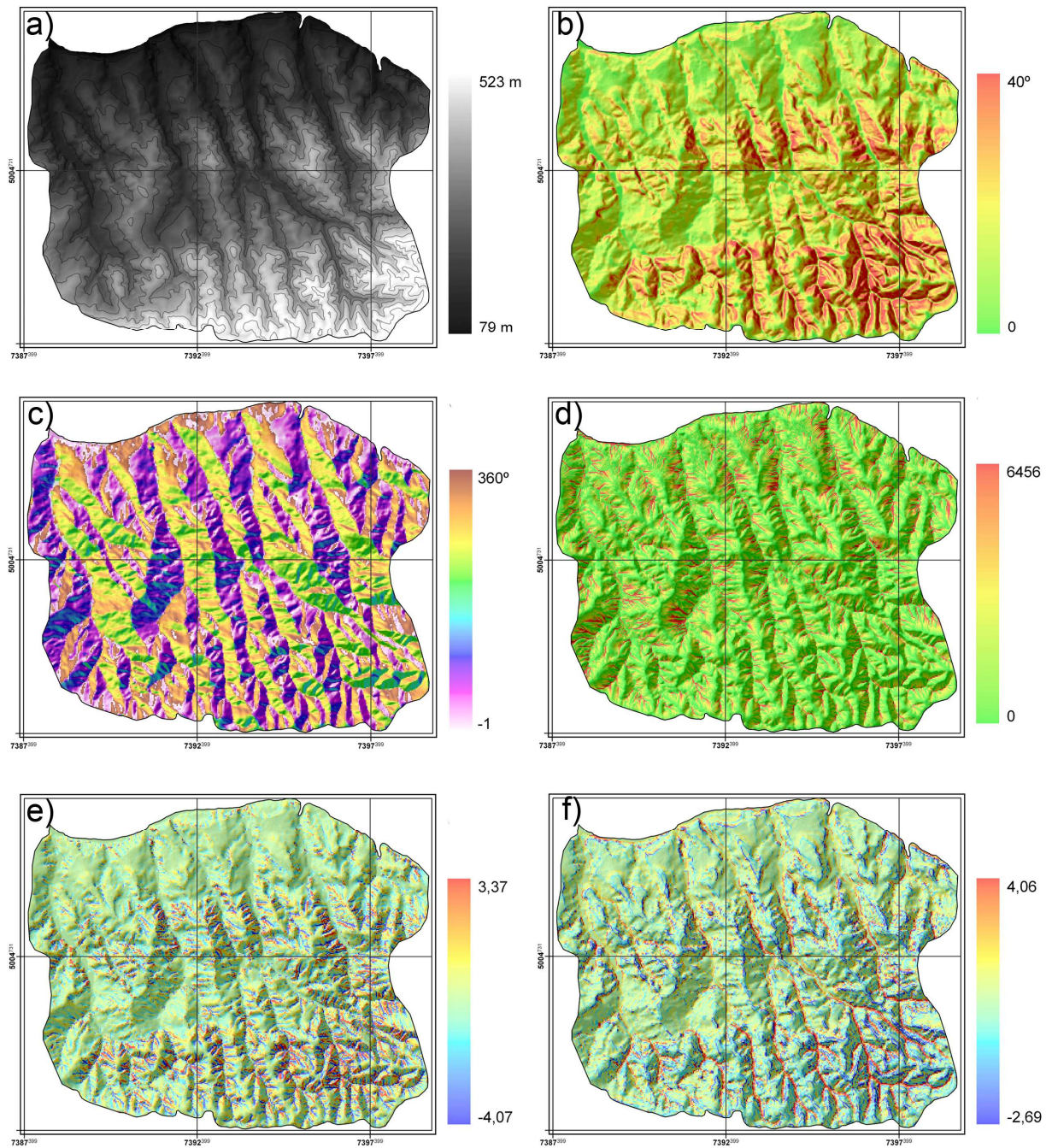
### 6.1.2.1 Conditioning Factors

The dataset part regarding the Conditioning Factors thus included morphometric, hydrological, geological and environmental Conditioning Factors and the Landslide Inventory.

#### 6.1.2.1.1 Morphometric Data

Topographic data have been obtained from the standard topographic maps of Serbia at 1:25000 scale in Gauss-Krüger Projection (Zone 7, ellipsoid Bessel 1841, sheets Bačka Palanka 377-4-1 and 377-4-2), where contour lines generated from a conventional photogrammetric restitution are given, with the equidistance of 10 m. For generation of the Digital Elevation Model (DEM), these contour lines were first digitized, and then decomposed to point data, extracted to DEM by TOPOtoRaster module, with 30 m cell resolution. Subsequently, DEM has been rectified for the erroneous pits and picks and tested for the consistency and accuracy by 10% split Cross-Validation of the source point data. Root Mean Square error of the validation was in reasonable limits (RMS was lesser than 0.4) so the DEM has been acknowledged. Even though the topographic maps with a given scale are suitable for DEMs productions with denser resolutions (Carrara et al. 1997), a DEM resolution of 30 m was chosen for two reasons: as a proper DEM resolution for regional landslide modeling and as the adequate support grid size compatible with other data sources used in this study, such as a geological map at 1:50000 scale, and 30 m Landsat imagery (Dungan 2001). All DEM-related geo-morphometric and hydrological thematic maps kept the same (30 m) resolution. The latter involved derivation of numerous morphometric parameters (Wilson & Gallant 2000) listed as follows (Fig. 23a-f).

- *Elevation* ( $F_1$ ) – a float (continual) raster (Fig. 23a), suggesting that the linear increase in potential energy with altitude brings the higher susceptibility to the higher ground. It actually represents a Digital Elevation Model (DEM) of the terrain, described earlier.
- *Slope angle* ( $F_2$ ) – a float raster (Fig. 23b) which is considered important for slope stability because of the direct physical relation with the landslide phenomenology. Greater angles ( $\theta$ ) propose higher instability of the slopes and vice-versa, but with restriction to a particular rock type (e.g. in solid rock, the slopes are expected to be stable even with a steep slope angle, while slopes in clay do not need a steep angle to host instability). From morphometric point of view, this factor is a first-derivate-based morphometric feature computed directly from the DEM by Degree Polynomial (DP) slope algorithms, also called D8 algorithm, referring to a 3x3 window, where 8 pixels surround the central one and define its value (Olaya 2004).
- *Aspect* ( $F_3$ ) – a float raster (Fig. 23c) which refers to the spatial exposure of the ground element (its azimuth). It is also computed from DEM values by means of DP-D8 algorithms and ranges from 0° to 360° (counter clockwise), suggesting that susceptibility to landslides rises from SW to NW quadrant, since the diurnal solar path influences moisture in slopes and topsoil mantle thickness. Thus NW slopes are the most inconvenient (with the highest moisture content and the thickest mantle detritus) while SW are the least susceptible.
- *Slope length* ( $F_4$ ) – a float raster (Fig. 23d) of morphometric length of the slopes, also derived from DEM by means of DP-D8 algorithm. It suggests that longer slopes tend to be more susceptible due to the higher possibility of hosting a retrogressive (upslope) landslide development, as the most common form of landslide progression on the slope.



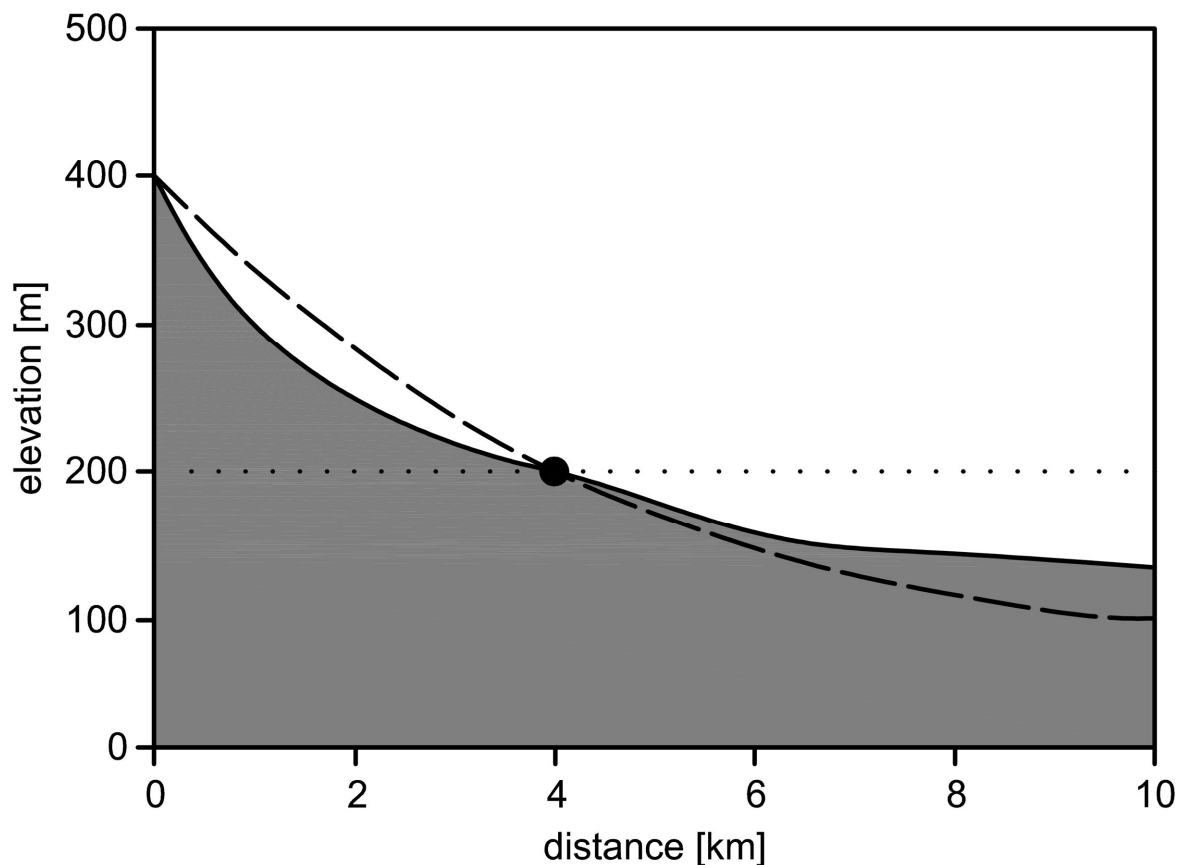
**Figure 23.** Various thematic inputs: **a)** elevation; **b)** slope angle; **c)** slope aspect; **d)** slope length; **e)** plan curvature; **f)** profile curvature.

- *Plan curvature* ( $F_5$ ) – a float raster (Fig. 23e) calculated on the basis of second order derivative from the elevation data (DEM), defining the convex/concave character of the surface in the direction perpendicular to the contour lines. Positive values suggest local relief depression (concavity); negative values represent positive (convex) forms, while zero values are reserved for the flat ground. Plan curvature is considered important as it reliably indicates convergence and divergence of slope surface in the depletion zone (concave forms of the landslide crowns, tension cracks and depressions, and also zones of the local water accumulation) and accumulation zones (convex forms of the landslide foot and toe) (Ohlemacher 2007).
- *Profile curvature* ( $F_6$ ) – also a float raster (Fig. 23f), practically the same as a plan curvature, different only in direction, which is here perpendicular to the contour lines

(in downslope direction). It also indicates the subtle morphology of the slope and reveal possible landslide elements, but also depicts the distribution of driving and resisting stresses along the landslide motion direction (Ohlemacher 2007).

### 6.1.2.1.2 Hydrological Data

Hydrographic network has been digitized from basic Topographic maps at 1:25000 scale. It has been upgraded by additional synthetic stream lines, computed by using morphometric operations over DEM (Compound Module in SagaGIS) and representing mainly small, higher-order tributaries. The vector of stream network has been used to generate the *distance from stream*. Another purely hydrological feature is given as *Topographic Wetness Index* (Wilson & Gallant 2000).



**Figure 24.** The cross-section of the Danube's tributary riverbed (solid line) in comparison to the theoretical curve of erosion basis (dashed line). Note that the interception point separates vertical and lateral erosion preference at approximately 200 m.

- *Distance from stream* ( $F_7$ ) – is a float raster (Fig. 25a) computed from vectorized stream network (by using Euclidean distance module in ArcGIS SpatialAnalyst). It depicts the influence of linear erosion on the slope stability, since deformation and failure processes develop regressively upslope under the vertical and lateral influence of the linear erosion. In narrow upper sections of the valleys vertical erosion dominates, steepening the slopes and destabilizing rock masses. On the other hand, lower sections tend to develop lateral erosion, widening the valley bed, once again pushing slopes off the balance. That critical point can be estimated by the analysis of the local erosion basis, and for the average profiles of the northern mountain slopes it rests at about 200 m (Fig. 24). In turn, combination of these two mechanisms erodes the slopes, whereas lateral erosion (below 200 m) leaves more prominent effects. For this reason, the factor  $F_7$  has been weighted, so that the lower ground has stronger

influence. Foregoing discussion suggests that areas closer to the stream lines are more affected than remote ones, thus buffering out the landslide susceptibility toward the ridges of local watersheds.

- *Topographic Wetness Index* or *TWI* ( $F_8$ ) – is another DP/D8 second order DEM-based float raster (Fig. 25b), and represents a morphometric parameter which defines the terrain retention (moisture distribution) in relation to the local topography, pinpointing the areas of higher water contents. This parameter roughly expresses the water retention distribution throughout the area, and due to the effective stress decrease in saturated slopes, it maps the areas with higher *TWI* values as relatively more prone to instabilities (Olaya 2004). It is a function of the local slope angle and upslope contributing area and can be formulated in different ways (Eq. 42) depending on the way  $a$  (upslope contributing area) is defined.

$$TWI = \ln \frac{a}{\text{tg}\theta} \quad (42)$$

### 6.1.2.1.3 Geological Data

Geological data were obtained from the digital geological map at 1:50000 scale, which has been compiled from a raw field geological map at 1:25000 and the Basic Geological map of Serbia at 1:100000 scale. Since it is not formational but chronostratigraphic one, the map was further simplified to meet the requirements of this case study<sup>23</sup>. Therefore, the generalization to a raster map with 30 m resolution was justifiable. The map was also used to derive the synthetic data, based on Euclidean distance calculations (ArcGIS SpatialAnalyst module). These included the buffer of geological structures and the buffer of hydrogeologically significant borders (Fig. 25c-e).

- *Lithology* ( $F_9$ ) – a discrete (categorical) raster (Fig. 25c) of present rock types derived after geological map as mentioned above. The map depicts 9 rock units, essentially different in their physical-mechanical behavior, thus differently prone to instabilities, i.e. (alluvium, terrace sediments, delluvium, loess, Pliocene clay and sand, Miocene marlstone and sandstone, Miocene limestone, Serpentinite and Paleozoic schist). It is common practice to range *lithology* (to reclassify it into different intervals) arbitrarily, by assigning scores or weights to each of the lithological classes (the scores can range from 0 to 1 for instance). However, in Machine Learning implementations (see Chapters 6.1.3.4-7) *lithology* needed to be treated differently, in order to avoid subjective quantification of categorical classes. The entire *lithology* factor has been actually isolated into  $m$  different sub-attributes, so called dummy variables, giving  $m$ -bit class codes for each unit ( $m=9$ ). In each sub-attribute a binary reclassification further applies, so that the given unit becomes class 1 and all remaining units become class 0. For instance, a delluvium class is the 3<sup>rd</sup> class of the *lithology* and it is hence coded as a sub-attribute 001000000. Within the 001000000 sub-attribute (new raster layer) instances corresponding to delluvium are assigned value 1 and all the rest is 0. By such intervention each rock unit has been treated by the proposed Machine Learning algorithm with no preference.
- *Distance from structures* ( $F_{10}$ ) – is a distance raster (Fig. 25d) which displays the distance from the geological structures such as faults and joints, obtained from the

<sup>23</sup> The map was divided into chronostratigraphic units, thus separating units by different age. This breaks the continuity of physically similar rock masses just because they are different in age, although they might belong to the same metamorphic/magmatic/sedimentation cycle. Formational map on the other hand follows the entire cycle of a rock mass creation, which in result keeps physically similar rock masses within the same units. Adopted guideline that similar rock masses express similar superficial dynamics better complies with formational map criteria, so the source chronostratigraphic map had to suffer some simplification and generalization of original units.

photogeological interpretation (See Chapter 6.1.2.5). Since faults and joints represent the zones of weakest shear resistance (limited only to a residual shear resistance) and also affected by the infiltrated water and fill material, it is logical to assume that instabilities are more prone in the areas closer to these structures. In more seismically active areas such parameter could be much more appreciated since the shear resistance faces further effects, related to the fault dynamics.

- *Distance from hydrogeological boundaries ( $F_{11}$ )* – is another buffer raster (Fig. 25e) which introduces the influence of geological boundaries defined by the change in hydrogeological function of the adjacent rock units. This function could change significantly as the porosity and permeability between two adjacent unit change. Any abrupt downslope decrease in permeability causes higher water pressures at the boundary, especially in characteristic structural disposition of the rock units. The boundaries that were suspected for such mechanism had been mapped and digitized and the appropriate buffer was created, suspecting that areas closer to the boundary are more susceptible to landslides.

#### 6.1.2.1.4 Environmental Data

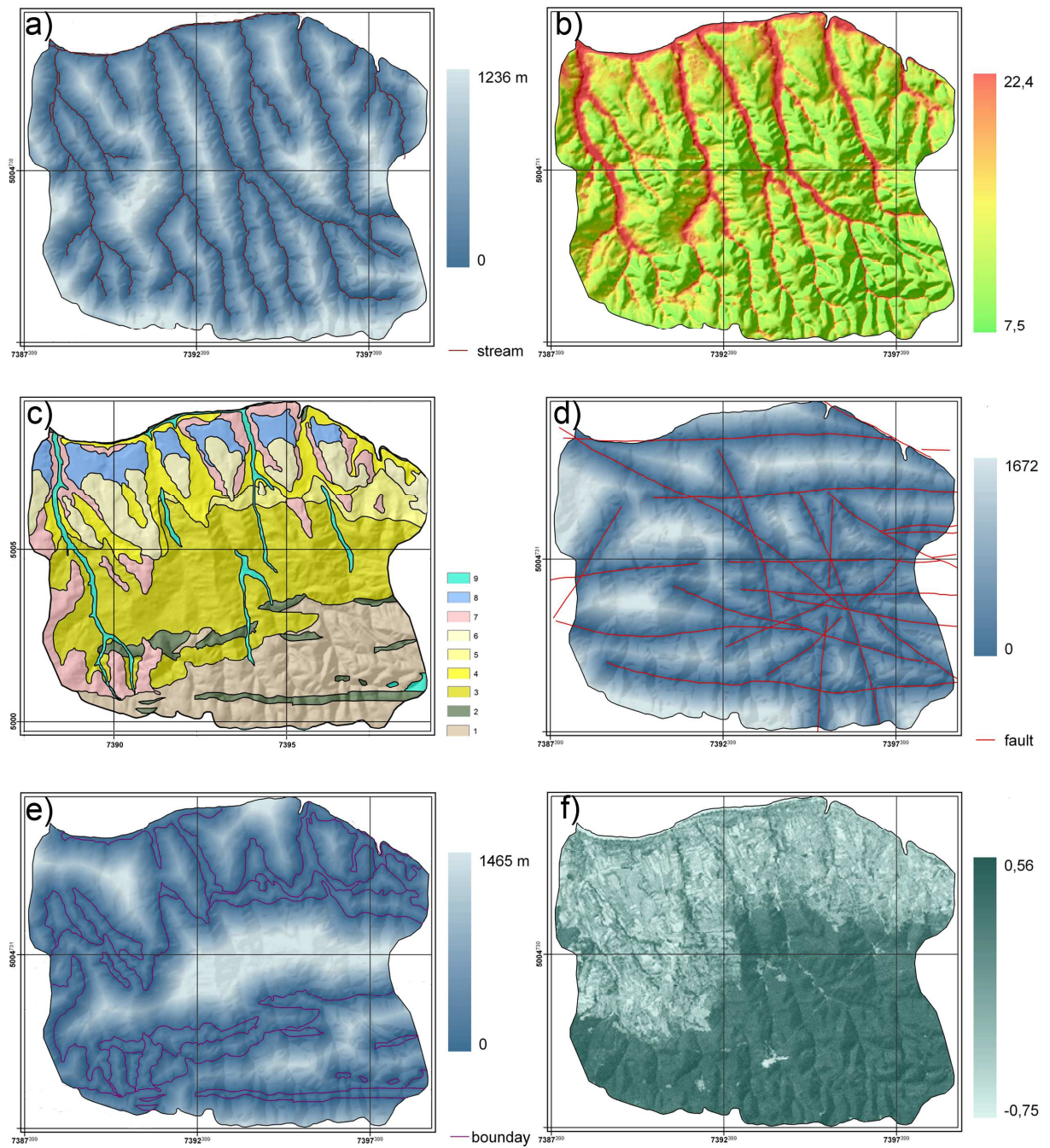
Environmental information particularly regarded the *vegetation cover*, due to a possible remediation that some vegetation types can provide for shallow landsliding (the influence of root cohesion and moisture retention). In the Multi-spectral Satellite Image Processing numerous spectral vegetation indices are proposed to estimate biomass and delineate different types of vegetation apart from bare soil, rock, wetlands or artificial surfaces (Glenn et al. 2008). After experimenting with several indices the *Normalized Difference Vegetation Index (NDVI)* has been acknowledged. *NDVI* basically explores the chlorophyll spectra using the multi-spectral (multi-band) Landsat TM image bands (Eq. 43).

$$NDVI = \frac{\text{NearInfraRed} - \text{Red}}{\text{NearInfraRed} + \text{Red}} \quad (43)$$

It particularly exploits the abrupt chlorophyll absorption difference in red and near-infra-red band (analogue to the chlorophyll's spectral signature trend). The range of the index falls from 1 to -1, wherein, vegetation has values close to 1, water bodies close to 0 and dry areas close to -1 (Ravi 2002), making the *NDVI* easy for reclassification (required in the most of the techniques, except for the ML-based ones). The raster has been created by combining two Image Processing platforms, IDRISI Taiga and ERDAS Imagine 8, resulting in a 30 m resolution raster map of *vegetation cover* (Fig. 25f).

- *Vegetation cover ( $F_{12}$ )* – is a discrete raster which separates heavily (value 1) and sparsely vegetated areas (value -1). Former are stabilizing loose topsoil to a certain extent, making it less susceptible than bare soil due to water retention and root cohesion. This attribute is created by processing Landsat 7 TM bands from summer 2008 (LE71870292008175EDC00 from USGS free on-line repository), via *NDVI*, which has proved to be more representative in the selected scale than more advanced biomass estimators, such as *Enhanced Vegetation Index (EVI)* (Glenn et al. 2008).





**Figure 25.** Various thematic inputs (continued): **a)** distance from stream; **b)** TWI; **c)** lithology (1=Paleozoic schist, 2=serpentine, 3=Miocene limestone/sandstone/marl, 4=Miocene marl/clay, 5=Pliocene clay, 6=loess, 7=delluvium, 8=river terrace, 9=alluvium); **d)** distance from structures; **e)** distance from hydrogeological boundaries; **f)** NDVI.

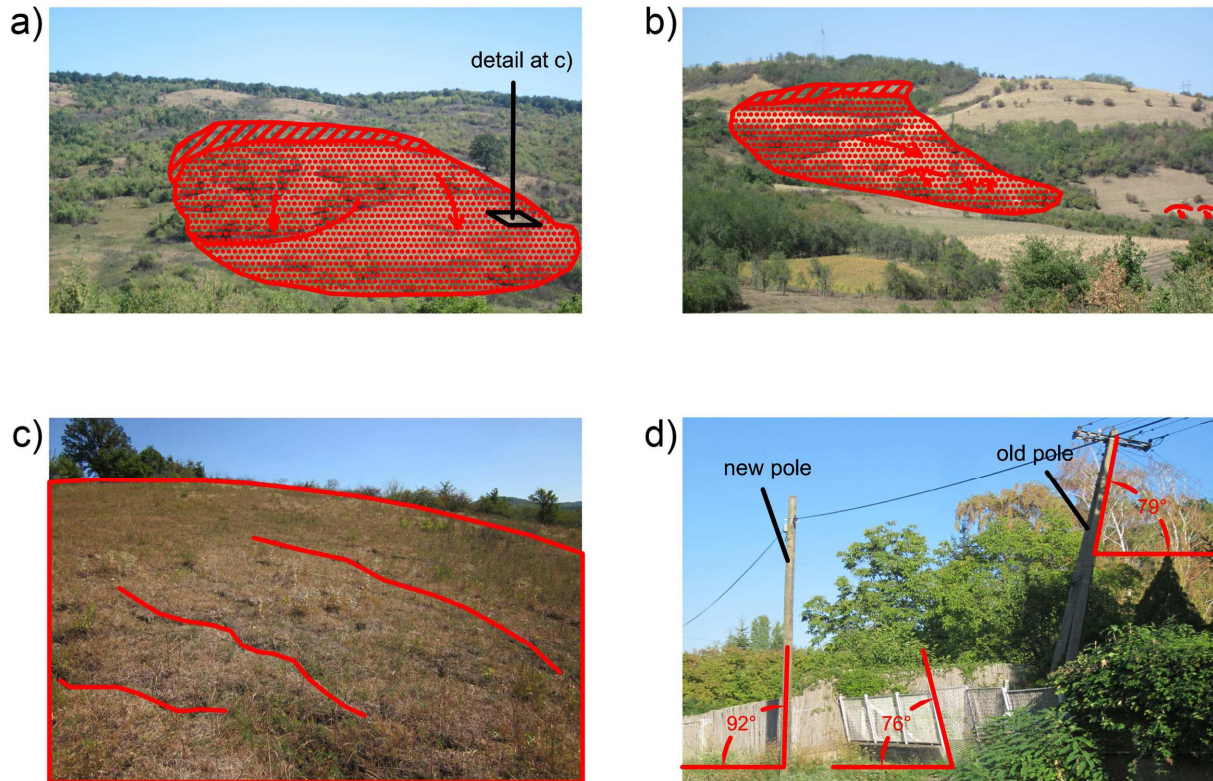
### 6.1.2.2 Landslide Inventory

A photogeological map (1:50000) represents a coupled geomorphological and structural map, that stresses the forms of the most current geological processes at play. It is compiled by adopting basic geomorphological units and appending the heuristic (RS-based) structural and stability interpretation. The latter was performed as an expert-based analysis of 30 m

Landsat TM imagery (LE71870292008175EDC00 from USGS free on-line repository<sup>24</sup>) and auxiliary derivatives (enhancements and processed images, i.e. various indices, Principal Components, Color Composites etc), as well as orthorectified aerial photographs at 1:33000 scale. At the final stage of its compilation the map has been verified on the field, by practicing conventional engineering-geological mapping methodology.

\*\*\*

Aforementioned verification methodology implies evidencing of the landslides, proposed by the photogeological interpretation (by means of RS methodology, stereoscopy of the aerial photographs in particular).



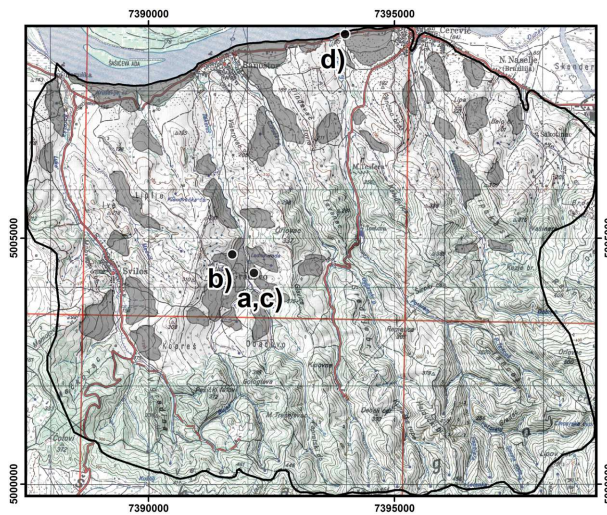
**Figure 26.** Augmented photo-documentation from the field investigation: **a)** reactivated landslide; **b)** dormant landslide with secondary scars; **c)** tension cracks (left flank of a); **d)** object deformations (photo by the author 2008-2012).

The first objective involved on-field check of the validity of the interpretation, especially for the landslides that have been mapped with lower certainty. These pre-defined locations have been checked by using conventional 1:25000 topographic maps and low-accuracy navigation device (still sufficient for reambulation at the 1:50000 source scale), and this type of field check had to be limited to the smaller landslides only, i.e. only those observable on the field. Several small polygons have been removed from the inventory, since being misinterpreted as landslides. Since the original interpretation had discerned landslides by their activity stage (active and dormant) the objective of this additional survey was also to find evidence which could support the activity estimation, as well as estimation of the triggering mechanism, landslide depth and type (in order to conform to the adopted classification system). Common indicators of the active sliding can be found in

<sup>24</sup> LE71870292008175EDC00 is the image repository name code meaning: LE7=ETM+ sensor was in operation, 187=path, 029=row, 2008=year, 175=day of the year (date of acquisition is hence 23 of June, 2008), EDC = Data held by EROS - Receiving station unknown, 00 = version 0.



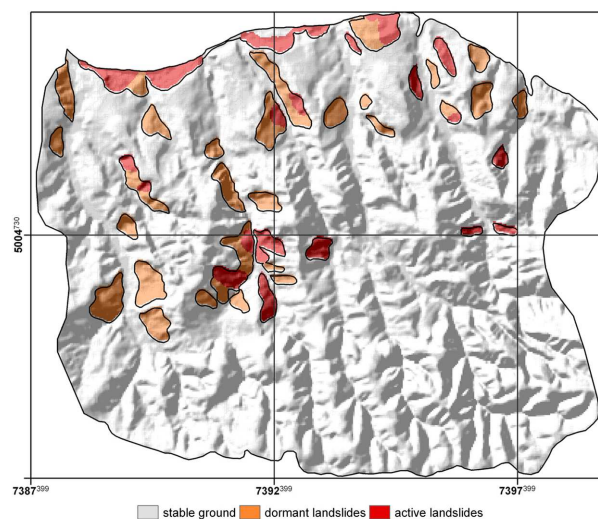
geomorphological, hydrogeological, botanical evidences, as well as in deformations of the man-made structures. For instance, fresh scars, opened tension joints locally filled with water, local ponds and hummocky topography are strong evidences of recent activities in the depletion and accumulation zones (Fig. 26a-c, 27). Also, fresh fissures in the buildings or paved roads, disarrangement of the fences and staircases (as most fragile constructions), and tilted tall objects such as poles or trees are further supporting the activity assumption (Fig. 26d, 27). Information from the members of local community are also appreciated, especially for dating of the landslide events, estimating the water table levels in aquifers, estimating the activity rate, estimating the trigger and assessing the damage produced by a single or multiple events.



**Figure 27.** Location of the photographs from Figure 26. (landslides are shaded polygons).

\*\*\*

Since the emphasis was on the stability analysis, only that part was extracted from the final photogeological interpretation, otherwise rich in other thematic geological content (geological structures and geomorphological processes). In this manner the Landslide Inventory raster map at 30 m resolution was obtained. It reveals the stability of the landscape by distinguishing two classes of landslides based on their activity stage, *dormant* and *active* landslides (Fig. 28).



**Figure 28.** Landslide Inventory.



According to the adopted classification (Varnes 1984), landslides fell into the category of rotational and translational earth slides, while other, minor occurrences of flows and falls were not taken into consideration. The map depicted the distribution of classes into the following categories: 3.6% *active landslides*, 5.6% *dormant landslides* and the remaining 90.8% conditionally stable ground (*non-landslides*).

**Table 7.** Raster thematic maps represented by thematic attributes of the input dataset.

<b>Conditioning Factor</b>	<b>source, scale/resolution</b>	<b>short description</b>
<i>elevation</i> ( $F_1$ )	topo-maps, 1:25000	DEM of the terrain surface
<i>slope angle</i> ( $F_2$ )	DEM, 30 m	angle of the slope inclination
<i>aspect</i> ( $F_3$ )	DEM, 30 m	exposition of the slope
<i>slope length</i> ( $F_4$ )	DEM, 30 m	length of the slope
<i>plan curvature</i> ( $F_5$ )	DEM, 30 m	index of concavity parallel to the slope
<i>profile curvature</i> ( $F_6$ )	DEM, 30 m	index of concavity perpendicular to slope
<i>distance from stream</i> ( $F_7$ )	DEM, 30 m	buffer of drainage network
<i>TWI</i> ( $F_8$ )	DEM, 30 m	ratio of contributing area $a$ and $\tan(F_2)$
<i>lithology*</i> ( $F_9$ )	geo-map, 1:50000	rock units
<i>distance from structures</i> ( $F_{10}$ )	geo-map, 1:50000	buffer of structures
<i>distance from h.g. boundary</i> ( $F_{11}$ )	geo-map, 1:50000	buffer of boundaries between rock units
<i>vegetation cover</i> ( $F_{12}$ )	Landsat images, 30 m	interpretation of vegetation, water bodies and bare soil, based on <i>NDVI</i>

\* nominal attribute

**Table 8.** General statistics of attribute layers.

<b>Conditioning Factor</b>	<b>maximum value</b>	<b>minimum value</b>	<b>mean value</b>	<b>standard deviation</b>	<b>IG ranking</b>
<i>elevation</i> (m)	523.19	79.83	241.64	94.61	1
<i>slope angle</i> (°)	40.16	0.00	11.77	6.62	5
<i>aspect</i> (°)	359.99	-1.00	173.26	116.47	10
<i>slope length</i> (m)	6456.41	0.00	103.77	152.19	11
<i>plan curvature</i>	3.3725	-4.0694	0.0085	0.0085	9
<i>profile curvature</i>	4.0607	-2.6968	0.0089	0.4060	8
<i>distance from stream</i> (m)	1225.60	0.00	319.59	224.89	6
<i>TWI</i>	22.49	7.56	11.51	2.83	4
<i>lithology*</i>	*	*	*	*	2
<i>distance from structures</i> (m)	1672.75	0.00	309.85	259.11	7
<i>distance from h.g. boundary</i> (m)	1465.09	0.00	306.32	295.45	12
<i>vegetation cover</i>	0.56	-0.75	0.17	0.21	3

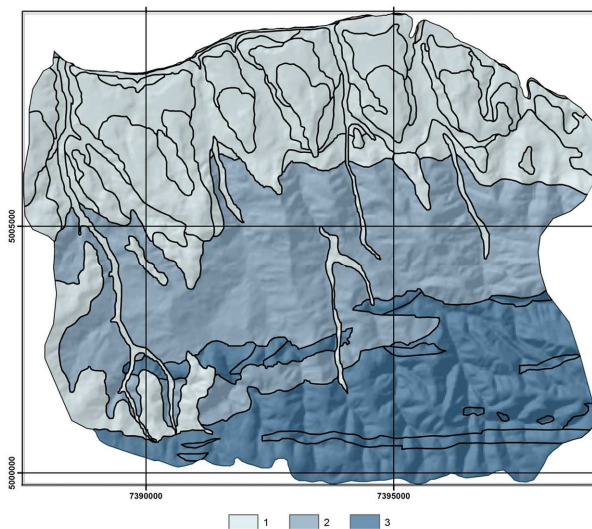
\* nominal attribute

### 6.1.2.3 Data Summary

The descriptions of used input data, their spatial resolution and sources are given in Table 7. The general statistics for chosen Conditioning Factors and *IG* ranking (Quinlan 1993) based on overall terrain data are shown in Table 8. In Appendix 1, a dataset variant used in some conventional methods is given. Note that all numeric inputs are ranged and nominal inputs are quantified there.

### 6.1.2.4 Deterministic Geotechnical Data

Deterministic analysis was somewhat specific, since it required a different type of data, from different resources. Apart from topographic data, i.e. DEM, these included physical-mechanical rock parameters, such as (dimensionless) cohesion  $c_{soil}$ , friction angle  $\varphi_{soil}$ , volumetric weight  $\gamma_{soil}$ , and recharge to soil transmissivity to precipitation ratio  $q/T$ . The most of these parameters are normally obtained from the laboratory tests of samples and apply for site-specific scaled investigations. Adopting and distributing these to regional scales is a very disputable issue for a number of reasons. The most obvious one is the proper sampling density which is unaffordable at regional scales, because each sample requires a borehole, sampling, and laboratory tests for each geological unit along the borehole. Appropriate sampling density and performing of a meaningful interpolation over the sampled values is therefore practically impossible. Hence some generalization has to be introduced, but one needs to be aware of the level of bias which is caused by generalization. Common practice (Pack et al. 2001) is to adopt the average values of required parameters per each quasi-homogeneous rock unit or to simply assign generally accepted values for a given rock mass (Bell 2007). Perhaps, the most objective estimation is the one regarding the  $q/T$  ratio, because it is obvious that the maximum precipitation is constant regardless to the rock units, while transmissivity  $T$  could be relatively precisely estimated. For instance  $q/T$  log-ratio of 1000 means that the ground received three times as more water than it can transmit, which is an extreme case and it is obviously more problematic for clayey, schistose or other relatively impermeable rock than for the sandy or calcareous units.



**Figure 29.** Different lithological domains (regions) of the area for determination of different geotechnical parameters: 1=clayey-marly soil; 2=calcareous-sandy soil; 3=solid bedrock.

With all this being said, the case study area has been subdivided into several quasi-homogeneous regions, by unifying rock units with similar consistency. It is obvious that lithological classes (Fig. 21, 25c) have been aggregated into 3 regions (Fig. 29). The according parameter values for these regions are given in Table 9. The most drastic generalization would be if just one region with a single set of physical-mechanical parameters would have been considered.

Finally, it is important to mention that such model and such data could not have been implemented without a GIS platform and although the regional deterministic modeling is disputable, it shows a certain potential, especially for the appropriate case studies (prevalent shallow landsliding scenarios). GIS thus emerges as a very powerful tool which easily administers such modeling procedures.

**Table 9.** Average parameters over different lithological domains (regions).

parameter	region 1 (clayey/marly soil)	region 2 (calcareous/sandy soils)	region 3 (solid bedrock)
$c_{soil}$	0.1–0.5	0.1–0.3	0.3–0.6
$\gamma_{soil}$ [kg/m <sup>3</sup> ]	2100	2300	2500
$\varphi_{soil}$ [°]	10–25	15–35	25–45
$T/q$ [log]	4000–5000	500–1000	6000–7000

### 6.1.3 Implementation, Results and Discussion

This particular (pilot) case study has had experienced all of the proposed modeling methods (See Chapter 5), and herein, the implementation of particular methods will be given in appropriate order.

#### 6.1.3.1 Model-1a

The model was deployed by the AHP heuristics (see Chapter 5.2.1.2) by using the arbitrary scores in 1–9 scale in pair-wise comparison matrix. These relational scores have been obtained through the scoring interview with colleagues that were familiar with the problematic. They have been scoring relations of proposed pairs of Conditioning Factors. There were several different configurations of factors because different examinees preferred different factor pairs, thus excluding factors and pairs which they could not confidently relate. For instance some morphometric features have not been taken into consideration. One of the initial elaborations is considering the following Conditioning Factors  $F_i$ : *lithology* ( $F_9$ ), *slope angle* ( $F_2$ ), *rainfall*<sup>25</sup> ( $F_{13}$ ), *erosion (distance from stream –  $F_7$ )*, *vegetation cover* (reclassified NDVI –  $F_{12}$ ), *elevation* ( $F_1$ ) and *aspect* ( $F_3$ ) (Tab. 7).

Since there has been almost no contradiction, i.e.  $CR=0.04$  (<10%) the eigenvector values were derived from the comparison matrix (Tab. 10) and gave the final weights  $w_i$  of the corresponding factors  $F_i$  (Tab. 11). The final distribution of weights is given in the last column of Table 11., suggesting that *lithology* and *slope angle* have the strongest influence on the process (with more than 20% in total score), unlike the *elevation* and *aspect* which have less than 10% of influence.

<sup>25</sup> *Rainfall* is actually a triggering factor, which is in this particular case represented by the amount of the weekly precipitation in the most extreme circumstances, i.e. summer storm months (the average weekly precipitation for the most extreme week of August, the month with the highest storm rates, has been used, and the data covered 1975–2000 period, by averaging and interpolating the data from the Hydro-Meteorological Survey of Vojvodina in Novi Sad). Rainfall variation turns insignificant for this relatively small study area (which has been approved by the following data-driven models). Thus the factor has been used only in Model-1a and has been abandoned since.

**Table 10.** AHP comparison matrix of the first model variant.

$F_i$	<i>lithology</i>	<i>slope angle</i>	<i>rainfall</i>	<i>erosion</i>	<i>vegetation cover</i>	<i>elevation</i>	<i>aspect</i>
<i>lithology</i>	1.00	1.00	3.00	2.00	4.00	6.00	9.00
<i>slope angle</i>	1.00	1.00	3.00	2.00	3.00	5.00	8.00
<i>rainfall</i>	0.33	0.33	1.00	2.00	2.00	5.00	4.00
<i>erosion</i>	0.50	0.50	0.50	1.00	3.00	3.00	4.00
<i>vegetation cover</i>	0.25	0.33	0.50	0.33	1.00	2.00	3.00
<i>elevation</i>	0.17	0.20	0.20	0.33	0.50	1.00	3.00
<i>aspect</i>	0.11	0.13	0.25	0.25	0.33	0.33	1.00
$\Sigma =$	3.36	3.49	8.45	7.91	13.83	22.33	32.00

**Table 11.** AHP weight derivation of the first model variant.

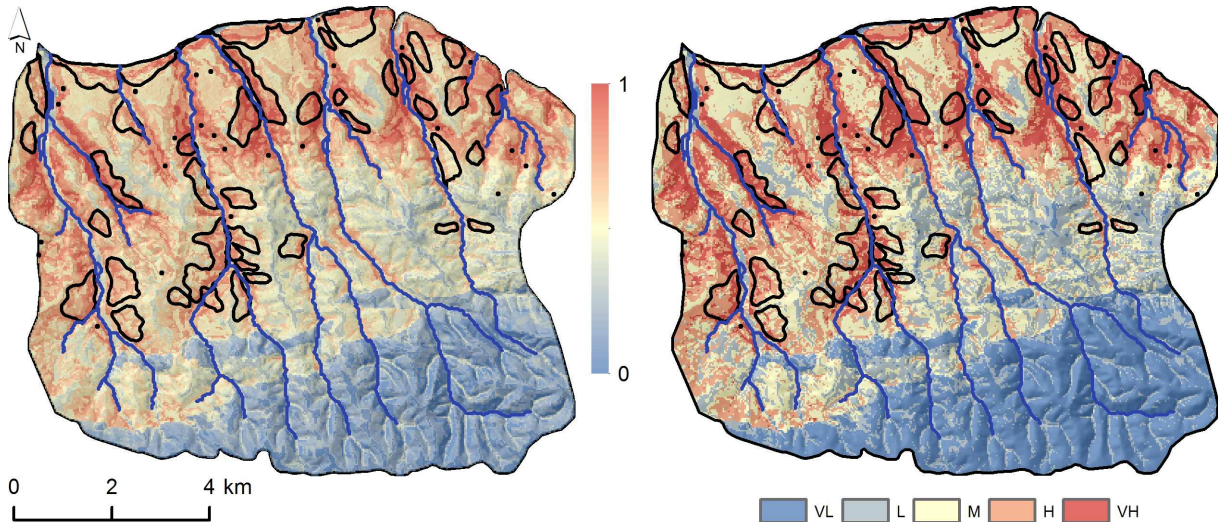
$F_i$	<i>lith.</i>	<i>sl.</i>	<i>rain.</i>	<i>er.</i>	<i>veg.</i>	<i>alt.</i>	<i>asp.</i>	$w_i$	%
<i>lith.</i>	0.29	0.29	0.35	0.25	0.29	0.27	0.29	0.29	29
<i>sl.</i>	0.29	0.29	0.35	0.25	0.22	0.22	0.26	0.27	27
<i>rain.</i>	0.10	0.09	0.12	0.25	0.14	0.22	0.15	0.15	15
<i>er.</i>	0.15	0.14	0.06	0.13	0.22	0.13	0.13	0.14	14
<i>veget.</i>	0.07	0.09	0.06	0.04	0.07	0.09	0.07	0.08	8
<i>alt.</i>	0.05	0.06	0.02	0.04	0.04	0.04	0.05	0.05	5
<i>asp.</i>	0.03	0.03	0.03	0.03	0.02	0.01	0.02	0.02	2
$\lambda_{max}=7.33$ ; $CI=0.05$ ; $RI=1.32$ ; $CR=0.04$							$\Sigma =$	1.00	100

Another variant of AHP model is worth mentioning, since it included some less common Conditioning Factors, thus including the entire input dataset: *elevation*, *slope angle*, *aspect*, *slope length*, *profile curvature*, *planar curvature*, *distance from stream*, *TWI*, *lithology*, *distance from structures*, *distance from hydrogeological boundary* and *vegetation cover*, i.e.  $F_{1-12}$  in the respective order. By using a consistent criteria, with  $CR=0.09$ , the final weights of the corresponding factors are obtained by the same procedures as described before (Tab. 12). The highest weights are once again assigned to *lithology* and *slope angle* (around 20%) and the lowest for *planar* and *profile curvatures* and *distance from structures* (equal or less than 2%).

The final AHP model (Model-1a) is defined as a combination of the two. By averaging common and adding unique Conditioning Factors a continual raster model has been generated and 0–1 normalized, to meet the quasi-probability range (Fig. 30). Visually, it could be inferred that the model has revealed some unstable zones, but tends to overestimate the probability of failure outside the existing landslides, while it underestimates within the existing landslides.

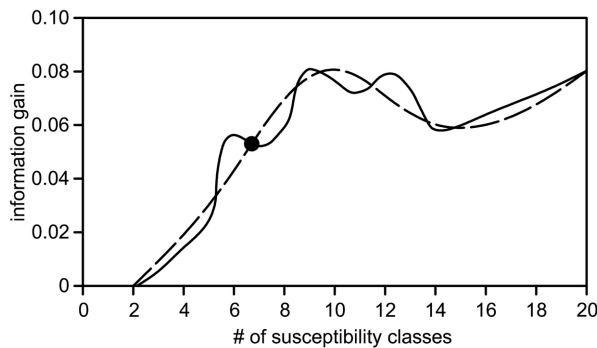
**Table 12.** Final AHP weights of the second model variant.

$F_i$	$F_1$	$F_2$	$F_3$	$F_4$	$F_5$	$F_6$	$F_7$	$F_8$	$F_9$	$F_{10}$	$F_{11}$	$F_{12}$
$w_i$ (%)	16.9	19.9	3.3	3.4	2.0	1.2	6.0	8.0	21.8	1.0	5.4	11.1



**Figure 30.** Model-1a (AHP model) in continuous quasi- $p$  scale (left) and in 5-leveled susceptibility scale (right). Black contours represent actual landslides. Scale bar shows 0–1 quasi-probability range.

However, the quantitative estimations of the result awakens several issues, primarily regarding which metrics fits best and more importantly, how to reclassify the model to make it optimal for comparison against the available Landslide Inventory (Fig. 30, 31).



**Figure 31.** Information Gain in function of number of classes (solid line) and its polynomial trend – entropy function (dashed line). Note that the highest  $IG$  and the steepest curvature of entropy function falls between classes 4–9, whereas the inflection point falls around class 6.

The latter has been approached by a conventional fashion of classifying the landslide susceptibility to 3–5 classes, i.e. low–high (Fell et al. 2008) as endorsed by the international communities. Still, the problem of actual reclassification remains. Several typical proposals could be made. Natural breaks (Jenkins breaks), which exploit the proportionality of data distribution per each class, could be a plausible option. The data could be thus reclassified to 3–5 desired classes and compared to the Landslide Inventory which also contains 3–5 classes. However, such redistribution is not entirely justified, because the Landslide Inventory classes (*active landslide*, *dormant landslide* and *non-landslide*) are not based on the proportional distribution. Some justification could be proposed, and it involves the entropy approach, by estimating *Information Gain*. The  $IG$  has been calculated for a wide range of classes (number of classes ranged from 1 to 255 classes with the Natural breaks intervals)

and it showed that 3–5 classes seem appropriate for expressing the landslide susceptibility (Fig. 31), since the inflection point of the IG function is close to 5.

This is generally unappreciated task, since it is very difficult to find the optimal metrics to express the model. For all spatially predictive models it is difficult to choose which metrics suppresses their predictability the least, while penalizing them for misclassification. Model-1a in particular had been penalized for overestimating, i.e. for exhibiting type 1 error (false alarms), which is acceptable in the landslide susceptibility framework. It might not be considered as a real error at all, because landslides might be absent at present, but could easily develop in time, hence confirming the model and its predictability. A brief quantitative report, regarding the adopted metrics (kappa and ROC measures) follows (Tab. 13). Some additional kappa statistics has been calculated for better understanding of class performance variations in relation to their population (Tab. 14).

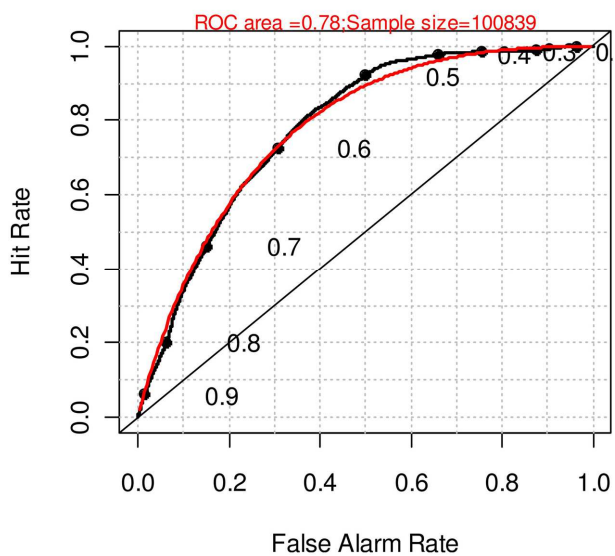
**Table 13.** Model-1a performance metrics.

Model	$\kappa$ -index	AUC	$fp_{rate-0}$	$\kappa_{fuzzy}$
Model 1a (AHP)	0.15	0.78	0.73	0.40

**Table 14.** Model-1a performance per class.

susceptibility class	Very Low	Low	Moderate	High	Very High
$\kappa_i$	0.34	0.51	0.28	0.32	0.35

The AUC of the AHP model (0.78) suggests a plausible result, which is not the case for its  $\kappa$ -index value (0.15) and also not directly evident from the map (Fig. 30). Thus ROC performance (Fig. 32) rather goes in its favor. The performance curve is moderately right-skewed, meaning that the model was not equally successful in discerning *landslides* from *non-landslides*. Such trend usually entails more liberal estimation than it suppose to. AHP tends to underestimate the landslides by claiming that there is low probability of occurrence on an actual landslide site (type 2 error), which is an important drawback of the model.



**Figure 32.** ROC curve of the Model-1a.

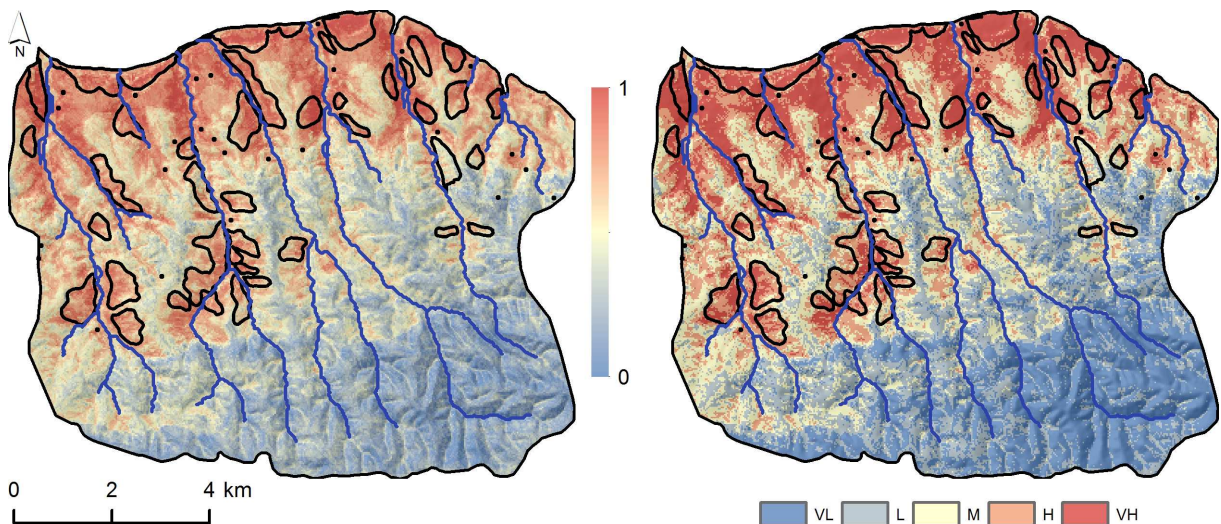


In conclusion, Model-1a is a plausible but not too confident way for landslide susceptibility mapping in the study area. The subjectivity of the model is substantial, and heavily depends on the expert's familiarity with the area. It is quite simple, appropriate for the GIS-based analysis on regional scales (independent AHP modules are already available in many GIS platforms), with low computational demands and time-consumption, but requires a significant amount of cooperation and advisement of other experts in the field. It could be convenient for a preliminary regional assessment and planning.

### 6.1.3.2 Model-2a

This conditional probability-based model has been deployed by direct Weights of Evidence technique, wherein positive, negative and zero weights have been considered. By excluding the negative and zero weights the procedure has been further simplified. The former is a measure which goes in favor of safety, since the stabilizing role of a factor is therein neglected, while exclusion of near-zero weight is a standard procedure (factor is indifferent to the dependent variable).

Conditioning Factors have herein been treated as independent variables, and the most of them (all of those with ordinal data type) have had to be adjusted to meet the requirements of the modeling technique<sup>26</sup>. It involved the reclassification of ordinal data (*elevation, slope angle, aspect, slope length, profile curvature, planar curvature, distance from stream, TWI, distance from structures and distance from hydrogeological boundary*) into appropriate ranges of intervals, based on Natural breaks method. This is where the subjectivity is being introduced to the model, because choosing the number of intervals and ranging method is arbitrary and depends on the practitioner. It has been particularly difficult to range and reclassify certain factors, such as *slope angle* and *aspect*, which have values that are not straightforward in relation to landslides. For instance, NE aspect is adjacent to SE, but has drastically higher susceptibility, while slopes higher than e.g. 30° rarely host landslides, but gentle slopes below 5° are likewise rarely hosting them. Alternatively, the model could have used a different unit area, for instance watershed units or some other topographic discrete parameter, and to compute zonal statistics instead of point/pixel-based statistics.



**Figure 33.** Model-2a (CP model) in continuous quasi- $p$  scale (left) and in 5-leveled susceptibility scale (right). Black contours represent actual landslides, scale bar shows 0–1 quasi-probability range.

<sup>26</sup> Nominal data such as *lithology* and *vegetation cover* did not need to be reclassified because their classes are unique and predefined by the nature of the factor itself.

**Table 15.** Model-2a performance metrics.

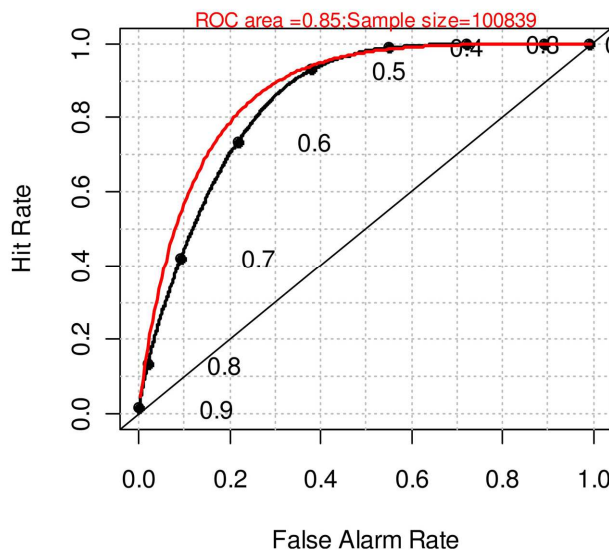
Model	$\kappa$ -index	AUC	$fp_{rate=0}^*$
Model-2a (CP)	0.36	0.85	0.78

\* $tp_{rate}$  of the non-landslide class

**Table 16.** Model-2a performance per class.

susceptibility class	Very Low	Low	Moderate	High	Very High
$\kappa_i$	0.48	0.69	0.33	0.57	0.62

The model (Fig. 33) has revealed similar, but somewhat better performance in comparison to the Model-1a. The susceptibility model has been also first evaluated by AUC metrics as a probabilistic one, and then reclassified in the same fashion and by the same criteria as in Model-1a. The latter involved kappa-statistics-based evaluation, including also the conditional  $\kappa_i$  per each of the five susceptibility classes (Very Low, Low, Moderate, High and Very High). According to the performance parameters (Tab. 15, 16), VL and L classes have been mapped better than the rest, but there is still slight overestimation in the VH class. On the other hand, probabilistic (continual) model has mapped higher classes quite well, but also slightly overestimates VH class. ROC curve (Fig. 34) shows similar trend as Model-1a (Fig. 32) with liberal estimations of landslides, which also contributes to the safety but entails conservative models. AUC is even higher (0.85) which does not surprise considering that the element of subjectivity is present, but very subtle and very useful for the fine tuning of the model.



**Figure 34.** ROC curve of Model-2a.

So far the 0–1 continual (probabilistic score) Model-2a has been discussed. When subjected to the same low–high susceptibility reclassification scheme as in Model-1a, the model also yields a visible improvement, where two most susceptible classes (H and VH) comply with the *active* and *dormant landslide* classes of the inventory. The  $\kappa$ -index of 0.36 is fairly improved in comparison to Model-1a, but still entails overestimation of landslides (high susceptibility classes).

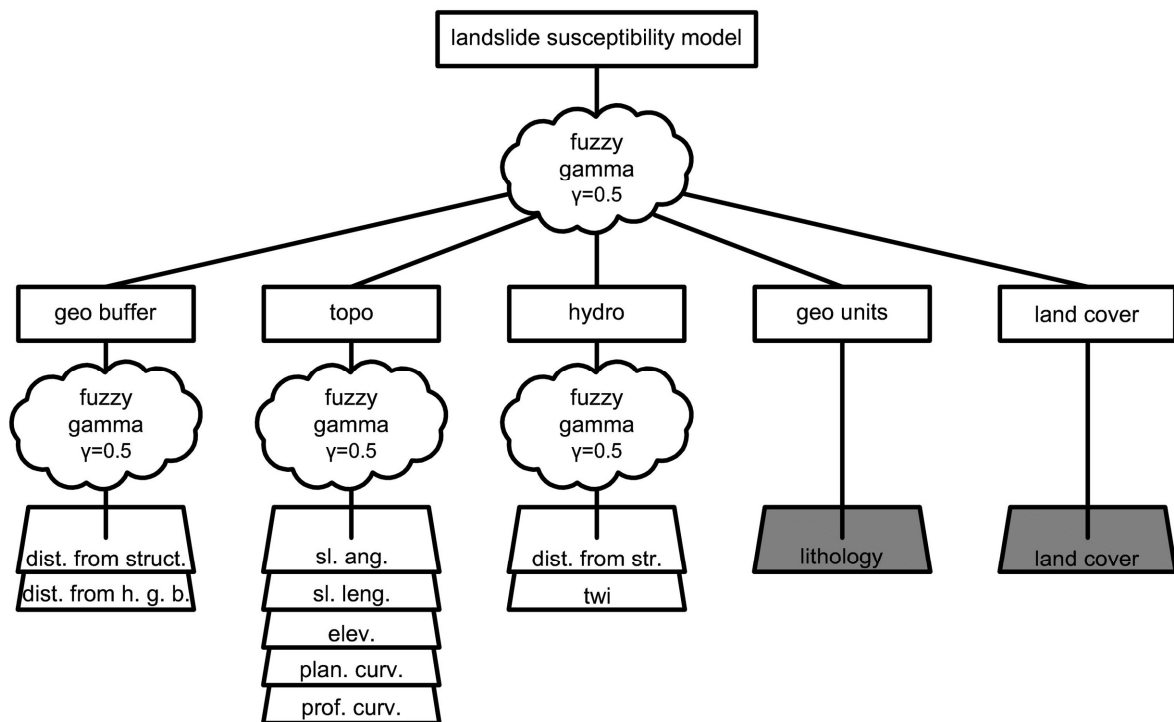


In conclusion, the model is producing relatively reliable prognosis, with slight overestimations, while still being reasonably simple and GIS-integrated. Evident performance improvement (Tab. 15, 16) indicates that the model is superior to Model-1a in both modes, probabilistic and reclassified.

### 6.1.3.3 Model-3a

Fuzzy model is a further step toward performance improvement, because it is somewhat similar to the statistical (objective) approach but allows fine tuning by introducing custom fuzzy combinations. In other words, it thrives on benefits of both, statistical and heuristic elements.

Model-3a suffers from the same limitation as Model-2a when it comes to data. It also requires ranging and reclassification of all continual ordinal Conditioning Factors as described for Model-2a (Appendix 1). Given the categorized (ranged) raster attributes and the referent Landslide Inventory map, the memberships of each class in each factor have been defined. Two parallel variants of the experiment were driven: Model-3a-CA used Cosine Amplitude, while Model-3a-FR used Frequency Ratio<sup>27</sup> to obtain the memberships. Both experiments had exactly the same course, thus the following manipulations took place in each.



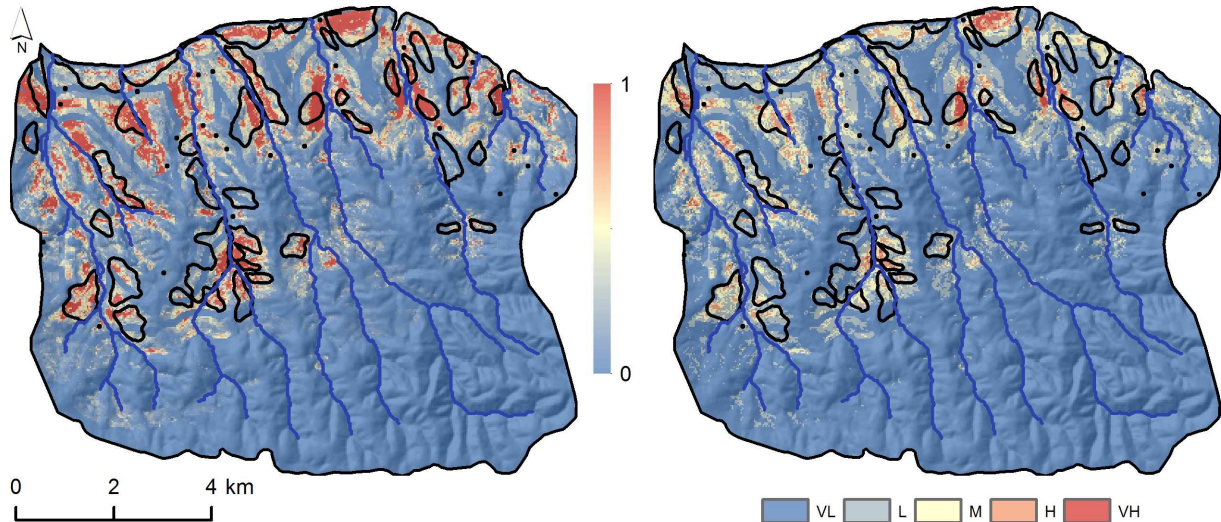
**Figure 35.** The structure of the fuzzy model. Nominal (categorical) data sources are shaded, and their level 1 models are the weighted reclassifications of themselves.

In order to combine memberships by different operators a small intervention has been undertaken to exclude too many extreme membership values (0 and 1) by replacing them with close approximations (0.0001 and 0.9999)<sup>28</sup>. Two-levelled fuzzy combination based on a

<sup>27</sup> Cosine Amplitude and Frequency Ratio are both Conditional Probability-based estimators, very similar to Weights of Evidence, except that they do not consider positive and negative weights, but the overall density/frequency of occurrences (landslides in this case).

<sup>28</sup> Extreme membership values are causing quick convergence of intermediate models in fuzzy combinations.

*priori* knowledge of the phenomena has been proposed (Fig. 35). In this way, the pairs of Conditioning Factors of similar origin were grouped together (e.g. *topo* sub-model included all derivatives of DEM). Continual susceptibility model with probability scores (0–1) was obtained in 2-level fuzzy combination (Fig. 36). Reclassified susceptibility model has been generated by ranging the continual values into five standard categories of relative susceptibility as described for Model-1,2a. Only the highest susceptibility class (Very High) was regarded for performance evaluation (*AUC*) against the referent Landslide Inventory (Tab. 17). This was inspired by the fact that existing landslides should be marked as a priority zone (preferably as a Very High susceptibility class).



**Figure 36.** Model-3a-FR-g (fuzzy model) in continuous quasi- $p$  scale (left) and in 5-leveled susceptibility scale (right). Black contours represent actual landslides, scale bar shows 0–1 quasi-probability range.

**Table 17.** Model-3a performance metrics.

Model	<i>AUC</i>	$tp_{rate}$
Model-3a-CA-wa* (weighted average)	0.65	0.37
Model-3a-CA-g* (gamma operator, $\gamma=0.5$ )	0.70	0.53
Model-3a-FR-wa** (weighted average)	0.71	0.56
Model-3a-FR-g** (gamma operator, $\gamma=0.5$ )	0.82	0.58

\*Cosine Amplitude memberships

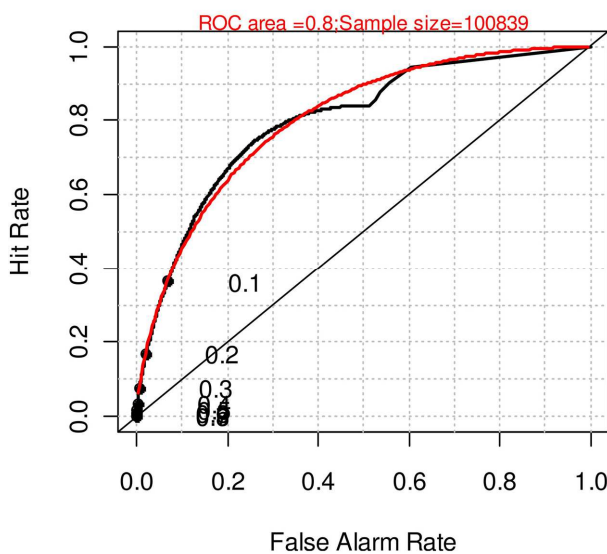
\*\*Frequency Ratio memberships

To remain consistent, the same type of the operator at both combination levels has been kept. Initial results in both experiments gave preference to Fuzzy Gamma Operator over the other, less subtle operators (see Chapter 5.2.2.2). Hence, the further model fitting has been directed toward the optimization of fuzzy parameter  $\gamma$ . Cases of  $\gamma=0$  (Fuzzy Product) and  $\gamma=1$  (Fuzzy Sum) were already regarded and dismissed as unsuitable, because of producing fast convergences and extremes. Several other choices within specified interval (0.25, 0.5 and 0.75) have been tested instead. It turned out that the best performance (*AUC*) was achieved by  $\gamma=0.5$ , making it a parameter of choice for the final susceptibility model. Finally, Model-3a-FR gave slightly better performance over Model-3a-CA, meaning that Frequency Ratio could be preferred over Cosine Amplitude for assigning memberships.

In the reclassified model of Model-3a-FR-g the distribution of relative susceptibility classes goes as follows: VL – 53%, L – 14%, M – 12%, H – 11%, and VH – 10%. Dominance of the VL class characterizes terrain as mostly stable, while similarly as in the referent inventory map, the most adverse zones occupy about 10% of the area. Furthermore, a

majority of the actual *landslide* instances fall into the VH and H classes (37% and 23% of all *landslides*, respectively), while M, L and VL classes occupy mostly *non-landslide* instances (75% of *non-landslide* instances in total for all three classes).

The highest overall performance of continual model in Model-3a-FR-g ( $AUC=0.80$ ) could be acknowledged and model accepted as plausible, which is also supported visually, since VH class corresponds very well with the spatial trends of the actual *landslides* (Fig. 36). Apparent influence of intermediate layer GeoBuffer caused several outliers by underestimating some *landslide* instances. A considerable drawback is relatively low  $tp_{rate}$  in both experiments (Tab. 17) which is inconvenient for any hazard-related analysis, since the model tends to underestimate actual *landslide* instances<sup>29</sup>. However, the actual performance is somewhat better, since only VH class has been regarded for cross-tabulation. Thus, H or even M class could be fair replacements for VH class, as they buffer-out around it, which might reduce the number of *False Negatives* if included in cross-tabulation (increasing the  $tp_{rate}$ ).



**Figure 37.** ROC curve of Model-3a-FR-g.

The continual Model-a3-FR-g has a ROC curve area of 0.8 and it is left-skewed with a sub-maximum approaching the peak maximum value, but more importantly its relative probability thresholds range is concentrated around the lower left corner of the plot (conservative model, strict for *False Negative* errors), which is an important preference for the model, even though it is learned by a pure qualitative description of the curve (Fig. 37). It is yet another fact that goes in favor of the model over the preceding ones.

In comparison to Model-1a and Model-2a (while still regarding only the VH class for comparison) fuzzy approach turned practically as successful as statistical model (Model-2a based on CP), but with more subjectivity involved in the modeling procedure (in ranging the input intervals, but also in selecting the operators and numbers of combination levels). It outperformed AHP model, not as much in the overall performance ( $AUC$ ) as in considerably higher  $tp_{rate}$ , granting a slight preference to the model.

Subjectivity in ranging input attributes was inevitable, due to incapability of the approach to handle the continual numerical variables (in the stage of assigning memberships). Another subjective intervention regarded proposing of the number of levels for fuzzy combination, and grouping the attributes with similar origin to level 1. Further

<sup>29</sup> Actually, the *False Negative* represents the most dangerous type of error, because it claims that there is no *landslide* while the *landslide* actually exists.

refinement of the assessment, leads to more elaborated techniques which surpass this issue easily, yet with considerable computational effort and time consumption. To conclude, fuzzy approach came up with a suitable model, while the modeling procedure remained simple, semi-automated and re-operable in a GIS environment. The resulting map could suit preliminary levels of risk or disaster management, landscape (regional) planning, route selection, insurance management and so forth.

### 6.1.3.4 Model-4a

From this model on, the character of the landslide assessment in the Fruška Gora case study becomes more predictive, as more sophisticated methods are being introduced. These allow the model to be built on a smaller portion of the area and extended to the rest of it (training/testing concept of Machine Learning techniques), while the input data remain more-or-less the same.

However it is necessary to mention that the ordinal Conditioning Factors did not have had to be retooled, only normalized, while ordinal data have had to be segregated (binarized) to multiple binary inputs<sup>30</sup>. Described protocols applied not only for Model-4a, but also for Model-5a, Model-6a and Model-7a. For all these models also applies the fact that the performance has been computed directly, because these classifiers, unlike the preceding models, produce a discrete output which is already classified into target classes. In this case the Landslide Inventory with three classes (*active*, *dormant*, *non-landslide*) has been used as susceptibility reference (H, M and L, respectively).

Model-4a is actually a *k*-Nearest Neighbor (*k*-NN) based model. This is an unsupervised, actually semi-supervised classifier, primarily intended for clustering and pattern recognition. Otherwise, it is computationally extremely demanding as explained before (see Chapter 5.2.3.2), which is the reason for several reductions in the *k*-NN experiments.

Firstly, the usage of the entire dataset would not be too meaningful in this case. Since *k*-NN is capable of recognizing and clustering spatially induced similarities throughout a given spatial domain (2D geographical space of a map), the most of the additional attributes could be discarded. Practically, the only information which is required is the spatial location (geographical coordinates) and landslide class label (from the referent Landslide Inventory) of the training instances. Secondly, the training area has had to be much smaller than usual, and the sampling strategy has had to be based on random but uniform dispersion of the instances. Only two sampling sizes have been considered, containing 1% and 5% of the area, respectively<sup>31</sup>. Reducing the training size is however one of the milestones of Machine Learning in general, but finding an optimal training sample size requires less computationally limiting environment. Even with such restrictions, the computational effort was significant. As a final reduction step, the original algorithm has been supplemented with a more advanced sorting function<sup>32</sup>, which partly suppressed the time consumption.

Further model optimization was limited for the same reason, which is why only several alternatives of the number of NNs have taken place. After 5-fold Cross-Validation in the training area, the optimal results have been achieved when the number of neighbors equaled 3, although further rising of neighbors resulted in similar but slightly lower performance. It could be speculated that such balance indicates the consistency of the

<sup>30</sup> In addition, since the processing did not take place in a GIS environment, the data have had to be filtered for no-data values, and the spatial reference have had to be temporarily removed.

<sup>31</sup> 1% equaled about 400 instances for training, and nearly 2000 in the 5% case.

<sup>32</sup> Sorting here included ordering of the neighbors from the closest to the furthest ones. Faster sorting resulted in a quicker finding of the 3 nearest neighbors and thus eased the calculation of their weights, making the classification also quicker.

chosen model (3-NN). The accuracy of the model in the bigger training sample equals just around 57%, which is even lesser than in the smaller (1%) sample (Tab. 18).

Theoretically speaking, the spatial pattern recognition which is provided by the  $k$ -NN algorithm could have led to a significant accuracy, but it is also apparent that sampling strategy plays a crucial role. The method explores the closest environment of an instance, which does not necessarily mean that it will yield results for any spatially correlated phenomenon, not just landslide distribution. For these reasons the expectations were not too optimistic, which has been proven right. The model rather served as a demonstration of the sampling strategy and its effects on the modeling, which have proven valuable for some later Machine Learning implementations.

Model 4a has proven that even sparse inputs (small portion of training data and reduced dataset) could lead to some prediction. However, its robustness which have caused numerous limitations have caused abandoning of this approach, with no further examination of its performance, nor its visualization, and turning to more efficient techniques.

**Table 18.** Model-4a performance metrics.

<b>Model</b>	<b>accuracy (%)</b>
3-NN (training sample=1%)	57.5
3-NN (training sample=5%)	56.5

### 6.1.3.5 Model-5a

Implementation of a Decision Tree algorithm C4.5 had taken place in this model. For this particular case study, the experimenting was not as detailed as in the next case study for instance. The experiment design involved balanced and unbalanced training datasets<sup>33</sup>, and it also included multiple classes of landslides for classification. Thus both, the predictive power of the model and its capability of discerning among different landslide types have been challenged.

As mentioned before, the experiment has not been too detailed, which has also reflected the optimization of the modeling parameters. Practically, all default settings have been adopted from the offered J48 (Weka software) implementation of C4.5 algorithm. *Confidence Level* of C4.5 has been set to 0.25 and adopted, without introducing alternative values. *Minimum Number of Objects in Leaf* has been also accepted from the default setting and has been set to 2. This is obviously the softest choice, since it is natural that at least two alternatives exist when conditioning instances of an attribute (Conditioning Factor).

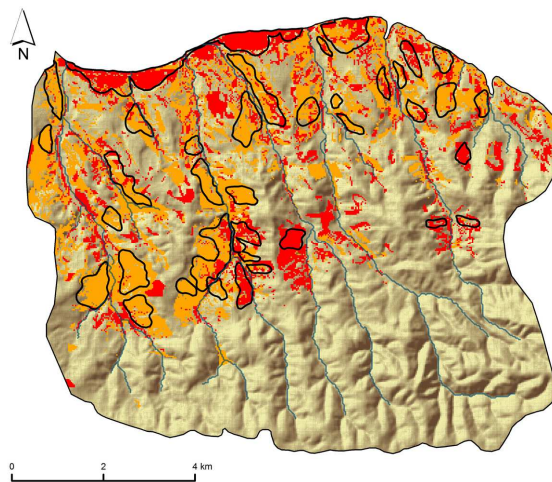
Three variants of the experiment have been designed, making actually six different results if one has in mind that each of the three has been run twice, once for the balanced and once for the unbalanced training dataset. It is important to mention that the training area has been selected by sampling instances randomly and uniformly throughout the area, which means that the predictive modeling has not been the central objective of the Model-5a. Instead, the principal objective was to handle the multi-class scenario and to challenge the ability of recognizing different classes of landslides correctly. In order to avoid the overfit, a 10-fold Cross-Validation was chosen in the training/testing procedure. The only difference

<sup>33</sup> Balanced training set should contain an approximately equal proportion of all target classes, in this case *non-landslide*, *dormant landslide* and *active landslide*. However, it is very common in landslide assessment that the *non-landslide* class dominates over the other(s) especially if the training areas are outlined as continuous spatial entities. An example of an unbalanced set for this area would be 4% of *active*, 6% of *dormant landslides* and 90% of *non-landslide* instances within the training area, while in the balanced set each category would occupy one third of the training area.

throughout the experimenting protocol was a successive decrease of the training sample size, i.e. 15% of the total area, 10% of the total area and 5% of the total area (6156, 12313 and 18496 of instances out of 123134 total instances, respectively). Accordingly, the model variants could be denoted as Model-5a-U/B-5–15%, where U/B stands for unbalanced/balanced variant of the experiment and the percentage in the end designates the training sample size.

The results were encouraging, as they reached decent performance in all performance measures (Tab. 19, 20). As expected, better performance has been achieved in the balanced experiments, especially regarding the  $fp_{rate}$ . The total variation of performance is minute in all models (except for the  $fp_{rate}$ ), which might be an indicator of the models' consistency and stability. *False Positives*, i.e.  $fn_{rate}$  for *non-landslide* class are kept in relatively low levels, particularly in the balanced variants of the model (which is their additional quality). The model also successfully discerns between all three given classes (*active landslide*, *dormant landslide* and *non-landslide*), suggesting that multi-class environment does not confuse the algorithm, even after the pruning of the tree. However, it is apparent from both, evaluation metrics and visual representation of the model (Fig. 38) that there is a tendency to overestimation of both landslide classes, but this finding goes in favor of the model in the landslide assessment framework. Obviously, the modeling performance rises with the rise of the training sample size. The model thus approves of the sampling strategy. On the other hand, such strategy is not practical for predictive modeling, because the real situations would require continuous training domains rather than scattered.

In conclusion, the model seems to be very successful in predicting *landslide* instances (with some overestimation), and for discerning among two different landslide types. The best representative of all model variants would be Model-5a-B-10%, which has the performance parameters close to the average of all models. In addition, the practical implementation has not been too time-consuming, as Machine Learning algorithms usually turn to be, but it is important to remind that there has not been any optimization prior to the modeling. The model is a superior to the all preceding models thus far.



**Figure 38.** Model-5a-B-10%. Orange areas represent *dormant landslides* and equal moderate susceptibility ( $p=0.5$ ), while red areas represent *active landslides* and equal high susceptibility ( $p=1$ ). Contours represent actual landslides.



**Table 19.** Performance of Machine Learning models (unbalanced training set).

Model	$\kappa$ -index	AUC	$fp_{rate-0}$
Model-7a-U-5%	0.57	0.79	0.4
Model-6a-U-5%	0.08	0.86	0.94
Model-5a-U-5%	0.42	0.76	0.56
Model-7a-U-10%	0.67	0.82	0.32
Model-6a-U-10%	0.08	0.86	0.94
Model-5a-U-10%	0.42	0.76	0.56
Model-7a-U-15%	0.73	0.84	0.32
Model-6a-U-15%	0.06	0.86	0.96
Model-5a-U-15%	0.54	0.80	0.43

**Table 20.** Performance of Machine Learning models (balanced training set).

Model	$\kappa$ -index	AUC	$fp_{rate-0}$	$K_{fuzzy}$
Model-7a-B-5%	0.38	0.85	0.11	-
Model-6a-B-5%	0.25	0.85	0.23	-
Model-5a-B-5%	0.28	0.82	0.18	-
Model-7a-B-10%	0.42	0.89	0.08	0.43
Model-6a-B-10%	0.25	0.85	0.24	0.36
Model-5a-B-10%	0.32	0.82	0.17	0.16
Model-7a-B-15%	0.43	0.90	0.06	-
Model-6a-B-15%	0.23	0.86	0.20	-
Model-5a-B-15%	0.34	0.85	0.16	-

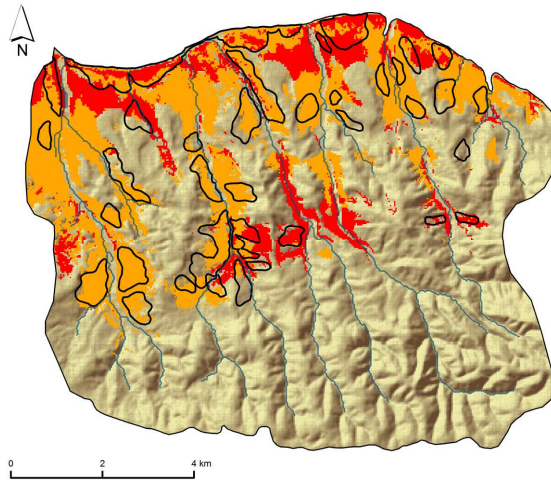
average adopted models are shaded

### 6.1.3.6 Model-6a

Herein, the implementation of Logistic Regression model has been performed. It is sufficient to say that the experimenting procedure has been exactly the same as in the previous model, so that no further details are necessary to explain the experimenting procedure.

The model is obviously troubled with overestimation (Fig. 39). Nearly one third of the total area is classified either as *active* or *dormant landslide*, which goes in favor of stability but does not serve the model's actual purpose, since the model is intended to be used by experts and planners as well as decision makers, who would like to have more room for maneuvering, i.e. less restrictions caused by landslide prognosis. The figures also support this view (Tab. 19, 20) where low  $\kappa$ -index (as low as 0.06 for the unbalanced variants of the model) in combination with very high  $fp_{rate}$  (up to 0.96) indicate that there is a lot of *False Positive* type of error, i.e. overestimation. This particularly affects the unbalanced variant of the model, while the balanced variants have considerably lower error rates, but still low  $\kappa$ -index. The AUC of their ROC curves are in the rank of the preceding models. Growing of the training set size has little or no effect on this model, at least in the selected 5-10-15% range. The model's ability to successfully distinguish between two given types of landslides has been obscured by the overestimating character of the prediction.

For all these reasons the model could be characterized as plausible, but it did not meet the expectations or top the performance of Model-6a. The model has proven to be very sensitive to the sampling strategy regarding the class balance in the training sample, and therefore inconsistent and unstable in this configuration with relatively small training set.



**Figure 39.** Model-6a-B-10%. Orange areas represent *dormant landslides* and equal moderate susceptibility ( $p=0.5$ ), while red areas represent *active landslides* and equal high susceptibility ( $p=1$ ). Contours represent actual landslides.

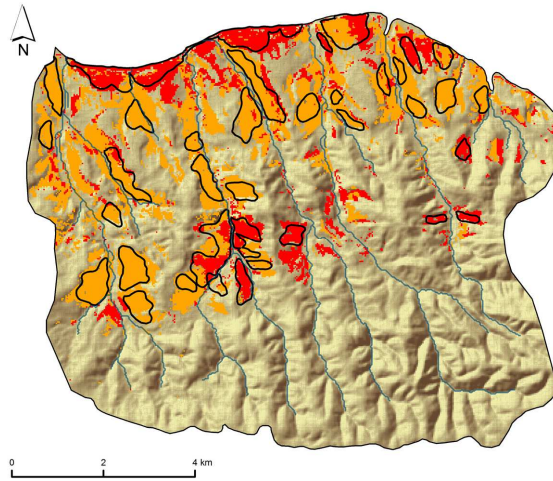
### 6.1.3.7 Model-7a

This model implemented SVM algorithm and has initially given promising results, which is why it has been elaborated in greater detail with different experimenting procedures. It is also the reason why this SVM technique has been favored for the other case studies.

There were basically two different experimenting settings in respect of the spatial distribution of training and test data. The first included a random training subset uniformly distributed over the whole case study area, and it has been further separated into unbalanced and balanced variant. The second involved adjacent train-test splits selected manually. As in the most of the preceding models, a multinomial landslide classification also applied. Hence, the model has been challenged for its predictability of landslides and the ability to sub-classify different landslide types, but from a more practical aspect. Since random uniform sampling of training instances is more targeted at satisfying the theoretical conditions for designing the best sampling strategy, the second type of experiment has been devoted to the practical aspects of having a genuine training sample as it is (with no sampling strategy alternatives) and predicting the landslides to adjacent area. Such scenario is but the simulation of the actual situation, where one part of the terrain might have the Landslide Inventory coverage, while the other, possibly adjacent part seeks one.

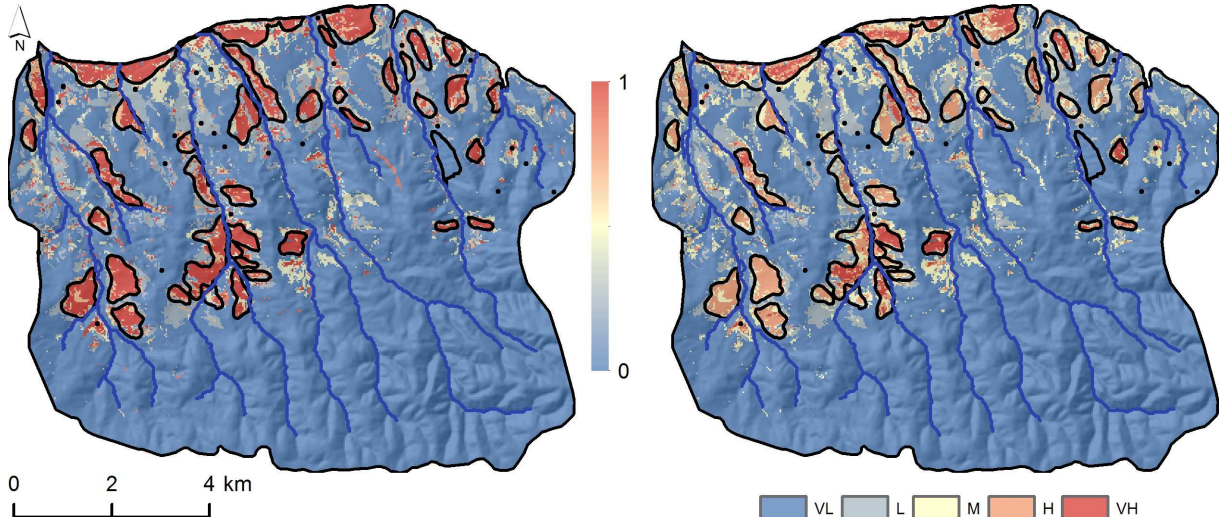
Regarding the first experiment type, i.e. the two subtypes involving unbalanced and balanced training sets, three different models were run. In each, a random training subset had different training sample size, i.e. 5% (6156 instances), 10% (12313 instances) and 15% (18496 instances) of the whole case study area available for training. Following the notation from the preceding models, these two variants could be denoted as Model-7a-U/B-5–15%. Accordingly, the second experiment type could wear a label Model-7a-33%.





**Figure 40.** Model-7a-B-10%. Orange areas represent *dormant landslides* and equal moderate susceptibility ( $p=0.5$ ), while red areas represent *active landslides* and equal high susceptibility ( $p=1$ ). Contours represent actual landslides.

In the unbalanced variants, the sampling instances were selected randomly, but equally spread over the area, and they contained a referent class proportion similar to the original landslide model (3.6% of *active landslides*, 5.6% of *dormant landslides*, and 90.8% of *non-landslide*). In the balanced variants (Fig. 40, 41), the sampling was the same, but the shares of all classes were equal (one third of all training instances). In fact, each training set for Model-7a-B-5–15% was obtained from the corresponding Model-7a-U -5–15% set by retaining all *landslide* points and selecting an equal number of stable points that are spatially uniformly distributed over the whole area.



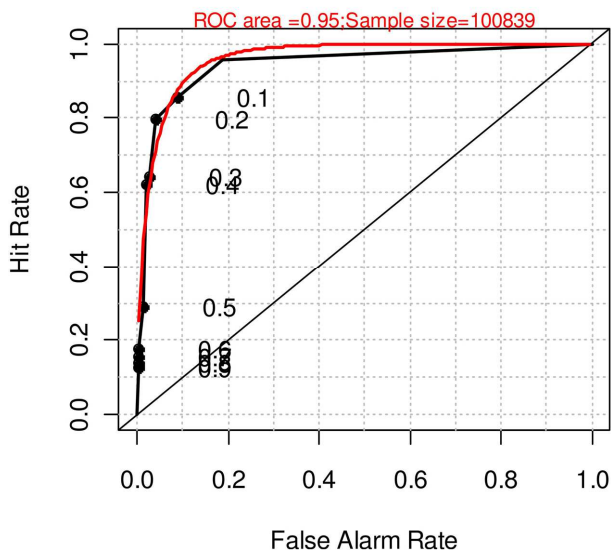
**Figure 41.** Model-7a-B-10% in continuous quasi- $p$  scale (left) and in 5-leveled susceptibility scale (right). Black contours represent actual landslides, scale bar shows 0–1 quasi-probability range. Note that the continuous scale variants are not predictive, but interpretative models of landslide susceptibility, obtained by averaging iteratively generated interpretative models of Model-7a-B/U type.

In order to illustrate experimenting procedure, the course of the Model-7a-U-5% experiment will be described, while the remaining experiments have completely analogous procedures. The negative effects of randomization (poor estimate of the model variance)

were minimized by creating 20 different random splits<sup>34</sup>, each containing 5% data points for training and 95% for the testing set. It is obvious that these splits are spatially overlapping to some degree, but full separation would not be feasible since some values of some Conditioning Factors would appear only in the test set.

Since there is a lack of information in most related papers concerning the adjustment of SVM learning parameters (Brenning 2005, Yilmaz 2009), a procedure of their estimation was conducted. The parameter estimation focused on the penalty factor  $c$  and Gaussian kernel width  $\gamma$ . Parameters  $c$  and  $\gamma$  took many different values ( $c = \{1, 10, 100, 1000, 10000\}$  and  $\gamma = \{0.1, 0.5, 1, 2, 4\}$ ). For each combination of  $(c, \gamma)$  a 10-fold Cross-Validation procedure was performed on a particular training set and the evaluation measures ( $\kappa$ -index and  $AUC$ ) were averaged and recorded for each combination of  $(c, \gamma)$ . The results showed that in all 20 splits of the experiment, optimal  $c$  and  $\gamma$  were the same (100 and 4, respectively). These values turned optimal for the Model-7a-U-10% and Model-7a-U-15%, as well.

Given the optimal parameters, the SVM classifier was trained over the entire particular training set and tested over the related test set. The evaluation measures included  $\kappa$ -index,  $AUC$  and the *False Positives for non-landslide class* ( $fp_{rate=0}$ ). The above procedure was repeated 20 times for each train-test split and the final measures for the experiment were averaged over all 20 splits. The same evaluation procedure was applied as in Model-5a and Model-6a, so that they could be compared (Tab. 19, 20.).



**Figure 42.** ROC curve of Model-7a-B-10%.

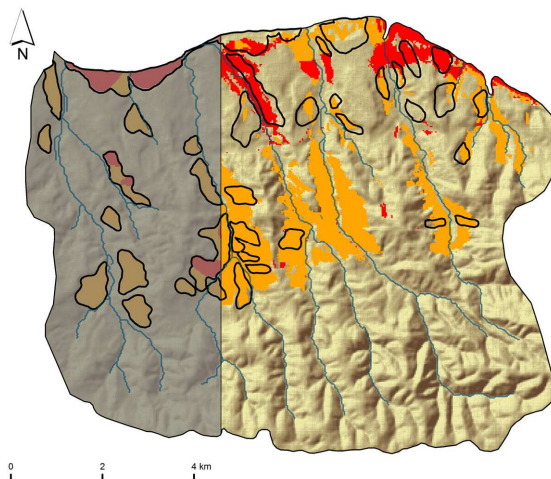
When the number of training points increases, the Model-7a and Model-5a improve their performance on the  $\kappa$ -index and  $AUC$ , while the Model-6a performance remains the same, as noticed before. Concerning the  $\kappa$ -index, the Model-7a significantly outperformed other models, followed by the Model-5a and Model-6a in respective order (because of very low  $\kappa$ -index Model-6a is put behind despite better  $AUC$ ). The Model-6a yielded the best  $AUC$  (0.86) when compared to the other two methods, but the Model-7a came close after increasing the size of the training set (0.84 nearly equals 0.86 of Model-6a). In addition to the previous explanations of the Model-6a results, such a big discrepancy between the  $\kappa$ -index (even in  $\kappa_{fuzzy}$ ) and the  $AUC$  can be explained by presence of multiple landslide classes and difficulties to weight the average class probability (Fawcett 2006), while for discrete

<sup>34</sup> Instead of 20, Model-7a-U-10% had 10, while Model-7a-U-15% had 7 iterations, in order to achieve a statistically meaningful procedure.

classifiers such as those of Model-5a and Model-7a the problem comes down to the simple averaging of individual binary cases (3 cases, one per each class).

As indicated before, the second type of the experiment was a practical simulation of the landslide prediction, where approximately one third of the area has been used for training and it was spatially separated as a meaningful entity from the rest of the area. The sampling has been done manually, with particular precaution to include all classes of the nominal Conditioning Factors (all binarized classes), Lithology ( $F_9$ ) in particular. In turn, a meridian separator, splitting the map to the west and east part, has given a fixed and continuous training and testing area (Fig. 9, 43). It consisted of approximately 41000 of instances, making it roughly one third of 123134 instances in total. Regarding the landslide distribution, it resembled the proportions of the landslide distribution over the entire area, meaning that it could be characterized as unbalanced training set. However, there has been no intervention to balance the set, because it is intended that simulation turns as realistic as possible.

After the same experimenting procedure as in the other models, Model-7a-33% gave relatively good results. Spatial trends of landslide bodies are being followed, while differences in landslide typology have been successfully modeled (Fig. 43). Overestimation of *landslide* instances is also apparent but there is not too much dispersion (spatial trends of the landslide bodies are well traced), so potential post-processing (e.g. majority filtering) would not give substantial improvements (up to a couple of percentages). Consequently, the  $\kappa$ -index values are low, while  $AUC$  of 0.71 is not as high as in Model-U/B-10–15% but still represents a fair prediction. *False Positive* rate is supporting this stand, but  $\kappa_{fuzzy}$  indicates that there is a considerable dissimilarity between the model and the inventory (Tab. 21).



**Figure 43.** Model-7a-33%. Orange areas represent *dormant landslides* and equal moderate susceptibility ( $p=0.5$ ), while red areas represent *active landslides* and equal high susceptibility ( $p=1$ ), training area is shaded. Contours represent actual landslides.

**Table 21.** Model-7a-33% performance metrics.

Model	$\kappa$ -index	$AUC$	$fp_{rate-0}$	$\kappa_{fuzzy}$
Model-7a-33%	0.17	0.71	0.39	0.03

Finally, it is convenient to comment on the strengths and weaknesses of the SVM model family, Model-7a. SVM does not need any feature selection technique as opposed to some other methods such as Decision Trees. This fact enables richer data representation, bigger number of inputs, and bigger variety of inputs. This aspect of the research is left to the future work. In addition, since the solution for the SVM separating hyper-plane is found from the convex quadratic programming optimization problem, it is guaranteed that the solution is

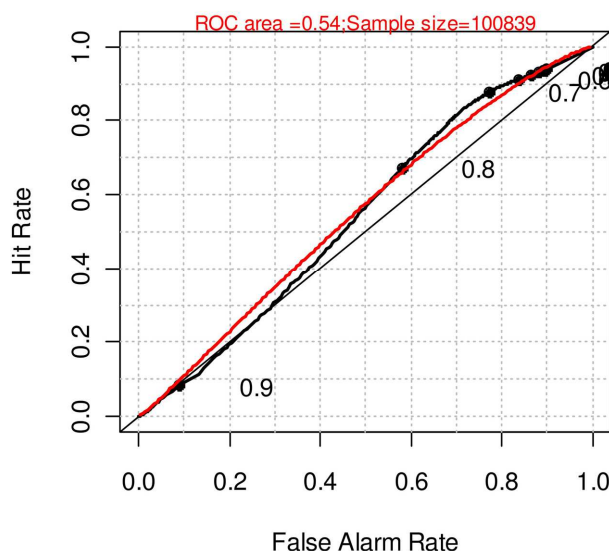
globally<sup>19</sup> optimal. Therefore, the SVM is a good replacement for Artificial Neural Networks which are usually stuck at local optima and are very difficult to train. On the other hand, the SVM does not output an interpretable model like Decision Trees, and hence could not be rewritten in the form of expert rules, which usually limits their application to one case study at a time, or retooling the input dataset if the study areas are extremely similar. When compared to Logistic Regression, they are much more memory and time consuming during the training phase, but that is probably not of great importance for the task of predictive landslide mapping, outside the disaster management framework (where very quick but plausible solutions are needed).

A small web-map created in the R environment, using “plotGoogleMap package” (see Appendix 2) presenting the area and related result of Model-7c-40% is available at:

<http://milosmarjanovic.pbworks.com/w/file/fetch/63738284/MyMapFruskaGora.htm>.

### 6.1.3.8 Model-8a

Implementation of a simplified deterministic model has not placed the performance expectations too high, for at least one reason. The case study area is actually dominated by deep-seated and larger landslides, while the simplified LEM stability model theoretically applies only for the smaller and shallower slides. The model has thus been only demonstrated in a GIS environment, and unfortunately the results have met the expectations.

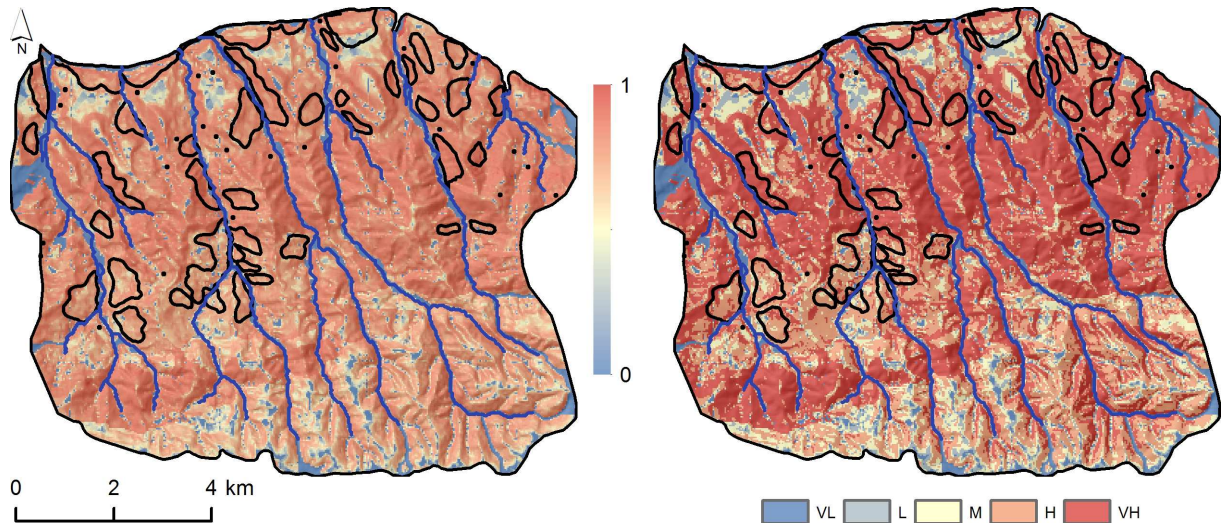


**Figure 44.** ROC curve of Model-8a.

Model-8a has been fed with the topographic data and geotechnical parameters as indicated before (see Chapter 6.1.2.7). The implementation in a GIS environment has been direct and relatively quick. There have been a lot of conveniences for customizing this stochastic model. For instance, each of the geotechnical parameters could have been regionalized, and both, upper and lower boundaries have been required. The sensitivity of the model has also been regarded (by slightly varying the inputs in a series of experiments, and calculating  $S/I$  per each) so that the average inversed  $S/I$  was generated using the average parameter intervals. Despite of laborious model fitting that has taken place, Model-8a has not managed to perform beyond the level of random guess ( $AUC \approx 0.5$ ), while the other performance indicators were nonetheless discouraging (Fig. 44).

Visually (Fig. 45), the model overestimates the landslides over the entire area, making it impossible to speculate whether some trends have been caught or not. The model could therefore be entirely discarded as unsuitable for a given study area.





**Figure 45.** Model-8a in continuous quasi- $p$  scale (left) and in 5-leveled susceptibility scale (right). Black contours represent actual landslides, scale bar shows 0–1 quasi-probability range.

## 6.2 Starča Basin (Croatia)

Regarding this study area, previous investigations of the landslide-related issues have included various engineering-geological endeavors, predominantly medium-scaled mapping, or large scale investigations related to the individual practice, reports for specific construction projects and so forth, but unlike in the previous case study there have been valuable exceptions.

Firstly, there have been some more detailed engineering-geological mapping projects entailing several sheets of synthetic engineering-geological maps at 1:10000 scale for Zagreb City and its surroundings, with a particular focus on mapping various landslide types as defined in the international classification WP/WLI 1995. By the courtesy of the National Institute for Geology and Geological Engineering of Croatia, the above mentioned repository has been at disposal as a principal data resource in preparing the second case study of this thesis. It was thus one of the main challenges of this case study to observe how proposed methodology behaves with the finer-scaled data to exploit the upscale of the input data, wherein the most of the data came with mentioned 1:10000 scale, while some have been collected at even finer scale (1:5000).

Finally, one research published by Mihalić et al. 2008 should be particularly outlined, since it mostly complies with the concept followed in this thesis. The authors approached the landslide assessment by implementing bivariate statistical techniques over unique-condition-defined unit areas and have retrieved a good correlation with the existing landslide records and contributed the understanding of landsliding process in the Starča Basin (Mihalić et al. 2008). However, they have proposed further enhancements, especially regarding the modeling techniques, and one part of this research has been directed toward that common goal (Marjanović et al. 2011b).

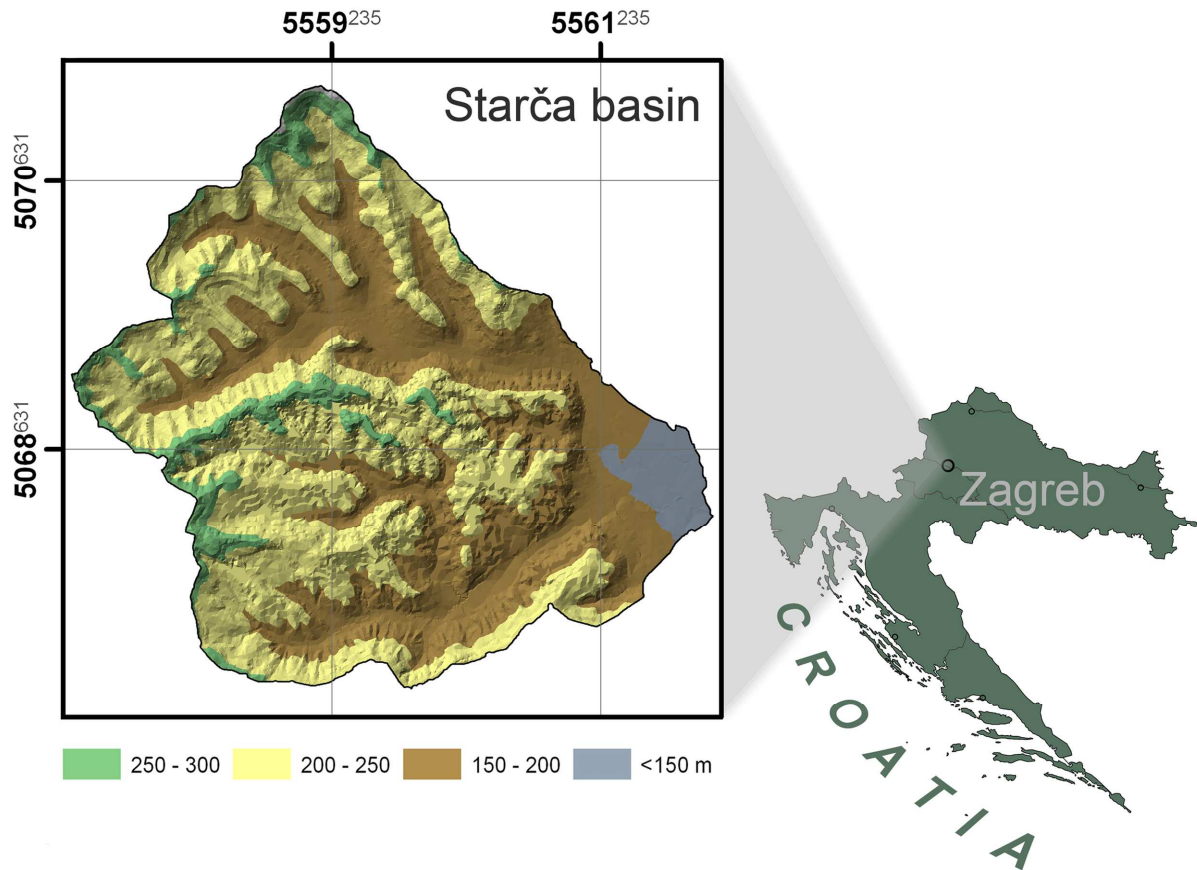
### 6.2.1 Setting

Starča Basin encompasses 12.25 km<sup>2</sup> of a hilly landscape (up to 300 m in elevation) stretching between N 45°44'35", E 15°44'15" – N 45°47'05", E 15°47'55", flattening in from the east and giving way to the outskirts of the Samobor Mountains, which represent the western border of the City of Zagreb, Croatia (Fig. 46).

Geotectonically, the basin is set on a verge of several major regional units, where Pannonian, Alpine and Dinaric sub-units meet. Their spatial relations and boundaries are not entirely defined, but it is most likely that the area is primarily shaped under the influence of Pannonian basin development, i.e. Transdanubian unit and its sub-structures, Zagreb-Cemplin lineament in particular (Dimitrijević 1997). Its NE–SW strike has been replicated in trends of the local structures, primarily normal faults with SE vergence, i.e. relatively subsided hanging walls toward SE. This gravitational faulting occurs along the edge of Transdanubian unit and logically follows the graben of the Sava river valley, which dominates the landscape. These older structures are traversed by younger, sub-vertical structures of meridian strike. However, a large part of the area is covered with Quaternary sediments, leaving the structures concealed, except in the higher grounds.

The basin is composed of the Upper Miocene and Plio-Quaternary clastics (Fig. 48f, 49). The landslides are mainly hosted in Pannonian marl and silt in the northern part, as well as in Plio-Pleistocene coarse-grained sand. They are moderately compressible for the upper Miocene units, but very sensitive to the groundwater variation (especially sand). Inconveniently, due to the locally shallow water tables, the units are often saturated from half a meter down from the ground level (Fig. 48e). These two units combined cover nearly 40% of the area, making a suitable ground for development of slope instabilities, which are particularly indicative on the flanks of gullies/valleys, even though these are sloping relatively gently and their local erosion basis never exceeds more than 150 m of elevation difference (maximal altitude is about 300 m and minimal around 150 m). Typical landslides are shallow, hosted in marl and clay mantle which can locally become considerably thick, with enforced

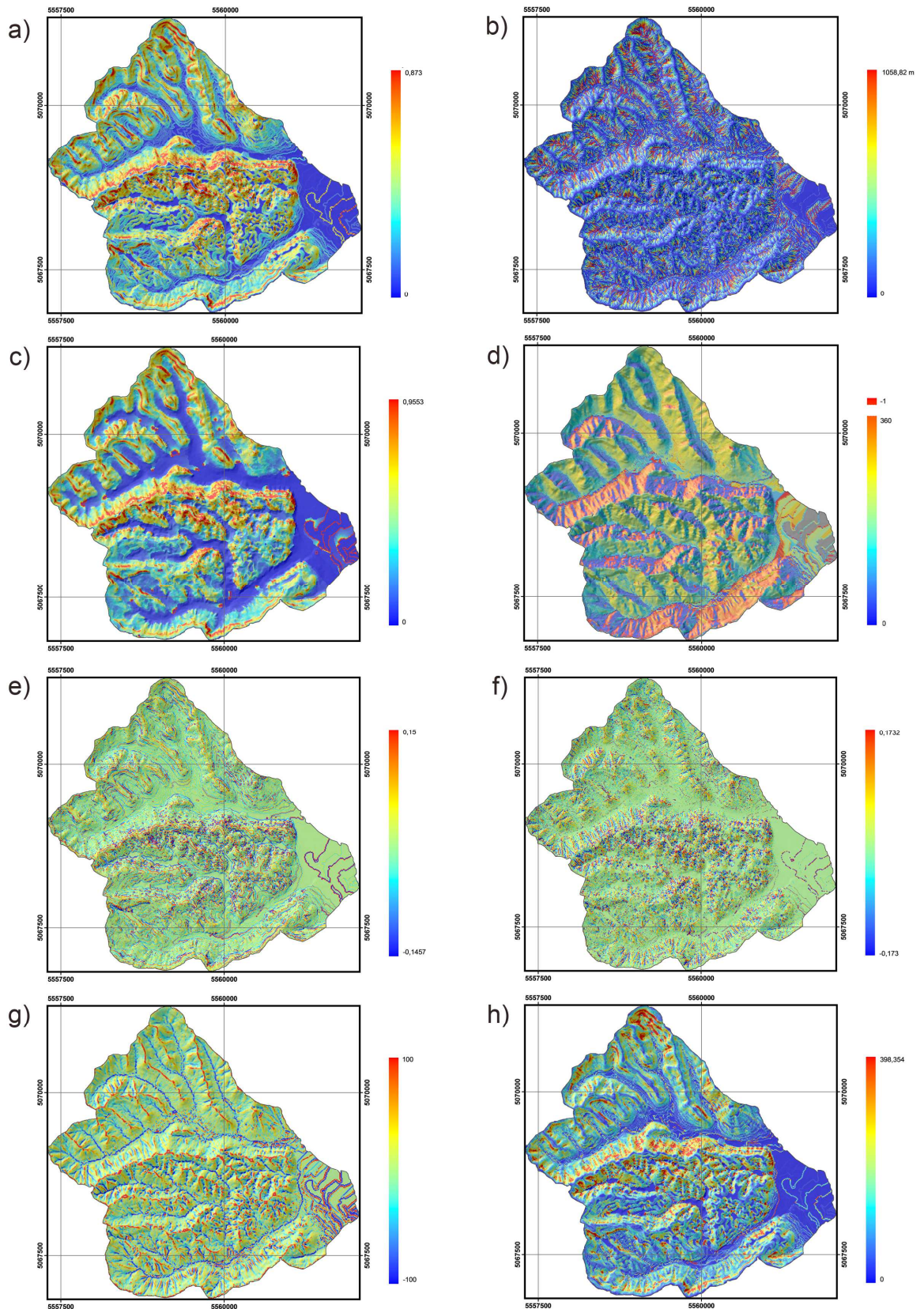
consequent character (upslope progression), but there are also landslides which are transiting to deep-seated, and these are nested in marl units. There are various activity stages to be witnessed, but most of the landslides are slow-moving, and in that regard, they impose no threat for the population, apart from the material damage on buildings and infrastructure.



**Figure 46.** Geographical setting of the study area (projection: Gauss Krüger – zone 5, Bessel 1841).

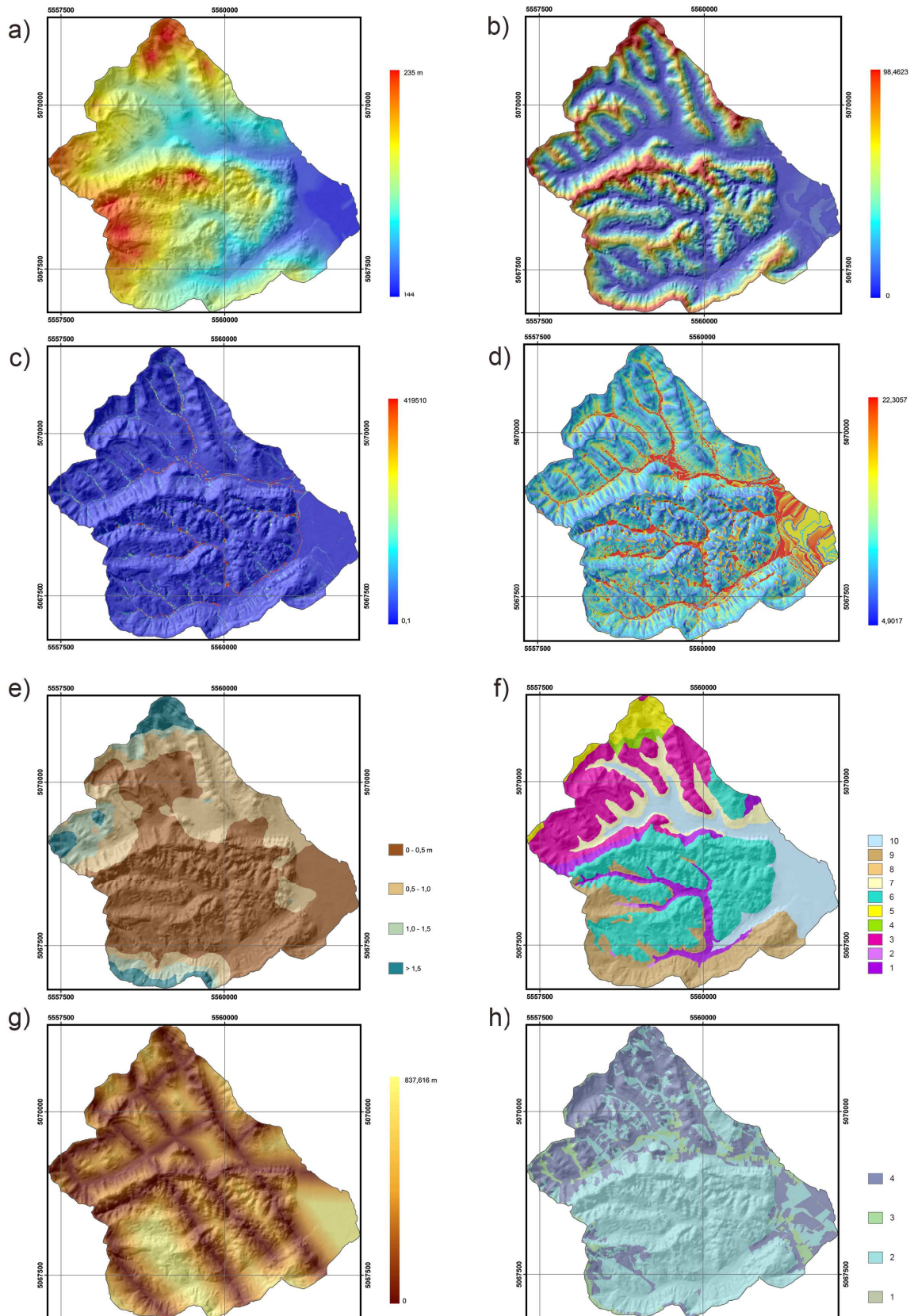
Given the geomorphological and geological features, the main triggers of landslides could be recognized in linear erosion and intensive rainfall connected with the groundwater variations. It is probably the combination of the two which ministers both, the long-term and short-term activation and reactivation cycles throughout the area. Apart from these, urbanization of the area could be considered as one of the primary causal factors for numerous shallow and relatively small landslides triggered by a combination of natural (e.g. intense, short period rainfall) and man-made processes. Settlements are generally small, but dense (Mala Gorica, Falašćak, Molvice) while the population is agriculturally active, which additionally destabilizes the slopes. On the other hand, there is a significant vegetation cover, particularly forests in the central part, which suppress the spreading of the shallow landslides. Indeed, these central parts are composed of a different type of sediments (Plio-Pleistocene gravel and silts) which are less prone to landslides.





**Figure 47.** Various Conditioning Factors: **a)** slope angle; **b)** slope length; **c)** downslope gradient; **d)** aspect; **e)** profile curvature; **f)** plan curvature; **g)** convergence index; **h)** LS factor.





**Figure 48.** Various Conditioning Factors (continued): **a)** channel base elevations; **b)** altitude above channels; **c)** SPI; **d)** TWI; **e)** groundwater depth; **f)** lithology (1=silt, 2=sandy silt, 3=marl, 4=hard marl, 5=laminated marl, 6=gravel, 7=eluvial clay, 8=sand, 9=clay, 10=alluvium); **g)** distance from structures; **h)** Land Cover (1=water body, 2=forest and semi-natural area, 3=artificial surface, 4=agricultural area).

## 6.2.2 Data

Resources for generating the input dataset of Starča Basin included: Landslide Inventory; Digital Terrain Model – DTM; geological map; hydrogeological map; and a *Land Cover* map. From the above mentioned resources, the input dataset was generated as an assembly of Conditioning Factors and the Landslide Inventory. By advantages of GIS software platforms (ArcGIS and SagaGIS) input data were processed, i.e. referenced and normalized (where applicable) and stored in form of a raster image format so that every pixel (every center node of the pixel to be more precise) represents one instance. Every attribute within the input dataset contained 122513 instances, 10 m set apart from each other (10 m cell resolution).

### 6.2.2.1 Conditioning Factors

Landslide Conditioning Factors involved a variety of input layers, some being directly digitized from the original thematic maps, others derived from additional spatial calculations and modeling. In effect, 16 input raster layers, with the same 10 m cell resolution, were available for further analysis. These could be divided into four thematic groups: morphometric, hydrological/hydrogeological, geological and environmental. The factors that turned more dominant or their type has not been included in the previous case study are briefly described, while the others have been only listed, because the same analogy from the previous case study (where all the types have been described in greater detail) applies herein. Graphical representations of all Conditioning Factors are given in Figure 47–48.

#### 6.2.2.1.1 Morphometric Data

A high precision ( $\pm 1$  m) terrain surface model has been generated by photogrammetric technique, in the framework of the orthophoto mapping project of the Zagreb City area, at 1:5000 scale. Point-based terrain model, i.e. triangulated network model, has been subsequently transformed to a Digital Terrain Model (DTM) by means of TIN-to-raster conversion. After standard (hydrological) corrections, a rectified DTM has been obtained. A host of morphometric parameters with proven relevance for landslide assessment (van Westen et al. 2003) have been derived from the DTM:

- *slope angle*,
- *slope length*,
- *downslope gradient* (a ratio of *slope angle* and *elevation* per point),
- *aspect*,
- *profile curvature* (terrain curvature in the steepest slope direction),
- *plan curvature* (terrain curvature along the contour of the slope),
- *convergence index* (*slope angle* convergence),
- *LS factor* (a ratio of the *slope length* and the length standardized by the Universal Soil Loss Equation – USLE),
- *channel base elevations* (values calculated as a vertical difference between a real DEM elevations and elevations of interpolated channel network, which provides the information on how far off a local flow each cell lies by interpreting the higher differences as more remote than lower ones, since in channel cells the factor's value is zero, while in non-channel cells the value is increasing with the distance),
- *altitude above channels* (another standard morphometric terrain attribute, yet sometimes important for determination of the relief energy based on potential energy differences, i.e. height differences between each cell and its local erosion basis, which is basically a DTM downshifted for the values of channel cells elevations).

### 6.2.2.1.2 Hydrological/Hydrogeological Data

Hydrological data inputs have been derived also by morphometric calculations from DTM, but in combination with the manually adjusted drainage network vector. These included:

- *Stream Power Index* or *SPI* (a potential power of the flows given by a relation of the local drainage area and local slope gradient),
- *Topographic Wetness Index* or *TWI* (topographic water retention potential given by a relation of upslope drainage area and slope gradient).

Hydrogeological information has been provided by a relatively dense piezometric groundwater pressure sampling. Piezometric map has been generated by a simple interpolation of the maximal piezometric pressure heads, measured in rainy period of 2004. Herein, it has been labeled as:

- *groundwater depth* (depths from the measurements of minimal water levels in wells, interpolated by nearest-neighbor method, ranged by 4 classes with 0.5 m intervals, i.e. 0–0.5, 0.5–1, 1–1.5 and >1.5 m).

### 6.2.2.1.3 Geological Data

Geological factors included layers derived from a geological map 1:5000, indicating the main geological units in the area and approximately located faults (Mihalić et al. 2008). By decompiling these from the original map the following factors have been acquired:

- *lithology* (representing 10 rock units as categorical classes<sup>35</sup> as follows: eluvial clay and silty clay with gravel (Quaternary), alluvial gravel with silty clay (Quaternary), gravel with silty clay (Plio-Pleistocene), coarse-grained sand (Plio-Pleistocene), sandy silt and silt (Pontian), marl with silt and calcareous siltstone (Pannonian), laminated marl with calcareous sandstone (Sarmatian) and marl (Badenian); considering relatively high proportion of clayey and marly units, lithological model suggests that shallow to deep-seated landslides could be hosted throughout the study area (Fig. 48f),
- *distance from structures*.

### 6.2.2.1.4 Environmental Data

Land Use map was prepared by direct visual interpretation of 1:5000 orthophoto according to CORINE classification (Nestorov & Protić 2009). The map was generalized into a 1<sup>st</sup>-level CORINE map and labeled:

- *Land Cover* (categorical attribute with 4 thematic classes<sup>35</sup>, including: agricultural areas 30%, artificial surfaces 4%, forests and semi-natural areas 65%, water bodies 1%).

---

<sup>35</sup> In order to give equal preference to every class, categorical attributes have been broken into  $m$  binary attributes, coding  $m$  different initial values (e.g. class 1 and 4 of *lithology* are coded as 1000000000 and 0001000000, respectively, while the same classes for *Land Cover* were 1000 and 0001).

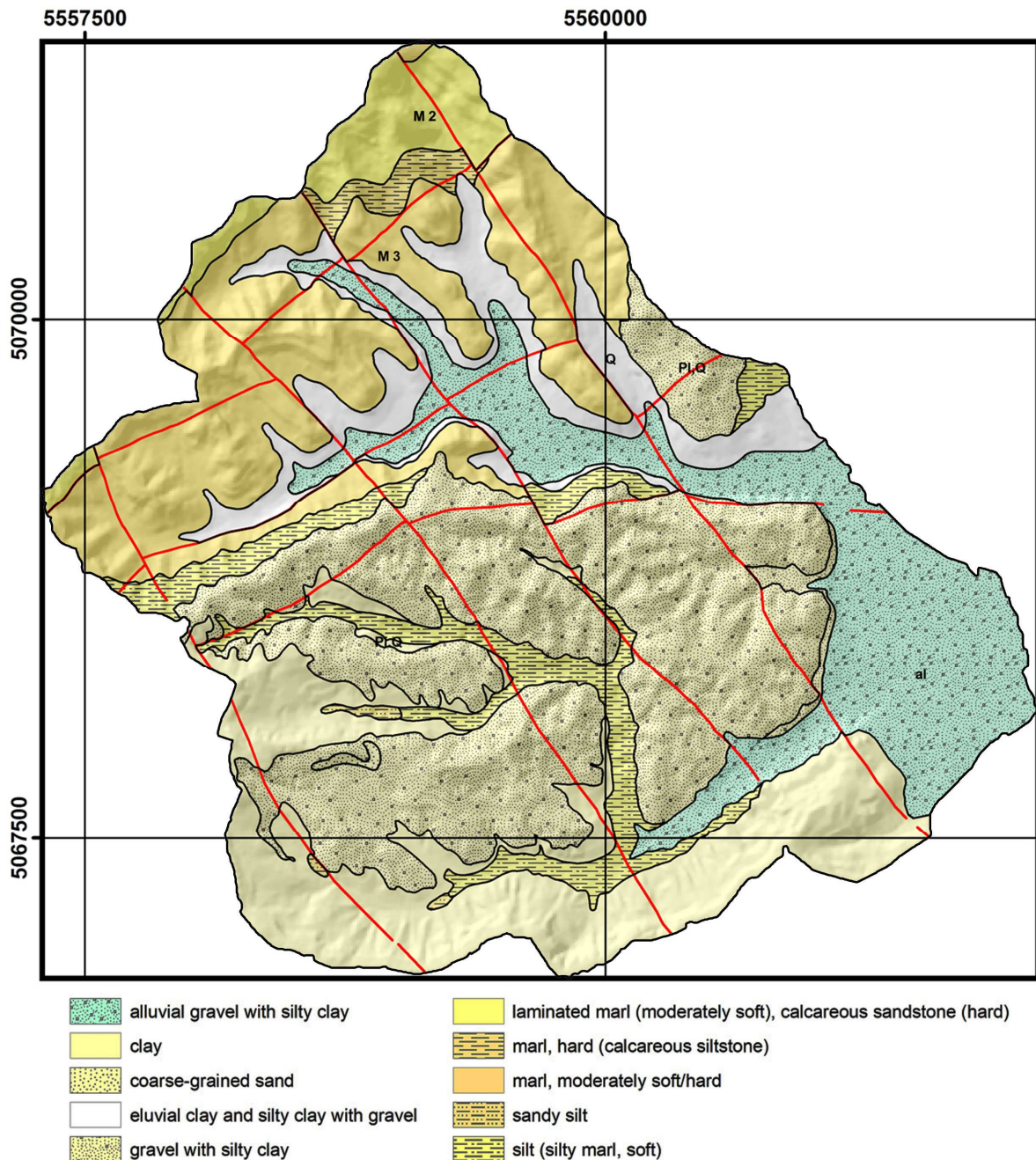


Figure 49. Geological map of the study area. Red lines represent faults.

### 6.2.2.2 Landslide Inventory

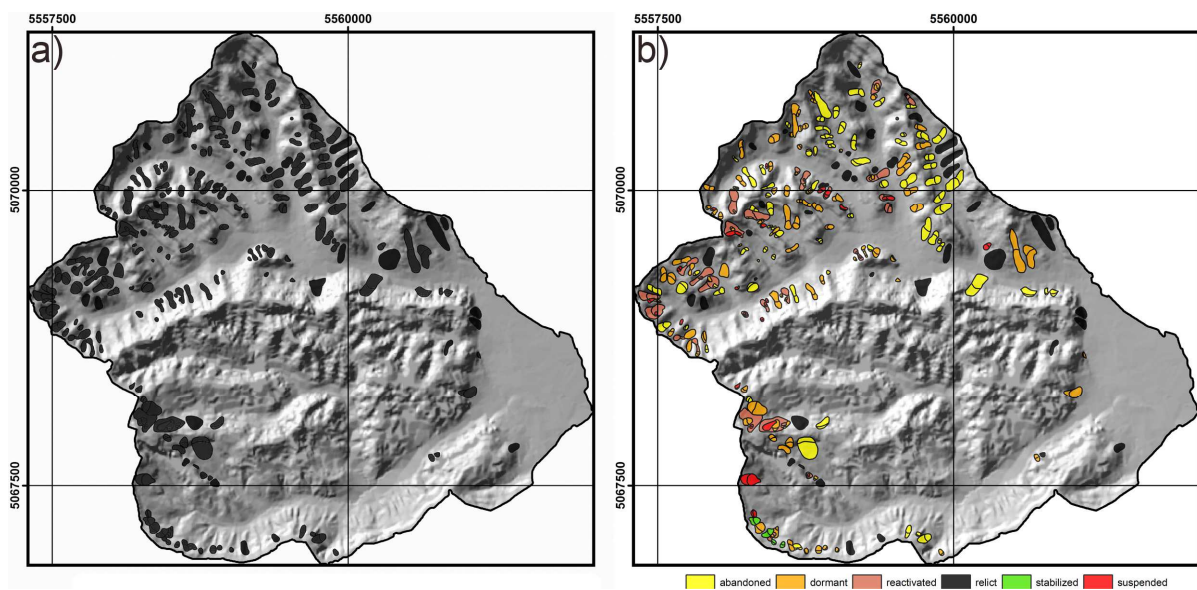
A detailed landslide map has been prepared through the systematic field survey (in the period of March–April 2004) at 1:5000 scale (Fig. 50). Total mapped landslide area reached only 0.87 km<sup>2</sup> (or 7.1% of the study area, which is statistically speaking, an undesirable proportion), with the density of about 0.1 slope failures per km<sup>2</sup>. Landslide Inventory is prepared in the form of a GIS database in which information on location, features and abundance of 230 mapped landslides are archived (Mihalić et al. 2008). Main landslide characteristics were described according to standard WP/WLI 1995 recommendations. Landslides were classified as (shallow) slide type according to Cruden and Varnes Classification (Varnes 1984), with age and state of activity determined according to the morphological indicators.



*Active, suspended and reactivated landslides* have had clearly recognizable fresh scars, without any vegetation cover, because of the movement within the last few years (59 slides). Most of the landslides were inactive and they have been classified as: *dormant landslides* (95 slides). These have had recognizable scars covered by vegetation during the period of inactivity. *Abandoned landslides* (72 slides) are characterized by hummocky surface topography and relicts of scars completely smoothed during the period of inactivity. *Stabilized landslides* included slopes remediated by different engineering measures (4 slides). *Relict landslides* (40 slides) were difficult to recognize, because the only indicator of their movement is typical roughly undulating slope morphology: concave depletion zone in the upper part and convex accumulation zone in the lower part.

The size of the landslides varies from 270 m<sup>2</sup> to 25073 m<sup>2</sup>, but most of the landslides range from 400 m<sup>2</sup> to 1600 m<sup>2</sup>. Regarding the activity style, there are single movements (150 slides) as well as complex, composite, successive and multiple movements (120 slides). Parent-child relationships have been also defined during the mapping. The relict slides have been excluded from the further analysis because of their mapping uncertainty.

For the purpose of this research, the Landslide Inventory has been used only in a raster image form and has been somewhat simplified in order to enhance the statistical representativeness of *landslide vs. non-landslide* categories. In this context, original landslide classes (WP/WLI classification) have been unified (Fig. 50a) in a basic variant of the inventory, and disaggregated on the basis of their activity stage in the second variant (Fig. 50b).



**Figure 50.** Landslide Inventory: **a)** simplified; **b)** source inventory with all (activity-based) classes.

### 6.2.3 Implementation, Results and Discussion

It has been already indicated that particular modeling techniques, which generally comply with the methodology followed in this research, have been already performed by other researchers. Weight of Evidence model, analogue to the Model-2a (except for the way the unit areas are defined) has been used for generation of a susceptibility map, containing three relative susceptibility classes (Mihalić et al. 2008). The achievements of that model have been assessed per class. High susceptibility class is for instance well correlated for reactivated, suspended and inactive slides (88.2% accuracy), but poorly with relict slides (34.6). Since the results of this study could be generally accepted, there was no need for generating Model-1b, Model-2b and Model-3b, pairing Model-1a, Model-2a and Model-3a,

respectively<sup>36</sup>. In addition, Model-4b has been rejected as inconvenient for the reasons explained for Model-4a, as well as Model-6b. Model-8b has been calculated and initially, good performance was expected from, since the theoretical conditions have been met (shallow landslides, hollow valleys, rainfall trigger). Unfortunately, after all the fitting and adjusting the performance has not topped 0.5 *AUC*, which has already been labeled as implausible in Model-8a. For this reason, the model will not be presented hereinafter. Thus, all the attention has been driven toward Model-5b and Model-7b, involving advanced Machine Learning techniques, C4.5 and SVM, respectively.

### 6.2.3.1 Model-5b

Since it has been proven suitable for susceptibility modeling in the previous case study, the Decision Tree algorithm C4.5 has been more thoroughly optimized during the experimenting procedure. The optimization has involved only *Confidence Level* parameter, while *Minimum Number of Objects in Leaf* remained 2 by default, since it is the most desirable to explore this most tolerant option. *Confidence Level* has been ranged in each of the sub-models of Model-5b from 0.05 to 0.95. Thus, only several trials in which only one parameter has been changing have been sufficient to optimize the model.

Experiment design was driven by the characteristics of the dataset, particularly the unbalanced distribution of Landslide Inventory classes. Since *non-landslide* class turned predominant over all landslide classes combined, the sampling strategy had been accordingly tuned. Two different dataset cases were induced, hence two Model-5b variants, labeled Model-5b-1 and Model-5b-2 are to be distinguished.

Model-5b-1 was trained on a generalized Landslide Inventory (Fig. 50a), containing only *non-landslide* ( $c_1$ ) and *landslide* ( $c_2$ ) class. The training sample was unbalanced (randomly selected and uniformly distributed samples) and contained 20% of the original dataset or 24500 out of 122513 instances.

Model-5b-2 was trained only over *landslide instances* and did not include *non-landslide* instances in the training nor testing. On the other hand it had multiple landslide classes to train upon: *dormant and abandoned* ( $c_1$ ), *stabilized* ( $c_2$ ), and *reactivated and suspended* ( $c_3$ ) *landslides* (only 10500 instances in total).

The first classifier was thus used to locate the landslides throughout the area, while the second was used to discern between three different landslide sub-types.

Both of the models/classifiers underwent the identical experimenting protocol discussed hereinafter. By varying the size of the training set in different experiments additional variants have generated.

Model-5b-1/2-100%: training has been performed on the whole training set (24500 instances for Model-5b-1-100% and 10500 instances for Model-5b-2-100%). Training and testing has been performed with no iteration (in a single run) by a 10-fold Cross-Validation (10-CV). Model-5b-1-100% reached  $\kappa$ -index of 0.52, and Model-5b-2-100% 0.82 (Tab. 22), meaning that the model better discerns between the landslides sub-types than between *landslides* and *non-landslides*.

Model-5b-1/2-20%: dataset has been randomly divided into 20%–80% splits. Training was performed on 20% of data (5000 instances Model-5b-1-20% and 2000 instances in Model-5b-2-20%). In order to obtain statistically relevant results, five different 20%–80%

<sup>36</sup> These models rely on a similar principle and use similarly preprocessed inputs (binarization of nominal and normalization of numeric data, filtering, Attribute Selection), and in the previous case study it has been shown that they tend to reach similar levels of performance. Furthermore, these models are not predictive in their nature, since they just re-interpret the existing landslide susceptibility, and therefore gave way to the advanced methods.

splits have been generated and the median among the obtained  $\kappa$ -index values was considered as representative (Tab. 22). As expected, the performance dropped significantly, especially in Model-5b-1-20%.

Model-5b-1/2-15%: seven 15%–85% splits have been generated (3800 instances in Model-5b-1-15% and 1500 instances in Model-5b-2-15% for training). Otherwise, the experiment is analogue to the previous one. Further decrease of average  $\kappa$ -index is noticeable (Tab. 22).

Model-5b-1/2-10%: ten 10%–90% splits have been generated (2500 instances in Model-5b-1-10% and 1000 instances in Model-5b-2-10%) and processed as in the previous experiments. The dropping trend continues as  $\kappa$ -index values became rather temperate for both models (Tab. 22).

**Table 22.** Performance evaluation of C4.5 and SVM classifiers by  $\kappa$ -index.

Experiment	Y=1		Y=2	
	X=5	X=7	X=5	X=7
Model-Xb-Y-100%	0.52	0.58	0.82	0.82
Model-Xb-Y-20%	0.38	0,47	0.63	0.65
Model-Xb-Y-15%	0.33	0.44	0.58	0.60
Model-Xb-Y-10%	0.31	0.40	0.48	0.55

**Table 23.** Information Gain (IG) ranking of the input layer attributes.

Conditioning Factor	IG	rank
<i>lithology</i>	0.06157	1
<i>channel base elevations</i>	0.04034	2
<i>groundwater depth</i>	0.02680	7
<i>SPI</i>	0.03038	4
<i>aspect</i>	0.02828	5
<i>altitude above channels</i>	0.03078	3
<i>TWI</i>	0.02789	6
<i>Land Cover</i>	0.02129	10
<i>downslope gradient</i>	0.02413	8
<i>LS factor</i>	0.02241	9
<i>slope</i>	0.02100	11
<i>convergence index</i>	0.01723	12
<i>plan curvature</i>	0.00800	14
<i>distance from structures</i>	0.00938	13
<i>profile curvature</i>	0.00605	15
<i>slope length</i>	0.00301	16

Viewing the experimenting results altogether, it could be inferred that the model is rather sensitive to reduction of the training sample size. In all experiments the algorithms exhibit better generalization of different landslide classes than landslides itself, meaning that they are better in categorizing landslides than actually mapping them (concerning this particular study area and chosen sampling strategy). Preliminary results suggest that it would be interesting to impose the algorithms over adjacent areas using the same inputs. This would be particularly interesting because these adjacent areas represent urbanized and densely populated environments, which have very similar terrain properties. Given the same

kind of Conditioning Factors as inputs, it could theoretically suggest to the experts which types of landslides are present prior to the actual field mapping.

Since a classifier based on *Information Gain* values (C4.5) has been in operation, it is suitable to present the ranking of the input features according to their *IG* value (Tab. 23). Apparently, the most informative layers are *lithology*, *channel base elevations*, *altitude above channels*, while surprisingly *slope* turned out mediocre to low, hand in hand with terrain *convergence index* and *Land Cover* for instance. One possible way to explain this is exaggeration of geological and to some extent hydrogeological influence on the landslide occurrence, so that they obscure the effects of slope steepness and *Land Cover*.

### 6.2.3.2 Model-7b

Finally, the implementation of the SVM technique which has been anticipated with the highest expectations has taken place. A more laborious experimenting has been undertaken, on top of the scheme inherited from Model-5b. It regarded more thorough feature selection, and a more realistic sampling strategy, which underpinned the predictive nature of the model.

Since the identical protocol has been in operation (similar optimization and same sampling strategy), it is possible to distinguish the same model sub-types as in Model-5b. For that reason, these variants are only to be listed and briefly commented hereafter.

Model-7b-1/2-100%: training and testing has been performed with no iteration (in a single run) by a 10-fold Cross-Validation (10-CV). Model-7b-1-100% reached  $\kappa$ -index of 0.58, and Model-7b-2-100% 0.82 (Tab. 22). Thus, the model also better discerns between the landslides types than among *landslides* and *non-landslides*.

Model-7b-1/2-20%: training was performed on 20% of data, and thus, five different 20%–80% splits were generated and the median among the obtained  $\kappa$ -index values was considered as representative (Tab. 22). The performance dropped significantly, especially in Model-7b-1-20%.

Model-7b-1/2-15%: seven 15%–85% splits have been generated, and the rest of the experiment is analogue to the previous one. Further decrease of average  $\kappa$ -index is noticeable (Tab. 22).

Model-7b-1/2-10%: ten 10%–90% splits have been generated and processed as before. The dropping trend continues as  $\kappa$ -index values became rather temperate for both sub-models (Tab. 22).

It is evident (Tab. 22) that the SVM has been a slightly better choice for modeling, given the same circumstances. Despite the slightly higher performance parameters than in concurrent C4.5-based models, impression of insufficient precision in mapping landslides remains. Instead, the SVM-based model is also showing considerably better use in discerning among different landslide classes.

Finally, a supplementary experiment, generating the last model, labeled Model-7b-40% has been performed. It has been preceded by a thorough Attribute Selection. The main idea behind it was to implement leave-one-out technique to generate the optimal model. *Information Gain* estimator has been employed (Tab. 23). In the optimization stage, there has not been any significant decrease in performance after leaving last 5 Conditioning Factors (*convergence index*, *plan curvature*, *distance from structures*, *profile curvature* and *slope length*). It is here disputable whether the binarized factors, such as *lithology* and *Land Cover* should have been observed as a whole or disaggregated into their binary sub-features. In result, there was no significant computational time saving or performance enhancement by introducing leave-one-out, which all brings one back to the discussion on justification of the feature selection (see Chapter 5.1).

The training in Model-7b-40% has thus commenced with dimensionally reduced (feature space-wise) training set, wherein approximately 40% of the total area has been split



for training. The training has been based on a simplified inventory variant which aggregates all *landslide* instances together (Fig. 50a). Once again, a cautious manual sampling had to be done in order to define the training domain, so that each of the nominal class (in all of the nominal inputs and the inventory) has been taken into account. This time the optimization has been done by 5-fold CV wherein nine pairs of  $c$ ,  $\gamma$  combinations have been tested (Tab. 24). According to the performance parameters and time consumption, the optimal combination was  $c=10$  and  $\gamma=10$  (even though it was not the best, it turned out to be much faster than the next best choice).

**Table 24.**  $c$ ,  $\gamma$  parameter combinations for optimization.

$(c, \gamma)$	<i>AUC</i>	$\kappa$ -index	$fn_{rate}$	train/test time-elapsed [s]
1,1	0.57	0.21	0.00	158.55/10.94
10,1	0.62	0.33	0.01	198.49/10.56
100,1	0.68	0.46	0.02	435.26/10.06
1,0.1	0.50	0.02	0.00	148.87/10.59
10,0.1	0.53	0.10	0.00	159.95/10.13
100, 0.1	0.55	0.17	0.00	242.82/9.99
1,10	0.81	0.63	0.04	886.08/7.80
10,10	0.77	0.61	0.03	317.29/8.50
100,10	0.69	0.48	0.01	236.71/12.05

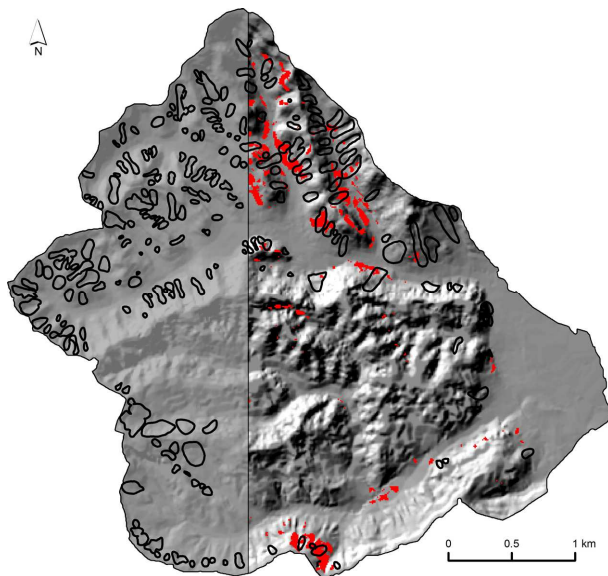
Subsequently, the algorithm has been applied to the remaining 60% of the area and some interesting results have been achieved (Fig. 51). The trends of some landslides have been well portrayed, but the general impression is that the model underestimates the *landslide* instances. Performance parameters values are unambiguously underpinning such impression (Tab. 25). Overall accuracy of 92.6% seems very promising, but it obscures the actual successfulness of the model, which is better revealed by the other measurements, broken down per classes. The  $K_{fuzzy}$  as a relatively novel parameter (for which no fitting had taken place, only the defaults provided by the MCK package have been considered), seems more appealing and slightly smoothens a rigorous indication given by the other parameters.

**Table 25.** Performance of the Model-7b-40%.

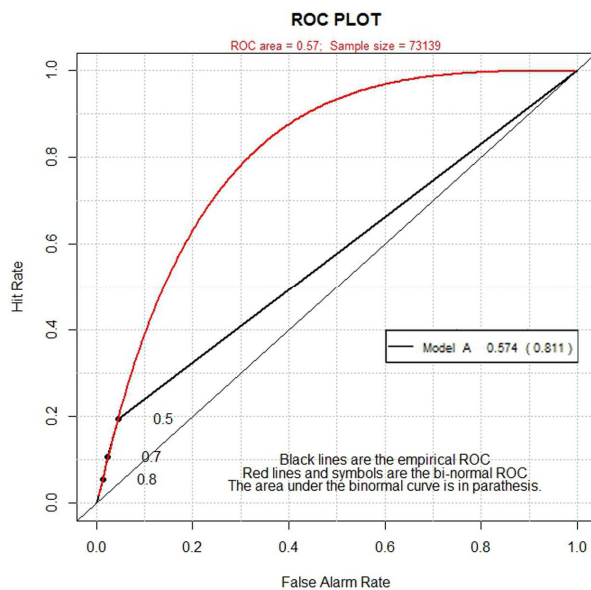
Model class	<i>AUC</i>	$\kappa$ index	accuracy	average $K_{fuzzy}$
<i>landslide</i>	0.54	0.21	24%	0.97
<i>non-landslide</i>	0.54	0.48	95%	

The prognosis is meaningful (following the logic of geomorphological or geological entities of the terrain), but scarce and incomplete. The model is obviously overfitted since a considerable discrepancy appears between training and testing performance. For instance, average *AUC* drops from 0.77 to 0.54, which is considered as a poor performance (Fig. 52). If it is for any consolation, the ROC curve is left-skewed, meaning that the output is conservative (strict for *False Negative* type of error) and that there is some room for further fitting. Also, it is apparent that fitting of the curve (red contour in Fig. 52) suggests better fitting of the ROC curve. One of the reasons is probably the nature of the landslides in Starča basin, with their relatively small size, locally high density of occurrence and strongly exhibited variation in activity. Generally, there have not been too many *landslide* instances to work with despite slightly larger training area size (40% instead of usual 33%). In statistical terms the sample was too small to give good results. Perhaps increasing of the resolution of the dataset could improve the results. On the other hand, similar proportions of instances worked well for the previous case study. In comparison to the previous investigation based on simple

bivariate statistics (Mihalić et al. 2008), this research elaborates further details on possibilities of handling the landslide Conditioning Factors and exploiting different aspects of the model, thus contributing to the overall comprehension of the landslide phenomena in the study area.



**Figure 51.** Model-7b-40% (predicted landslides are in red, while the actual ones are contoured; the training area is shaded).



**Figure 52.** ROC curve of Model-7b-40% (fitted ROC curve is given in red).

A small web-map, created in the R environment, using “plotGoogleMap package” (Appendix 2) presenting the area and related result of Model-7c-40% is available at:

<http://milosmarjanovic.pbworks.com/w/file/attach/63741247/MyMapStarca.htm>.

### 6.3 Halenkovice Area (Czech Republic)

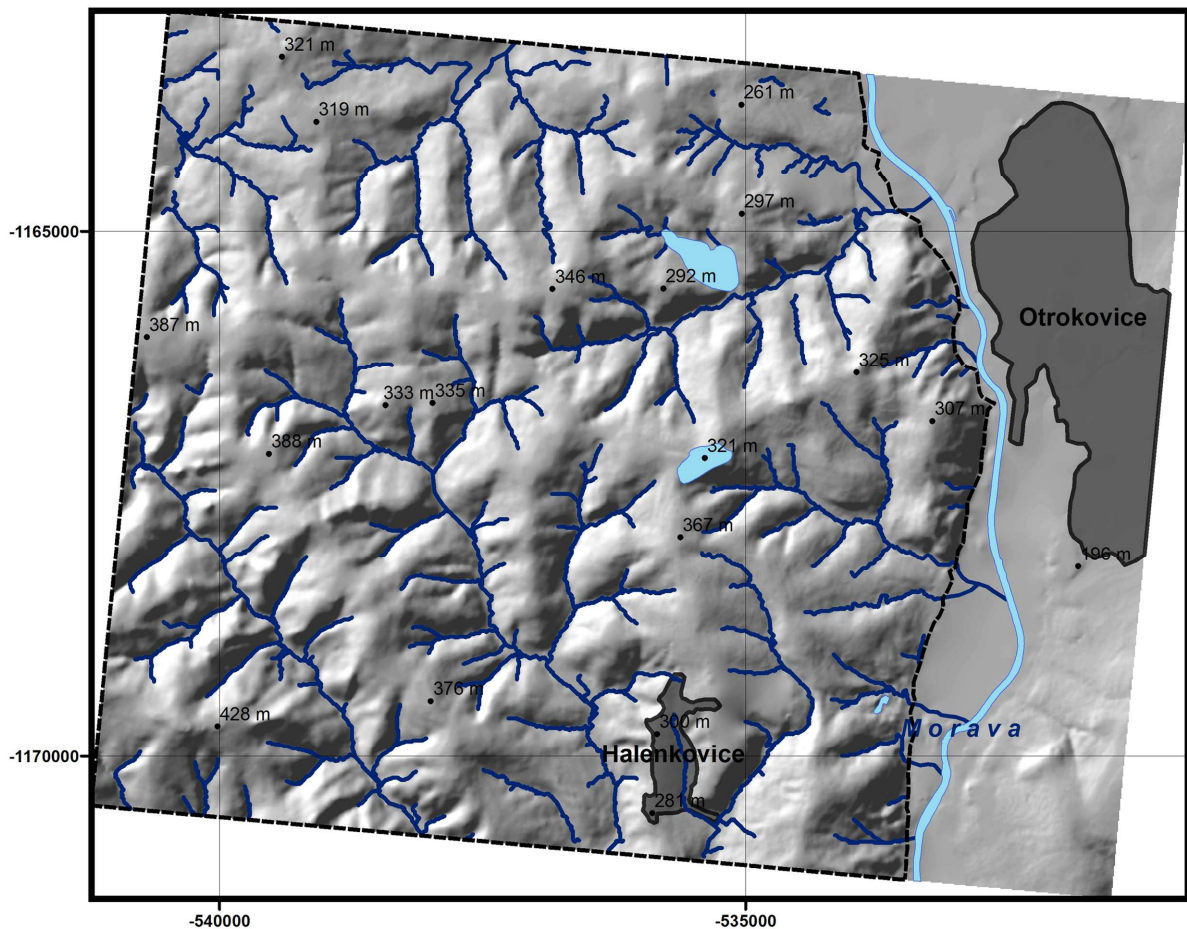
This last case study is actually a work in progress, which means that only preliminaries are to be given hereinafter, as well as further notices and ideas. It is further important to emphasize that the study site Halenkovice underwent a series of various landslide-related investigations in the past decade. There has been a number of landslide mapping investigations regarding the Western Carpathians in Moravia, particularly in districts of Zlín and Vsetín, thus encompassing this relatively small study area near Halenkovice. The researchers of these terrains coped with different aspects of landsliding process at regional scale, mainly including engineering-geological mapping (defining the metrics, geometry, mechanisms and typologies of the landslides) (Kircher et al. 2000, Pánek et al. 2011a), but there were also several susceptibility assessments attempts (Kianička & Čapková 2005, Klimeš 2008a,b, Klimeš et al. 2009, Klimeš & Novotný 2011) using Susceptibility Potential Index (SPI) as a final assessment output, as well as estimations of the triggering events in relation to the landslide occurrence (Bíl & Müller 2008). These researchers acted independently or jointly, within the national projects, such as slope deformations documenting and mapping, conducted by the Czech Geological Survey (CGS) and completed in 2011<sup>37</sup>. Such consistency in research comes along with the acute and realistic motifs for landslide investigation, since a host of slope failures has been recently witnessed in Moravian flysch Carpathians (Kircher et al. 2000, Bíl & Müller 2008, Klimeš et al. 2009, Pánek et al. 2011b). Multiple occurrences have been recorded in 1997, 2006 and 2010, which are to be recalled as years with unusually intensive precipitation. It is therefore objective to assume the dominant role of precipitation in triggering of landslide events, but also a combination if precipitation and floods (Rybář & Novotný 2005). Another apparent reason for research consistency lays in a good cooperation of geoinformatic and applied geo-science communities and appreciation of applied GIS in academic circles (which is not entirely true for Serbian or Croatian communities), as well as better financial support (primarily through independent projects).

Conveniently, some parts of the study area have been surveyed for over a decade throughout the different projects of Department of Geoinformatics at Palacký University Olomouc and diploma projects of its graduates, leaving some unpublished and unused materials which have contributed to the research in this case study<sup>38</sup>. Some related work (in terms of landslide assessment), such as bivariate and multivariate statistical approaches, as well as some basic deterministic modeling, have already been involved. It is herein considered that the logical extension of these investigations lays in the application of the most advanced Machine Learning techniques (SVM in particular), but the deterministic model has also been revisited. The latter is additionally inspired by the character of landslides and terrain features, which theoretically seem suitable for such approach.

---

<sup>37</sup> The project name (verbatim et literatim) "Creating an interactive slope stability and rock avalanches risk map of Czech Republic", in original: "Vytvoření interaktivní mapy rizika porušení stability svahů a skalního řízení v České republice", VaV SP/1cp/157/07. The subproject which was mostly engaged with this goal was (verbatim et literatim), "Documenting and mapping of slope movement in CZR", in original "Dokumentace a mapování svahových pohybů v ČR", ISPROFIN č. 215124-1.

<sup>38</sup> By the courtesy of the Department of Geoinformatics from Olomouc all related repositories have been made available for this research.

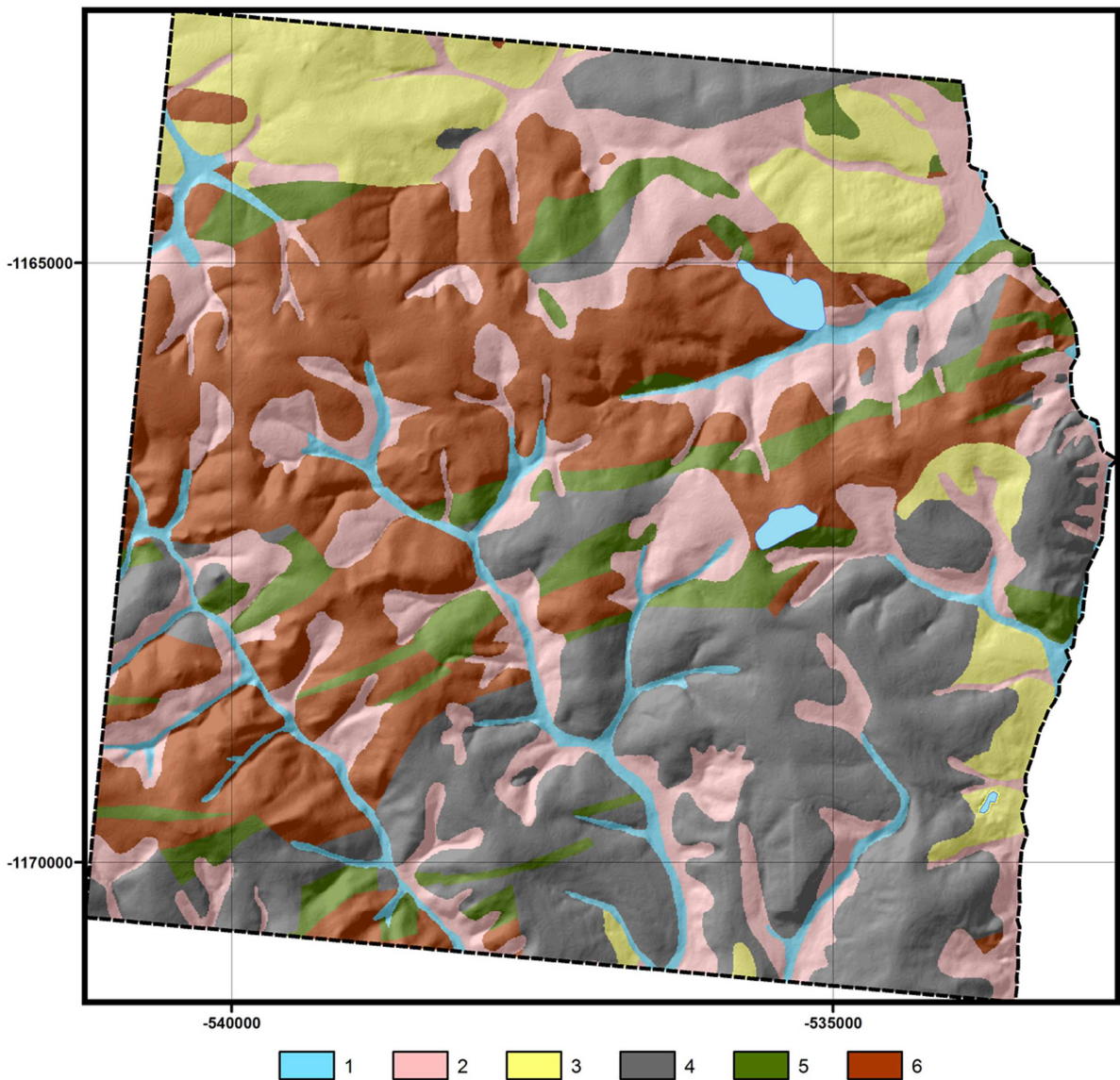


**Figure 53.** Geographical location of the study area (map projection: Krovak, JTSK).

### 6.3.1 Setting

The study area (Fig. 53) is situated near Halenkovice village, in SE Moravia (Czech Republic), occupying roughly 60 km<sup>2</sup> on Halenkovice Plateau, between Bohemian Massif and Carpathians. The territory is located in the outer part of the Outer Western Carpathians (Carpathian Foredeep) which are made of Mesozoic and Tertiary flysch formations, so-called Flysch Carpathians, segmented locally by Paleogene basins or grabens which are spatially linked and typically have marine and lacustrine evolution. In terms of tectonic regionalization, the area belongs to the Račajská nappe of the Magura flysch group, characterized by a very complex nappe system, stretching NE–SW and verging toward NW. From NW to SE propagates a composite graben of Morava River, linking smaller surrounding structures (regional normal faults), hence traversing the Magura group perpendicularly (Kircher et al. 2000, Demek et al. 2012). Geological composition of Račanská unit includes the stratified sandstone, alternating with conglomerate and inter-layered clay-slate. These differ in thickness, hydrogeological function and mechanical characteristics, which enables different types of instabilities to occur. The unit can be subdivided into Zlín, Belovež and Solaň units (Fig. 54, Tab. 26). Quaternary units are presented by fluvial sand, sandy gravel, loam and clay, and locally it can be distinguished as eluvial and delluvial soil mantle, where the landslides are usually hosted. The most typical are the instabilities in the rock slope faces, but this particular area (where the soil mantle thickens significantly) is also susceptible to shallow slides.





**Figure 54.** Geological setting of the area (1=alluvium, 2=delluvium, 3=loess, 4=Zlín subunit, 5=Belovež subunit, 6=Solaň subunit, see Table 26).

According to the international landslide classification (Varnes 1984) predominant types of the slope processes in these terrains are earth-slides, earth-flows and rock-falls (Kircher et al. 2000, Bíl & Müller 2008), wherein shallow landslides predominate over deep-seated ones (Kircher et al. 2000). Displacements occur when the bodies of these shallow landslides suffer abrupt changes in the pore pressure regime, which is chiefly triggered by a heavy rainfall/snow thaw in combination with the undercutting linear erosion. These and some other conditions enable slow shallow movements with annual reactivation dynamics.

The area is sparsely populated, thus the landslides do not pose major threat, as perhaps some other hazardous phenomena, but Halenkovice village is partly surrounded by potentially dangerous landslides, which can affect the infrastructure and cause primarily material damage, due to the slow displacement rates.

**Table 26.** List of lithological units of the study area.

#	Lithological unit	Lithological composition
1	Alluvium	organogenic sediments sandy-clay, sandy-loam, sand sand and sandy-gravel of fluvial terraces, sandy-gravel of alluvial fans (middle Pleistocene, Riss)
2	Delluvium	sandy-loam and loamy-sand of delluvial loess sand and sandy-clay (fluvial/delluvial) sandy-loam and loamy-sand loamy-sediments with rocky fragments
3	Loess	loamy-loess loess and loamy-loess
4	Zlín subunit	Magura flysch formation, Račajská nappe unit: siltstone and sandstone (upper Eocene, lower Oligocene) Magura flysch formation, Vsetín unit: alternating flysch with marly-claystone and glauconitic sandstone (upper Eocene, lower Oligocene)
5	Belovež subunit	Magura flysch formation, Račajská nappe unit: alternating flysch with green-gray and red-brown claystone (upper Eocene, lower Oligocene)
6	Solaň subunit	Magura flysch formation, Lukov unit: alternating flysch with sandstone and conglomerate (Paleocene) Magura flysch formation, Račajská nappe unit: non-segmented flysch (Campagne - Paleocene)

### 6.3.2 Data

A set of thematic attributes very similar to the one already used in the previous case studies has been assembled from several sources:

- topographic maps at 1:10000, sheets 25-31-22 and 25-33-02, and partly sheets 25-31-23 and 25-33-03 from ZABAGED database (map server) – Český Úřad Zeměměřičský a Katastrální (ČUZK) as vectorized contour lines,
- orthophoto (5 m resolution) from ČZUK,
- geological map at 1:50000, sheets 25-33 Uherské Hradiště and 25-31 Kroměříž, from 1994, a repository of (CGS) and on-line web-service at <http://mapy.geology.cz/website/geoinfo/viewer2.htm>, and recently (2012) added on-line repository<sup>39</sup> for sheet Otrokovice at 1:25000, which covers a smaller part of the study area, available at [http://mapy.geology.cz/geocr\\_25/](http://mapy.geology.cz/geocr_25/),
- slope instability map at 1:10000 from 2011, sheets ZM10 25-31-23, ZM10 25-31-22, ZM10 25-33-02 and ZM10 25-33-03, available at CGS online repository [http://mapy.geology.cz/svahove\\_nestability/](http://mapy.geology.cz/svahove_nestability/),
- CGS Relational Database Management System (RDBMS) Oracle database (authorized access only), containing detailed documentation on each mapped landslide, including (ID number, map sheet number, region, GPS coordinates of the scarp, author of the map, mapping date, type of deformation, landslide class, dimensions – length/width, area, depth, slope angle, trigger, lithological composition,

<sup>39</sup> Nation-wide geological mapping project of CGS at 1:25000 has commenced at 1999, and till present several sheets are completed and released on-line for open usage.

activity stage and class, remediation measures, Land Use, endangered objects, degree of risk, comments, photo-documentation, miscellaneous).

As usual, the inputs derived from these sources can be divided into a Landslide Inventory and a group of Conditioning Factors. Additionally, there is a group of deterministic inputs, containing principal geotechnical parameters for a simplified deterministic modeling of the landslide susceptibility. Most of the inputs have been processed in ArcGIS and SagaGIS software.

### 6.3.2.1 Conditioning Factors

The group of these factors has subsumed morphometric, hydrologic, geological, and environmental data. They have all been rasterized with 10 m cell resolution, which has seemed optimal regarding the quality of the data and the support problem of mixed scales (1:5000, 1:10000, 1:25000 and 1:50000), as well as the computational cost reduction (the number of pixels/instances is reciprocal to the resolution). In turn, the total area has been presented by total of 577931 instances (pixels), which could be already characterized as a robust dataset, especially if one has in mind that these instances had 24 dimensions, i.e. 24 different Conditioning Factor attributes attached to it. Once assembled these inputs have been preprocessed by some standard procedures (binarization of nominal, normalization of ordinal data etc), and subsequently underwent two Attribute Selection procedures, based on *Information Gain* and *Gain Ratio*, in order to rank the factors properly for the leave-one-out Machine Learning scheme, which will be discussed later on. Similarly as in the previous case study, there was no need to explain the acquisition and processing of the input data in detail, because these are very similar to procedures described in the first case study (see Chapter 6.1.2).

#### 6.3.2.1.1 Morphometric Data

Terrain surface model (DEM) has been obtained from the vectorized contour lines at 1:10000 scale. During the process, not only the intermediate (basic) contour lines, but also supplementary and depression contours, and individual spot heights have been included, in order to reach a higher precision. Two methods were used, TOPOtoRaster and TINtoRaster, and the latter turned out to be more reliable, as in both previous case studies. The following morphometric parameters (Fig. 55, 56a-d) have been derived from DEM:

- *elevation*,
- *slope angle*,
- *downslope gradient* (a ratio of *slope angle* and elevation per point),
- *aspect*,
- *convergence index* (*slope angle* convergence),
- *profile curvature* (terrain curvature in the steepest slope direction),
- *plan curvature* (terrain curvature along the contour of the slope),
- *LS factor* (a ratio of the slope length and the length standardized by the USLE),
- *channel base elevations* (values calculated as a vertical difference between real DEM elevations and elevations of interpolated channel network),
- *altitude above channels* (a DTM downshifted for the value of channel cells elevations).

### 6.3.2.1.2 Hydrologic Data

Using morphometric calculations from DTM in combination with the manually adjusted drainage network vector, two factors (Fig. 56e) were derived:

- *Topographic Wetness Index* or *TWI* (topographic water retention potential given by a relation of upslope drainage unit area and slope gradient),
- *distance from stream* (Euclidean buffer of drainage network).

### 6.3.2.1.3 Geological Data

Digitized (raster) geological map has been reclassified according to the units described in Table 26., and *lithology* (Fig. 54) has been obtained. Subsequently, the raster with six classes has been (binarized) disaggregated into six separate binary factors, similarly as in the previous case studies. The factors represented the following:

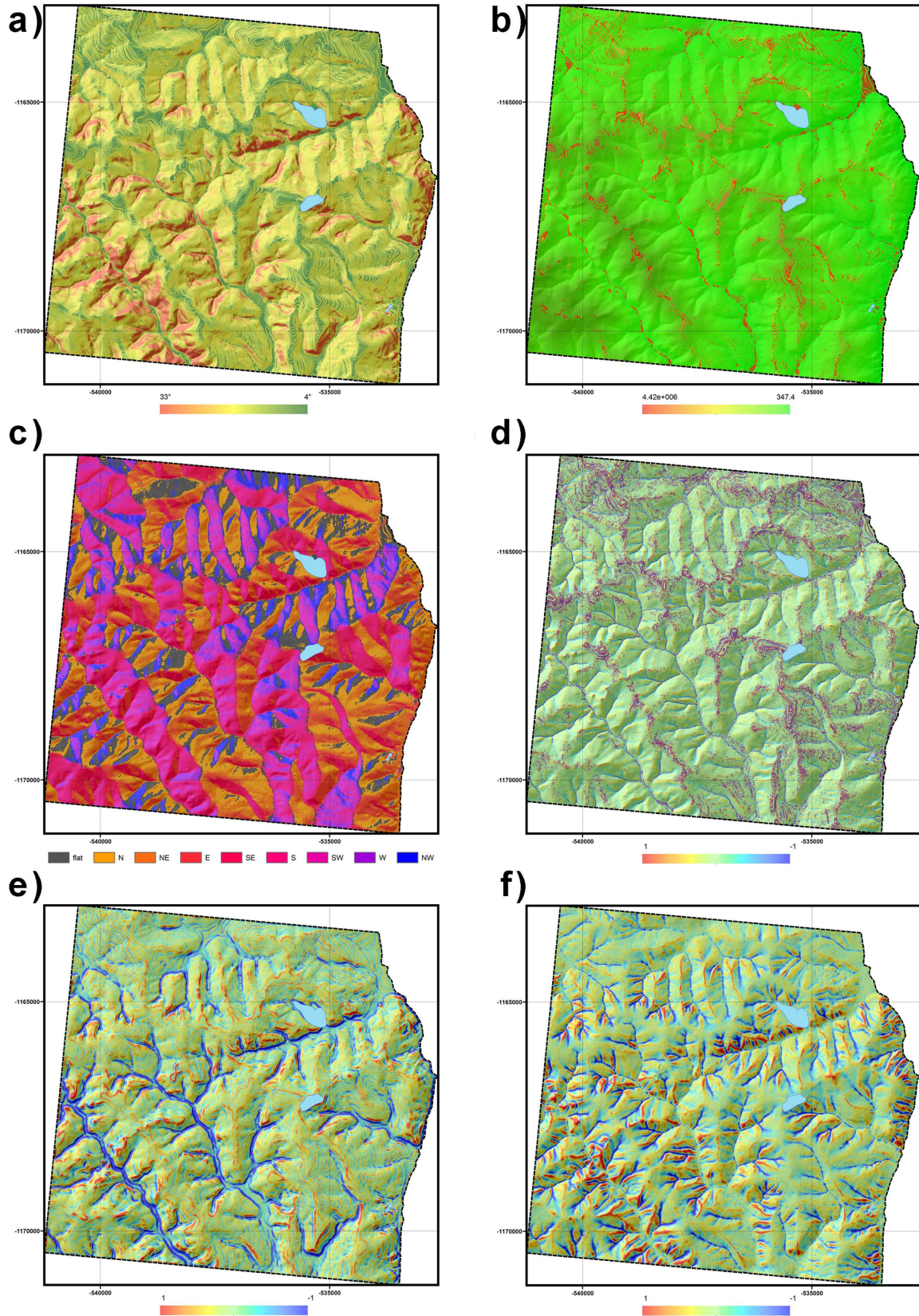
- *lithology=1* is Alluvium,
- *lithology=2* is Delluvium,
- *lithology=3* is Loess,
- *lithology=4* is Zlín subunit,
- *lithology=5* is Belovež subunit,
- *lithology=6* is Solaň subunit.

### 6.3.2.1.4 Environmental Data

Environmental influence has been featured by *Land Cover*, another nominal factor with seven classes (Fig. 56f). Likewise, seven different binary factors have been derived, and each represents one of the predefined *Land Cover* classes. The classes have been interpreted by vectorizing the orthophoto (5 m resolution) and field survey of the area. The classes are generally improvised classes of CORINE standard and can be listed as follows:

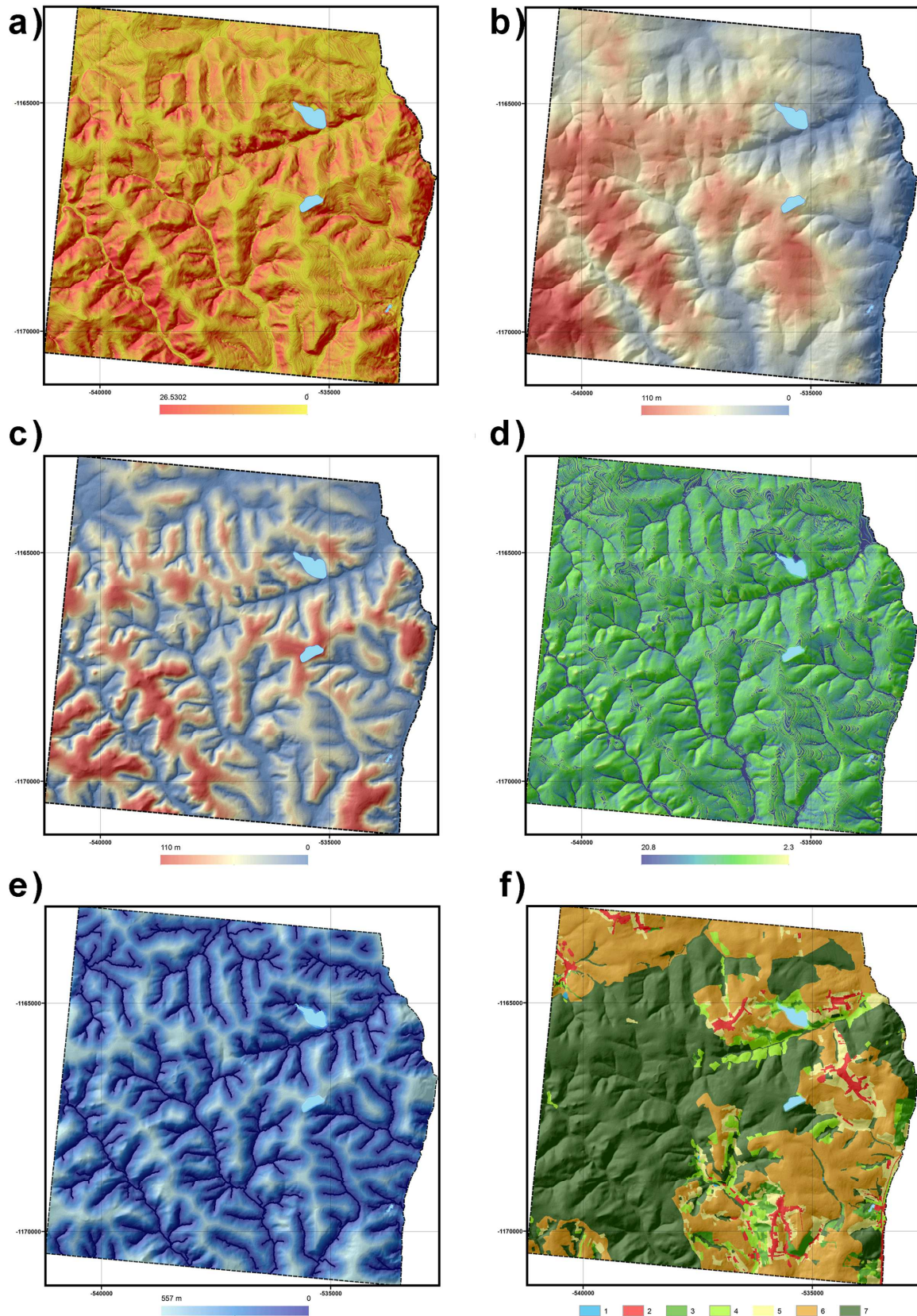
- *Land Cover=1* is water body,
- *Land Cover=2* is built-up area,
- *Land Cover=3* is sparsely forested areas,
- *Land Cover=4* is grasslands,
- *Land Cover=5* is orchards and gardens,
- *Land Cover=6* is arable land,
- *Land Cover=7* is forest.





**Figure 55.** Various Conditioning Factors: **a)** slope angle; **b)** downslope gradient; **c)** aspect; **d)** convergence index; **e)** profile curvature; **f)** plan curvature.

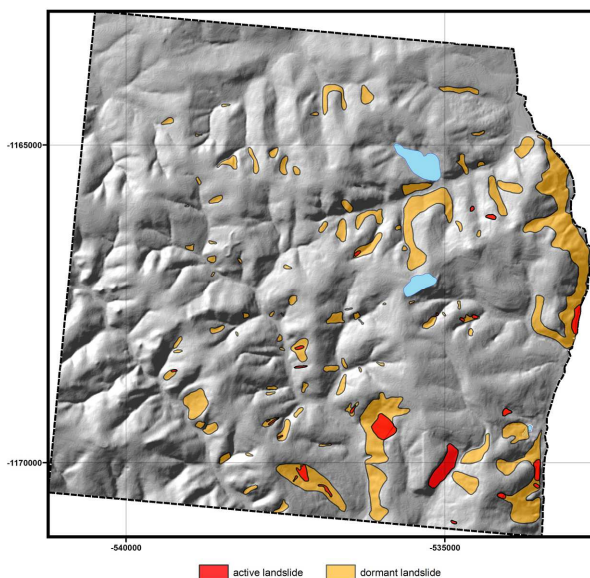




**Figure 56.** Various Conditioning Factors (continued): **a)** LS factor; **b)** channel base elevations; **c)** altitude above channels; **d)** TWI; **e)** distance from stream; **f)** Land Cover (1=water body, 2=built-up area, 3=sparse forest, 4=grassland, 5=forest).

### 6.3.2.2 Landslide Inventory

The inventory has been originally digitized in multiple landslide classes discerning not only among different activity stages of landslides, but also different mechanisms. Although shallow earth-slides dominate throughout the area, several earth-flows are also present, particularly in the northern part with the steeper slopes and narrower valley channels. Since these two types have entirely different phenomenology (geometry, dynamics and mechanism) it is logical to assume that different Conditioning Factors will have different roles in both, which have led to two separate investigations. In the previous case studies only the slide failure types have been elaborated, and proposed methodology required analogous type of subject to model. In order to remain consistent with previous case studies, the flow types have thus been excluded from the inventory, and only slide type of failure has remained. Furthermore, the original activity classes of slide failures do not entirely match the international (Varnes 1984) classification. They rather follow the local (national) classification (Záruba & Mencl 1987), although they could be approximated as: *active*, *suspended* and *dormant*. For this reason, and in order to reduce the computational costs, the landslide classes have been unified, so that the final inventory contained only *landslide* and *non-landslide* instances (Fig. 57). The landslides vary in size from 100–10000 m<sup>2</sup> or locally even bigger if regarded as composite slides. According to the most recent sheets of engineering-geological maps, there are over 20 active slides bigger than 100 m<sup>2</sup> among the total of 125 mapped landslides in the study area.

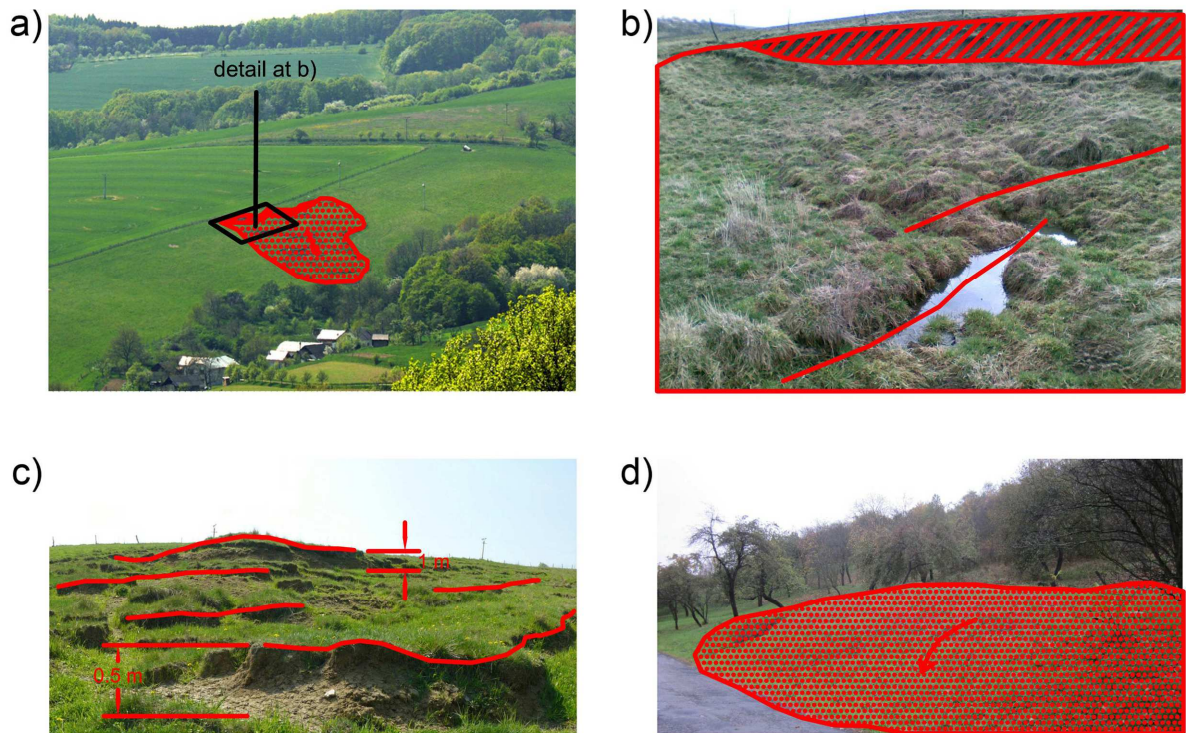


**Figure 57.** Landslide Inventory.

\*\*\*

During the field reconnaissance and survey, standard observations and measurements have been carried out (landslide morphology and metrics, depth estimation, tension cracks, object deformations etc.). The survey has not been systematic and the data have been collected in sequences in 2008–2012, because field revisits served only to ensure some particular occurrences with accent on the training area (Fig. 58). Annual revisits allowed visual monitoring of particular occurrences.





**Figure 58.** Augmented photo-documentation from the field investigation: **a)** shallow landslide in the outskirts of Halenkovice (photo by L. Marek 2009); **b)** detail of the landslide crown, showing the main scarp and tension cracks filled with water (photo by L. Marek 2008); **c)** measuring of the heights of the scarp cascades for depth estimation, where sliding depth equals  $\sim 2.5$  times the scarp height above the ground level (photo by L. Marek 2009); **d)** toe of a shallow landslide (photo by the author 2012).

\*\*\*

### 6.3.2.3 Deterministic Geotechnical Data

Specific geotechnical parameters were required for the deterministic modeling, analogue to the Model-8a. Since the area varies in lithological composition, the regionalization of the parameters has been performed, even though such generalization is not entirely justified from the geotechnical point of view. Basically, two regions were sufficient for this purpose (and also limited by the availability of geotechnical laboratory records): flysch and non-flysch units, i.e. flysch vs. Quaternary incoherent units. According to the internal CGS database and sampling which has been carried out for the most prominent landslides (Krejčí et al. 2008) and some indirect consultations from the CGS researchers, these regions are generalized as presented in Table 27. The model also required a standard DEM and a sample of existing *landslide* instances (for calibrating the classification of *S*). The same SinMap 2 GIS (via MapWindow) package has been used as before.

**Table 27.** Average parameters over different lithological domains (regions).

parameter	region 1 (flysch)	region 2 (non-flysch)
$c_{soil}$	0–0.24	0–0.13
$\gamma_{soil}$ [kg/m <sup>3</sup> ]	2732	2100
$\varphi_{soil}$ [°]	19–20	9–13
$T/q$ [log]	1000–3000	500–1000

### 6.3.3 Implementation, Results and Discussion

Previous research of Halenkovice area have been already involving some basic modeling techniques, such as bivariate, even multivariate, but based on different unit area and with non-predictive approach. Therefore, the most of the basic techniques, seen in previous case studies have been skipped, and modeling has thus been limited to the particular Machine Learning and deterministic techniques.

**Table 28.** Attribute Selection and ranking.

#	IG Ranking	IG	GR Ranking	GR
1.	<i>elevation</i>	0.0630	<i>Land Cover=3</i>	0.2075
2.	<i>Land Cover=3</i>	0.0560	<i>Geology = 2</i>	0.0290
3.	<i>channel base elevation</i>	0.0481	<i>Land Cover=6</i>	0.0272
4.	<i>LS factor</i>	0.0389	<i>Land Cover=5</i>	0.0235
5.	<i>elevation above channels</i>	0.0389	<i>lithology=6</i>	0.0161
6.	<i>slope angle</i>	0.0338	<i>elevation</i>	0.0158
7.	<i>Land Cover=6</i>	0.0254	<i>channel base elevation</i>	0.0152
8.	<i>aspect</i>	0.0246	<i>Land Cover=4</i>	0.0133
9.	<i>lithology=2</i>	0.0195	<i>Land Cover=1</i>	0.0127
10.	<i>convergence index</i>	0.0192	<i>elevation above channels</i>	0.0122
11.	<i>distance from stream</i>	0.0132	<i>LS factor</i>	0.0116
12.	<i>plan curvature</i>	0.0122	<i>slope angle</i>	0.0078
13.	<i>lithology=6</i>	0.0105	<i>Land Cover=7</i>	0.0060
14.	<i>Land Cover=5</i>	0.0092	<i>aspect</i>	0.0046
15.	<i>profile curvature</i>	0.0076	<i>convergence index</i>	0.0046
16.	<i>Land Cover=7</i>	0.0059	<i>distance from stream</i>	0.0045
17.	<i>TWI</i>	0.0054	<i>plan curvature</i>	0.0039
18.	<i>Land Cover=4</i>	0.0038	<i>profile curvature</i>	0.0031
19.	<i>lithology=4</i>	0.0009	<i>TWI</i>	0.0026
20.	<i>lithology=1</i>	0.0003	<i>lithology=1</i>	0.0009
21.	<i>lithology=5</i>	0.0002	<i>lithology=4</i>	0.0009
22.	<i>Land Cover=2</i>	0.0001	<i>lithology=5</i>	0.0007
23.	<i>Land Cover=1</i>	0.0001	<i>Land Cover=2</i>	0.0006
24.	<i>lithology=3</i>	0.0001	<i>lithology=3</i>	0.0003

#### 6.3.3.1 Model-7c

The curiosity of this model in comparison to the previous case studies lies in leave-one-out learning scheme, where only one, predictive experimenting protocol has been used. As leave-one-out scheme implies, prior to the learning itself, Attribute Selection has had to take place.

Herein, two filters have been used (Tab. 28), but the choice came down to the ranks provided by the *Information Gain*, since it turned out to be more balanced (with lesser extremes) than the *Gain Ratio*, and also provides more logical scenario. Given *IG* ranks have been used as a criterion for the leave-one-out procedure, by eliminating one last-ranked

factor iteratively, after each training cycle, until an obvious convergence of the model performance metrics is achieved<sup>40</sup>.

**Table 29.**  $c$ ,  $\gamma$  parameter combinations for optimization.

$(c, \gamma)$	<i>AUC</i>	$\kappa$ -index	$fn_{rate}$
1,1	0.72	0.56	0.01
10,1	0.78	0.65	0.01
100,1	0.73	0.60	0.01
1,10	0.62	0.33	0.01
10,10	0.69	0.43	0.01
100,10	0.70	0.51	0.01

**Table 30.** Performance of Model-7c-40% with  $c=50$  and  $\gamma=1$  in leave-one-out learning scheme.

training set	accuracy
initial	84.86%
- <i>lithology</i> =3	85.16%
- <i>lithology</i> =3- <i>Land Cover</i> (LC)=1	85.24%
- <i>lithology</i> =3-LC=1-LC=2	85.36%
- <i>lithology</i> =3-LC=1-LC=2- <i>lithology</i> =5	86.77%
- <i>lithology</i> =3-LC=1-LC=2- <i>lithology</i> =5- <i>lithology</i> =1	86.89%
- <i>lithology</i> =3-LC=1-LC=2- <i>lithology</i> =5- <i>lithology</i> =1- <i>lithology</i> =4	86.97%
- <i>lithology</i> =3-LC=1-LC=2- <i>lithology</i> =5- <i>lithology</i> =1- <i>lithology</i> =4-LC=4	86.99%
- <i>lithology</i> =3-LC=1-LC=2- <i>lithology</i> =5- <i>lithology</i> =1- <i>lithology</i> =4-LC=4-TWI	87.01%

The experimenting protocol was based on the Model-7a-33% or Model-7b-40% (predictive modeling scheme). The training split has again been sampled so that it included all the necessary classes of nominal factors (*lithology* and *Land Cover*). This time, parallel direction was more convenient than meridian for splitting the training and testing area manually, because of the spatial propagation of the mentioned categorical inputs. Training area occupied approximately 40% of the total area or 233037 instances. The model could thus be conventionally labeled as Model-7c-40%.

The optimization (over training area) of the Model-7c-40% was very extensive and very time-consuming. It lasted for over 77 hours on conventional machine (i5 Intel Processor on 3.3 GHz and 16 GB RAM of which 3 GB was available for Java emulation of Weka software due to 64-bit OS). Similar pairs of  $c$ ,  $\gamma$  parameters as in Model-7b-40% have been taken into consideration for training, which has been realized by a 10-fold Cross-Validation technique. Combinations with  $\gamma=0.1$  have been excluded due to the poor performance which has been witnessed in a prompt test of the software (Weka and R) over this data. Thus, the parameters have been varied successively for one order of magnitude (Tab. 29). The values of all three evaluators (*AUC*,  $\kappa$ -index and  $fn_{rate}$ ) are rather balanced, but the optimal parameter pair should belong to range  $c=10-100$  and  $\gamma=1-10$ . Some further adjustments revealed that  $c=50$  and  $\gamma=1$  gives the best results, and has been adopted as optimal.

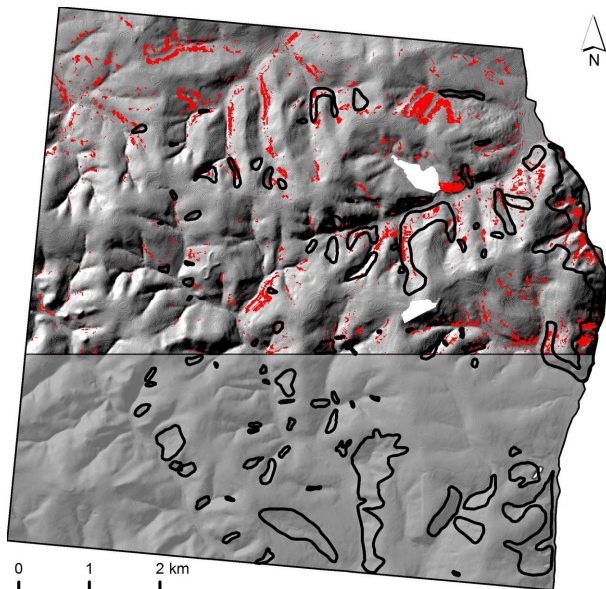
<sup>40</sup> *IG* unlike *GR* favors natural conditions over artificial influence, i.e. morphometric and geological parameters over *Land Cover* classes. However there is some regularity between the ranks since the most of the factors' ranks do not vary that much. Exceptions are some *Land Cover* and *lithology* classes.

The experimenting procedure has been continued by training with adopted  $c$ ,  $\gamma$  over the training area while using a leave-one-out protocol. In each iteration the lowest-ranked Conditioning Factor has been removed from the training set and the according model built over the testing area. Such procedure has been continued until the performance parameters converged to a similar value (close to the highest performance), i.e. until the performance parameters started to drop down. In this particular case, eight iterations, up to  $TWI$  as the last of the removed factors, have been sufficient to reach a convergence (Tab. 30). Hence, the last variant has been adopted as representative for Model-7c-40%.

Other performance parameters did not prove as convincing, since again  $fp/fn_{rates}$  have reached considerable shares (Tab. 31). This can be confirmed by confusion matrix and ROC curve (Fig. 60), where *False Positives* predominate, with 26706 out of 322000 instances misclassified. Visually however, the Model-7c-40% map (Fig. 59) nicely and logically indicates the landslide trends, placing them along potentially unstable slopes of small and deep valleys, along which the landslides (even earth-flows which are not considered in analysis) have been recorded. It gets us back to the old discussion on how should such predictive models be evaluated. It remains unrevealed issue whether a model has been overfitted or it is just that the predictions reach beyond the present Landslide Inventory to some nearer future state. It would take a landslide hazard analysis to further elaborate this question. One possible solution for softening the performance error is  $K_{fuzzy}$ , but even more advanced and customized schemes are required.

**Table 31.** Performance parameters of Model-7c-40%.

Model class	AUC	$\kappa$ -index	accuracy	average $K_{fuzzy}$
landslide	0.57	0.11	15%	0.85
non-landslide	0.57	0.38	94%	



**Figure 59.** Model-7c-40% SVM landslide map (predicted landslides are in red, while actual ones are contoured; the training area is shaded).

One additional very prompt experiment had been carried out. It involved the same set only the roles of training and testing areas have been inverted. Accordingly, a Model-7c-60% has been built over the remaining 40% of the area. The purpose of this test was to challenge the training sample strategy and to establish whether the increase of training sample size yields better results. There has been a very slight increase of performance

(accuracy=88.07%,  $AUC=0.58$ ,  $\kappa$ -index=0.2) leading to a conclusion that inversion, i.e. enlarging of the training area does not make a significant change.

A small web-map, created in the R environment, using “plotGoogleMap package” (Appendix 2) presenting the area and related result of Model-7c-40% is available at:

<http://milosmarjanovic.pbworks.com/w/file/attach/63739326/MyMapHalenkovice.htm>

In conclusion, the method turned out to be relatively accurate, as expected from the experiences drawn from previous case studies, and it also tops the accuracies reached by some other methods, such as bivariate and multivariate statistics. The model can be characterized as underestimating in terms of *landslide* instances, but yet following logical trends of landslide occurrence. Apart from that point, the experiment design was valid (selection of the splits, optimization of the parameters, preprocessing of the inputs were apparently correct) as shown in the inversed test in Model-7c-60%. A serious drawback of the method is its time-consumption, as optimization alone lasts for several days, while the model implementation on training area had required more than 4 hours. Another drawback is the poor data compatibility of the modeling software and GIS platform, where CSV text files had to be communicated throughout. Since the arrangement and preparation of such files at some points had to be manual, the room for user error is introduced. For further notice, it would be interesting to challenge the algorithm with multi-class (multinomial) scenario (e.g. *active*, *dormant*, *suspended*, *fossil*, and other landslide classes could be specified and related to the High-Low susceptibility scale). Also, the study is in its beginning and it might be interesting to extend it methodologically and to compare the results. It is also intended to perform some of the filtering techniques in the post-processing, because illogical errors (such as pixel islands) are apparent and could be easily avoided in this manner.

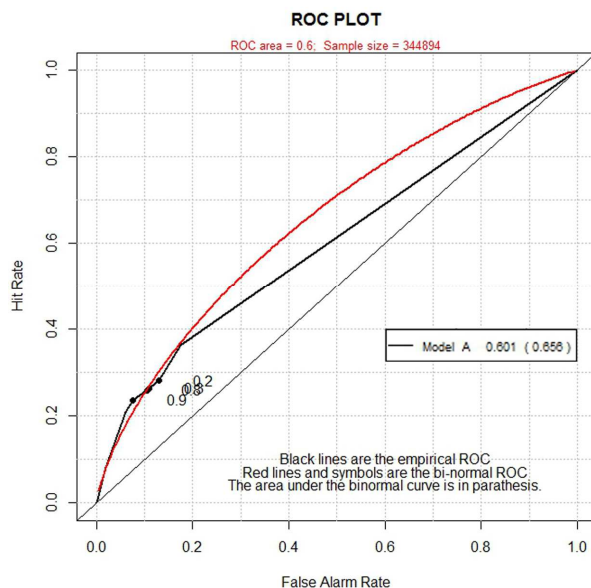
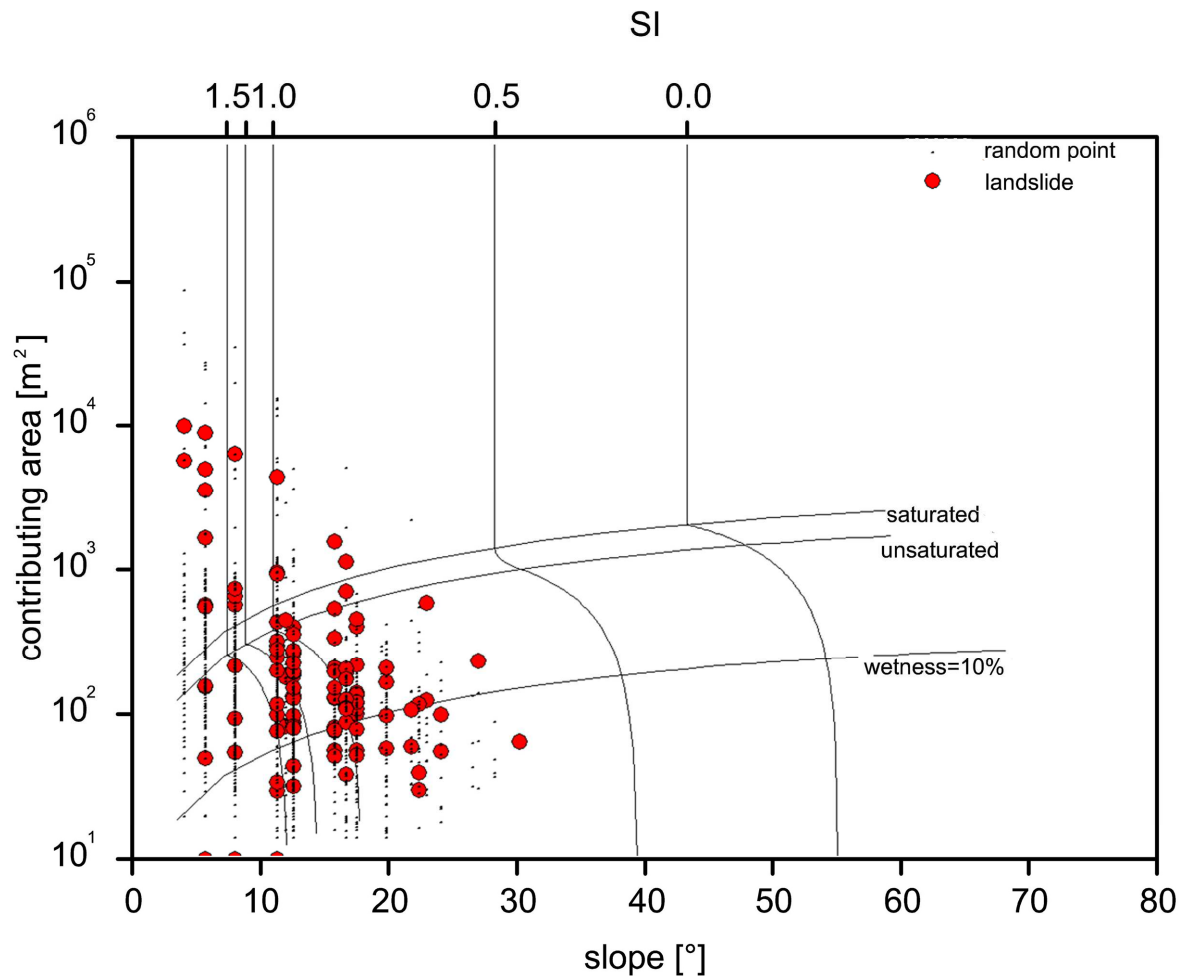


Figure 60. ROC curve of the Model-7c-40%.

### 6.3.3.2 Model 8-c

A deterministic model, which has been estimated suitable for a given case study, has been generated in a GIS environment, using the available geotechnical and other data. The procedure was exactly the same as for building Model-8a (and Model-8b) and it shall not be repeated hereinafter. The only difference is that only two domains have been considered for regionalization (generalization) of geotechnical data (see Chapter 6.3.2.3).



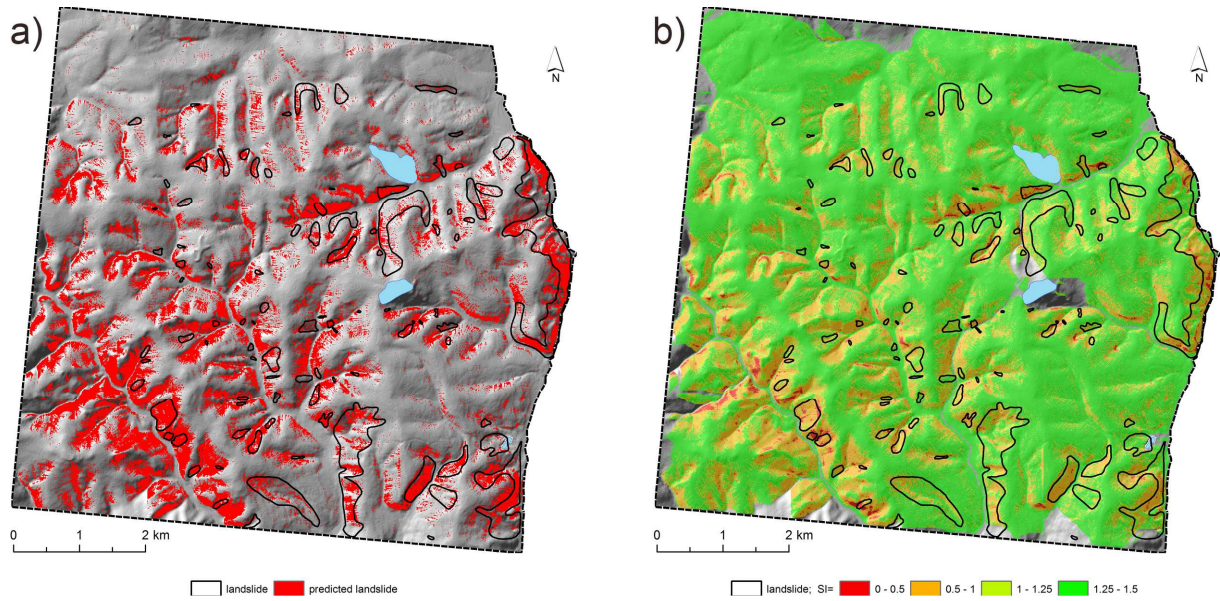


**Figure 61.** SA plot for region 1 (flysch).

Although the data have been based on realistic, laboratory tested samples, some fitting has been carried out, mostly regarding the different  $q/T$  conditions, because these have the highest variances within a single region. For the final model, a classification scheme which represents a result of a particular geotechnical parameter combination has been used (Fig. 61). From the Figure 61. it is obvious that numerous *landslide* instances (red points on the plot) have reached beyond stability limit ( $SI > 1.0$ ), which should not be the case (landslides should have  $SI < 1.0$ ), but after fitting several combinations of available geotechnical parameters and several  $q/T$  regimes, this combination seemed to be the most representative for the model.

The model have had some issues with underestimation and overestimation of the *landslide* instances, assuming that all instances with  $SI$  less than 1.0 represent landslides (the highest susceptibility class). They are logically located along the slope valleys (Fig. 62), but in some valleys they are abundant and in others missing. The performance measurement suggests that there are considerable misclassification errors of both  $fp$  and  $fn$  (very low  $\kappa$ -index values, while  $AUC$  is better balanced), but also a relatively good overall accuracy (Tab. 32).

In conclusion, the model does not seem too promising, but it is not to be underestimated. The prognosis might be rather realistic, in fact it partly agrees with the disputable prognosis made in Model-7c-40% and Model-7c-60%, suggesting that it could be viable after all. Model-8c could hardly be accepted as a stand-alone model. The author recognizes its potential use as an additional input factor for the Machine Learning-based models, which will be exploited in the future work.



**Figure 62.** Model-8c: **a)** Reclassified Model-8c by unifying all *landslide* instances where  $SI < 1$  (actual landslides are contoured); **b)** Original Model-8c (actual landslides are contoured, susceptibility classes with  $SI < 1$  are in red and orange tones.).

**Table 32.** Average performance parameters of Model-8c.

<b>AUC</b>	<b><math>\kappa</math>-index</b>	<b>accuracy</b>	<b><math>K_{fuzzy}</math></b>
0.65	0.14	79.38%	0.62

## 7 Main Achievements

Since the previous chapter has been rather voluminous and the information turned abundant and very detailed, some essential achievements and their relation to the initial research objectives are to be clarified in the following paragraphs. The objectives (structured as in Chapter 2) have been complied by the following achievements:

1. *Exploiting only low-cost data resources (available or open-source topographic, geological, satellite imagery and other repositories) and open source software packages.*

In all of the case studies presented in Chapters 6.1–3.2, the data that have been used were obtained for free. Topographic information, CORINE classification maps, even geological maps, orthophotos and Landslide Inventories are freely available up to a certain scale (1:50000). It is also the case of LANDSAT and other similar multispectral images, which are sometimes available in several time series, but the quality of these images might not always satisfy the requirements. These scales turned sufficient for conducting proposed methodology and fulfilling the Objective 1 of this thesis. In addition, open source solutions, such as SagaGIS, MapWindow, Weka, R, MapComparisonKit and others, have been fully exploited in processing and modeling of these data, which utterly rounds-up the Objective 1.

2. *Inspecting of the phenomena from different case-studies, including similar, but sufficiently different terrains (in order to compare the modeling results and test the capabilities of proposed methodological solutions).*

Selected case studies have been somewhat similar, but still different enough to challenge the proposed methodology from different aspects. For instance, the first and the second case studies are geologically similar, since their landslide occurrences are mostly linked to Tertiary formations of similar type, while the triggers are also somewhat similar (linear erosion dominates in the first case study, while precipitation dominates in the second, but in total they are both tightly related to the landslide occurrence in both study areas). Still, the typology of the landslides is entirely different, since the first study area is typical for the deep-seated earth slides, while in the second study area shallow earth slides and flows dominate. In that respect the second case study better complies with the third case study area, but the geological ambient is entirely different. Such diversity, turned out to be challenging for proposed methods, which is why some techniques have better success than the others. Therefore, it could be said that the Objective 2. has been appreciated consistently throughout this thesis.

3. *Standardizing the data acquisition regarding the data type, scale, preprocessing procedures and so forth (in order to have fully comparable models from different case-studies) using GIS.*

Although the datasets did not contain exactly the same inputs, it is possible to perceive some standard pattern. It implies that each case study must have had several morphometric Conditioning Factors, and at least one hydrological, geological and environmental Conditioning Factor. The only exception was with the deterministic models, which required specific (geotechnical) data inputs that have had to be arbitrarily adjusted within certain limits in order to suit the model. Furthermore, all of the inputs (except the latter) underwent the same processing procedure, as demanded by Machine Learning methods or other used methods, for that matter. These procedures are clearly explained for all of the case studies (see Chapters 6.1–3.2), which leads to a conclusion that the Objective 3 (see Chapter 2) has been fully perceived throughout the thesis.

4. *Implementing a variety of well-known modeling approaches, but also experimenting with the state-of-the-art techniques, advanced methods and unprecedented solutions for landslide assessment using GIS. Resulting models are to present transient relative values over the area, pinpointing landslide-endangered zones and safe zones (which shall be further elaborated).*

The first, pilot case study has been the most extensively elaborated, since there has been no similar investigation performed over this area before. Thus, the entire gamut of proposed methods has been involved, while in the last two case studies, the methods have been intentionally reduced to those which might have led to some new discoveries, which would supplement the previous investigations, conducted by other practitioners. In this sense, the fulfilling of the Objective 4 has been asserted.

5. *Evaluating the results, i.e. the models performance in the most appropriate fashion, obtaining qualitative and quantitative descriptors of the models performance using GIS in combination with statistical tools.*

The evaluation of the individual models in all of the case studies has been always given by several performance parameters, such as accuracy, several types of  $\kappa$ -indices, different error rates, ROC curves and *AUC*, all based on contingency tables (confusion matrices). Nevertheless, the evaluation of the modeling performance has remained problematic, especially for the predictive models (generated by advanced, Machine Learning methods), for a number of problems addressed in discussions that followed every single model (Chapters 6.1–3.3). The most appropriate method for model comparison (see Chapter 8) turned out to be the ROC curve, because it allows qualitative and quantitative evaluation of the model. It needs to be mentioned that some parameters such as  $K_{fuzzy}$  for instance, seemed very promising, but have remained relatively unexploited, mainly because they have not been widely accepted in the community. Having such minor drawbacks in mind, it still could be inferred that the most appropriate fashion for model evaluation has been followed, i.e. that Objective 5 has been practically fulfilled.

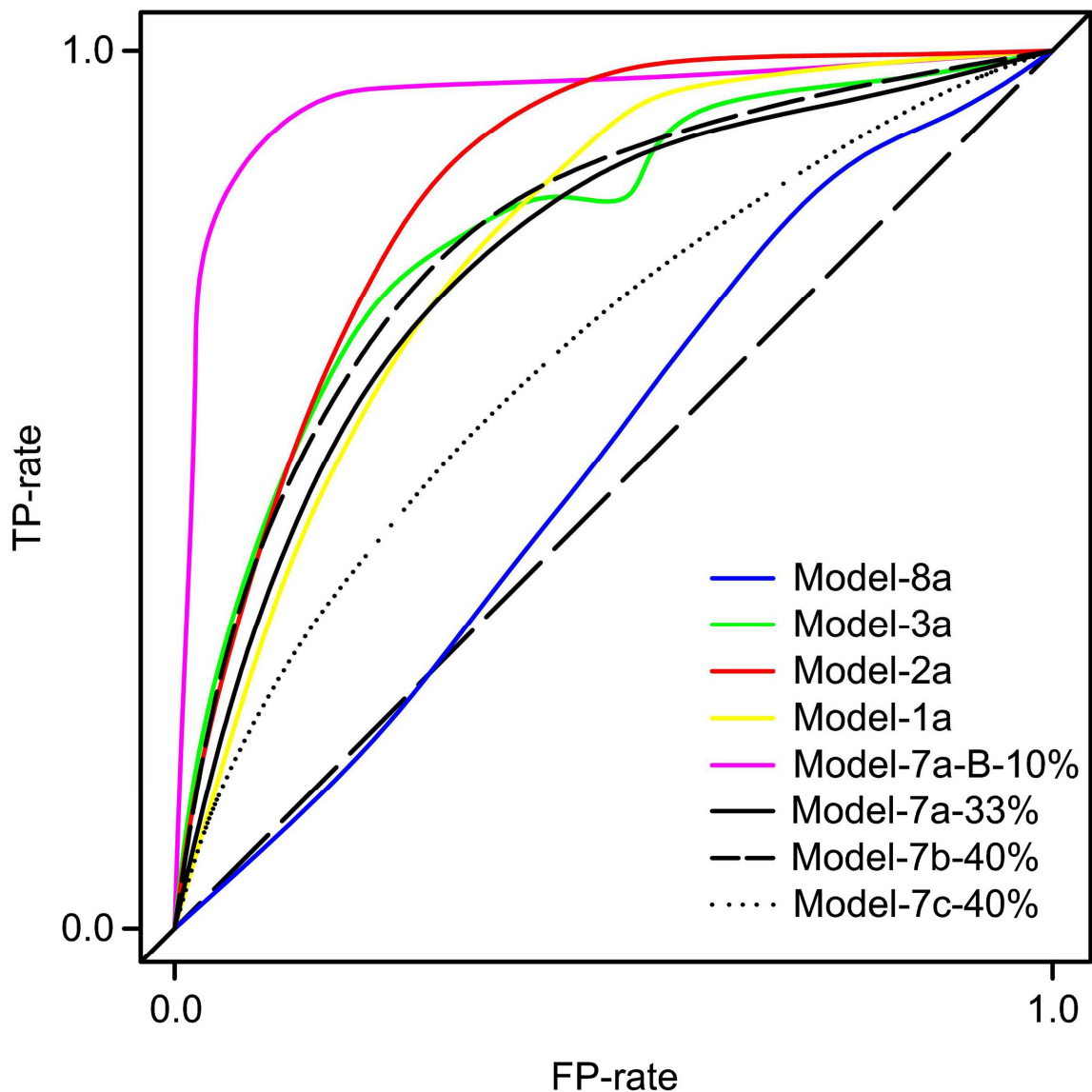
6. *Visualizing and publishing the results in the form of generic maps per each case-study using GIS, and web-GIS and estimating their applicability.*

Visualization of the most of the models has been given by separate maps (see Chapters 6.1–3.3), while some of the insignificant results have not been visualized on purpose. The most interesting models (Appendix 3), i.e. predictive models based on Machine Learning techniques, have been additionally featured as interactive web-maps, and made publically available (Appendix 2). Their applicability is left for discussion of those who find them appealing or useful/useless for their particular needs (planning, modeling, mapping, managing, etc.). Therefore, the final objective of this thesis, the Objective 6, has also been completed.

## 8 General Discussion

Before a final conclusion, it would be appropriate to re-discuss some of the issues and to round-up general impressions driven from the author's experiences related to the presented research. It is also an opportunity to revisit the comparative analysis of presented modeling methods.

Advanced models, based on Machine Learning techniques, have proven their supremacy over more common approaches (heuristic, statistic etc.) for a number of reasons.



**Figure 63.** ROC curve comparison.

Firstly, they tend to perform much better in the case of interpretative models, i.e. landslide susceptibility models, of course, when appropriate sampling strategy for training/testing is set. Continual, 0–1 range false probability models are returned by using intermediate models, created after each iteration, and averaging and normalizing them, not directly as in the case of conventional models. In all of the case studies it turned out that they easily outperformed conventional models (Fig. 63). Their ROC curves tend to reach the peak

performance with relatively low probability thresholds (Fig. 63), meaning that they give more conservative (safer) outputs. They have also shown a good generalization capacity when challenged with multinomial tasks (several landslide classes to discern).

As for the second type of models, i.e. predictive models such as Model-7a-33%, Model-7b-40% and Model-7c-33% (Appendix 3), the comment is not as straightforward. These are all predictive models in their nature, and since they are based on the training/testing protocols, they perfectly simulate potential scenarios (e.g. the situations in which an area does not have Landslide Inventory, but adjacent area does). Thus, simulated models over the testing area represent the actual prediction of the landslide distribution over that area. As previously mentioned, the outcomes are not so clear and the general comment is therefore less meaningful. Some of them have been very successful and very applicable (Model-7a-33%), while the others have been troubled with the overfit problem. Several factors have been recognized as possible causes of the overfit. The training sample size and the way the sample has been selected are crucial, because the overfit appears most likely due to the overabundant *non-landslide* class. The algorithm (SVM, C4.5, Logistic Regression etc.) learns too many wrong relations between *non-landslide* instances and the inputs, tending to produce considerable amount of *False Negative* errors. The algorithms are further challenged when switching from binary case (*landslide vs. non-landslide*) to multi-class case. It is highly likely that the performance would drop even further in such cases, which is why it is advisable to perform both binary and multinomial cases. Furthermore, one needs to consider the processing capacity while searching for the optimal *landslide* population size. It affects the number of instances and indirectly, the scale of the research. From the experiences of the author, it is recommendable that the *landslide* instances occupy at least 10% of the instances, as long as the total number of instances is kept below  $10^6$  points, since it has been shown that upscaling does not necessarily bring better modeling performance. It is also noticed that the Attribute Selection have had minor influence in preventing the overfit (as shown by various *Chi-Square*, *IG* and *GR* examples in this research), so the key for avoiding the overfit remains with the inventory and the sampling strategy of the training/testing split. In general, there are some improvements that might be made to reduce the overfit directly (e.g. by limiting the *non-landslide* class by some additional criteria, in order to make a more balanced training samples or by limiting to specific landslide types if not in discrepancy with the aforementioned *landslide* class population size) or indirectly, by improving the results through the postprocessing filtering.

For more details on the first case study on Fruška Gora Mountain (Serbia), the author suggests the publications which address particular problems related to this research, ranging from common to advanced models and evaluation metrics (Marjanović 2009, Marjanović et al. 2009, Marjanović 2010a, Marjanović 2010b, Marjanović & Caha 2011, Marjanović et al. 2011a, Marjanović et al. 2011c, Marjanović 2013). The second case study, Starča Basin (Croatia), is not as elaborate, but there are some significant references where one can find particular details, which might not be presented in this thesis entirely (Marjanović et al. 2011b). Finally there is the last case study, Halenkovice area (Czech Republic), which is still under research. Some reference has been already made (Marjanović 2012), but the final results are prepared to be published in a special Springer Book Series on Computational Approaches for Urban Environments (Geotechnologies and the Environment series) in 2013 (the author has been invited to contribute to the issue).



## 9 Conclusions

This thesis rounds-off a detailed methodological proposal for mapping landslide susceptibility in non-conventional ways, using some simple and advanced modeling methods. These have been tailored by the according research motifs and objectives, which have been consistently followed. The thesis is savored by three case studies on which the proposed methodology has been fully employed, tested and discussed (Chapter 6.). It outputs a dozen of different interpretable models, which have their drawbacks and benefits and different practical relevance.

### 9.1 Benefits and Drawbacks

Presented models have an apparent linkage between their complexity and their GIS integration possibilities. More complex models are naturally difficult to implement, while simple ones are readily GIS supported, which could be their general benefit at present, because data manipulation outside GIS requires either some manual data handling, either some additional programming effort. Nonetheless it can become a root of some systematic errors made during this kind of handling or while communicating external and GIS software. In addition, complex models, such as Machine Learning-based ones are much more demanding, computationally, i.e. time-wise, thus unsuitable when a quick prediction is needed. Nowadays it is possible to follow-up the hardware technology for affordable price, while software solutions are henceforth directed towards parallel and cloud computing, which should maximize the performance and shortens the processing time. It draws one to a logical conclusion that what are being the complex models at present would eventually become easily deployed models, but then even more complex models will take over with new demands and new challenges posed to the hardware and software solutions. In this particular research it has been inferred that: MS Windows OS is not as computational-friendly as it is user-friendly (unlike Linux for instance); ArcGIS is the most robust GIS platform, but fails to follow up the module development as fast as its open source counterparts; R is very customizable and very flexible, plus it is practically GIS-integrated, but not too user-friendly and not so robust for handling large datasets like Weka does. In brief, a combination of various solutions is still necessary, but holistic solutions are perceivable and R is one solid example of it.

Another issue which has been indicative in the most of the models is the evaluation, so it could be discussed as one of the drawbacks. It is hard to evaluate predictions in the landslide susceptibility scenario (unlike in hazard or risk scenario where prognosis relates to the specified time series) because only present (and past) landslides can be witnessed. Future cannot be accounted for, but it is obvious that all *False Positives* are not necessarily erroneous (having in mind that the spreading of landslide body follows a logical spatial trend) and that the predictive nature of the model should not be suppressed by the strict performance metrics. On the other hand, *False Negatives* should be strictly penalized, because they represent unacceptable error in the landslide assessment framework. Some of the performance measurements (e.g.  $K_{fuzzy}$ ) are smoothing the errors down by taking into account the size of the particular *landslide* class. Since these classes are usually much smaller in size than a *non-landslide* class, it could re-endorse the model which has been underestimated. It is probably the most objective evaluation parameter thus far.

Furthermore, there is an important benefit from downscaling to be discussed. It has been shown that too detailed data (such as data from the last two case studies) can make some problems in predictive modeling. This particularly affects the Machine Learning-based models, which tend to overfit on too detailed data. It is therefore necessary to find the optimal scale (or resolution or level of detail) for a given dataset in order to reach the best possible performance. It is usually recommended to adopt the scale of the inputs, but even more usually these are not coming with the same level of detail. It is also noticed that in some case

studies where the landslides are smaller the performance of all proposed models has suffered an apparent decline. In this case, downscaling would be additional drawback of the model. On the other hand, too much of upscaling would endanger the limits of computing capacity of the hardware/software. Alternative solution would be the tiling of the area into several sub-areas, but it is important to mention that some of the methods (Machine Learning-based ones) would be affected and compromised by such solution. Experiences drawn from this research suggest that the area with one million of points (pixels) is a fair upper limit for the size of the area, while the lower limit could be 100000. These limits apply only to the particular circumstances (particular hardware and software solutions).

Finally it is desirable to once again underline that high quality of input data can guarantee a plausible result, even by using the simplest modeling solutions, while on the other hand, no model, no matter how sophisticated cannot help if the input data are poor in quality.

## 9.2 Applicability

Each kind of the presented models can find some purpose at some level of assessment, in a wide gamut from preliminary to detailed research. Furthermore, a special case of assessment, involving for instance detailed landslide mapping, can substantially benefit from such models. Actually these models are not intended for replacing the conventional mapping, but to supplement it and to be used in the preliminary stage of map development, i.e. in the early stage of research planning. They are thus semi-products of landslide assessment, i.e. intermediate models which are used by the experts to compile a final map. Most of the models can successfully pinpoint the critical areas and guide the practitioners towards more efficient mapping.

In practice, these kinds of models can easily find their purpose in regional, small scale planning, urban planning, strategic planning, but also some preliminary insurance analysis, planning of detailed research or sampling, updating the inventories, tracking changes and so forth.

## 9.3 For Further Notice

The most of the advanced models in this research have shown that much more needs to be done to achieve reliable semi-automated landslide mapping and landslide susceptibility assessment. It particularly concerns SVM or DT-based Machine Learning models, which are far from becoming operative and scrutinized in landslide assessment framework, for now. There are several directions for further improvements and possibilities which unfold from the experience gathered through this research.

First, there is a fundamentally different concept of defining the unit area, i.e. alternating from pixel-based to areal-based approach. Two possibilities are commonly imposed:

- generating a Unique Condition Area or UCA (quasi-homogeneous area) by successively parsing of the input data classes (using for instance raster/vector combine modules),
- using slope or watershed units (generated by various morphometric modules) which has been considered as the most logical in the landslide assessment framework (Guzzetti et al. 2012, van Westen et al. 2006), because landslides are indeed elements of slope configuration.

Since they are both area-based, they could allow creation of the additional synthetic inputs, such as statistical parameters (variances, means, standard deviations, etc.) of other inputs. This would offer the whole new source of relations between the landslide occurrence and the input data. On the other hand, they tend to generalize just as any other choropleth map, and may not be as suitable as smoothly distributed raster models. They are also both

rather subjectively segmented, (especially UCA) since they depend on the ranging choices made for the (ordinal, continuous) inputs.

Finding new resources of inputs is thus one of the milestones for further model refinement. Expert-based, inputs could significantly contribute, especially if they are focusing on geological or engineering-geological terrain features. For instance, inputting geological domains as quasi-homogeneous areas in terms of stratification character (spatial and sedimentological) has been proven useful in landslide assessment (Guzzetti et al. 2012). Unfortunately, such inputs require additional engagement of the experts and resources which can turn insurmountable problem (e.g. generating of geological quasi-homogeneous domains require extensive RS and field techniques and qualified experts to generate it, although there are some trends toward creating simple domains automatically). New inputs could be found by changing the unit area definition, as explained above, but in addition, the inputs can reflect the dependency of neighboring values, i.e. they can emphasize the actual critical points or zones, which are the most indicative in the landslide scenario (landslide crown or toe). Change of the slope morphology is one such example, wherein it is indicative that sudden changes of slope angles are related to the landslide occurrence. Principal Components of input variables might be an effective way of emphasizing the changes in trends of the input variable values. Finally, new inputs can be found as some intermediate models, for instance simple AHP-based models or deterministic models can be fed as additional inputs of the more advanced e.g. Machine Learning-based models. Multi-temporal inputs are also desirable, but rarely available (historical repositories, especially on dating slope displacements, aerial and satellite images in monitoring context, terrestrial monitoring techniques, such as surveying, LiDAR and Radar scanning). They would enable upgrading susceptibility assessment to a hazard or risk framework. Such integrated approach does sound optimal, and with present development of RS systems it is realistic to expect that in couple of decades from now it will be much easier to model landslide hazard and risk.

Another idea for more precise modeling is to include only the landslide source areas as a reference, i.e. to discern between the source and accumulation areas of the landslide body at inventory level, and to train the model only over the areas which have suffered the conditions leading to failure. Accumulation areas do not necessarily face these same conditions, particularly in cases of flows, which are linearly elongated and have relatively long runout distances, hence conditions can change drastically. This would require upscaling to at least 10 m resolution (1:5000–1:10000 scales) and appropriate pilot areas, with relatively large and clearly discernable landslide bodies (preferably flows).

As for the most advanced techniques used in this research (SVM, DT, LR) it is noticeable that multinomial tasks (multiple landslide classes) are not as unfeasible as it first sounds. In fact, it turned out that these complex models perform better in the multinomial than in binary (*landslide vs. non-landslide*) environment. Their generalization power is getting fully exploited and one should not hesitate to challenge the technique if a reasonable population of multiple landslide classes is present in the area (if some of the classes are coming down to a single or a few examples the training would be statistically unlevelled).

Another important challenge would be the using of classifier chains, by combining different techniques and perhaps fuzzifying their combination. More robust and readily post-processed models should be expected therein. It would lead to the fusion of discrete classifiers such as SVM and probabilistic, generative classifiers, such as LR. In fact, there is already a trend for fusing these two branches in Relevance Vector Machines (RVM), which are based on SVM but with the probabilistic output (Tipping 2001).

All these comments are proposing the ideas for improvements of the susceptibility or spatial landslide prediction. Assuming that at one point, the most optimal solution for susceptibility framework will be reached, it would then be entirely new challenge to deal with the hazard and risk frameworks, which is the author's remote objective, from the current stand-point.

## 10 References

- Abe S. (2005). Support Vector Machines for pattern classification. Springer, London, UK, pp. 471.
- Alcántara-Ayala I. (2002) Geomorphology, natural hazards, vulnerability and prevention of natural disasters in developing countries. *Geomorphology*, vol. 47, pp. 107–124.
- Aleotti P., Chowdhury R. (1999) Landslide hazard assessment: summary review and new perspectives. *Bulletin of Engineering Geology and the Environment*, vol. 58, pp. 21-44.
- Bai S.B., Wang J., Lü G.N., Zhou P.G., Hou S.S., Xu S.N. (2010) GIS-based logistic regression for landslide susceptibility mapping of the Zhongxian segment in the Three Gorges area, China. *Geomorphology*, vol. 115, pp. 23-31.
- Barredo J.I., Benavides A., Hervás, J., van Westen C.J. (2000) Comparing heuristic landslide hazard assessment techniques using GIS in the Tirajana basin, Gran Canaria Island, Spain. *International Journal of Applied Earth Observation and Geoinformation*, vol. 2000/1, pp. 9-23.
- Bell F.G. (1999) *Geological Hazards their assessment, avoidance and mitigation*. E & FN Spon, London, UK, pp. 648.
- Bell F.G. (2007) *Engineering Geology*. Elsevier, Oxford, UK, pp. 581.
- Belousov A.I., Verzakov S.A., Von Frese J. (2002). Applicational aspects of support vector machines. *Journal of Chemometrics*, vol. 16/8-10, pp. 482-489.
- Bíl M., Müller I. (2008) The origin of shallow landslides in Moravia (Czech Republic) in the spring of 2006. *Geomorphology*, vol. 99, pp. 246-253.
- Bogdanović Ž., Bugarski D. (1984) The Krčedin river islands. Review of research – Institute of Geography, vol. 14, pp. 33-39.
- Bonham-Carter G. (1994) *Geographic information system for geosciences: Modeling with GIS*. Pergamon, New York, USA, pp. 398.
- Böhner J., McCloy K.R., Strobl J. (2006) SAGA – Analysis and Modelling Applications. Göttinger Geographische Abhandlungen, Göttingen, Germany, pp. 130.
- Brabb E.E., Pampeyan E.H., Bonilla M.G. (1972) Landslide susceptibility in San Mateo County, California. US Geological Survey Miscellaneous Field Studies, Map MF-360, scale 1:62500.
- Brenning A. (2005) Spatial prediction models for landslide hazards: review, comparison and evaluation. *Natural Hazards and Earth System Sciences*, vol. 5, pp. 853-862.
- Brenning A. (2012) Improved spatial analysis and prediction of landslide susceptibility: Practical recommendations. In: *Landslides and Engineered Slopes: Protecting Society through Improved Understanding*, Eberhardt et al. (eds), Taylor & Francis Group, London, UK, pp. 789-794.
- Brunetti M.T., Peruccacci S., Rossi M., Luciani S., Valigi D., Guzzetti F. (2010) Rainfall thresholds for the possible occurrence of landslides in Italy. *Natural Hazards and Earth System Science*, vol. 10/3, pp. 447-458.
- Burges C.J.C. (1998) A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, vol. 2/1, pp. 121-167.
- Caniani D., Pascale S., Sdao F., Sole A.: Neural networks and landslide susceptibility: a case study of the urban area of Potenza. *Natural Hazards*. vol. 45, pp. 55-72.
- Carrara A., Cardinali M., Detti R., Guzzetti F., Pasqui V., Reichenbach P. (1991) GIS techniques and statistical models in evaluating landslide hazard. *Earth Surface Processes & Landforms*, vol. 16/5, pp. 427-445.
- Carrara A., Bitelli G., Carla R., (1997) Comparison of techniques for generating digital terrain models from contour lines. *International Journal of GIS*, vol. 11/5, pp. 451–473.
- Carrara A., Pike R. (2008) GIS technology and models for assessing landslide hazard and risk. *Geomorphology*, vol. 94, pp. 257-260.
- Cascini L. (2008) Applicability of landslide susceptibility and hazard zoning at different scales. *Engineering Geology*, vol. 102/3-4, pp. 164-177.
- Cascini L., Calvellido M., Grimaldi G.M. (2010a) Groundwater modeling for the analysis of active slow-moving landslides. *Journal of Geotechnical and Geoenvironmental Engineering*. vol. 136/9, pp. 1220-1230.
- Cascini L., Cuomo S., Pastor M., Sorbino G. (2010b) Modeling of rainfall-induced shallow landslides of the flow-type. *Journal of Geotechnical and Geoenvironmental Engineering*, vol. 136/1, pp. 85-98.
- Cascini L., Cuomo S., Della Sala M. (2011) Spatial and temporal occurrence of rainfall-induced shallow landslides of flow type: A case of Sarno-Quindici, Italy. *Geomorphology*, vol. 126/1-2, pp.148-158.

- Chacón J., Irigaray C., Fernández T., El Hamdouni R. (2006) Engineering geology maps: landslides and geographical information systems. *Bulletin of Engineering Geology and the Environment*, vol. 65, pp. 341-411.
- Chamaptiray P. K., Dimri S., Lakhera R. C., Sati S. (2006) Fuzzy-based method of landslide hazard assessment in active seismic zone of Himalaya. *Landslides*, vol. 4, pp. 101-111.
- Corominas J., Copons R., Moya J., Vilaplana J.M., Altimir J., Amigó J. (2005) Quantitative assessment of the residual risk in a rockfall protected area. *Landslides*, vol. 2/4, pp. 343-357.
- Corominas J., Moya J. (2010) Contribution of dendrochronology to the determination of magnitude-frequency relationships for landslides. *Geomorphology*, vol. 124/3-4, pp. 137-149.
- Cotecchia F., Santaloia F., Lollino P., Vitone C., Mitaritonna G. (2010) Deterministic landslide hazard assessment at regional scale. *Geotechnical Special Publication*, vol. 199, pp. 3130-3139.
- Cristiani N., Shawe-Taylor J. (2000) Support Vector Machines and other kernel-based learning methods. Cambridge University Press, Cambridge, UK, pp. 189.
- Demek J., Hradecký J., Kirchner K., Pánek T., Létal A., Smolová I. (2012) Recent landform evolution in the Moravian-Silesian Carpathians (Czech Republic). In: *Recent Landform Evolution*, Lóczy, Stankoviansky, Kotarba (eds.), Springer, Dordrecht, The Netherlands, pp. 103-140.
- Dungan J.C. (2001) Scaling up and scaling down: The relevance of the support effect on remote sensing of vegetation. In: *Modelling and scale in Geographical Information Science*, Tate & Atkinson (eds.), John Wiley & Sons, New York, USA, pp. 121-235.
- Dimitrijević M.D. (1997) *Geology of Yugoslavia*. Geological Institute GEMINI, Belgrade, Serbia, pp. 187.
- Einstein H.H. (1988) Special lecture: landslide risk assessment procedure. In: *Landslides – Proceedings of the 5<sup>th</sup> International Symposium on Landslides*, vol. 2, 10-15 July, Lausanne, Switzerland, pp. 1075-1090.
- Ercanoglu M., Gokceoglu C. (2006) Use of fuzzy relation to produce landslide susceptibility map of a landslide prone area (West Black Sea Region, Turkey). *Engineering Geology*, vol. 75, pp. 229-250.
- Ercanoglu M., Kasmer O., Temiz N. (2008) Adaptation and comparison of expert opinion to analytical hierarchy process for landslide susceptibility mapping. *Bulletin of Engineering Geology and the Environment*, vol. 67/4, pp. 565-578.
- Erener A., Düzgün H. (2010) Improvement of statistical landslide susceptibility mapping by using spatial and global regression methods in the case of More and Romsdal (Norway). *Landslides*, vol. 7/1, pp. 55-68.
- Ermini L., Catani F., Casagli N. (2005) Artificial neural networks applied to landslide susceptibility assessment. *Geomorphology*, vol. 66, pp. 327-343.
- Falaschi F., Giacomelli F., Fedrici P.R., Pucinelli A., Avato G.D'A., Pochini A., Ribolini A. (2009) Logistic regression versus artificial neural networks: landslide susceptibility evaluation in a sample area of the Serchio River valley, Italy. *Natural Hazards*, vol. 50, pp. 551-569.
- Fawcett T. (2006) An introduction to ROC analysis. *Pattern Recognition Letters*, vol. 27, pp. 861-874.
- Fell R. (1994) Landslide risk assessment and acceptable risk. *Canadian Geotechnical Journal*, vol. 31, pp. 261-272.
- Fell R., Glastonbury J., Hunter G. (2007) Rapid landslides: The importance of understanding mechanisms and rupture surface mechanics. *Quarterly Journal of Engineering Geology and Hydrogeology*, vol. 40/1, pp. 9-27.
- Fell R., Corominas J., Bonnard C., Cascini L., Leroi E., Savage W. (2008) Guidelines for landslide susceptibility, hazard and risk zoning for land use planning. *Engineering Geology*, vol. 102, pp. 83-84.
- Ferretti A., Monti-Guarnieri A., Prati C., Rocca F., Massonnet D. (2007) *InSAR Principles: Guidelines for SAR Interferometry Processing and Interpretation (TM-19)*. ESA Publications, Noordwijk, The Netherlands, pp. 226.
- Fielding A.H., Bell J.F. (1997) A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environmental Conservation*, vol. 24, pp. 38-49.
- Frattini P., Crosta G., Carrara A. (2010) Techniques for evaluating performance of landslide susceptibility models. *Engineering Geology*, vol. 111, pp. 62-72.
- Geniest C., Rivest L.P. (1994) A statistical look at Saaty's method of estimating pairwise preferences expressed on a ratio scale. *Journal of Mathematical Psychology*, vol. 38, pp. 477-496.
- Gerath R., Jakob M., Mitchell P., Van Dine D. (2010) Guidelines for legislated landslide assessment for proposed residential developments in BC. Association of Professional Engineers and Geoscientists of British Columbia (APEGBC), British Columbia, Canada, pp. 75.

- Glenn E.P., Huete A.R., Nagler P.L., Nelson S.G. (2008) Relationship between remotely sensed Vegetation Indices, canopy attributes and plant physiological processes: what Vegetation Indices can and cannot tell us about the landscape. *Sensors*, vol. 8, pp. 2136-2160.
- Goh T.C.A., Goh S.H. (2007) Support vector machines: Their use in geotechnical engineering as illustrated using seismic liquefaction data. *Computers and Geotechnics*, vol. 34, pp. 410-421.
- Gokceoglu C., Sezer E. (2009) A statistical assessment on international landslide literature (1945–2008). *Landslides*, vol. 6, pp. 345-351.
- Guzzetti F., Carrara A., Cardinali M., Reichenbach P. (1999) Landslide hazard evaluation: a review of current techniques and their application in a multi-scale study, Central Italy. *Geomorphology*, vol. 31, pp. 181-216.
- Guzzetti F., Cardinali M., Reichenbach P., Carrara A. (2000) Comparing landslide maps: A case study in the upper Tiber River basin, central Italy. *Environmental Management*, vol. 25/3, pp. 247-263.
- Guzzetti F., Galli M., Reichenbach P., Ardizzone F., Cardinali M. (2006) Landslide hazard assessment in the Collazzone area, Umbria, central Italy. *Natural Hazards and Earth System Science*. vol. 6/1, pp. 115-131.
- Guzzetti F., Peruccacci S., Rossi M., Stark C.P. (2007) Rainfall thresholds for the initiation of landslides in central and southern Europe. *Meteorology and Atmospheric Physics*, vol. 98/3-4, pp. 239-267.
- Guzzetti F., Peruccacci S., Rossi M., Stark C.P. (2008) The rainfall intensity-duration control of shallow landslides and debris flows: An update. *Landslides*, vol. 5/1, pp. 3-17.
- Guzzetti F., Manunta M., Ardizzone F., Pepe A., Cardinali M., Zeni G., Reichenbach P., Lanari R. (2009) Analysis of ground deformation detected using the SBAS-DInSAR technique in Umbria, Central Italy. *Pure and Applied Geophysics*, vol. 166/8-9, pp. 1425-1459.
- Guzzetti F., Mondini A.C., Cardinali M., Fiorucci F., Santangelo M., Chang K.T. (2012) Landslide inventory maps: New tools for an old problem. *Earth-Science Reviews*, vol. 112, pp. 42–66.
- Hagen, A. (2003) Fuzzy set approach to assessing similarity of categorical maps. *International Journal of Geographical Information Science*, vol. 17/3, pp. 235-249.
- Hungr O., Leroueil S., Picarelli L. (2012) Varnes classification of landslide types, an update. In: *Landslides and Engineered Slopes: Protecting Society through Improved Understanding*, Eberhardt et al. (eds), Taylor & Francis Group, London, UK, pp. 47-58.
- Hwang S.G., Guevarra I.F., Yu B.O. (2009) Slope failure prediction using a decision tree: A case study of engineered slopes in South Korea. *Engineering Geology*, vol. 104, pp. 126-134.
- Janjić M., (1962) *Engineering geology Characteristics of Terrains of National Republic of Serbia* (in Serbian). Nučna knjiga, Belgrade, Serbia, pp. 352.
- Jiang H., Eastman J. R. (2000) Application of fuzzy measures in multi-criteria evaluation in GIS. *International Journal of Geographical Science*, vol. 14, pp. 173-184.
- Kanevski M., Pozdnoukhov A., Timonin V. (2009) *Machine Learning for Spatial Environmental Data: Theory, Applications and Software*. EPFL Press, Lausanne, Switzerland, pp. 368.
- Kanungo D. P., Arora M. K., Sarkar S., Gupta R. P. (2006) A comparative study of conventional, ANN black box, fuzzy and combined neural and fuzzy weighting procedures for landslide susceptibility zonation in Darjeeling Himalayas. *Engineering Geology*, vol. 347-366.
- Kanungo D. P., Arora M. K., Gupta R. P., Sarkar S. (2008) Landslide risk assessment using concepts of danger pixels and fuzzy set theory in Darjeeling Himalayas. *Landslides*, vol. 5. pp. 407-416.
- Kanungo D. P., Arora M. K., Sarkar S., Gupta R. P. (2009) A fuzzy set based approach for integration of thematic maps for landslide susceptibility zonation. *Georisk*, vol. 3. pp. 30-43.
- Kecman V. (2005) Support Vector Machines – an introduction. In: *Support Vector Machines: Theory and applications*, Wang (ed.), Springer, Berlin, Germany, pp. 1-48.
- Kianička J., Čapková I. (2005) Using SAR Interferometry for Detecting Landslides and Subsidence in the Coal Basin in Northern Bohemia. In: *Proceedings of GIS Ostrava 2005*, 23-26. January, Ostrava, Czech Republic, pp. 1-11.
- Kilibarda M., Bajat B. (2012) PlotGoogleMaps: The R-based web-mapping tool for thematic spatial data. *Geomatica*, vol. 66/ 1, pp. 283-291.
- Kircher K., Krejčí O., Máčka Z., Bíl M. (2000) Slope deformations in eastern Moravia, Vsetín District (Outer Western Carpathians). *Acta Universitatis Carolinae*, vol. 35, pp. 133-143.
- Klimeš J. (2008a) Use of the deterministic approach for the landslide susceptibility mapping, Vsetínské vrchy Highland, Czechia. *Geografie-Sbornik CGS*, vol. 113/1, pp. 48-60.
- Klimeš J. (2008b) Analysis of preparatory factors of landslides, Vsetínské Vrchy Highland, Czech Republic. *Acta Research Reports*, vol. 17, pp. 47-53.



- Klimeš J., Baroň I., Pánek T., Kosačik T., Burda J., Kresta F., Hradecký J. (2009) Investigation of recent catastrophic landslides in the flysch belt of Outer Western Carpathians (Czech Republic): progress towards better hazard assessment. *Natural Hazards and Earth System Sciences*, vol. 9, pp 119-128.
- Klimeš J., Novotný R. (2011) Landslide susceptibility assessment in urbanized areas: Example from Flysch Carpathians, Czech Republic. *Acta Geodynamica et Geomaterialia*, vol. 8/4, pp. 443-452.
- Komac M. (2003) Geohazard map of the Central Slovenia – the mathematical approach to landslide prediction. In: *Proceedings of the 3rd International Conference on GIS for Earth Science Applications*, 26-29. May, Ljubljana, Slovenia, pp. 1-5.
- Komac M. (2005) Napoved verjetnosti pojavljanja plazov z analizo satelitskih in drugih prostornih podatkov. *Geološki Zavod Slovenije*, Ljubljana, Slovenia, pp. 232.
- Komac M. (2006) A landslide susceptibility model using the Analytical Hierarchy Process method and multivariate statistics in perialpine Slovenia. *Geomorphology*, vol. 74/1-4, pp. 17-28.
- Krejčí O., Kycl P., Baroň I., Šikula J., Krejčí Z., Kašperáková D. (2008) Přehled provedených prací. *Czech Geological Survey*, Prague, Czech Republic, pp. 114.
- Laininen P., Hämäläinen R. (2003) Analyzing AHP-matrices by regression. *European Journal of Operational Research*, vol. 148, pp. 514-524.
- Landis J., Koch G.G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, vol. 33/1, pp. 159-174.
- Lee E.M., Jones D.K.C. (2004) *Landslide risk assessment*. Thomas Telford Publishing, London, UK, pp. 454.
- Lee S., Ryu J.H., Won J.S., Park H.J. (2004) Determination and verification of weights for landslide susceptibility mapping using an artificial neural network. *Engineering Geology*. vol. 71, pp. 289-302.
- Lee S., Ryu J.H., Kim L.S. (2007) Landslide susceptibility analysis and its verification using likelihood ratio, logistic regression, and artificial neural network models: case study of Youngin, Korea. *Landslides*, vol. 4, pp. 327-338.
- Le Roueil S, Vaunat J, Locat J, Lee H, Faure R (1996) Geotechnical characterization of slope movements. In: *Landslides*, vol. 1, Senneset (ed.), Balkema, Rotterdam, The Netherlands, pp. 53-47.
- Lynn M., Bobrowsky P. (2008) *The landslide handbook – A guide to understanding landslides*. U.S. Geological Survey, Reston, Virginia, USA, pp. 129.
- Malamud B.D., Turcotte D.L., Guzzetti F., Reichenbach P. (2004) Landslide inventories and their statistical properties. *Earth Surface Processes and Landforms*, vol. 29/ 6, pp. 687-711.
- Marinoni O. (2004) Implementation of the analytical hierarchy process with VBA in ArcGIS. *Computers and Geosciences*, vol. 30, pp. 637-646.
- Marjanović M. (2009) Landslide susceptibility modelling: a case study on Fruška Gora Mountain, Serbia. *Geomorphologia Slovaca et Bohemica*, vol. 1/2009, pp. 29-42.
- Marjanović M., Bajat B, Kovačević M. (2009) Landslide susceptibility assessment with machine learning algorithms. In: *Proceedings of International Conference on Intelligent Networking and Collaborative Systems (INCoS) 2009*, 4-6 of November, Barcelona, Spain, pp. 273-278.
- Marjanović M. (2010a) Machine learning methods for landslide susceptibility modeling. In: *Geospatial Crossroads@GI\_forum'10*, Car, Griesebner, Strobl (eds) – *Proceedings of 2010 GI\_forum*, 6-9 July, Salzburg, Austria, pp. 150-159.
- Marjanović M. (2010b) Regional scale landslide susceptibility analysis using different GIS-based approaches. In: *Geologically Active*, Williams et al. (eds), *Proceedings of 11th IAEG Congress*, 5-10 September, Auckland, New Zealand, pp. 435-442.
- Marjanović M., Caha J. (2011) Fuzzy approach to landslide susceptibility zonation. In: *Proceedings of the Annual International Workshop on Databases, TExts, Specifications and Objects (DATESO) 2011*, 20. April, Písek, Czech Republic, pp. 181-195.
- Marjanović M., Kovačević M., Bajat B, Voženílek V. (2011a) Landslide susceptibility assessment using SVM machine learning algorithm. *Engineering Geology*, vol. 123, pp. 225-234.
- Marjanović M., Kovačević M., Bajat B, Mihalić S., Abolmasov B. (2011b) Landslide assessment of the Starča basin (Croatia) using machine learning algorithms. *Acta Geotechnica Slovenica*, vol. 2011/2, pp. 45-55.
- Marjanović M., Kovačević M., Bajat B, Voženílek V., Marek L. (2011c) Využití klasifikačních algoritmů metod strojového učení pro účely prostorového modelování. In: *Metody umělé inteligence v geoinformatice*, Voženílek, Dvorský, Húsek (eds), *Univerzita Palackého v Olomouci*, Olomouc, Czech Republic, pp. 185.

- Marjanović M. (2012) Advanced landslide assessment of the Halenkovice experimental site. In: Proceedings of the 1st InDOG Doctoral Conference 2012, 29th October – 1st November, Olomouc, Czech Republic, pp. 38-41.
- Marjanović M. (2013) Comparing the performance of different landslide susceptibility models in ROC space. In: *Landslide Science and Practice – Volume 1: Landslide Inventory and Susceptibility and Hazard Zoning*, Margottini, Canuti, Sassa (eds), Springer, Berlin, Germany, pp. 607.
- Maskrey A., Juneja S., Peduzzi P., Schaerpf C. – ISDR group of authors (2009) Global assessment report on disaster risk reduction (2009): Risk and poverty in a changing climate. United Nations, Geneva, Switzerland, pp. 207.
- Mavrouli O., Corominas J., Wartman, J. (2009) Methodology to evaluate rock slope stability under seismic conditions at Solá de Santa Coloma, Andorra. *Natural Hazards and Earth System Science*, vol. 9/6, pp. 1763-1773.
- Mihalić S., Oštrić M., Vujnović, T. (2008) A Landslide susceptibility mapping in the Starca Basin (Croatia, Europe). In: Proceedings of the 2nd European Conference of International Association for Engineering Geology (EUROENGE), 15-20. September, Madrid, Spain, pp. 1-7.
- Mitchell T.M. (1997) *Machine Learning*. McGraw Hill, New York, USA, pp. 414.
- Mondini A.C., Guzzetti F., Reichenbach P., Rossi M., Cardinali M., Ardizzone F. (2011) Semi-automatic recognition and mapping of rainfall induced shallow landslides using optical satellite images. *Remote Sensing of Environment*, vol. 115, pp. 1743-1757.
- Montgomery D. R., Dietrich W. E. (1994) A physically-based model for topographic control on shallow landsliding. *Water Resources Research*, vol. 30/4, pp. 1153-1171.
- Nadim F., Kjekstad O., Peduzzi P., Herold C., Jaedicke C. (2006) Global landslide and avalanche hotspots. *Landslides*, vol. 3, pp. 159-173.
- Nefeslioglu H.A., Gokceoglu C., Sonmez H. (2008) An assessment on the use of logistic regression and artificial neural networks with different sampling strategies for the preparation of landslide susceptibility maps. *Engineering Geology*. vol. 97, pp. 171-191.
- Nestorov D., Protić I. (2009) CORINE Land Cover Mapping in Serbia. *Građevinska Knjiga*, Belgrade, Serbia, pp. 184.
- Olaya V. (2004) *A gentle introduction to SAGA GIS*. Olaya Victor, Madrid, Spain, pp. 202.
- Ohlemacher G.C. (2007) Plan curvature and landslide probability in regions dominated by earth flows and earth slides. *Engineering Geology*, vol. 91, pp. 117–134.
- Pack R.T, Tarboton D.G, Goodwin C.N. (2001) Assessing Terrain Stability in a GIS using SINMAP. In: Proceedings of 15th annual GIS conference, 19-22 February, Vancouver, Canada, pp. 56-68.
- Pánek T., Tábořík P., Klimeš J., Komárková V., Hradecký J., Šťastný M. (2011a) Deep-seated gravitational slope deformations in the highest parts of the Czech Flysch Carpathians: Evolutionary model based on kinematic analysis, electrical imaging and trenching. *Geomorphology*, vol. 129, pp. 92–112.
- Pánek T., Brázdil R., Klimeš J., Smolková V., Hradecký J., Zahradníček P. (2011b) Rainfall-induced landslide event of May 2010 in the eastern part of the Czech Republic. *Landslides*, vol. 8, pp. 507–516.
- Pavlović R., Lokin P., Trivić, B. (2005) Geological conditions of rational usage and preservation of the Fruška Gora Mountain area (in Serbian), unpublished, Department for Remote Sensing and Structural Geology, University of Belgrade, Serbia.
- Petley D. (2012) Global patterns of loss of life from landslides. *Geology*, vol. 40/10, pp. 927–930.
- Quinlan J.R. (1993) *C4.5: Programs for Machine Learning*. Morgan-Caufman, San Mateo, USA, pp. 302.
- Ravi G. (2002) *Remote sensing Geology*. Springer, Berlin, Germany, pp. 655.
- Regmi N. R., Giardino J. R., Vitek J. D. (2010) Assessing susceptibility to landslides: Using models to understand observed changes in slopes. *Geomorphology*. vol. 122/1-2, pp. 25-38.
- Rossi M., Guzzetti F., Reichenbach P., Mondini A.C., Peruccacci S. (2010) Optimal landslide susceptibility zonation based on multiple forecasts. *Geomorphology*, vol. 114, pp. 129-142.
- Rybář J., Novotný J. (2005) Vliv klimatogenních faktorů na stabilitu přirozených a antropogenních svahů. *Zpravodaj Hnědé uhlí*, vol. 3, pp. 13-28.
- Saaty T.L. (1980) *Analytical Hierarchy Process*. McGraw-Hill, New York, USA, pp. 287.
- Saaty T.L. (2003) Decision-making with the AHP: Why is the principal eigenvector necessary. *European Journal of Operational Research*, vol. 145, pp. 85-91.
- Saito H., Nakayama D., Matsuyama H. (2009) Comparison of landslide susceptibility based on a decision-tree model and actual landslide occurrence: The Akaishi Mountains, Japan. *Geomorphology*, vol. 109, pp. 108-121.

- Samui P., Lansivaara T., Kim D. (2011) Utilization relevance vector machine for slope reliability analysis. *Applied Soft Computing*, vol. 11/5, pp. 4036-4040.
- Santangelo M., Cardinali M., Rossi M., Mondini A.C., Guzzetti F. (2010) Remote landslide mapping using a laser rangefinder binocular and GPS. *Natural Hazards and Earth System Science*, vol. 10/12, pp. 2539-2546.
- Savvaidis P.D. (2003) Existing landslide monitoring systems and techniques. In: *From Stars to Earth and Culture*, Dermanis (ed.), The Aristotle University of Thessaloniki, Thessaloniki, Greece, pp. 242-258.
- Smirnov A., Boisvert E., Paradis S. J. (2008) Support vector machines for 3D modeling from sparse geological information of various origins. *Computers and Geosciences*, vol. 34, pp. 127-143.
- Sorbino G., Sica C., Cascini L. (2010) Susceptibility analysis of shallow landslides source areas using physically based models. *Natural Hazards*, vol. 53/2, pp. 313-332.
- Sorbino G., Migliaro G., Foresta V. (2011) Laboratory investigations on static liquefaction potential of pyroclastic soils involved in rainfall-induced landslides of the flow-type.
- Srivastava V., Srivastava H. B., Lakhera R. C. (2010) Fuzzy gamma based geomatic modeling for landslide hazard susceptibility in a part of Tons river valley, northwest Himalaya, India. *Geomatics, Natural Hazards and Risk*, vol. 1, pp. 225-242.
- Süzen M.L., Doyuran V. (2004) A comparison of the GIS based landslide susceptibility assessment methods: Multivariate versus bivariate. *Environmental Geology*, vol. 45/5, pp. 665-679.
- Tangestani M. H. (2004) Landslide susceptibility mapping using the fuzzy gamma approach in a GIS, Kakan catchment area, southwest Iran. *Australian Journal of Earth Sciences*, vol. 51/3, pp. 439-450.
- Tipping M.E. (2001) Sparse Bayesian Learning and the Relevance Vector Machine. *Journal of Machine Learning Research*, vol. 1, pp. 211-244.
- Turner K., Shuster L.R. (1996) Landslides investigation and mitigation, special report 247. Transportation Research Board, Washington, USA, pp. 675.
- van Westen C.J., Rengers N., Soeters R. (2003) Use of geomorphological information in indirect landslide susceptibility assessment. *Natural Hazards*, vol. 30/3, pp. 399-419.
- van Westen C.J., van Asch T.W.J., Soeters R. (2006) Landslide hazard and risk zonation—why is it still so difficult?. *Bulletin of Engineering Geology and the Environment*, vol. 65, pp. 167-184.
- Varmuza K., Filzmoser P. (2009) Introduction to multivariate statistical analysis in chemometrics. Taylor & Francis, Boca Raton, USA, pp. 321.
- Varnes D.J. (1984) Landslide Hazard Zonation: A Review of Principles and Practice. International Association for Engineering Geology, Paris, France, pp. 63.
- Wang W., Xie C., Du X. (2009) Landslides susceptibility mapping in Guizhou province based on fuzzy sets theory. *Mining Science and Technology*. vol. 19/3, pp. 399-404.
- Wilson J.P., Gallant G.C. (2000) *Terrain analysis: Principles and applications*. John Wiley & Sons, New York, USA, pp. 479.
- Witten I.H., Frank E., Hall M.A. (2011) *Data Mining Practical Machine Learning Tools and Techniques*. Elsevier, Burlington, USA, pp. 929.
- WP/WLI — International Geotechnical Societies' UNESCO Working Party on World Landslide Inventory (1995) A suggested method for describing the rate of movement of a landslide. *International Association Engineering Geology Bulletin*, vol. 52, pp. 75-78.
- Xu C., Dai F., Xu X., Hsi Lee Y. (2012a) GIS-based support vector machine modeling of earthquake-triggered landslide susceptibility in the Jianjiang River watershed, China. *Geomorphology*, vol. 145-146, pp. 70-80.
- Xu C., Xu X., Dai F., Saraf A.K. (2012b) Comparison of different models for susceptibility mapping of earthquake triggered landslides related with the 2008 Wenchuan earthquake in China. *Computers & Geosciences*, vol. 46, pp. 317-329.
- Yao X., Dai F.C. (2006) Support vector machine modeling of landslide susceptibility using GIS: A case study. In: *Proceedings of the 10th IAEG conference*, 6-10 September, Nottingham, UK, pp. 1-12.
- Yao X., Tham L. G., Dai F. C. (2008) Landslide susceptibility mapping based on support vector machine: A case study on natural slopes of Hong Kong, China. *Geomorphology*, vol. 101, pp. 572-582.
- Yilmaz I. (2009) Comparison of landslide susceptibility mapping methodologies for Koyulhisar, Turkey: conditional probability, logistic regression, artificial neural networks, and support vector machine. *Environmental Earth*, vol. 61/4, pp. 821-836.
- Yuan L., Zhang Y. (2006) Debris flow hazard assessment based on support vector machine. *Wuhan University Journal of Natural Sciences*, vol. 14/4, pp. 897-900.
- Zadeh L. A. (1965) Fuzzy sets. *Information and Control*, vol. 8/3, pp. 338-353.

- Záruba Q., Menci V. (1987) Sesuvy a zabezpečování svahů. Academia, Prague, Czech Republic pp. 340.
- Zhao H. (2008) Slope reliability analysis using a Support Vector Machine. Computers and Geotechnics, vol. 35, pp. 459-467.

## Complete List of Author's Publications

- Marjanović M.** (2009) Landslide susceptibility modelling: a case study on Fruška Gora Mountain, Serbia. *Geomorphologia Slovaca et Bohemica*, vol. 1/2009, pp. 29-42, ISSN: 1337-6799.
- Marjanović M.** (2009) Landslide susceptibility modelling: Case study on Fruška gora mountain. In: *Proceedings of Svahové deformace a pseudokarst 2009*, 13-15 of May, Vsetín, Czech Republic, (CD only).
- Marjanović M.**, Bajat B, Kovačević M. (2009) Landslide susceptibility assessment with machine learning algorithms. In: *Proceedings of International Conference on Intelligent Networking and Collaborative Systems (INCoS) 2009*, 4-6 of November, Barcelona, Spain, pp. 273-278, ISBN: 978-0-7695-3858-7.
- Marjanović M.** (2010a) Machine learning methods for landslide susceptibility modeling. In: *Geospatial Crossroads@GI\_forum'10*, Car, Griesebner, Strobl (eds) – *Proceedings of 2010 GI\_forum*, 6-9 July, Salzburg, Austria, pp. 150-159, ISBN: 978-3-87907-496-9.
- Marjanović M.** (2010b) Regional scale landslide susceptibility analysis using different GIS-based approaches. In: *Geologically Active*, Williams et al. (eds), *Proceedings of 11th IAEG Congress*, 5-10 September, Auckland, New Zealand, pp. 435-442, ISBN: 978-0-415-60034-7.
- Abolmasov B., Mihalić, S., Hadži-Niković G., **Marjanović M.**, Krkač M. (2010) Socioeconomic influence of natural disasters in the Western Balkan countries. In: *Proceedings of 19<sup>th</sup> Congress of the Carpathian Balkan Geological Association (CBGA 2010)*, 23-26 of September, Thessaloniki, Greece, ISBN: 978-960-9502-02-3.
- Marjanović M.**, Kovačević M., Bajat B. (2010) Artificial Intelligence-based models for landslide mapping. In: *Proceedings of 1<sup>st</sup> Project Workshop (Risk Identification and Land-Use Planning for Disaster Mitigation of Landslides and Floods in Croatia)*, 22-24 of November, Dubrovnik, Croatia, ISBN: 978-953-6953-27-1.
- Bajat B., Kovačević M., **Marjanović M.** (2010) Some experiences with utilization of Support Vector Machines in landslides susceptibility mapping. In: *Proceedings of 1<sup>st</sup> Project Workshop (Risk Identification and Land-Use Planning for Disaster Mitigation of Landslides and Floods in Croatia)*, 22-24 of November, Dubrovnik, Croatia, ISBN: 978-953-6953-27-1.
- Marjanović M.**, Čaha J. (2011) Fuzzy approach to landslide susceptibility zonation. In: *Proceedings of the Annual International Workshop on Databases, TExts, Specifications and Objects (DATESO) 2011*, 20. April, Písek, Czech Republic, pp. 181-195, ISBN: 978-80-248-2391-1.
- Marjanović M.**, Kovačević M., Bajat B, Voženílek V. (2011a) Landslide susceptibility assessment using SVM machine learning algorithm. *Engineering Geology*, vol. 123, pp. 225-234, ISBN: 0013-7952.
- Marjanović M.**, Kovačević M., Bajat B, Mihalić S., Abolmasov B. (2011b) Landslide assessment of the Starča basin (Croatia) using machine learning algorithms. *Acta Geotechnica Slovenica*, vol. 2011/2, pp. 45-55, ISSN: 1854-0171.
- Marjanović M.**, Kovačević M., Bajat B, Voženílek V., Marek L. (2011c) Využití klasifikačních algoritmů metod strojového učení pro účely prostorového modelování. In: *Metody umělé inteligence v geoinformatice*, Voženílek, Dvorský, Húsek (eds), Univerzita Palackého v Olomouci, Olomouc, Czech Republic, pp. 185, ISBN: 978-80-244-2945-8.
- Marjanović M.** Djurić U., Petrović R. (2012) Modelovanje hazarda od klizišta različitim metodama u GIS-u. In: *Proceedings of the 14<sup>th</sup> Symposium of Engineering Geology and Geotechnics (DGEITS)*, 27-28 of September, Belgrade, Serbia, ISBN: 978-86-89337-01-3.
- Marjanović M.** (2012) Advanced landslide assessment of the Halenkovice experimental site. In: *Proceedings of the 1<sup>st</sup> InDOG Doctoral Conference 2012*, 29th October – 1st November, Olomouc, Czech Republic, pp. 38-41, ISBN: 978-80-244-3260-1.
- Marjanović M.**, Burian J., Miřijovský J., Harbula J. (2012) Urban Land Cover Change of Olomouc City Using LANDSAT Images. *World Academy of Science, Engineering and Technology* vol. 71, pp. 525-531, eISSN: 2010-3778.
- Marjanović M.** (2013) Comparing the performance of different landslide susceptibility models in ROC space. In: *Landslide Science and Practice – Volume 1: Landslide Inventory and Susceptibility and Hazard Zoning*, Margottini, Canuti, Sassa (eds), Springer, Berlin, Germany, pp. 607. ISBN: 978-3-642-31324-0.
- Djurić U., **Marjanović M.**, Šušić V., Petrović R., Abolmasov B., Zečević S., Basarić I. (2013) Land-use suitability analysis of Belgrade city suburbs using machine learning algorithm. In: *Proceedings of Symposium GIS Ostrava 2013 - Geoinformatics for City Transformations*, 21-23 of January, Ostrava, Czech Republic, ISBN: 978-80-248-2952-4.

# List of Figures

<b>Figure 1.</b>	Global (per continent) distribution of different hazard types.....	2
<b>Figure 2.</b>	People exposure to landslides.....	3
<b>Figure 3.</b>	A simplified illustrative landslide classification after Varnes.....	5
<b>Figure 4.</b>	Landslide elements .....	6
<b>Figure 5.</b>	Spans of different measures of landslides .....	7
<b>Figure 6.</b>	Landslide activity stages .....	8
<b>Figure 7.</b>	Schematic illustration of landslide types .....	9
<b>Figure 8.</b>	An example of Logistic Regression classification.. ..	39
<b>Figure 9.</b>	An example of manual training-testing area split .....	40
<b>Figure 10.</b>	The overfit problem. ....	41
<b>Figure 11.</b>	A 4-fold Cross-Validation scheme. ....	42
<b>Figure 12.</b>	<i>k</i> -NN classification principle.....	43
<b>Figure 13.</b>	An example of a simple tree structure on landslide assessment example .....	44
<b>Figure 14.</b>	The illustration of the Gain Ratio on the decision tree node.....	45
<b>Figure 15.</b>	Architecture of SVM .....	47
<b>Figure 16.</b>	A general binary classification example.....	48
<b>Figure 17.</b>	The infinite slope model. ....	50
<b>Figure 18.</b>	Stability Index in relation to stability classes in Area-Slope space .....	51
<b>Figure 19.</b>	An example of a ROC curve.....	54
<b>Figure 20.</b>	A flowchart of the research.....	55
<b>Figure 21.</b>	Geological setting of the study area .....	59
<b>Figure 22.</b>	Schematic models of instabilities present in the case study area.....	61
<b>Figure 23.</b>	Various thematic inputs .....	64
<b>Figure 24.</b>	The cross-section of the riverbed .....	65
<b>Figure 25.</b>	Various thematic inputs.....	68
<b>Figure 26.</b>	Augmented photo-documentation from the field investigation.....	69
<b>Figure 27.</b>	Location of the photographs from Figure 26.....	70
<b>Figure 28.</b>	Landslide Inventory .....	70
<b>Figure 29.</b>	Different lithological domains (regions).....	72
<b>Figure 30.</b>	Model-1a (AHP model).....	75
<b>Figure 31.</b>	Information Gain in function of number of classes .....	75
<b>Figure 32.</b>	ROC curve of the Model-1a.....	76
<b>Figure 33.</b>	Model-2a (CP model). ....	77
<b>Figure 34.</b>	ROC curve of Model-2a.....	78
<b>Figure 35.</b>	The structure of the fuzzy model. ....	79
<b>Figure 36.</b>	Model-3a (fuzzy model).....	80
<b>Figure 37.</b>	ROC curve of Model-3a.....	81
<b>Figure 38.</b>	Model-5a-B-10%. ....	84
<b>Figure 39.</b>	Model-6a-B-10%. ....	86
<b>Figure 40.</b>	Model-7a-B-10%. ....	87
<b>Figure 41.</b>	Model-7a-B-10%. ....	87
<b>Figure 42.</b>	ROC curve of Model-7a-B-10%.....	88
<b>Figure 43.</b>	Model-7a-33%.....	89
<b>Figure 44.</b>	ROC curve of Model-8a.....	90
<b>Figure 45.</b>	Model-8a .....	91
<b>Figure 46.</b>	Geographical setting of the study area .....	93
<b>Figure 47.</b>	Various Conditioning Factors .....	94
<b>Figure 48.</b>	Various Conditioning Factors. ....	95



<b>Figure 49.</b> Geological map of the study area.....	98
<b>Figure 50.</b> Landslide Inventory .....	99
<b>Figure 51.</b> Model-7b-40%.....	104
<b>Figure 52.</b> ROC curve of Model-7b-40% .....	104
<b>Figure 53.</b> Geographical location of the study area .....	106
<b>Figure 54.</b> Geological setting of the area.....	107
<b>Figure 55.</b> Various Conditioning Factors. ....	111
<b>Figure 56.</b> Various Conditioning Factors (continued).....	112
<b>Figure 57.</b> Landslide Inventory. ....	113
<b>Figure 58.</b> Augmented photo-documentation from the field investigation.....	114
<b>Figure 59.</b> Model-7c-40%.....	117
<b>Figure 60.</b> ROC curve of the Model-7c-40%.....	118
<b>Figure 61.</b> SA plot for region 1 (flysch). ....	119
<b>Figure 62.</b> Model-8c. ....	120
<b>Figure 63.</b> ROC curve comparison for several models of a sub-branch.....	123

## List of Tables

<b>Table 1.</b>	Updated Varnes landslide classification .....	8
<b>Table 2.</b>	Contemporary RS systems .....	13
<b>Table 3.</b>	Attribute ranking for Starča Basin case study.....	29
<b>Table 4.</b>	An example of AHP comparison matrix.....	33
<b>Table 5.</b>	An example of AHP weights derivation .....	33
<b>Table 6.</b>	Confusion matrix and appropriate error measures .....	52
<b>Table 7.</b>	Raster thematic maps represented by thematic attributes of the input dataset.....	71
<b>Table 8.</b>	General statistics of attribute layers .....	71
<b>Table 9.</b>	Average parameters over different lithological domains (regions) .....	73
<b>Table 10.</b>	AHP comparison matrix of the first model variant.....	74
<b>Table 11.</b>	AHP weight derivation of the first model variant.....	74
<b>Table 12.</b>	Final AHP weights of the second model variant .....	75
<b>Table 13.</b>	Model-1a performance metrics .....	76
<b>Table 14.</b>	Model-1a performance per class.....	76
<b>Table 15.</b>	Model-2a performance metrics .....	78
<b>Table 16.</b>	Model-2a performance per class.....	78
<b>Table 17.</b>	Model-3a performance metrics .....	80
<b>Table 18.</b>	Model-4a performance metrics .....	83
<b>Table 19.</b>	Performance of Machine Learning models (unbalanced training set).....	84
<b>Table 20.</b>	Performance of Machine Learning models (balanced training set).....	85
<b>Table 21.</b>	Model-7a-33% performance metrics. ....	89
<b>Table 22.</b>	Performance evaluation of C4.5 and SVM classifiers by $\kappa$ -index .....	101
<b>Table 23.</b>	Information Gain (IG) ranking of the input layer attributes .....	101
<b>Table 24.</b>	$c$ , $\gamma$ parameter combinations for optimization. ....	103
<b>Table 25.</b>	Performance of the Model-7b-40% .....	103
<b>Table 26.</b>	List of lithological units of the study area.....	108
<b>Table 27.</b>	Average parameters over different lithological domains (regions) .....	114
<b>Table 28.</b>	Attribute Selection and ranking .....	115
<b>Table 29.</b>	$c$ , $\gamma$ parameter combinations for optimization. ....	116
<b>Table 30.</b>	Performance of Model-7c-40% .....	116
<b>Table 31.</b>	Performance parameters of Model-7c-40%.....	117
<b>Table 32.</b>	Average performance parameters of Model-8c .....	120

# Appendices

**Appendix 1.** Table of ranged Conditioning Factors and their fuzzy memberships for Model-3a.

Conditioning Factor (type, group) categories	$\mu_{FR}$	$\mu_{CA}$	$\chi^2$ ( $\chi^2_{critical}$ )
<i>distance from structures</i> (continual, geo-buffer)			1949.6
0–134	0.051	0.781	(27.9)
134–276	0.092	0.770	
276–426	0.141	0.805	
426–582	0.260	1	
582–755	0.261	0.812	
755–942	0.178	0.445	
942–1159	0.024	0.077	
1159–1418	0	0	
1418–1758	0.262	0.197	
1758–2305 m	1	0.568	
<i>distance from h.g. boundary</i> (continual, geo-buffer)			306.3
0–94	0.275	1	(27.9)
94–218	0.038	0.630	
218–342	0.107	0.499	
342–458	0	0.372	
458–589	0.040	0.295	
589–726	0.484	0.429	
726–878	0.579	0.313	
878–1050	1	0.332	
1050–1244	0.458	0.105	
1244–1749 m	0.682	0	
<i>distance from stream</i> (continual, hydro)			2381.6
0–94	0.550	0.643	(27.9)
94–212	0.900	1	
212–324	0.780	0.809	
324–432	0.453	0.431	
432–543	0.346	0.305	
543–660	0.218	0.165	
660–797	0	0	
797–966	0.024	0.002	
966–1173	0.362	0.100	
1173–1542 m	1	0.214	
<i>TWI</i> (continual, hydro)			4947.6
7.5–9.3	0	0	(26.1)
9.3–10.3	0.172	0.261	
10.3–11.4	0.696	1	
11.4–12.8	1	0.955	
12.8–14.3	0.811	0.575	
14.3–16.2	0.821	0.442	
16.2–18.3	0.781	0.320	
18.3–20.8	0.506	0.176	
20.8–22.5	0.131	0.079	

<i>aspect</i> (categorical, topo)			1091.0
flat	0	0	(26.1)
N	0.594	0.701	
NE	0.552	0.688	
E	1	1	
SE	0.889	0.490	
S	0.278	0.140	
SW	0.645	0.638	
W	0.494	0.571	
NW	0.414	0.433	
<i>elevation</i> (continual, topo)			7515.7
78–102	0.660	0.619	(27.9)
102–138	1	1	
138–173	0.828	0.838	
173–209	0.530	0.499	
209–248	0.158	0.147	
248–287	0.141	0.118	
287–329	0.018	0.013	
329–376	0	0	
376–426	0	0	
426–540 m	0	0	
<i>slope angle</i> (continual, topo)			4453.1
0–4.2	0.300	0.243	(18.5)
4.2–9.5	1	1	
9.5–14.8	0.473	0.403	
14.8–21.1	0.119	0.086	
21.1–40.1°	0	0	
<i>slope length</i> (continual, topo)			1346.8
0–60	0.435	1	(27.9)
60–181	0.591	0.964	
181–353	0.937	0.960	
353–602	1	0.667	
602–981	0.650	0.261	
981–1506	0.301	0.080	
1506–2196	0.178	0.033	
2196–3094	0.427	0.061	
3094–4392	0.187	0.019	
4392–6499 m	0	0	
<i>plan curvature</i> (continual, topo)			989.4
concave	0	0	(18.5)
-	0.657	0.333	
flat	1	1	
-	0.626	0.419	
convex	0.149	0.059	
<i>profile curvature</i> (continual, topo)			1214.0
concave	0	0.009	(18.5)
-	0.414	0.287	
flat	1	1	
-	0.741	0.366	
convex	0.081	0	

<i><b>lithology</b></i> (categorical, geo-units)			8319.3
al' - Danube's inundation plane	0.100	0.099	(29.6)
al - aluvium	0.211		
dl - delluvium cover	0.807	1	
t - terrace sediments	1	0.785	
l - loess	0.338	0.334	
Pl - clay	0.858	0.847	
M <sub>2</sub> - marlstone	0.133	0.083	
M <sub>1</sub> - limestone, sandstone	0.469	0.880	
Se - ultra-mafic rocks	0	0	
J - limestone	0	0	
Pz - schists	0.002	0.003	
<i><b>Land Cover</b></i> (categorical, land cover)			6316.2
water	0	0	(16.2)
arable land	1	1	
grass land	0.992	0.852	
forest	0.132	0.168	

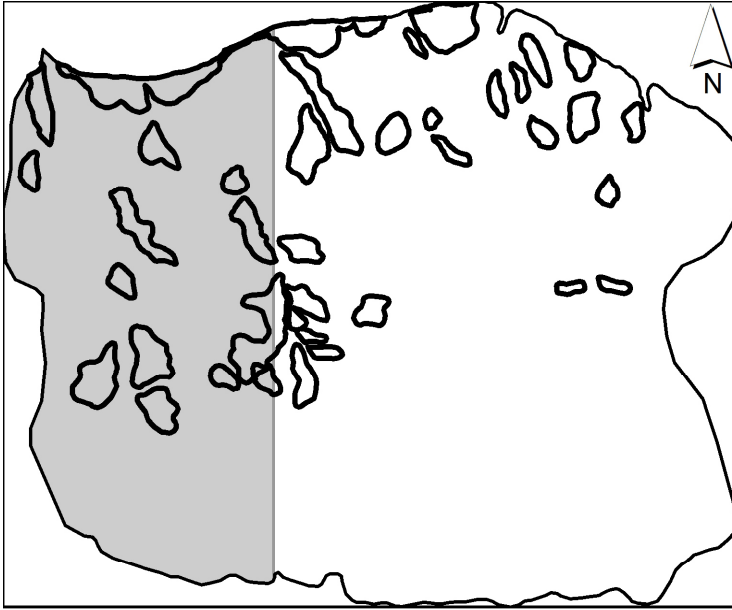
## Appendix 2. R code for plotGoogleMap package.

```
#-----  
#Halenkovice                                landslides  
#-----  
setwd("g:/PhD/tmp/_Halenkovice/Sesuv/PlotGM")  
getwd()  
system("dir")  
#-----  
library(rgdal)  
library(plotGoogleMaps)  
#-----  
study.area<-readOGR(".", "study_area1")  
str(study.area @data)  
proj4string(study.area)<-CRS("+proj=krovak +lat_0=49.5 +lon_0=24.83333333333333 +k=0.9999  
+x_0=0 +y_0=0 +ellps=bessel +units=m +no_defs")  
  
training.area<-readOGR(".", "training_area")  
#str(training.area)  
str(training.area @data)  
proj4string(training.area)<-CRS("+proj=krovak +lat_0=49.5 +lon_0=24.83333333333333 +k=0.9999  
+x_0=0 +y_0=0 +ellps=bessel +units=m +no_defs")  
  
result1<-readOGR(".", "svm_result1")  
str(result1 @data)  
proj4string(result1)<-CRS("+proj=krovak +lat_0=49.5 +lon_0=24.83333333333333 +k=0.9999 +x_0=0  
+y_0=0 +ellps=bessel +units=m +no_defs")  
  
landslides<-readOGR(".", "Landslides U")  
str(landslides @data)  
proj4string(landslides)<-CRS("+proj=krovak +lat_0=49.5 +lon_0=24.83333333333333 +k=0.9999  
+x_0=0 +y_0=0 +ellps=bessel +units=m +no_defs")  
  
map1=plotGoogleMaps(study.area,  
                    zcol = "label",  
                    add=T,  
                    colPalette = 'black',  
                    strokeColor = 'black',  
                    strokeWeight = 5,  
                    filename='MyMapHalenkovice.htm',  
                    layerName='study area',  
                    mapTypeId = "TERRAIN")  
map2=plotGoogleMaps(training.area,  
                    zcol = "label",  
                    add=T,  
                    colPalette = "black",  
                    strokeColor = 'black',  
                    strokeOpacity = 1,  
                    strokeWeight = 3,  
                    previousMap=map1,  
                    filename='MyMapHalenkovice.htm',  
                    layerName='training area',  
                    mapTypeId = "TERRAIN")  
map3=plotGoogleMaps(result1,  
                    zcol = "label",  
                    add=T,  
                    strokeOpacity = 0,  
                    strokeWeight = 0,  
                    colPalette = "red",  
                    previousMap=map2,
```

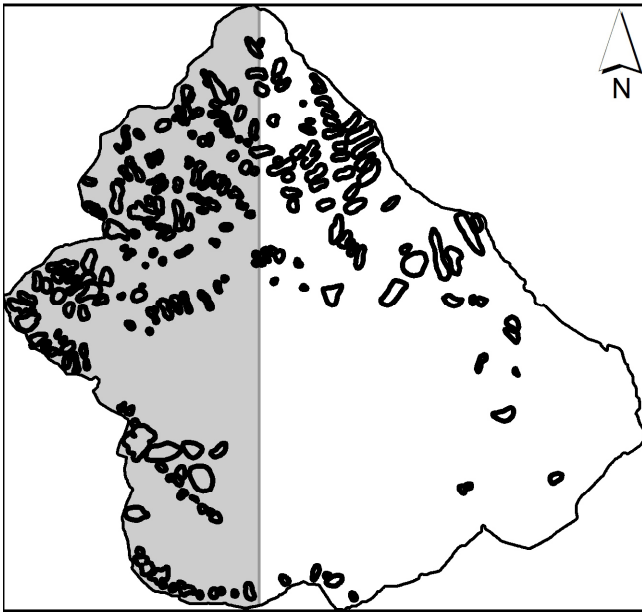
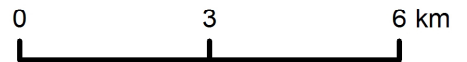


```
filename='MyMapHalenkovice.htm',
layerName='SVM model',
mapTypeId = "TERRAIN")
map4=plotGoogleMaps(landslides,
zcol = "label",
add=F,
colPalette = "orange",
strokeColor = 'orange',
strokeOpacity = 1,
strokeWeight = 3,
previousMap=map3,
filename='MyMapHalenkovice.htm',
layerName='landslides (active & dormant)',
mapTypeId = "TERRAIN")
```

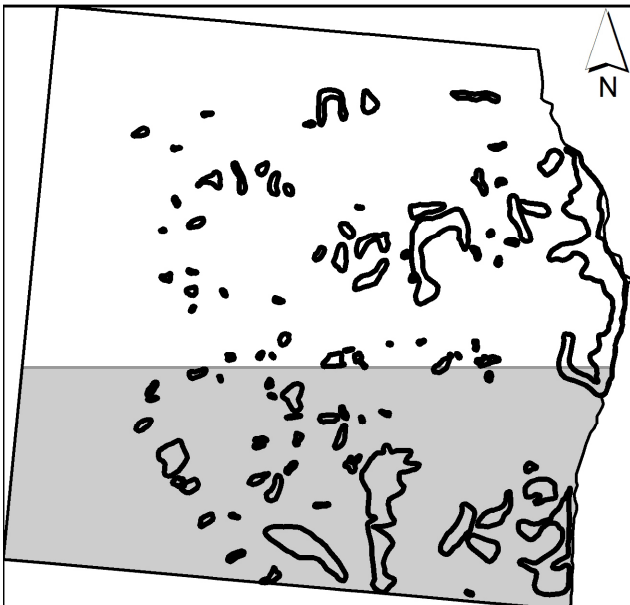
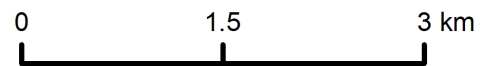
Appendix 3. Maps of predictive models.



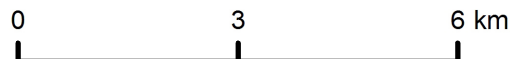
training area is shaded  
actual landslides are contoured

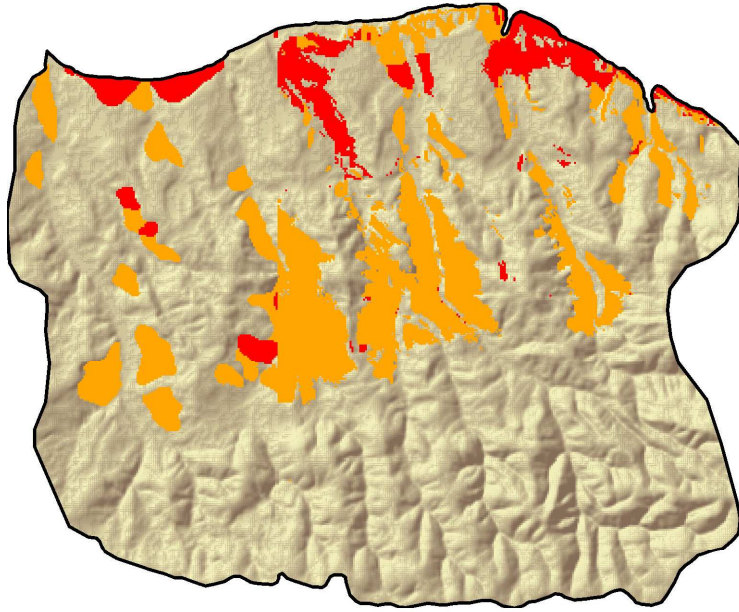


training area is shaded  
actual landslides are contoured





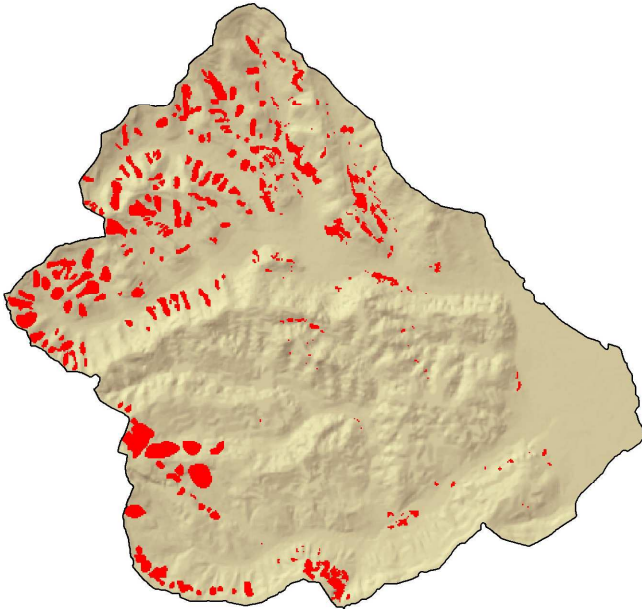
training area is shaded  
actual landslides are contoured





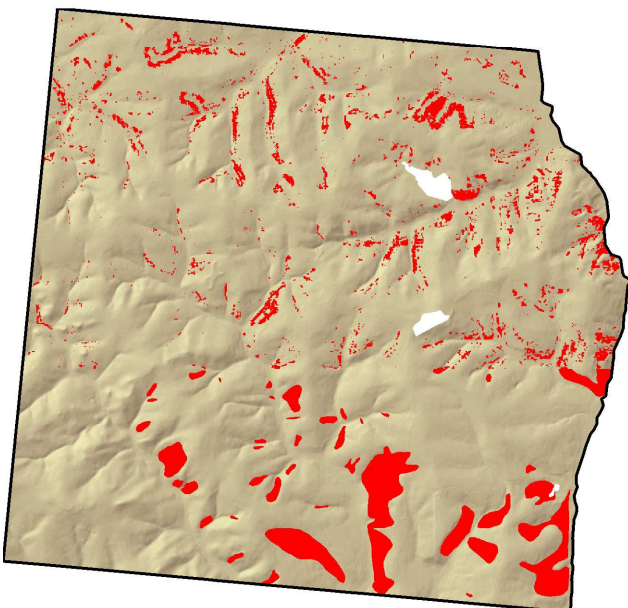
Model-7a-33%  
(SVM predictive model)

-  dormant landslides
-  active landslides



Model-7b-40%  
(SVM predictive model)

-  landslides



Model-7c-33%  
(SVM predictive model)

-  landslides

## Shrnutí

Disertační práce řeší pokročilé metody předpovědi půdních sesuvů, od teoretických základů po konkrétní praktické příklady ve třech zájmových územích. Předmět výzkumu představuje velmi komplexní a heterogenní přírodní fenomén, jehož kvantitativní prognózy se obvykle popisují náchylností, nebezpečím nebo rizikem. Autor klade důraz na náchylnost terénu ke klouzání, tj. prostorovou pravděpodobnost výskytu půdních sesuvů. Tento přístup vychází většinou z nedostatku vhodných časoprostorových dat potřebných pro analýzu nebezpečí nebo rizika. Na druhou stranu maximálně využívá všechna ostatní dostupná prostorová data, včetně geologických, geomorfologických, hydrologických, hydrogeologických a dalších dat o vlastnostech životního prostředí, která jsou v praxi často označována jako podmíněné faktory podmiňující půdní sesuvy.

Hlavní cíle této disertace jsou:

1. použití dostupných, bezplatných dat a softwarových produktů s cílem prokázat, že využíváním stávajících dostupných zdrojů lze provést hodnotnou analýzu předpovědi půdních sesuvů,
2. testování původní metodologie v několika zájmových územích (navzájem dostatečně podobných, avšak i dostatečně rozlišných), aby bylo možné objektivně diskutovat o úspěšnosti navrhované metodologie,
3. standardizace vstupních dat z hlediska jejich objemu, typu, kvality a poměru, a jejich před-zpracování pomocí GIS,
4. použití řady metod modelování náchylnosti k půdním sesuvům, od jednoduchých až po pokročilé, s cílem je objektivně a podrobně porovnat,
5. použití co nejrelevantnějších metod pro evaluaci modelů předpovědi sesuvů s cílem co nejobjektivnějšího kvalitativního a kvantitativního porovnání těchto modelů,
6. vizualizace a publikování výsledků použitím GIS a prostředků webové kartografie.

Navrhovaná metodologie zahrnuje řadu metod, které lze rozdělit na metody předzpracování, modelování náchylnosti a metod hodnocení. Největší důraz byl kladen na metody modelování náchylnosti k sesuvům půdy. Byly použity jak metody nejpoužívanější, tak i nejjednodušší, tak i nejpokročilejší a nejsložitější metody, a to:

- heuristické (na základě subjektivní zkušenosti autora, který se zabývá problematikou sesuvů),
- deterministické (na základě známých fyzikálních principů týkajících se sesuvů půdy, které jsou do značné míry aproximovatelné),
- statistické (na základě statistické závislosti na vlastnostech různých vlastností terénu a sesuvů půdy),
- metody strojového učení (na základě logicky-matematicko-statistických algoritmů, které poloautomaticky nacházejí vztahy mezi vlastnostmi terénu a projevy půdních sesuvů).

K poslední skupině patří metody k-nejbližší sousedství (nearest neighbor, k-NN), logistická regrese (LR), rozhodovací stromy (Decision Trees, DT) a Support Vector Machines (SVM), jejichž modely jsou předmětem disertační práce.

Je potřeba zdůraznit, že autorem navržená metodologie, která je stejným nebo podobným způsobem použita ve všech třech zájmových územích a usilující o standardizaci, může být aplikována na zcela jiná území, která splňují určitá kritéria a mají k dispozici odpovídající údaje. Návrh metodologie začíná od problematiky výběru konkrétního typu

(mechanismu) sesuvu, který je přítomen ve vybraných územích, dále pokračuje přes výběr vstupních dat osesuvech půdy, která slouží jako podklad pro hodnocení modelu. Po definování základních kritérií metodika navrhuje použití řady metod pro předzpracování, modelování náchylnosti a/nebo predikci sesuvů, po kterém následuje představení, evaluace a srovnávání výsledků. Závěrem metodologie vrcholí v diskusi o výhodách a nevýhodách modelu a diskusi o nejvhodnějším modelu pro konkrétní účel použití.

Výzkum probíhal ve třech územích a byl realizován v období čtyř let díky podpoře GAČR projektu *Metody umělé inteligence v GIS (Methods of artificial intelligence in GIS) (205/09/0793)*. Výzkum zahrnoval sběr dat vybraných lokalit ještě před použitím navržené metodiky. Tyto údaje byly prostřednictvím GIS připraveny v souladu s požadavky těchto metod.

První zájmové území zahrnuje severozápadní svahy pohoří Fruška Gora (Srbsko) podél břehu Dunaje s rozlohou cca 100 km<sup>2</sup>, při čemž asi 10 % území je ovlivněno sesuvnými procesy. Většinou jde o projevy hlubokých rotačních a kompozičních sesuvů vyvinutých v neogenních pánvích. Vzhledem k velikosti území a podrobnosti dostupných vstupních dat byla pro analýzu zvolena 30metrové prostorové rozlišení a rastrový formát, což znamená, že za základní jednotku byl pixel o rozměru 30x30 m. Území bylo reprezentováno rastrovou vrstvou s 100 000 buňkami, které nesly informace o  $n$ -různých tematických vlastnostech území, takže každý pixel mohl být považován za vektor o  $n$  souřadnicích. Ve vstupních datech jsou zahrnuty především popis půdních sesuvů (získaný terénními metodami a metodami dálkového průzkumu Země, na kterém jsou odděleny pouze případy stejného typu, tj. hluboké sesuvy půdy typu *earth slide* podle přijaté klasifikace a pro které byly definované fáze aktivity) a jim odpovídající podmíněné faktory:

- sklon svahu, délka svahu, expozice, elevace, planární (horizontální) a profilové (vertikální) křivost svahu, TWI a vzdálenost od drenážní sítě (získané z digitálního modelu reliéfu, který byl modelován z topografické mapy v měřítku 1 : 25000 se základním intervalem vrstevnic 10 m),
- litologické jednotky, vzdálenost od zlomu a vzdálenost od významných geologických hranic (vyznačených na geologických mapách v měřítku 1 : 50000),
- vegetační pokryv (získaný z LANDSAT snímků s rozlišením 30 m a zpracovaný podle vegetačních indexů).

Použitím navržené metodologie pro daný soubor vstupních dat byly odvozeny různé modely náchylnosti a pokročilými metodami byly sestaveny i modely prostorové predikce sesuvů. Úspěšnost modelů je definovaná několika parametry, z kterých je nejdůležitějším ukazatelem ROC křivka, protože umožňuje kvalitativně-quantitativní hodnocení modelu. Lze konstatovat, že použitím pokročilých metod (LR, DT a SVM) jsou jednoznačně nejvhodnější modely náchylnosti, které mají relativně vysokou přesnost, při které je negativní typ chyby (false negative) minimální. Nicméně některé modely, např. fuzzy model získaný vícevrstevnou fuzzy kombinací, vykazují určitý potenciál i přes mírně nižší přesnost a minimální nežádoucí chyby. Na druhé straně, prostorové predikce sesuvů u modelů založených na LR, DT a SVM technikám lze hodnotit jako úspěšné, oproti deterministickému modelu, který může být zcela ignorován a považován za nevhodný pro dané území. Pokročilé modely byly úspěšně použity i v případech s více než jednou kategorií sesuvy (aktivní a nečinné).

Druhé zájmové území se nachází v povodí řeky Stareč u Záhřebu (Chorvatsko), rozloha kolem 15 km<sup>2</sup> s asi 10 % území ovlivněného sesuvnými procesy s tím, že mechanismus a typologie sesuvů jsou zcela odlišné. Jsou zde mělké sesuvy v terciárních a kvartérních ložiscích, jejichž hlavní hnací silou je eroze v kombinaci se srážkami. Použit byl soubor vstupních dat podle výše popsaného. Dále byla použita rastrová reprezentace o rozlišení 10 m (kvůli menší rozloze zájmového území a menším rozměrům sesuvů), takže celé území bylo vyjádřeno rastrovou vrstvou s 100 000 buňkami. Byla použita podobná metodologie jako výše, ale v o něco menším objemu, protože některé podobné analýzy se

základními metodami už na daném území proběhly. Proto byl důraz kladen na pokročilé metody, přesněji DT, resp. SVM techniky, a to pro jednu, resp. pět kategorií sesuvů (definovaných na základě jejich aktivit). Uvažovány byly také modely náchylnosti a predikce obou technik. Výsledky ukázaly o něco slabší úspěšnost v modelech náchylnosti a ještě menší v predikci samých sesuvy půdy. Zajímavé je, že lepší hodnocení vykázaly modely s několika kategoriemi sesuvů než jednodušší modely s jednou kategorií sesuvu. Tyto výsledky jsou pravděpodobně způsobeny nízkým prostorovým rozlišením rastru, ale i samou rozlohou zájmového území, přítomností velké řady stejných jevů (pět kategorií), které měly za následek tzv. overfit, tj. špatné naučenou relaci v procesu trénování algoritmu.

Poslední zájmové území je v okolí města Halenkovice ve Zlínském kraji (Česká republika) o rozloze přibližně 50 km<sup>2</sup> s mělkými půdními sesuvy vyvinutými v terciárních flyších. V analýze byl použit soubor vstupních dat podobný ve výše popsaném textu, prostorové rozlišení gridu bylo 10 m, což vytvořilo rastrovou vrstvu o 500 000 buňkách. Důraz byl kladen na použití pokročilých metod, hlavně SVM techniky, a byl testován i i deterministický model s ohledem na to, že zde vyskytující se mělké sesuvy jsou teoreticky vhodné pro takový typ modelu. Modely založené na SVM technikách byly omezeny jen na predikční modely s tím rozdílem, že SVM model je omezený dodatečným optimalizačním postupem *leave-one-out*, zatímco deterministický model měl obě varianty (náchylnosti a predikce). SVM model lze ohodnotit jako průměrný, ale stále vykazující určitý potenciál v predikci. Deterministický model je v tomto případě nejednoznačný, protože některé části terénu modeluje velmi dobře, zatímco některé velmi špatně, a to i po komplexní optimalizaci, což omezuje model jen na určité geologické prostředí v zájmovém území.

Některé z modelů jsou prezentovány prostřednictvím nástrojů internetového mapování a jsou k dispozici na adresách:

<http://milosmarjanovic.pbworks.com/w/file/fetch/63738284/MyMapFruskaGora.htm>

<http://milosmarjanovic.pbworks.com/w/file/fetch/63741247/MyMapStarca.htm>

<http://milosmarjanovic.pbworks.com/w/file/fetch/63739326/MyMapHalenkovice.htm>

Závěrem lze konstatovat, že vytýčené cíle disertační práce byly splněny, a že navrhovaná metodologie podala dobré výsledky – v případech některých modelů méně úspěšné než v jiných. Modely náchylnosti (zejména modely získané použitím pokročilých metod) mohou najít uplatnění v různých aspektech plánování a projektování v regionálním měřítku, ale také pro regulaci ochrany, systémy včasného varování i pro pojišťovny. Zvláštní přínos mají predikční modely, na jejichž zdokonalování se stále může pracovat. Jejich aplikace se může navázat na uplatňování modelu náchylnosti, zatímco predikční modely mohly být použity pro účely mapování sesuvů a tvorbu jejich databází v regionálním měřítku.