University of Zurich^UZH

Department of Geography

Approach to the perception of wilderness based on user generated open data

Examining and extending wilderness information by combining GIS-based wilderness information with social media data

GEO 511 Master's Thesis

Author

Markus Baumann
12-715-256

Supervisors: Prof. Dr. Ross Purves, Dr. Nicole Bauer and Olga Chesnokova

Faculty member: Prof. Dr. Ross Purves

25.01.2018
Department of Geography, University of Zurich

Contact

Author

Markus Baumann

Zugerstrasse 108
8810 Horgen – Switzerland
baumann_markus2@hotmail.ch

Supervisors

Prof. Dr. Ross Purves

Head of Geocomputation Unit
Department of Geography
University of Zurich
Winterthurerstrasse 190
8057 Zurich – Switzerland
ross.purves@geo.uzh.ch

Dr. Nicole Bauer

Eidg. Forschungsanstalt WSL
Zürcherstrasse 111
8903 Birmensdorf
nicole.bauer@wsl.ch

Olga Chesnokova

Department of Geography
University of Zurich
Winterthurerstrasse 190
8057 Zurich – Switzerland
olga.chesnokova@geo.uzh.ch

Acknowledgement

The master's thesis represents the end of my 5.5 years studying at the University of Zurich, at the Department of Geography. This time represents the period of my largest intellectual development and possibly also the period with largest scope of freedom in my life. The great time I spent at the GIUZ and all the amazing people I met during that time formed me to the person I am right now. Great thanks to the whole department which offered me challenges, amusement and orientation but most of all, fascination. Geography will accompany me for all coming steps in my life.

Special thanks are appropriate relating to this work for my supporting supervisors:

- Prof. Dr. Ross Purves, for giving me general orientation and fascination as well as valuable hints about methodological processes and content during my work;
- Dr. Nicole Bauer, for offering me her knowledge, her books and important dialogues about the wilderness concept;
- Olga Chesnokova, who advised me during struggling periods, gave me orientation where confusion came up and calmed me down when necessary.

Thank you much. Your support cannot be put into words.

At the concept presentation, Nicole offered me a place to study at WSL where I was integrated into the department of economic and social sciences (WISOZ). Great thanks to the whole WISOZ-team which included me into their team, was open-minded and helpful all the time. Special thanks to Sarah Radford and Aline von Atzigen who provided me with the wilderness GIS-model applied in this work and also to Sebastian Moos of Mountain Wilderness for his interests in my work. Some last thanks to Carmen Rinaldo, who still supported me with patience and endurance, as well as to Angie Burch, Felicita Riesow and Tiziana Speckert, for proof reading and application support.

Markus Baumann

January 2018

Abstract

Human-nature interaction and the broader context of wilderness became increasingly important in recent years. Many stakeholders and decision makers request solutions to detect, analyse and visualize this interaction. Large-scale approaches considering applications like the geographical information system (GIS) attempt to assess the wilderness phenomenon on a spatial base. Since the wilderness concept is a cultural concept of a perceptually defined phenomenon these technical approaches have been criticised to not accurately respect the perceptual nature of this phenomenon. The acquisition of perceptual information is generally related to large temporal and also financial effort. User generated content represents a new open source of available perceptual data which has been generated in a social context.

By retrieving social metadata from the open photo-sharing web platform Flickr and applying them to a GIS-based wilderness model, this work addresses the critics to GIS-based evaluations and also the temporal and financial effort required for gathering appropriate data. This pioneer project evaluates the aptitude of Flickr photograph metadata to the wilderness research context by evaluating the influence of various characteristics of such data. Since the wilderness concept has a social but also a spatial context, the spatial features of Swiss wilderness quality defined by a GIS-model are accessed and combined with the information generated by tag-based evaluations. The output of those evaluations is used to reveal further wilderness information to the applied GIS-model. General methodological tools proposed by the information retrieval research field have been applied and extended in order to fit the purposes of this work.

The evaluations within this work have illustrated that social media data suit the requirements for scientific wilderness research, although several biasing characteristics need to be considered and handled. The tag-based evaluations have revealed that wilderness features and characteristics defined by a GIS-model can also be determined in the metadata of Flickr photographs. Furthermore, the combination of this perceptual information with the technical GIS-approach allowed further characterization of the GIS-based wilderness information. Finally, Flickr photograph metadata was evaluated to be appropriate for generating new insights into wilderness conditions and human-nature interaction, despite limits regarding social media characteristics.

Zusammenfassung

Die Interaktion zwischen Mensch und Natur und das Konzept von Wildnis hat in den letzten Jahren an Bedeutung gewonnen. Viele interessierte Akteure und Entscheidungsträger verlangen nach Lösungsansätzen, um dieses Zusammenleben besser zu ergreifen, zu analysieren und darstellen zu können. Grossräumige Untersuchungen mit Systemen wie dem Geographischen Informationssystem (GIS), untersuchen das Wildnis-Phänomen auf räumlicher Basis. Da das Wildnis-Konzept kulturell geprägt und stark mit der Wahrnehmung verknüpft ist, wurde bei Untersuchungen des Wildnis-Konzeptes mit GIS die Vernachlässigung des wahrnehmungsbezogenen Charakters von Wildnis kritisiert. Die Generierung grosser Mengen an wahrnehmungsbezogenen Daten ist aber zeitlich und finanziell aufwändig und arbeitsintensiv. Doch durch die Digitalisierung sind alternative Datenquellen entstanden, die grosse Quantitäten an wahrnehmungsbezogenen Daten bereitstellen und in einem sozialen Kontext generiert wurden.

Durch die Nutzung sozialer Medien in Form von Foto-Metadaten der Web-Plattform Flickr und deren Anwendung auf ein GIS-Modell, wird einerseits auf die genannte Kritik reagiert und andererseits Bezug zur finanziell und zeitlich aufwandsgeringen Beschaffung von wahrnehmungsbezogenen Daten genommen. Durch die Kombination von GIS- und Flickr-Daten wird untersucht, ob diese Daten generell für den wissenschaftlichen Ansatz zu Wildnis geeignet sind und ob man damit die GIS-basierten Informationen erweitern kann. Dies wird hauptsächlich anhand von Tag-basierten Evaluationen umgesetzt. Diese Untersuchung ist neu im wissenschaftlichen Wildnis-Kontext und bezieht methodische Ansätze aus dem Forschungsfeld der Information Retrieval (IR), die an die Kriterien dieser Arbeit angepasst wurden.

Die Untersuchungen haben ergeben, dass Flickr Foto-Metadaten den wissenschaftlichen Voraussetzungen zum Ansatz von Wildnis gerecht werden, wenn auch einige wichtige Charakteristiken beachtet und behandelt werden müssen. Die Tag-basierenden Untersuchungen haben ergeben, dass Unterschiede in GIS-basierten Wildnis-Modellen auch in den Flickr-Metadaten aufgefunden werden können. Diese festgestellten Unterschiede können genutzt werden, um die GIS-Informationen mit zusätzlichen wahrnehmungsbezogenen Informationen der Flickr Gemeinschaft zu erweitern. Damit kann neues Wissen und Einblicke in die wahrgenommene Wildnis gewonnen werden, die das Verständnis der Mensch-Natur Interaktion und somit den Schutz und die Erhaltung von Wildnisgebieten unterstützen können.

Table of contents

List of figures

List of tables

List of abbreviations

GIR	Geographic Information Retrieval
GIS	Geographical Information System, application for spatial analyses and visualizations
IDE	Integrated Development Environment
IR	Information Retrieval
MCE	Multi Criteria Evaluation
UGC	User Generated Content, applied as synonym to social media data
User	contributor or member of a social media platform
Tf-idf	Term frequency, inversed document frequency
WSL	Wood, Snow and Landscape, federal institute in Switzerland

1 Introduction

"The same fate, soon or later, is awaiting them [wild zones] all, unless awakening public opinion comes forward to stop it." (Muir, 1898, p. 17)

The wilderness concept is a cultural construct which has been declared to be endangered in certain regions of the world since many years. As John Muir mentioned, it is up to the public opinion to preserve and protect it in order to keep a balanced human-nature interaction. Scientific organizations like the federal institute Wood, Snow and Landscape (WSL)[1] as well as non-governmental organizations like Mountain Wilderness[2] advocate the continuous advancement of evaluations and theories to promote such balanced interaction. The increased governmental ambition for rewilding in the past two decades in Switzerland indicates the necessity of wilderness also in Switzerland (Bauer, Wallner, & Hunziker, 2009). Various approaches assess the population's attitude towards wilderness through questionnaires (Bauer, 2005) while others implement mathematical models to determine the wilderness phenomenon in Switzerland (Radford et al., unpublished). Both kinds of approaches are necessary to optimally sensitise public and prevent wilderness zones from their fate.

The digitalization has yielded to new sources of information which can serve to increase the quality and the expressiveness of such approaches. Social media platforms represent a new alternative source of information compared to conventional sources. This work attempts to use this source of data to extend the information of an already existing approach. Related work has detected important characteristics of user generated content which need to be considered, such as differences in user-specific behaviour and differences in contributor activity. (Radford et al., unpublished) have initialized a technical approach methodologically referring to the wilderness concept initiated by (Carver et al., 2012) to the area of Switzerland. The wilderness information of that model is extended by the spatial and semantic information of Flickr photographs to optimize the wilderness information according to perceptual social media information.

1.1 Research aim

The general research aim of this work can be split up into three main challenges. First, to demonstrate that user-specific behaviour of a social media platform does not restrict the aptitude of user generated content to wilderness research. Second, that spatial variations in wilderness specified by a spatially explicit model can also be detected in public web-shared geotagged social media data. Third, to assess if wilderness can be determined and further characterized by analysing tags of location-based user generated content. By combining a spatially explicit model like a GIS-model with social media data, the GIS-model is extended by perceptual information, generated in a social context.

[1] https://www.wsl.ch/de.html

[2] http://mountainwilderness.ch/

The following research questions are addressed:

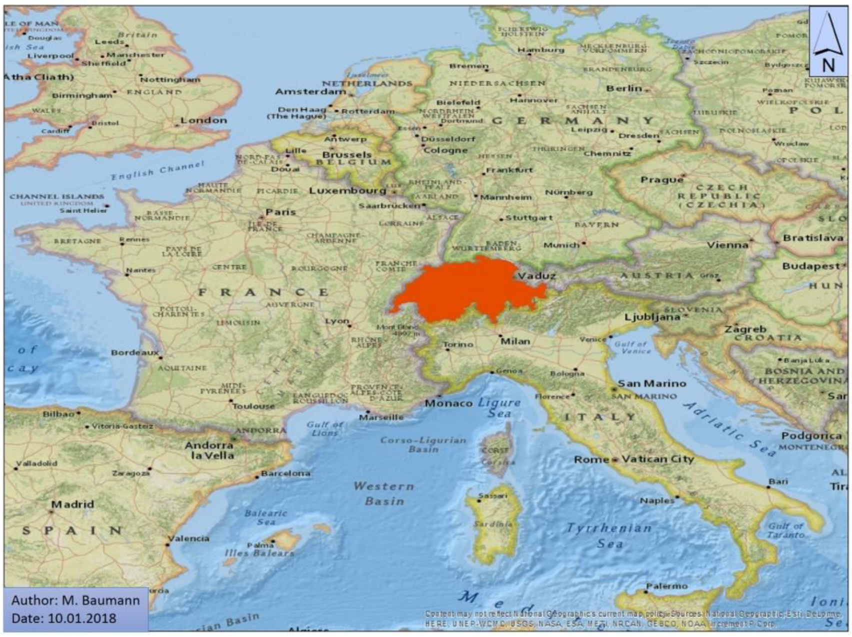
1. *Is the user-specific behaviour on a social media platform relevant to the aptitude of user generated content for scientific wilderness research?*

2. *Are variations in wilderness as quantified by a spatially explicit model reflected in the spatial distribution of user generated content?*

3. *Can GIS-based wilderness information be further characterized by consulting tags of user generated content?*

1.2 Study area

The research area concentrates on the political borders of Switzerland, as coloured in red, visualized in Figure 1.1. Since the research area is oriented at the applied GIS data, and the applied GIS-model has been evaluated for Switzerland, the research area is also restricted to that area. The evaluation does not imply the political area of Liechtenstein. Switzerland is topographically particularly interesting for wilderness research, since it covers many mountain regions with large potential for wilderness areas which are affected by hiking or skiing tourism.



Author: M. Baumann
Date: 10.01.2018

Figure 1.1 – Research area (red) in Europe (source of basemap: ESRI ArcMap 10.4.1)

1.3 Structure of the thesis

The second section presents a theoretical introduction to the wilderness concept and how this concept is embedded in science. While the third section specifically addresses the two applied main datasets, the fourth section describes how the theoretical knowledge has been connected and applied to the two datasets methodologically. Section five refers to the results evaluated through the methodological steps and section six opens the discussion for the results where their content and interpretations will be analysed critically and strengths and weaknesses will be examined. Section seven finally concludes the findings and optional future evaluations are proposed.

2 Theoretical background

This section introduces the whole research area on a theoretical base and aims to highlight where this approach is placed in broader research context. A historical overview of the wilderness debate is illustrated in the first subsection. In the second, the challenges of the vague and subjective nature of the term wilderness and the influence on its definition will be assessed. How research deals with these challenges methodologically will be described in the third subsection whereas subsection four shortly concludes the theoretical findings particularly important to this work.

2.1 Wilderness in history

The origins of wilderness as a concept reach back to the two philosophers and poets Ralph Waldo Emerson (1803 – 1882) and Henri David Thoreau (1817 – 1862) (Stremlow & Sidler, 2002). Their perspective and fascinations for the environment inspired many activists such as John Muir (1838 – 1914) who became a key personality in wilderness protection and preservation debate (Nash, 2014). Muir was one of the first conservationists, naturalist and environmental philosopher who wrote about the protection of wilderness and the beauty of the western nature (Muir, 1898). Inspired by the mentor of the idea for national parks, Henry David Thoreau he was the first who promoted scientific interest to the Yosemite Park area in western USA (1870) and is very famous for the activism he did during his lifetime (Bauer, 2005). Muir wrote several books and scientific articles where he points out the importance of wilderness and the consciousness to preserve the beauty of nature. As protection and assistance for land and environmental planning are key goals of the wilderness debate, John Muir was one of the first activists representing these values. While in America the concept of national parks and the willingness to protect nature spread in the 19th century, Europe remained untouched by this trend for the most part (Habron, 1998a). Thus, the wilderness concepts of America and Europe developed differently and have to be separated. Bauer (2005) describes the different meanings in western culture between the United States and Europe. While in American culture the concept of wilderness as a positive contrast to urban life had a stronger influence and developed earlier, a similar development came up in Europe much later and less powerful. Starting in the mid-19th century the western "wilderness-spirit", initialized by R.W. Emerson and H.D. Thoreau had a large influence on the Western perception of wilderness and also on the attitude of the Western population to protect nature (Bauer, 2005; Nash, 2014). In recent years, the differences between the American and the European meaning of wilderness began slowly to merge (Habron, 1998a). Thus, the term can nowadays be applied in science for both continents. Since the activism of the aforementioned personalities, the way how the population and science perceived wilderness has remarkably shifted (ÖBF und WWF - Österreichische Bundesforste and World Wide Found For Nature, 2012). Although the wilderness concepts of these two continents have developed differently, this shift in wilderness perception happened in both continents and has the same initial situation. In both continents, originally men had many negative associations to wilderness. Stremlow and Sidler (2002) specifies multiple varying associations people made with wilderness which were varying over time. Stremlow & Sidler split these associations up into three temporally distinguishable perspectives.

Wilderness as space of myths

For cultures at early stages, myths and legends were symbolically telling about the contrast between cultural control against wild uncontrolled nature. Spatially, this contrast was reflected by already cultivated, managed land against wild, unknown regions. These myths comprised cultural identity and a sense for meaning and order to build cultural stability and identity even if their message was negatively associated to wilderness. This mythological perspective was strengthened by the symbolic battles between the expansive human against a seemingly overpowering nature and the final conquest of wild zones like the Alps or the colonization of the Middle West.

Wilderness as space of scariness

As a following development, people were scared about the unknown and wild character of wilderness. Beauty and positive associations were according to (Stremlow & Sidler, 2002) only made to cultured land which could be identified as "beneficial". Dangerous and misanthropic places were avoided and especially mountains like the Alps were seen as fearful areas of scare. These negative associations can also be recognized in art where wild landscapes where rarely put into focus. It is obvious that landscape aesthetics as John Muir has characterized it, did not yet exist at these times but came up in the 18th century when a broader audience developed enthusiasm to the beauty of the Alps or the wild, untouched landscapes of America (Habron, 1998a).

Wilderness as space of idyll

In Europe, the negative association of wild land changed at the end of the 18th to the beginning of the 19th century when romanticization and literary revaluation of mountains and wild natural landscapes established. Especially this development took place in America much earlier. While a strong negative contrast was made between urban and rural life before, the contrast tended to become smaller and smaller until an increasing tendency of positive associations turned the perception about wilderness. While words like misanthropic or fearful represented the wild landscape characteristics in literature before, new views called it friendly and pristine. Wilderness was still seen as a contrast to the city life but had no longer a negative association (Stremlow & Sidler, 2002). Rather the woods and mountains became interesting as places to escape the stressful city life in order to enjoy the solitude, silence and the fresh air. This literary change had a strong influence on European and American culture so that the access to wild land became increasingly important and the landscape aesthetics, which John Muir has been convinced it is worth to be protected, reached a broader audience.

These perspectives illustrate a change in how people perceived wilderness over a large time while negative associations were replaced by positive ones. This shift indicated more human interest for these wild regions so that the term wilderness became increasingly important. While in the United States the term wilderness has already been defined in the Wilderness Act in 1964, it has been established in Europe as a classification of the IUCN protected area categories in 1994 (ÖBF und WWF ‑ Österreichische Bundesforste and World Wide Found For Nature, 2012). The integration of the term into governmental and international institutions increased its interest for science, whereby nowadays, wilderness reaches an interdisciplinary field of varying interested parties. As the interest in wilderness protection and preservation increased, the requirement for official definitions arose.

2.2 Defining wilderness

"Wilderness is so heavily frightened with meaning of a personal, symbolic and changing kind as to resist easy definition." (Nash, 2014, p. 1)

The ongoing problem with its definition is as old as the wilderness debate itself. Robert Marshall (1930) has debated about the definition of wilderness already close to 90 years ago and Aplet et al. (2000) have claimed that the definitions have not much changed since seventeen years. But in the meantime, there was a requirement for analysing wilderness in its detail in order to initialize an accurate definition. The aforementioned citation of (Nash, 2014) illustrates how difficult such a definition is and that many factors influence the perception and therefore the definition. The following subsection attempts to reveal these factors in order to demonstrate the complexity of wilderness definition while subsection 2.2.2 refers to definition variabilities detectable in literature and institutions.

2.2.1 Factors influencing the perception of wilderness

The three historical wilderness perspectives mentioned in subsection 2.1 have demonstrated that the perception of the population has changed in a strong degree over time. The observation of the trend of increasing positive associations to wilderness has also been detected by Cordell, Tarrant, and Green (2003) who examined verifiable shifts of the wilderness perception of American population between the years of 1994 and 2000. Thus, the trend is an ongoing process until today. But which factors effectively influence human perception of wilderness?

Since nature still surrounds human environments, every person most likely has some interaction and association to nature. Swanwick (2009) examines in her paper how important access to a natural environment for the population in their daily life is, and for which reasons people like to interact with nature. With her work she supports the findings of Aplet et al. (2000), who illustrates that wilderness perception varies from person to person and make it therefore very difficult to define in a general way. However, many have asserted that multiple factors influence the way how people perceive wilderness and which associations they connect to it (Coeterier, 1996; Habron, 1998a; Kliskey & Kearsley, 1993; Lindemann-Matthies et al., 2014; Stremlow & Sidler, 2002; Swanwick, 2009). To show the variety of individual wilderness perception, the most important factors noted by these authors will be described now. In order to position the following paragraphs, it is necessary to mention that wilderness perception is a more specified sub-research field of the broader field of landscape perception whereas many approaches and methods are similar.

Culture and ethnic

Culture is one of the main factors and multiple researches have already demonstrated the relevance of the cultural background to the perception of wilderness (Cordell et al., 2003; Habron, 1998a; Kliskey & Kearsley, 1993; Lindemann-Matthies et al., 2014; van Zanten et al., 2016a; van Zanten et al., 2016b). At this point it requires returning again to the perspectives of Stremlow and Sidler (2002) in subsection 2.1 to add the information that their work was based on European and American literature. The trend, that people nowadays see wilderness as a natural feature being worth to be protected and not to be influenced by human activity, depends on the cultural background. Harris (2006) has demonstrated significant cultural differences in the perception of nature between Swiss and Chinese population. They mentioned that Chinese people are still convinced that nature is alien and worthy to be improved by human manipulation (Harris, 2006). This example shows that depending on the culture in which people live they associate different values to wilderness and also the willingness to protect it.

Background associations

Another factor is the background association a person has to wilderness. Childhood experiences in nature and memories are relevant influences to perception and can have strong effects on both sides, positive or negative (Habron, 1998b). A Swedish case study is mentioned by van Zanten et al. (2016b) that demonstrates, that individual landscape preferences depend much on varying landscape experiences (Adevi and Grahn (2012) in: van Zanten et al. (2016b)). Albeit a person who never left the city or has never been into a forest, he has some good or bad associations which influence perception. (Habron, 1998a) points out, that differences in the perception of wilderness can especially occur comparing different sample groups. Depending on where people live, rural or urban. Depending on the reason why and how people interact with nature and wilderness, the perception varies individually. People may do these associations during their work, when practicing leisure activities or simply when living or practicing tourism activities (Swanwick, 2009). A woodman for instance may have a different association to wild land than a banker because of differing professional environments.

Differences in the perception of wilderness have also been detected regarding the age or the educational level of the perceiver. Bauer (2005) illustrate significant differences between people younger than 39 years compared to people older than 65 years in the criteria for an area to be classified as wilderness. Where younger people put more weight on how pristine and untouched from human influence a certain landscape is, do older people relate wilderness landscape to high vegetation density. Splitting up the population into sample groups, like according to their age, is a common way to compare attributes of the population. The sample groups Habron (1998a) defined in his work, were separated according to their interaction with wilderness and their living environment. He differentiated between mountaineers, rural inhabitants, rural outdoor workers and conservation managers where for each group the educational level has been determined. His work shows that for example mountaineers have in general a higher educational level and therefore are more aware about the impacts on landscapes by human influences as for example littering. Thus, the awareness of these impacts influences the perception of this sample group. The influence of the educational level has been confirmed by many (Habron, 1998b; Swanwick, 2009; van Zanten et al., 2016b) and is therefore an important factor influencing wilderness perception as well.

Landscape attributes and the number and type of human artefacts

As Stremlow and Sidler (2002) has pointed out, the term wilderness does primarily have a spatial relevance which means, that it describes a spatial area of nature first of all. The classification of this area depends on multiple landscape attributes and characteristics. Therefore the term wilderness cannot be associated to one single landscape characteristic (Habron, 1998a, 1998b). In literature diverse different landscape forms are associated to the term wilderness, like forests, canyons, deserts, tundra regions, high-mountain regions or jungles (Stremlow & Sidler, 2002). These characteristics are defined by different physical landscape attributes amongst others like climate or temperature. Multiple approaches have analysed these attributes in order to determine the most relevant ones for human perception (Coeterier, 1996; van Zanten et al., 2016b). Both of these approaches focus on the same objective but analyse the influence of different landscape attributes, which is not unusual for wilderness research. However, van Zanten et al. (2016b) concludes that the different influences of each attribute cannot be determined because some attributes are beneficial in one place but have negative effects in other places. This finding underlines the complexity of landscape perception. So at this point it can be summarized, that the perception depends on the perceiver on the one hand, and on features within the perceived landscape on the other.

According to the previous paragraphs, the perception of wilderness is influenced by a broad variety of factors, which change individually. Changing perception means changing definitions. Thus, finding a definition accurate for multiple individuals already seems difficult so finding one for an international area is even more challenging.

2.2.2 Variabilities in wilderness definitions

The previous section has pointed out the wide range of varying perceptions which may influence the definition of wilderness. Individual experiences and memories, cultural associations as well as physical attributes in the landscape take part to landscape preferences and influence personal perception and landscape classification. But how to transform all these perceptual influences into a general acceptable and representative definition? Is it even possible to generate a representative definition for all kinds of wilderness regions? *"An international definition of wilderness does not and, more importantly, cannot exist."* as Habron (1998a, p. 13) argues. But some definitions are close to a universal validity. The complexity of the term, the multi-scaled features influencing it and the different landscapes across the globe complicate providing a representative definition (Aplet et al., 2000; Bauer, 2005; Stremlow & Sidler, 2002). A simple example illustrates how different the term wilderness can be defined: The population of Western civilization perhaps sees a deep African jungle as absolute wilderness whereas the indigenous folk living in this jungle define our civilized cities as wilderness, as described by Gomez-Pompa and Kaus (1992).

„A wilderness, in contrast with those areas where man and his own works dominate the landscape, is hereby recognized as an area where the earth and its community of life are untrammelled by man, where man himself is a visitor who does not remain. [...]"

US Wilderness Act 1964: in (ÖBF und WWF ‐ Österreichische Bundesforste and World Wide Found For Nature, 2012).

This definition embodies wilderness as the idea of an area with no human influence, which is an argument that still takes part in current scientific definitions. But the wilderness attribute of having no human influence is by far not the only one. Although Aplet et al. (2000) has pointed out that the wilderness definitions did not change a lot since 1930, with ongoing research process, the number of detected relevant attributes increased much. While only two attributes, namely remoteness and primitiveness, have been seen as the essential attributes to define wilderness in 1985, the current definitions depend on many more (Lesslie & Taylor, 1985). These wilderness attributes will be the focus of the following paragraph.

The mentioned two attributes, remoteness and primitiveness, have been initialized by Lesslie and Taylor (1985) in combination with the wilderness continuum concept. This concept takes respect to the vague nature of wilderness and represents it as a continuum rather than a phenomenon with strict borders. In 1988, Lesslie, Mackey, and Preece (1988) applied one of the first technical approaches based on a geographical information system (GIS) in order to analyse and visualize wilderness attributes. This sort of technical approach opened new opportunities to handle the wilderness debate. Computational calculation capacity enabled fast measurements and automatic visualization methods so that the attribute remoteness could be calculated in a much easier way. Wilderness mapping, a new part of the wilderness debate, has been initialized. Lesslie et al. (1988) generated wilderness maps of the state Victoria in Australia, which were based on the attributes remoteness from access, remoteness from settlements, aesthetic naturalness and biophysical naturalness. Three out of these four attributes were calculated by a simple distance function. In order to generate these wilderness maps, Lesslie et al. (1988) had to specify a clear wilderness definition. As in all GIS-based approaches, their definitions are based on the attributes they include into their model, which in case of Lesslie et al. (1988) were the aforementioned four ones. Thus, each GIS-based approach requires some quantified attributes that build the definition for the terminal model or map. Various approaches followed the base of this technical approach but most initialized their own definition for their wilderness models

(Aplet et al., 2000; Fritz, Carver, & See, 2000). These new approaches concentrated generally on quantified wilderness data and on simple distance functions while in combination with remoteness, also accessibility and solitude became important attributes (Fritz & Carver, 1998). In 2000, Aplet et al. (2000) argued that remoteness and primitiveness stated by Lesslie and Taylor (1985) in combination with solitude would not be enough to describe wilderness. Lesslie & Tayler demonstrated, that previous wilderness debates would have focused too much on uncontested wilderness areas like national parks but would neglect the option that wilderness could appear everywhere depending on naturalness and freedom from human control. They initialized an alternative wilderness continuum that contains the attributes naturalness and freedom from human control (Figure 2.1).

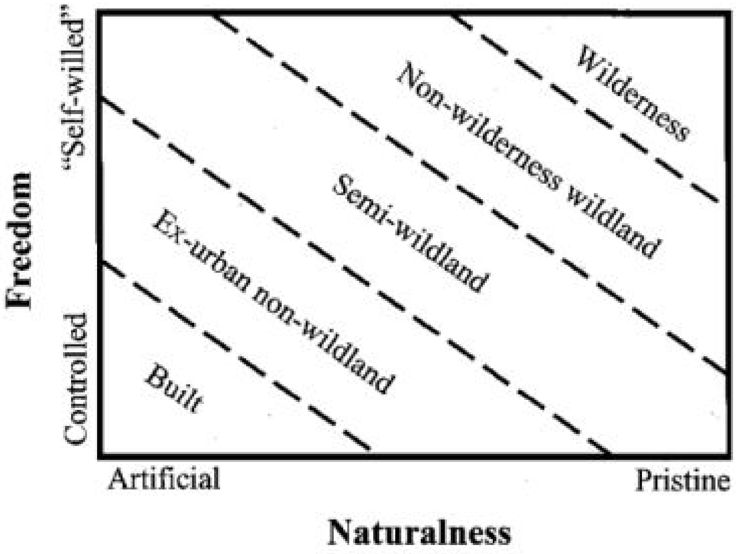


Figure 2.1 – The "continuum of wilderness" by Aplet, Thomson, and Wilbert (2000).

Five different wilderness levels are visualized and are separated by dashed lines symbolizing the continuity and vagueness of wilderness phenomenon. The continuum shows that the larger naturalness and freedom from human control are, the wilder is a certain area. The attribute freedom has later been renamed into freedom from human impact and became an important wilderness attribute (Carver et al., 2012). In 2012, Steve Carvers selection of relevant attributes contained remoteness as initialized by Lesslie and Taylor (1985), naturalness (Lesslie et al., 1988), human impact (Aplet et al., 2000) and ruggedness represented by a digital elevation model (DEM). Carvers selection of attributes has proved its worth since it has been developed and successfully applied multiple times to varying regions like Scotland's national parks (Carver et al., 2012), Death Valley USA (Carver, Tricker, & Landres, 2013), Iceland (Tims, 2014) and Switzerland (Radford et al., unpublished). However, many wilderness approaches have applied different definitions for their models and none of them can be said to be wrong since the definition depends on subjective parameters and perception. Nevertheless, the selection of wilderness attributes and thus also the applied definition of a GIS-model is still subjective. This subjectivity has to be considered when interpreting results of a GIS-based evaluation like this approach. Thus, Carver's selection of relevant attributes for representing wilderness will be discussed in section 6.4.1 since the applied GIS-model in this work is based on Carver's attributes.

But why did the last paragraph concentrate on the importance of wilderness definitions and their attributes? The relevance of an accurate wilderness definition is high because decision makers and planning agencies orient their actions according to the results of such approaches. Their actions positively influence the environment, if the definition is accurate. The majority of the population in some way feels connected to nature and its surrounding landscape (Swanwick, 2009). Thus, the access to some kind of natural environment, unimportant if it means a distinct piece of wood or an urban park, is classified as highly relevant. Swanwick (2009) also states out that the majority of the population is convinced that the access to natural infrastructures improves life quality. The public interest of a healthy environment has increased in recent years in Switzerland. This can be recognized by current political debates and the emergence of NGO's like Mountain Wilderness[2]. Thus, public demand for decision makers and planners equipped with the best wilderness models increases.

The last paragraphs have illustrated that the wilderness definition is as individual as wilderness perception. In order to build a representative definition, many approaches have concentrated on different wilderness attributes to formulate such a definition. The research activities in the past thirty years show remarkably how complex the whole wilderness debate is and which challenges have to be faced. The following subsection refers to these challenges and demonstrates how science conceptually and methodologically deals with them.

2.3 Wilderness in research

John Muir and other activists initialized scientific interest to wilderness regions and propagated that wilderness is a necessity which needs protection (Muir, 1898). Since then, multiple methodological techniques to measure, analyse and visualize human-nature interaction have been developed. The following subsections give a brief overview about the different methodologies.

2.3.1 Social science approaches

When science began to develop interest in a proper definition of the wilderness phenomenon scientists were asking about how the population perceives it and which factors were relevant to them (Habron, 1998b). In order to collect information of the broader public, surveys and questionnaires in different forms have been applied concerning individual preferences and aversions to landscape and wilderness. Methodologically, the surveys can be divided into three main types: written questionnaires (Bauer, 2005; Bauer et al., 2009), photographic questionnaires (Habron, 1998a, 1998b; van Zanten et al., 2016b) and structured interviews (Coeterier, 1996). Approaches of the first type tend to ask the surveyed population theoretical questions about their perception and attitudes of wilderness and more general, nature and landscape. For the second type either real or manipulated photographs of varying landscapes have been applied to retrieve information about individual preferences and perceptions. The third type also uses photographs to represent landscapes, but rather asks for response orally than in written form. These three different methods allowed the scientists to generate information about different landscapes and how population judged the importance of recreation, preservation and protection activity. The generated social information supported local land managers and decision makers to better understand human-nature interaction and to establish recreation zones and protected wilderness areas (Bauer, 2005). The generation of that social information for the advancement of the general wilderness ideas and goals became increasingly important when digitalization allowed new technical methods to dispose the information and generate new output.

2.3.2 Quantifying wilderness by GIS-based models

Wilderness has a spatial meaning representing some kind of landscape area (Stremlow & Sidler, 2002). Wherever a spatial feature is set into scientific focus some kind of visualization, mostly in form of feature mapping is required. Therefore, Lesslie et al. (1988) developed a new digital mapping approach based on Geographical Information System (GIS) in 1988. Such a technical GIS-based approach normally requires quantified data, which can then be transformed into spatial attributes represented by spatial layers in the program. A weighted combination of these layers is called multi criteria evaluation (MCE) and results in a digitalized model representing the analysed region. In the case of Lesslie et al. (1988), the final map represents the region of Victoria in Australia, detecting wilderness zones to monitor the status of local wilderness resource. To specify which spatial features are relevant to the model, information generated by social approaches can be consulted. Thus, the information generated by social science approaches has high relevance also for GIS-based approaches although direct transformation into spatial information is challenging. Quantifying information gathered through large-scale perception questionnaires is not a simple process. But defining thresholds and determining key attributes, which can in a further step be used in a GIS approach makes it feasible.

The wilderness definition of a GIS approach is strictly connected to the attributes applied in the model, as described in section 2.2. The model generated by Lesslie et al. (1988) defines wilderness therefore by the attributes remoteness from access, remoteness from settlements, aesthetic naturalness and biophysical naturalness. Carver and Fritz (1995) refer to the approach of Lesslie et al. (1988) and worked on the evaluation of attributes concerning wilderness in the following years. These evaluations were also opened to a public audience in form of a web-based survey in order to collect perceptual information and refer to the perceptual nature of wilderness (Carver, Evans, & Fritz, 2002). Some years later Steve Carver initialized an approach which most of current GIS-based research refer to (Carver et al., 2012). Wilderness attributes as naturalness, human impact, remoteness and ruggedness build the base attributes for his MCE. But already before Carver started to work on GIS-based wilderness approaches, some critics about that method evoked. In 1993, Kliskey and Kearsley criticised that GIS-based approaches would be too mechanistic and would not take the perceptual nature and social aspects of wilderness into account (Kliskey & Kearsley, 1993). Fritz et al. (2000) encounter these critics by pointing on the strengths of GIS-based approaches as they are an effective and efficient way to analyse and visualize the wilderness phenomenon. This answer gives no adequate response to that critique, but advocates at least the benefit of GIS-based approaches. A more adequate answer would be to mention the practical usability of these approaches which no comparable method can achieve. However, modern technologies take note to that critique and offer new ways to combine GIS-models with social and perceptional data, so that the representation of the perceptual nature of wilderness is augmented.

2.3.3 Social media data / user generated content

With the initialization of the Web 2.0 social media became increasingly important (Antoniou et al., 2010). While the Web 1.0 was basically restricted to one-way communication, the Web 2.0 allows more flexible forms of communication. The presence on social media platforms provide their users new ways to interact with their communities, sharing content or posting photographs. Social media platforms build a large pool of data generated in a social context (Tenerelli, Demšar, & Luque, 2016). Using these large quantities of social data, or also called user generated content (UGC) offers new technical methods to deal with wilderness. In order to resume the arguments of the last section, these social media data provides the information requested by the critiques of Kliskey and Kearsley (1993). Since social media data most likely have been generated in a voluntary, social context, it suits to the critiques and can therefore be applied to augment a GIS-based approach. This has already been initialized by (Tims, 2014) who applied UGC for the wilderness attribute solitude. Comparable approaches have attempted to assess cultural ecosystem services (CES) by applying this data source (Gliozzo, Pettorelli, & Haklay, 2016; Tenerelli et al., 2016). The large potential lays in the open and free access of certain platforms, in the large quantities of available data and also in the partially or complete structured form these data can be accessed. Most of the platforms differ in one way or another and therefore the question arises, which of them matches most to the requirements of this work?

Different platforms, different purposes

The platforms offering free open data differ in multiple ways and a reflection about which might be most accurate for this analysis can only be suggested. Depending on the purpose, each platform offers different kind of data. While Twitter[3] generally focuses on tweets in form of text messages others like Flickr[4], Panoramio[5] or Geograph[6] concentrate on photographs. Many have compared the data provided by these platforms and have identified relevant differences like the scope of the platform and the supporting community (Gliozzo et al., 2016), spatial distribution (Antoniou et al., 2010; van Zanten et al., 2016a) or user participation (Antoniou et al., 2010; Purves, Edwardes, & Wood, 2011). Depending on the scope, the user's attitudes to upload content differs broadly. Purves et al. (2011) illustrate that the spatial distributions of photos on Geograph differs in a strong degree from the one of Flickr. According to the findings of van Zanten et al. (2016a), the users of Geograph are more spatially aware than Flickr users. This is valuable information for the decision, which platform should be used for the evaluation in this work. But depending on the evaluation, the different metadata provided by each platform have even more importance to this decision.

Accessible metadata

Each platform has its own collection of metadata they offer for open access. Most likely, all platforms provide text-based metadata like the title, tags, descriptions or comments of the shared content. Also more specific information, like the interest group, the user identification or very specific camera information can be accessed (Gliozzo et al., 2016). Others provide spatial information like the coordinates or the spatial accuracy. Valuable information can also be revealed by accessing temporal metadata like the dates when a specific post or upload has been generated (Antoniou et al., 2010). In order to retrieve wilderness information out of social media data, this work basically requires some textual information in form of tags and the spatial information in from of coordinates. But the broad selection of different metadata tempts to access more metadata than actually needed. This can lead to large data volumes which increase calculation time and effort and slow down the evaluation process. Thus, an adequate selection of required metadata is suggested to be predetermined in order to reduce downloading and processing time and storage. The knowledge about the different accessible metadata of different platforms is a valuable information source, which also supports the decision about the selection of a platform.

Working with social media data often means working with data generated by thousands of users showing different behaviours in activity and uploading quantity. These differences include much variability to the data which might be seen as an advantage in some cases but are

[3] https://twitter.com/

[4] https://www.flickr.com/

[5] https://www.panoramio.com/

[6] https://www.geograph.org.uk/

disadvantageous in others. The negative effects of these variabilities have been experienced, analysed and described by a few approaches and will stand in focus of the next paragraph.

Relevant characteristics of social media data

While considering social media data for an evaluation, certain characteristics need to be respected or at least considered. For research in general, these characteristics might cause relevant bias to their evaluations and need to be handled somehow (Purves & Mackaness, 2016). The differences in participation between the contributors of a platform have been detected as a very specific characteristic of varying social media platforms (Nielsen, 2006). According to Purves et al. (2011) this **participation inequality** can be described by a bimodal curve, whereas Flickr is less affected of that characteristic than for example the platform Geograph[6]. One of the reasons for such significant differences in user activity can be assigned to prolificness of certain users. A **prolific user** is a contributor that shows extreme activity which might affect the dataset (Purves et al., 2011). According to Purves et al. (2011), it has to be distinguished between most prolific users, which show a very high uploading activity and least prolific users, which are very inactive and have contributed only very few photos. Many users try a web service like a social media platform only once, and in case of Flickr, may upload only a single picture. Purves and Mackaness (2016) addresse this characteristic specifically and suggests to analyse the biasing effect of these prolific users to the whole evaluation. In case of a relevant effect, all data provided by prolific users need to be excluded in order to make sure the evaluation is not influenced by such bias. In case of no relevant effect, exclusion would only mean loss of valuable information and can be ignored.

Another specific characteristic which can cause significant bias to scientific researches with social media are contributors which upload large quantities of data at once. This phenomenon has been analysed by Hollenstein and Purves (2010) and called **bulk upload**. This special characteristic is not mentioned nor addressed in many approaches working with social media data. The few approaches who take it into consideration simply define their own rules which and how they manage this biasing characteristic. In order to minimize bias of bulk uploads Hollenstein and Purves (2010) excluded all data with same x- and y-coordinates and lower accuracy values than 9. A typical trait according to which data generated as bulk uploads can be identified is that certain parts of their metadata are identical. Most affected are identical tags, identical geolocation or even both, according to Hollenstein and Purves (2010). However, research has no clear procedural method which addresses bulk uploads. Due to the suggestion of Purves and Mackaness (2016), this work needs to consider the eventual bias caused by bulk uploads.

A broadly discussed research theme around georeferenced social media data is their **spatial accuracy**. Working with social media means to be aware of several sources for spatial errors as described by Hochmair and Zielstra (2012). The manual geo-referencing process is quite error-prone since it depends on the spatial knowledge of the user about the exact location when taking a photograph. Errors like the footprint mismatch or the similar object error described by Hochmair and Zielstra (2012) lead to spatial inaccuracy which might have strong influence to spatial analyses, especially by using a spatial model with high resolution as applied in this work. Rattenbury and Naaman (2009) describe place and landmark detection based on tags which could be seen from far away using the example of Golden Gate Bridge in San Francisco (USA). Many geotagged photos showing this bridge are not georeferenced at the location of the bridge itself but rather at the position where they have been taken. These kinds of errors reduce the spatial accuracy of social media data to a relevant degree. But in modern times, the geotagging process is no longer exclusively a manual process. According to Hochmair and Zielstra (2012) , many photographing devices automatically geotag their photos, as long as the devices get the signal of their position. This increases spatial accuracy and helps to prevent some of the errors described above. Improving such geotagging processes and increasing the accuracy is a major challenge in GIR research (Manning, Raghavan, & Schütze, 2008). Nevertheless, great care is required while interpreting the outputs of social media data analyses, as concluded by Purves and Mackaness (2016).

2.3.4 Retrieving information

The analysis purpose of this approach requires techniques to extract information from the social media data described in the last section. These techniques are provided by information retrieval (IR), or more specifically, geographic information retrieval (GIR). In general, these research fields develop retrieving systems for various purposes and attempt to increase the effectiveness of these systems Manning et al. (2008). Several research directions of GIR are relevant for this approach. Retrieving information from text generated in social contexts, such as tags, requires the consideration of several characteristics. On the one hand, the general characteristics of social media data described in the last section and on the other hand, tag-specific characteristics relevant to the information retrieval process. This work focuses on the retrieval of information from Flickr tags.

The tagging process is in general a manual process and requires a certain effort. Going to a place, taking a photograph and also tagging this photo, represents a social process that indicates certain associations of the user to the captured content. Tag analyses are applied in many research fields in order to retrieve these associations and detect patterns and trends within these social datasets (Abbasi et al., 2009; Gschwend & Purves, 2012; Hollenstein & Purves, 2010; Purves et al., 2011; Rattenbury, Good, & Naaman, 2007; Rattenbury & Naaman, 2009). But specific tag characteristics like the language, spelling mistakes but also term ambiguities challenge the retrieving processes. The following paragraphs briefly illustrate these characteristics of tags and their relevance to this work.

Language

Since the Web 2.0 can be accessed internationally, social media platforms are used by many language groups. Therefore, tags and other text metadata are principally generated in different languages. In order to retrieve information of such multi-language datasets these languages have to be considered (Purves & Mackaness, 2016). But not only different languages in general are challenging. Vernacular expressions have to be handled in regions like Switzerland or the Basque country, where the population at least partially speaks a native language broadly differing from the major languages like English, French or German (Purves & Derungs, 2015). These challenges are relevant to this work insofar as the study area refers to Switzerland, a country with four officially spoken languages and a native language in addition.

Semantic ambiguity

One of the major challenges in IR is the development of accurate disambiguation techniques (Manning et al., 2008). Multiple tags spelled in the same way may have different meaning and need to be distinguished in order to retrieve accurate information. Considering different languages within the analysed dataset make the disambiguation process even harder. Searching for specific tags is especially affected by ambiguity bias. Thus, for all interpretations concerning the semantic of tags, the potential risk for ambiguity has to be taken into account.

Toponyms

Toponyms or place names are tags assigned with spatial information and are therefore especially important in geographic information retrieval. They build the majority of all applied tags in platforms like Panoramio[5], Geograph[6] or Flickr[4] (Hollenstein & Purves, 2010; Purves et al., 2011; Rattenbury & Naaman, 2009). According to Zipf's law occur most of the terms in a social media dataset quite rarely and a few terms very often (Purves & Mackaness, 2016). This also counts for toponyms which has been further analysed by Hollenstein and Purves (2010) who describe in their work that city-, regional or and country-level toponyms are most frequently applied in their dataset. These unequal frequencies affect retrieving systems as certain tags are counted according to the frequency they have been applied to photos.

As all other types of tags are toponyms affected by the challenges of multilingualism and semantic ambiguity, as argued by Purves and Derungs (2015). But the spatial reference of toponyms leads to another challenge called toponym ambiguity. Smith and Crane (2001) describe toponym ambiguity by illustrating places assigned with multiple names or names assigned to different places while Brunner and Purves (2008) analysed spatial distances between ambiguous toponyms. These analyses show that the influence of ambiguities within datasets may influence also spatial evaluations and cause undesirable bias. Multiple approaches provide solutions to decrease such bias (Rattenbury et al., 2007; Rattenbury & Naaman, 2009). Another specific characteristic of toponyms is their spatial vagueness. This characteristic has already been researched by many (Humayun & Schwering, 2012; Jones et al., 2008; Montello et al., 2003; Purves & Derungs, 2015) and is seen as an elemental challenge in GIR. Hollenstein and Purves (2010) describes that the vagueness in human conceptualization of describing places is related to the quality and limitation of spatial knowledge on the one hand, and on the other to the continuous nature of geographic features, as for example wilderness.

The described characteristics have illustrated that spatial information retrieval from social media data might not be that trivial but includes many challenges. Although toponyms only have a minor role in this work, their relevance in this context is necessary to be mentioned since toponyms are a frequently accessed source in GIR. Nevertheless, GIR provides tools to retrieve information from social media data out of which one is particularly relevant to this work.

Tf-idf method

The basic origin of this method lays in general information retrieval which is interested in finding the most representative words for each document in order to classify the documents and their content thematically. Although a reference to the method description by Manning et al. (2008) would be sufficient, this paragraph attempts to give a brief overview about the method since it represents the methodological core of this work. A basic approach to evaluate this classification is the measurement of most appearing terms within the document which is called ***term frequency*** (tf). But not only is the pure number of term occurrence decisive to identify the specific thematic content of a document. Applying this method to a corpus of different sport magazines would most likely state the term *sport* as one of the most representative words for each magazine, even if this would be the case for all sport magazines. The classification of the term *sport* as a representative term for a football magazine would not allow distinguishing the football magazine from a golf magazine. The classification would be not specific enough. Thus, an additional parameter has to be applied which extends the information of tf with a document-specific information about each term (t). It requires analysing the number of documents which contain the specific term in order to retrieve the information of how specific a term matches to a document. This additional function is called *document frequeny* (df) and is measured inversely by dividing the total number of documents (N) by the df-value. Inversing the function gives it the name of ***inverted document frequency*** (idf). In order to strengthen the effect of the idf-value it will be logarithmized with the base of ten.

$$\mathrm{idf(t)} = \log(\frac{\mathrm{N}}{\mathrm{df(t)}})$$

Multiplying the number of times a term (tf) occurs with the information of the number of documents containing this tag (idf) represents the standard tf-idf function.

$$\mathrm{tfidf(t,d)} = \mathrm{tf(t,d)} \times \mathrm{idf(t)}$$

To explain the effects of this formula in words a short list according to Manning et al. (2008) helps out. Tf-idf assigns to term (t) a weight in document (d) that is:

1. highest when t occurs many times within a small number of documents;
2. lower when the term occurs fewer times in a document, or occurs in many documents;
3. lowest when the term occurs in virtually all documents.

A large interest has been set to the tf-idf-equation also in spatial analyses. Purves and Mackaness (2016) count the tf-idf method as a major methodological tool for exploring tag-based georeferenced photographs. Spatially oriented tag analyses have applied the function to their needs in order to identify most spatially representative tags for a specific region. Rattenbury and Naaman (2009) describe an approach to extract place semantics out of spatially referenced photographs assigned to clusters which represent the documents of the tf-idf method. Applying the formula to all photographs within a cluster for each occurring tag, a tf-idf score has been evaluated illustrating the representativeness of this tag to the specific spatial region. Others have applied the method for touristic reasons. For example Huang (2016) has applied the tf-idf function in order to identify the popularity of a location for touristic analyses based on social geotagged data. As concluded by Purves and Derungs (2015), their methods need to be used for answering research questions defined by domain experts. This work addresses their methods and applies tf-idf to wilderness research. This illustrates that the tf-idf method is applicable for many different research purposes and stands in focus of this work.

2.4 Concluding findings

The theory has shown that working with the thematic of wilderness is complex since its definition is still unspecified. The different aspects influencing wilderness perception and the vague nature of the phenomenon complicate an adequate definition. Practicing scientific research to wilderness requires a clear and strict definition which GIS-approaches implement in form of different wilderness attributes. GIS-based approaches have been criticised to not respect the perceptual nature of wilderness. Social media data include perceptual information as it is generated in a social context. Thus, combining a GIS-based wilderness approach with this social information addresses the critique on the one hand, but on the other might also generate valuable insights into wilderness which could not be evaluated in a different way. But also by considering social media data many influencing characteristics have to be respected. Since wilderness is a term with social and spatial aspects, the content as well as the spatial distribution of social media data is relevant to this work. Geographic information retrieval provides tools to work with social media data which methodologically stand in focus of this work. Retrieving wilderness information by combining, comparing and further analysing social media data with a GIS-based wilderness model describes the overall process of this work.

3 Data

This section describes the two main data sources applied in this work in more detail. On the one hand, the social media data accessed from the open port of the photo sharing platform Flickr and on the other the GIS-model generated by Radford et al. (unpublished) are explained.

3.1 Geotagged Flickr photos

Flickr[4] is a photo sharing web platform with the general goal to offer photo management and sharing services. The decision to concentrate on one single social media platform was taken in an early process step. Since Geograph is not available for Swiss regions and Panoramio is not openly accessible anymore because the service has been stopped only few opportunities for the area of Switzerland remained. Since this work concentrates on a spatial phenomenon, the geotagged photographs are an optimal way to verify, if the retrieved text-based content corresponds to the applied location. Thus, due to that reason and personal preference, Flickr has been chosen as base dataset for this work. Many have compared different social media platforms and identified various differences. This work does not contain relevant comparison to other data sources, though certain differences referring to their characteristics would be interesting to examine. However, the photos uploaded by the photo providers and users of Flickr can be accessed and downloaded over a web-based application programming interface (Flickr API)[7]. Each user has the opportunity to declare their uploaded photos as public. A public status of the photograph is a requirement for the downloading process over the API. Therefore, the base dataset of this work only contains photos with a public status. This work focuses on the metadata of each photograph rather than the photo in form of an image. As mentioned in the study area (section 1.3), only those photos taken inside Swiss borders are relevant to this work. Since Flickr does not necessarily request spatial referencing of the uploaded photographs only 4% of all photos on Flickr are geotagged as others have experienced (Hochmair & Zielstra, 2012). Thus, Flickr can be defined as a spatially implicit data source compared to related spatially explicit platforms like Geograph which urge their users to explicitly upload geo-tagged photographs (Antoniou et al., 2010). The retrieved base dataset containing metadata of 2’983’444 georeferenced pictures has been accessed at the 29th of April 2017 and reach back until 1970, although the Flickr service was initialized in 2004. But as default for temporally undefined pictures some temporal metadata is set to the year 1970 by Flick[8]. The metadata gathered by Flickr reaches from simple photo identification numbers over when they have been taken and uploaded to very specific camera features like brightness and contrast values. The relevant metadata for this approach are visualized in

Table 3.1. Downloaded as a comma-separated value file (CSV) and transformed to FlickrPhoto-objects in Java Eclipse Luna could each metadata specifically be accessed. Since every photograph needs a unique identifier Flickr applies a unique identification number (photo_ID) to all uploaded photographs. These unique number suites as a key attribute which is a requirement for a database but also very useful for Java programming.

[7] https://www.flickr.com/services/developer

[8] https://www.flickr.com/services/api/misc.dates.html

Metadata	data type
photo_ID	integer
accuracy	integer
longitude	float
latitude	float
user_ID	String
photo_title	String
tags	String
url	String

Table 3.1 – Retrieved metadata from the Flickr portal

Spatial accuracy differs between photo-sharing platforms as shown by Hochmair and Zielstra (2012). The Flickr platform relates their spatial accuracy level according to the applied zoom level during the geo-referencing process. Therefore, the approximate spectrum of the accuracy levels (see Table 3.2) reaching from zero to sixteen has been initialized by Flickr. For the base dataset a restriction to level 4 has been applied which excludes photos georeferenced at a zoom level of country size and smaller in order to reduce spatial accuracy bias. Tims (2014) decided to restrict his collection of Flickr photos to an accuracy level of 11 which would reduce the number of photographs in this work by approximately additional 3%. The distribution of the photos illustrated in Table 3.3 shows that the majority of the base dataset has an accuracy level above 10 and more than 50% of all photos have highest accuracy values. Therefore, it has been decided to not further restrict the base dataset in order to prevent losing valuable data.

Accuracy level	Accuracy number
World	1
Country	~3
Region	~6
City	~11
Street	~16

Table 3.2 – Flickr spatial accuracy levels

Accuracy number	Number of photos
4	1‘507
5	3‘024
6	7‘522
7	6‘584
8	7‘096
9	23‘687
10	39‘956
11	163‘000
12	307‘069
13	241‘645
14	299‘936
15	344‘780
16	1‘537‘638

Table 3.3 – Distribution of accuracy of base Flickr dataset

The user_ID represents the identification number of each user defined by Flickr. This ID can be used as unique identifier and allows user-specific analyses like tagging or uploading behaviour. Da Rugna, Chareyron, and Branchet (2012) have defined new unique identifiers for all user_ID’s accessed from Flickr for privacy reasons but since none of these ID’s will be published and the base dataset only contains photos declared as public this step has not been processed in this work.

Titles or tags of the photographs are seen as having large potential for Information Retrieval (IR). Combined with the spatial information of a GIS-model, valuable information can be gained according to the Geographic Information Retrieval (GIR) community. The second and the third research questions concentrate on tag-based analyses. This highlights the importance of tags to this work. Flickr tags have been applied by many other approaches for varying research purposes and have certain characteristics (Da Rugna et al., 2012; Di Minin et al., 2016; Schmitz, 2006). One of them is that Flickr has developed an automatic tagging functionality in 2015 which allocates auto-generated tags based on image recognition. This work concentrates to a considerable part on user-specific analysis where individual tagging behaviour is analysed. Therefore, auto-generated tags have not been considered in this work. Other characteristics of general social media data have been mentioned in the theoretical section (section 2.3.3) and some others will be discussed during this work.

3.2 Spatially explicit wilderness map

In order to connect the Flickr photographs and their metadata to wilderness phenomenon a second source of information has been accessed. The evaluation of the theoretical part has illustrated that scientific wilderness research requires a clear definition for the wilderness phenomenon. The definition applied in this work is based on a spatially explicit wilderness model, generated in a GIS environment by the federal institute *Wood, Snow und Landscape* (WSL). The different wilderness aspects and attributes are transformed into multiple criteria, which were then weighted by experts according to the multi-criteria analysis method. This GIS approach draws on the work of Steve Carver but is applied to the area of Switzerland and aims to spatially identify the potential of wilderness zones (Carver et al., 2012; Carver et al., 2013). As initialized by Carver et al. (2012), the WSL approach defines four base criteria: *naturalness, human impacts, remoteness* and *ruggedness*. These criteria contain multiple other aspects (see Appendix A) which makes the definition of wilderness quite complex. The work of Radford et al. (unpublished) has not been published but methodologically follows the rules of a leading wilderness scientist. Thus, the GIS-model can be used to represent Swiss wilderness regions. The final map applied in this approach (Figure 3.1) respects all experts asked for the criteria weighting process and distinguishes between 20 wilderness quality categories whereas the category 1 (red) represents the least, resp. category 20 (blue) the highest wilderness quality. The raster size of the map is 100x100 meters which stands for a high spatial resolution and makes a total of about four million raster cells. Compared to the approach of Carver et al. (2013) which included datasets with 1km^2 raster cells is the approach of WSL ten times more precise. But certain data sources for wilderness attributes have been aggregated according to Radford et al. (unpublished). The cells are not equally distributed across all twenty wilderness categories, which is important insofar as spatial analysis with these categories have to be normalized by the number of cells per wilderness category. The categories 1 to 3 are nearly not present in the dataset due to the multi criteria evaluations applied during the production of the map. Such low wilderness quality indices as classified for these three categories can hardly be achieved in Switzerland according to Radford et al. (unpublished). So the weighted criteria hardly result in a wilderness category classified lower than 4. Furthermore has to be taken into consideration that the ratio of these 20 wilderness quality categories represents the potential of wilderness quality zones within the Swiss borders. This ratio cannot be applied to another area outside of Switzerland without any adaptation. Depending on the region, the ratio would have to be extended or even compressed. However, all references to this wilderness GIS-model in this work either argue with the wilderness quality index applied by the model or their corresponding wilderness categories describing all hectares in Switzerland. So, talking about higher wilderness categories refer to regions applied with higher wilderness quality indices such as 15 or 20.

To come back to the wilderness map, obviously even with a few knowledge of Swiss topography can be determined, that most of the classified wilderness zones are in mountain regions. This fact has been confirmed to be applicable to whole Europe continent by the European Environment Agency (EEA) (2010). The weight and effect of this tendency to this work will be further discussed in section 6.4.1.

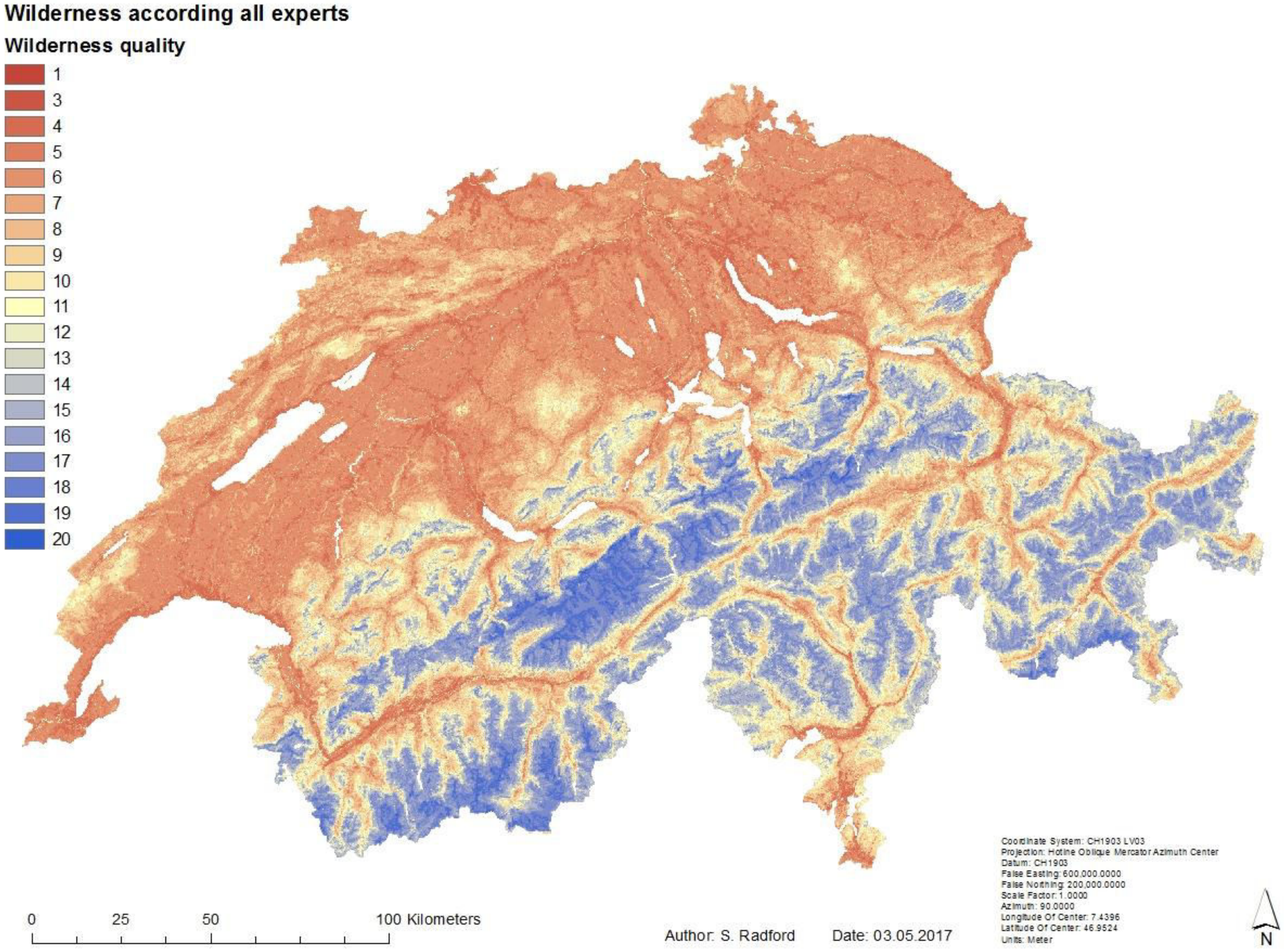

Wilderness according all experts
Wilderness quality
1
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
0
25
50
100 Kilometers
Author: S. Radford
Date: 03.05.2017
Coordinate System: CH1903 LV03
Projection: Hotine Oblique Mercator Azimuth Center
Datum: CH1903
False Easting: 600,000.0000
False Northing: 200,000.0000
Scale Factor: 1.0000
Azimuth: 90.0000
Longitude Of Center: 7.4396
Latitude Of Center: 46.9524
Units: Meter
N

Figure 3.1 – Wilderness quality map, initialized by Radford, Senn, and Kienast (unpublished), base dataset of this work

4 Methods

The methodological process of this work started with the data acquisition in form of retrieving the Flickr photograph metadata over the Flickr API and asking WSL to consign their GIS-model. As a first step, the Flickr data had to be tested for adequacy to fulfil the requirements for this approach. During these tests the wilderness map was related to the Flickr photographs as a second step. Third, with the spatial knowledge in which wilderness category each photograph is placed, tag-based tf-idf evaluation has been examined. And fourth, an attempt to further characterize wilderness according to the output of the third examination has been initialized.

4.1 Software and data structure

The majority of the work, including data acquisition, data management as well as most of the calculation steps and evaluation, has been processed in the Java-based integrated development environment (IDE) Eclipse Luna version 4.4. For some spatial analyses and for visualization purpose the ArcGIS 10.4.1 environment has supported the Java-based approach. Many steps done in the Eclipse environment could have been done alternatively in a database. Java has simply been chosen because of personal preference reason. Furthermore, Microsoft Excel of Office14 has been used for visualizing statistical data structures and tables.

4.2 Aptitude of Flickr photographs to scientific context

Scientific work with social media data requires the consideration of various specific characteristics according to the theoretical section (subsection 2.3.3). To determine if the data is convenient for this research and generally for wilderness research, certain tests are required in advance. Purves and Mackaness (2016) suggest that in order to get a quick overview over social media dataset the data has to be regarded from a global perspective. The applied tests are composed of analysis to spatial distribution and user-specific uploading behaviour analysis and attempt to generate this requested overview perspective. Based on the literature review four expectations to the Flickr dataset can be stated according to the typical characteristics of social media data. First, to address the findings of multiple evaluations (Antoniou et al., 2010; Purves et al., 2011; van Zanten et al., 2016a), the spatial distribution is expected to mainly concentrate on urban areas. Second, the Flickr data is expected to exhibit a bimodal activity curve, as described by Purves et al. (2011). In other words, few contributors are extremely active while the majority shows low activity. In order to prevent bias due to prolific users, Purves and Mackaness (2016) suggest evaluating the influence of them to the base dataset. Therefore, the influence of prolific users is expected to be high as third expectation. And fourth, certain bias can be expected coming from data generated in form of bulk uploads (Purves & Mackaness, 2016). The methods to evaluate these expectations are addressed by the following subsections.

4.2.1 Spatial distribution

Accessing the base Flickr dataset over the Flickr API returned a tab-separated values (TSV) file which built the base data file for all further research steps. Since wilderness is a spatial phenomenon the spatial distribution of the photos was analysed by extracting all coordinates and visualizing them in ArcGIS. In order to merge the Flickr data to the GIS-model, the number of photo per hectare was required. Therefore, a grid with 100 meter cell size covering the whole area of Switzerland has been initialized and each cell has been assigned by the number of photos on it. According to this grid, two separate maps have been generated.

Flickr photo density map

The grid described above has been transformed into points, assigned with the values of numbers of photographs. With these points, a point density evaluation considering their values has been initialized. Since a large contrast between the densest regions and the regions with less uploaded photographs can be observed, either a classification with geometric intervals or a classification respecting the neighbouring regions could have been applied. Figure 5.1 has been classified according to the latter by respecting an area of the surrounding three hectares or raster cells. This process has flattened the values of all map cells which explains the value range in the map legend.

Quantile map

Following the expectation that Flickr data hotspots focus on urban areas, the relation to population density may give further insights into where the Flickr community tend to upload photographs. In order to evaluate such tendencies a quantile map has been generated by O. Chesnokowa, C. Derungs and R. S. Purves at the University of Zurich. Since many hectares in Switzerland do not contain any Flickr photographs, the spatial information of the number of Flickr photos as well as the population had to be aggregated to two kilometres first. Then, the two data ranges have been normalized and classified into ten quantiles which have been subtracted in order to get a comparison between the number of Flickr photos and the population. Another way to analyse this phenomenon is to compare the Flickr density and population in form of surfaces by Chi-square test, as applied by Antoniou et al. (2010). Due to comparable results but less accurate output this step will not be further described nor illustrated within this work. A brief overview can be found in the appendix (Appendix B).

4.2.2 Relating social media data with GIS-based wilderness map

In order to analyse the aptitude of Flickr data to the wilderness debate the two information sources had to be combined. The technical steps which had to be applied to combine the GIS wilderness map with the Flickr photographs will not be explained in detail. In short, a scripted Java program has assigned the wilderness values of each photograph by accessing their coordinates and retrieving the corresponding wilderness value of the GIS-raster. Therefore, each photograph had a new attribute representing the wilderness quality index from the GIS-map with a range from one to twenty. In other words, each photograph has been assigned to a wilderness category, according to their location in the GIS-model. Some photographs could not be categorized by this assignment process due to the following spatial reasons and had to be excluded from the main dataset. First, the GIS-map has excluded all waterbodies from their multi criteria evaluation, which means that raster cells representing waterbodies have been classified with no data value. Accordingly, all photographs with coordinates referencing on waterbodies were out of interest for further wilderness-specific evaluation and were excluded. Second, some assignment errors occurred along the national border. Since the wilderness GIS map is represented in a raster of 100 meter grid-size and the national border is a vector-based polygon with very high resolution some small areas of Swiss territory have not been covered by the wilderness raster. That was the reason why some photographs not lying on waterbodies have not been assigned with a wilderness value, though they were lying within the national borders. In total, an exclusion of 138’299 (4.6%) photographs had to be taken into account which reduced the base dataset for wilderness evaluations to 2’845’145 photographs. With the applied steps the Flickr dataset was prepared for evaluations to the aptitude of Flickr data to wilderness research and the influence of user-specific uploading behavior.

4.2.3 User-specific uploading behaviour

According to the mentioned expectations, the aptitude of Flickr photographs is evaluated by analysing basic statistical values about the participation of all Flickr users, by analysing the influence of prolific users and the influence of bulk uploads. This section describes how this work has evaluated these influences methodologically.

Participation of Flickr users

The literature evaluation has shown that social media platforms aggregate data from users with varying uploading ambitions. Not only the variety in cultural background and individual preferences influence the data but also the reason why a person actually uploads a photograph. The influences of theses varieties have been described in previous sections (section 2.3.3) but still, this dataset requires more detailed insights into user behaviour, in order to confirm the data as adequate for this research. Therefore, some basic statistics have been evaluated in order to improve the understanding of the Flickr density map (Figure 5.1) and the uploading behaviour of the Flickr users. Relevant information about uploading behaviour can be found in general data characteristics like the number of contributors and their uploading activity which has been extracted from main dataset by an implemented Java code and visualized. Large differences in user uploading activity are expected, which is called participation inequality (Purves et al., 2011).

Bias due to prolific users

In order to analyse the influence of prolific users, conditions have to be compared between the dataset containing and not containing prolific users. According to two factors this comparison has been evaluated; the spatial distribution and the Flickr tags. Having a look at the distribution shows in which wilderness zones prolific users are most active. Evaluating differences in tags give insights into effects prolific users may have to wilderness-specific tag analyses. In general, if a relevant effect is identified, photos would have to be excluded from the base dataset according to Purves et al. (2011). If a relevant effect would be determined to the tags in wilderness areas, tag analysis could either not be recommended as adequate method in wilderness research or the prolific users would again have to be excluded from the main dataset. According to the knowledge of the author no threshold has yet been defined to set the relative number of uploaded photographs per user to classify him as prolific. Therefore, in this approach the threshold for high active users has been set to the upper 10% of all data which means the thirteen most active users would be classified as prolific. The lower threshold representing the barrier between normal activity and low activity users is set to less than five photographs uploaded per person. This threshold can be explained with the argument that contributors using Flickr service only once, most likely do not upload many pictures. Thus, they are very inactive and can be excluded for this research. Considering the fact that the research area has been defined to national borders of Switzerland, all photographs lying outside the borders have not been respected. This means that if a user generally uploads pictures in America for example but has visited Switzerland and uploaded one picture within the Swiss borders, only the latter picture will be considered. Thus, not all users with only few uploaded picture in the base dataset have used the Flickr service only once. However, excluding both, the most and the least prolific users according to the two defined thresholds would result in a reduction of 11.6% of all photographs as well as 47.9% of all users, visualized by Figure 5.6. According to Purves and Mackaness (2016), these prolific users and their generated data would be required to be excluded. But before doing so the effect of these prolific users has to be identified as relevant to the wilderness research context. An exclusion of photos having no influence to the output of the evaluation would be a waste of valuable data.

In general, Flickr's spatial distribution of uploaded photographs concentrates on urban areas compared to other social photo-sharing platforms like Geograph[6] (Gschwend & Purves, 2012). This could be confirmed by the analysis of the spatial distribution (Figure 5.4). This leads to the assumption that also prolific users basically concentrate their uploading procedure on urban areas. If that assumption would be correct, it would mean that the data generated by prolific users would only affect urban areas and wilderness zones would not be affected or at least only a little. To confirm this assumption, two evaluations have been initialized which both analyse the difference between the Flickr dataset with and without prolific users. For both evaluations, the GIS-model has been considered and all photographs have been classified according to their corresponding wilderness category, as described in the last subsection. The first evaluation analyses the difference in number of tags in total for each wilderness category. This evaluates, in which wilderness categories the prolific users have been most active, according to the GIS-model. The second evaluation concentrates on the difference of unique tags between the dataset with or without prolific users. The specification to unique tags per wilderness category gives further insights into the uploading behaviour of prolific users. Evaluating these differences show how significant the bias of prolific user data is and gives partial answers to the first research questions and to the question, if the prolific user data should be excluded from the main dataset or not.

Bias due to bulk uploads

The most active user has uploaded 57'504 or 1.9% of all photos. The question arises, which uploading behaviour this users has. Uploading such a quantity of data means much temporal effort. In order to reduce this effort, social media users sometimes tend to upload hundreds of photographs at once.

The Flickr platform does not restrict their contributors to a certain uploading quantity or time. Thus, users have the ability to upload large quantities of photographs at once, all with the same metadata like coordinates, tags or other text. This phenomenon is called bulk uploads according to section 2.3.3. One reason why bulk uploads are generated most likely is to reduce temporal effort of the uploading procedure but other reasons are possible. Bulk uploads can cause bias to the main dataset according to Purves and Mackaness (2016) and need to be excluded from the main dataset. Additionally, Hollenstein and Purves (2010) mentions that especially in spatially oriented analyses and tag semantic analyses the bias of bulk uploads have to be taken into account. One way to detect bulk-uploading users is to compare the metadata of the uploaded photographs on similarity, as applied by Hollenstein and Purves (2010). But as mentioned before, coordinates are not the only comparable kind of metadata for analysing metadata similarity. Alternatively, text-based metadata like tags could be compared. The question is which of these two show more accurately the similarity between photo metadata? Coordinates would be the perfect parameter to show similarity if they would not differ within a bulk upload. This work states the expectation that coordinates show large differences within the same bulk upload due to the following reason. Modern devices like cameras and mobile phones automatically assign their location, if accessible for the device, to the taken pictures according to Hochmair and Zielstra (2012). If a user uploads these photographs not directly but in a later step as bulk upload, the coordinates will already be assigned to the photo and no metadata similarity could be detected regarding the coordinates. Since the majority of all photos have highest spatial accuracy values, a relevant number of photos most likely have been geo-tagged automatically by the devices. Therefore, better representative than the coordinates may be tags, generated exclusively by manual user process. Since the mean number of tags per photo is 6.2 when all empty tags are excluded, the risk that a user applies the same combination of tags more than once for different uploads is seen as negligible. Thus, it is expected that tags differ less within a bulk upload than coordinates. In order to confirm this expectation, a non-representative data analysis has been initialized where two very prolific users have been analysed. All uploaded metadata of these users have been categorized according to their tag-combinations, so that for each tag-combination a list of all photos applied with that tag-combination has been generated. If the coordinates would be a better indicator for metadata similarity, as applied by Hollenstein and Purves (2010), the spatial distances between the photos within these lists would be minimal or even zero. But the evaluation has shown that the distances between the coordinates within a list of photos with same tag-combination of a specific user are sometimes many kilometres. The mean of both users for identical coordinate within the same list of photos with identical tag-combination reaches at 71%. Thus, tags or tag-combinations seem to be more adequate for analysing metadata similarity for the base Flickr dataset. Therefore, in contrast to Hollenstein and Purves (2010), this work considers the tag metadata to determine if a photograph has been generated in form of bulk uploads or not.

According to the theoretical introduction (section 2.3.3), all metadata generated by bulk uploads should be removed from the base dataset. But how many photographs need to be uploaded at once in order to be classified as bulk uploads? Is an upload of ten photos applied with the same metadata already a bulk upload? Again, literature has, according to the knowledge of the author, not defined any threshold which addresses this question. Accordingly, this work defines a threshold of about minimal 500 uploaded photos per user to identify a user as a bulk uploader. The value of this threshold has to be set according to the bias a bulk upload of this size would have to the main dataset. Worst scenario which could happen by setting this threshold to 500 photos would be that a user who has uploaded 499 photographs in total would have uploaded all of them at once. Spatially, this bulk upload would have a strong influence on a local scale which could especially in tag-analysis show disturbing bias. In order to reduce the risk for following analyses with tags from such a case, the number of contributors which have applied the specific tags would have to be taken into consideration. Additionally, taking into account that the total number of photos applied in this analysis is around three million pictures a local weight of 499 photos should not count too much on a national scale. Considering this threshold, a new approach to identify bulk uploading users has been initialized.

New approach to bulk uploads

In order to identify a bulk uploading contributor, the following criteria have to be fulfilled:

- The contributor has at least uploaded a certain quantity of photographs (500).
- The uploaded metadata of the contributor need to be highly similar, resp. show low percentage of unique metadata.

To measure these criteria, firstly, the base dataset had to be reduced to contain only the users which have uploaded more than 500 photos. For the remaining 885 users the uploading behaviour has been classified according to their tag metadata similarity as a second step.

For this bulk upload analysis, no single tags but the tag-combination of each photograph have been analysed. A **tag-combination** means the whole string of all tags separated each by a delimiter. This means that photographs with same tags but different order of tags are not seen as similar. Uploading pictures with same tags but different tag order requires additional effort than just copy-pasting the same tag-combination for a bulk of photos, which is the reason why this tag similarity analysis does not respect individual tags but only tag-combinations. These combinations of tags have been classified according to whether they have been applied only to a single photograph or to multiple ones. In the first case the tag-combination would be seen as unique. Therefore, for each user the percentage of unique tag-combinations (**UTC**) has been determined in this analysis. This value gives more insight into user-specific tagging behaviour and already gives information about how precise the user describes his photographs with tags. Basically, the smaller this value the more a user tends to bulk uploading.

$$\boldsymbol{UTC} = How\ many\ percent\ of\ all\ individually\ uploaded\ tag \\ -\ combinations\ are\ unique$$

But only using the UTC is not enough to determine a user to be a bulk uploader or not. At this point it has to be reflected that a user uploading thousand photos and applying a unique tag-combination to every second picture would have no unique tag-combination even if the user would have applied 500 different tag-combinations and would therefore be classified as bulk uploading user. To avoid this bias the UTC-value has been extended to a formula by an additional parameter. The most frequently used tag-combination (**MFTC**) of each user has been evaluated as well as how much percentage of all of the users posted photos this MFTC has been applied to (**MFTCa → MFTC-appliance**). The higher this value, the more a user tends to be a bulk uploader. As an example, the MFTCa-value stands at 35% if the corresponding user has applied his MFTC to 35% of all his uploaded photographs.

$$\boldsymbol{MFTCa} = \ How\ much\ percent\ of\ all\ individual\ uploaded\ photos\ have \\ been\ tagged\ with\ the\ MFTC$$

Since this value is relative and considers the total number of uploaded photographs by a user, the value range lies between hundred and zero. But the weighting within this range cannot be seen as linear. A MFTCa-value of 100 means that the user has only used one single tag-combination for all his photographs and can be seen as an extreme bulk uploader whereas a value of 50 means that this user has at least uploaded 50% of his data in form of bulk uploads. In the most extreme case the other 50% have been tagged with unique tag-combinations, which might rarely be the case in reality though. But even in this extreme case the user tends to bulk uploading because 50% of his data have not unique metadata. This shows that the range of this MFTCa-value cannot be weighted linearly or in other words, an exponential range is closer to reality.

It has also to be taken into consideration that these two values cannot be weighted equally because the MFTCa is more decisive to detect bulk uploading than UTC. This can be explained as follows. On the one hand, a user with a very low UTC value does not necessarily need to be a bulk uploader as described in the previous example whereas a user with a very high MFTCa absolutely needs to be a bulk uploader. On the other hand, UTC-value is affected by MFTCa insofar as the higher the latter the smaller the range of UTC becomes. A MFTCa of 75% does not allow UTC to be larger than 25%. Thus, weighting these two values as equal would not represent real conditions. This is why the MFTCa has to be classified as more decisive and requires an exponential range so that the MFTCa is squared by 2. The division of the UTC from a squared MFTCa build a new formula which represents the **bulk-index** (**bi**). This value refers directly to the question if the uploading behaviour of a user is classified as tending to bulk upload or not. The higher the value the more a user is classified as a bulk uploader.

$$\mathbf{bi} = \frac{MFTCa^2}{\mathrm{UTC}}$$

Squaring MFTCa, which can in most extreme case have a value of 100, leads to large variations in range of the formula, which complicates the interpretation of corresponding results. In order to increase the qualifying process, the value range had to be reclassified. Taking the logarithm of the current formula reduces the value range from -5 to 5, which augments the differentiation between bulk uploaders and non-bulk uploaders.

$$\mathbf{bi} = \log_{10}\left(\frac{MFTCa^2}{\text{UTC}}\right)$$

Two special cases have to be taken into account while applying this function. First, if the UTC value of a certain user is zero, which is conceivable, a dividing-by-zero exception has to be handled. This work has assigned a value of 0.1 for such cases. Second, many users tend to apply no single tag to their uploaded photographs which has not been respected in the downloading process of the main dataset. However, the so called ***empty tags*** are seen as normal tags and would all be classified as being generated within the same upload, which is most likely not true in most cases. Without respecting these empty tags, a weighty risk exists that all photographs applied with empty tags are counted to the MFTC. This would be a large bias to user-specific analysis since some users apply no tags to the majority of their photos, which does not imply that they have been uploaded as bulk uploads. This would lead to a high MFTC-value for this user even if he does not tend to bulk uploads. Thus, the photos with empty tags have to be excluded from the MFTC-evaluation and should neither be counted as unique for the UTC-value.

Applying the bulk index equation to this work can give valuable insights into user-specific uploading behaviour and help to determine if the data generated in form of bulk uploads need to be removed from the base dataset. Additionally, it provides a first opportunity to reveal which users tend to bulk uploading and classifies them according to their uploading behaviour.

4.3 Tag-based evaluations on wilderness

The theoretical background at the beginning of this work has revealed that the GIS-based approaches can be extended by data generated in a social context such as Flickr photographs. Considering the broad variety of different metadata the Flickr data contains, a variety of opportunities exist to combine the Flickr information with the GIS-model. In order to answer the second research question, especially the coordinate metadata and the tags of each Flickr photograph will be considered. This section describes the methodological steps applied to evaluate relations between the GIS-based approach and Flickr data.

4.3.1 Specific tags representing wilderness

The most straight forward approach to evaluate relations between the GIS-model and the Flickr dataset would be to verify the spatial distribution of the tag *wilderness* according to the twenty categories representing wilderness quality. Tims (2014) illustrate that the low number and the spatial distribution of photographs tagged by *wilderness* do not allow accurate analyses in Iceland. This is also the case in Switzerland where less than 600 geotagged photos within the bounding box of the political borders could be detected. Thus, another approach needs to be taken into consideration. One opportunity is to evaluate a collection of tags which most represent wilderness. This collection has been evaluated in two steps. First, a literature-based evaluation for terms which are semantically close to wilderness has been initialized. As a second step, the co-occurrence of all kinds of different tags with the tag *wilderness* has been measured. As suggested by Purves and Mackaness (2016) is the co-occurrence of social media tags a valuable method to evaluate the meaning of tags and can aid to understand potential ambiguities between them.

Representative terms according to literature
The term *wilderness* has many associations and facets as highlighted by the theoretical part of this work (section 2.2.1). Some associations can be used to represent wilderness by other terms which has been evaluated on the base of definitions and other literature. This evaluation is rather arbitrary and risks of bias due to subjectivity which need to be considered when interpreting the results.

Co-occurrence of tags with wilderness
Measuring the co-occurrence of tags is a simple way to evaluate the tags which most represent another specific tag (Purves & Mackaness, 2016). Retrieving all photographs containing the specific tag *wilderness* or *wildnis* and evaluating the tags most co-occurring with these tags returns a ranked co-occurrence list. This has been initialized by an implemented Java code. The tags with highest co-occurrence coefficients have been taken into consideration as being the most representative tags for wilderness within the main Flickr dataset. Although most photographs tagged with wilderness only lie inside the bounding box but not within the borders of Switzerland they have been used for co-occurrence analysis.

4.3.2 Detecting wilderness by tf-idf evaluation

A more developed way to identify correlations between tags and the wilderness concept is to apply a method called tf-idf, as discussed in the theoretical part (section 2.3.4). The combination of the co-occurrence analysis and the tf-idf method approaches the research question from two different perspectives which increase the quality of the whole tag-based evaluation of this work.

The aim of the tf-idf function is to evaluate and rank terms occurring in a predefined number of documents according to their representativeness for a specific document. In the case of this work the function has been applied to the wilderness concept, where the documents are represented by wilderness categories defined by the GIS-model. Applied to the thematic of this work the question answered by this tf-idf method can be formulated as follows: Which tags best represent each of the twenty wilderness categories?

The simplest way to detect the most representative tag within each category would be to have a look at the most frequent tags occurring in high wilderness categories. Working with social media data some characteristics have to be taken into account which obviously reveal that simply respecting tag frequency is quite error-prone and only shows the distribution of tags over all wilderness classes. As denoted in section 2.3.4 represent toponyms or place names a sizable proportion of the whole quantity of uploaded tags whereas toponyms of city-, regional- or and country-level are most frequently applied. This fact could also be measured in the main dataset of this work where tags like s*witzerland, schweiz* or *suisse* are the most frequent tags in most wilderness categories. But tags like *switzerland*, *zurich* or other regional toponyms do not really give valuable information about wilderness though. So toponyms do not seem to be the appropriate kind of tags for this analysis and therefore, only focusing on the frequency a tag occurs within a wilderness category gives no valuable information about the representativeness of this tag to that specific category. Additional to frequency, the information about how many times a tag occurs in the specific category compared to all the other categories has to be respected at this point. The stronger the quantity of same tags concentrates on a specific category the more representative it is for that category. This is what the tf-idf represents and it's the reason why tf-idf has been applied in this work instead of simple frequency. A more detailed definition of the equation can be found in section 2.3.4.

Nevertheless, the characteristics of tf-idf evaluation have been analysed according to the applied Flickr dataset. This stated out that the different factors of the tf-idf equation have strong influences on the output of the evaluation. Two factors and their influences will be described briefly in the next two paragraphs.

1. Number of photographs and tf-value

The tf-value represents the number of times a specific tag has been applied within a wilderness category. Especially toponyms as described above score very high tf-values since they are applied very frequently by many users. The weight of the tf-value is for those frequently tagged terms so high that their tf-idf values cannot be surpassed by any other terms. Thus, either is the weight of the tf-value too highly weighted for these terms or the terms simply cause too much bias to the dataset due to their frequent use.

2. Number of compared documents influences tf-idf value range

In this work only 20 documents or wilderness categories have been compared. Accordingly, the idf-variable, which is strongly related to the number of tested documents, is restricted to 20 categories, which is not a large volume of documents. This has a strong influence on the values returned by the method. Applying the idf-equation to these 20 documents the highest score this value can achieve is 1.3, scored if the tag occurs in only one of all documents. The idf-part of the equation should increase the rank of tags which are more and decrease others which are less representative. This method has been developed in general IR research where hundreds of documents are compared. The ideal number of documents to work with tf-idf is about hundred documents because in this case the idf-value promotes tags appearing in less than the half of all documents and decreases the ranks of tags appearing in more than the half of all documents. Thus, the breakpoint is in the middle of the scale. Working with 20 documents the breakpoint lies at two of twenty documents, so in this approach only tags occurring in two or one document are promoted. The influences of this factor affect the ranking of this evaluation and therefore require a critical reflection.

Three disturbing factors have been evaluated at this point biasing the result of the standard tf-idf evaluation. These factors have been transformed into challenges to improve results of this method and find the most adequate ranking system for tag representativeness to this approach. First, tags with very high frequency which are generated by most of the users, described as the first observed factor in this subsection, had to be managed. Second, a reflection about the effect of the second factor was required and third, a way to deal with the strong weight of high frequency tags generated by very few users had to be found. The reaction to these challenges was twofold. On the one hand, an exclusion of all toponyms has been initialized as a reaction to the first challenge, and an adaption to the tf-idf equation referring to the third challenge has been applied on the other. Reflections about the second challenge were only held shortly and no concrete reaction has been initialized therefore.

Tag exclusion

The strong presence of toponyms causes bias to the tf-idf equation, as explained previously. This bias influenced the results that some non-representative tags appeared in the top ranks of tf-idf evaluation which has to be avoided.

The exclusion encompasses toponyms on the one hand but also other tags which either have been generated automatically by the devices or applications the photographs have been taken or they simply give no valuable semantic information. In order to give some examples for these specific tags the terms *uploaded:by=flickrmobile* or *iphoneography* could be mentioned. Tags generated by the photographing devices cannot be taken into account in this evaluation because the tagging process was not a manual process but rather a digital one. A second reason to exclude these kinds of tags is that in this approach neither auto-generated tags of Flickr have been taken into account. Thus, this approach follows the strict rule to concentrate on manual tagging processes and to exclude tags generated by computer programs. The exclusion was managed manually after the tf-idf lists have been created. In other approaches this manual process has been done by applying a database connected to a gazetteer which checks for toponyms (Jones et al., 2008). Since the temporal effort to build such a database and exclude all these toponyms would have exceeded the effort of manual exclusion by far and specific tags like *swissalps* or *schweizervoralpen* might not have been caught by the database exclusion the manual way has been preferred in this case. Many of these lists contain more than hundred thousand tags which is the reason why not the whole list could have been modified. The focus of exclusion has been set to the top 50 ranks per wilderness category which is why the list of excluded tags attached in the appendix is not longer. Terms with lower ranks are never counted as representative tags and are therefore irrelevant for the discussion. However, the list of excluded tags is long and in order to have a clear structure a classification according to Hollenstein and Purves (2010) has been applied who structured the excluded toponyms according to their spatial extent from country- to local-scale. The complete, exact and classified list of all excluded terms can be found in Appendix C.

Reflections about the number of compared documents

The second described characteristic of the tf-idf evaluation refers to the challenge that the number of compared documents has an impact on the value scales of the evaluation. The literature review at the beginning of this work has shown that the actual tf-idf values are rarely discussed in research but rather their ranks within the output lists. Since this work follows that strategy and the actual values of the ranked lists will not be discussed, no reaction was required to that challenge as long as no comparison between the tf-idf-ranking of different data samples is examined. However, the number of documents compared in this work is restricted to the 20 wilderness categories anyway. So no additional reaction has been done referring to this second challenge.

Adaption of tf-idf formula

The third challenge refers to the strong presence of high frequency tags generated by just very few users. Previous steps have already shown that an adaption of the standard tf-idf equation would be required in order to avoid tags generated by very few users being in the top ranks of tf-idf. Adapting the equation to the requirements of research is not uncommon. Already the prolific user analysis and bulk uploads refer to user-specific characteristics and uploading behaviour and shown that a single user can have relevant impact on spatial distribution of photographs and hence also on tags. Rattenbury and Naaman (2009) as an example has applied an additional factor to the normal tf-idf method which respects the important information by how many users a specific tag has been uploaded within a category. This additional information is important insofar as the standard mode of tf-idf does not consider this factor and this approach requires respecting the effect of that factor. This idea has been applied for this work in form of an additional parameter here called **user frequency** (uf). For each tag (t) the number of users (U(t)) this tag has been generated by within the corresponding wilderness category has been evaluated and divided by the total number of users (U) who generated photographs within that category. This division builds the uf-value which is multiplied to the standard equation of tf-idf.

$$\mathrm{uf(t)} = \frac{\mathrm{U(t)}}{U}$$

The adapted equation looks like the following:

$$\mathrm{tfidf(t, d)} = \mathrm{tf(t, d)} \times \mathrm{idf(t)} \times \mathrm{uf(t)}$$

According to this adapted version, subsequently called tf-idf-uf-equation, the final ranked output lists have been generated in order to answer the second and third research questions. The toponym exclusion and the consideration of the uf-parameter append two new characteristics to all tags ranked by the outputs of the tf-idf-uf equation. Accordingly, the current state of all characteristics the ranked tags embody at the current state are enumerated here:

1. The tag has been applied frequently within the category
2. The tag has been registered in only very few wilderness categories
3. The tag has been generated by many users
4. The tag is neither a toponym nor a tag applied by the device or application

4.4 Characterization according to tf-idf-uf evaluation output

The results of the third research question have been generated on base of the same tf-idf-uf output lists as the ones for the second research question. Thus, no additional methodological steps were required to generate the results for the third research question. All characterizations are based on the ranked lists of the tf-idf-uf evaluation and the spatial distribution of these tags across the wilderness categories.

5 Results and interpretations

This major section objectively describes the outputs of their corresponding methods of section 4 while the structure is oriented according to the research questions (section 1.2). For each question a subsection is defined where the results are prepared for their discussion in the subsequent section (section 6).

5.1 Aptitude of Flickr photographs to scientific context

5.1.1 Spatial distribution

Density map

As mentioned in the methods (section 4.3.1), the Flickr density map (Figure 5.1) has been classified by geometric intervals and respects the neighbouring three hectares for each cell. The data ranges from 0 to 178, which does not represent the actual number of photographs per hectare but gives a relative overview. The actual value range is less important since the required information can be extracted also from relative values. On the map, most relevant hotspots are visualized by red circles and the colour range indicates high photo density with dark red colours. The hectare with highest photo density can be found in Therwil (BL) with 18’907 photos and the second highest in Romont (FR) with 14’442 photos. Although these two hectares have very high numbers of georeferenced photos, no large hotspot can be found on the map at their position. Respecting the neighbouring hectares flattens the very local hotspots so that they no longer reach highest values on the map. Long lines of medium photo densities indicate that either the contributors took photographs while driving on roads or when passing valleys. The main valleys are especially highlighted in the canton Wallis but also in Graubünden.

Quantile map

The quantile map (Figure 5.2) shows different information compared to the density map, since the quantiles of numbers of photos per cell are compared with the quantiles of population. Red colours refer to larger numbers and blue to smaller numbers of Flickr photos than expected. Although the resolution is two kilometres on the quantile map, certain tendencies can be detected. The midland, generally more populated than the mountain regions, shows more blue coloured areas while mountain regions principally have more red areas. Where the quantiles of the compared two datasets have approximately equal quantiles, the regions are coloured with a slight yellow. These areas mainly concentrate on the city hotspots, visualized in the density map. This makes sense due to the fact that these areas have high population and high Flickr photo density. The zones with highest values of Flickr quantiles compared to the population can be found in high mountain regions such as the Jungfraujoch, Matterhorn but also Säntis in the Appenzellerland and Pilatus in Luzern. The more populated places in the Alps like the Rhonetal in the canton Wallis or the Rheintal between Landquart and Chur stick out as having less Flickr photographs than expected.

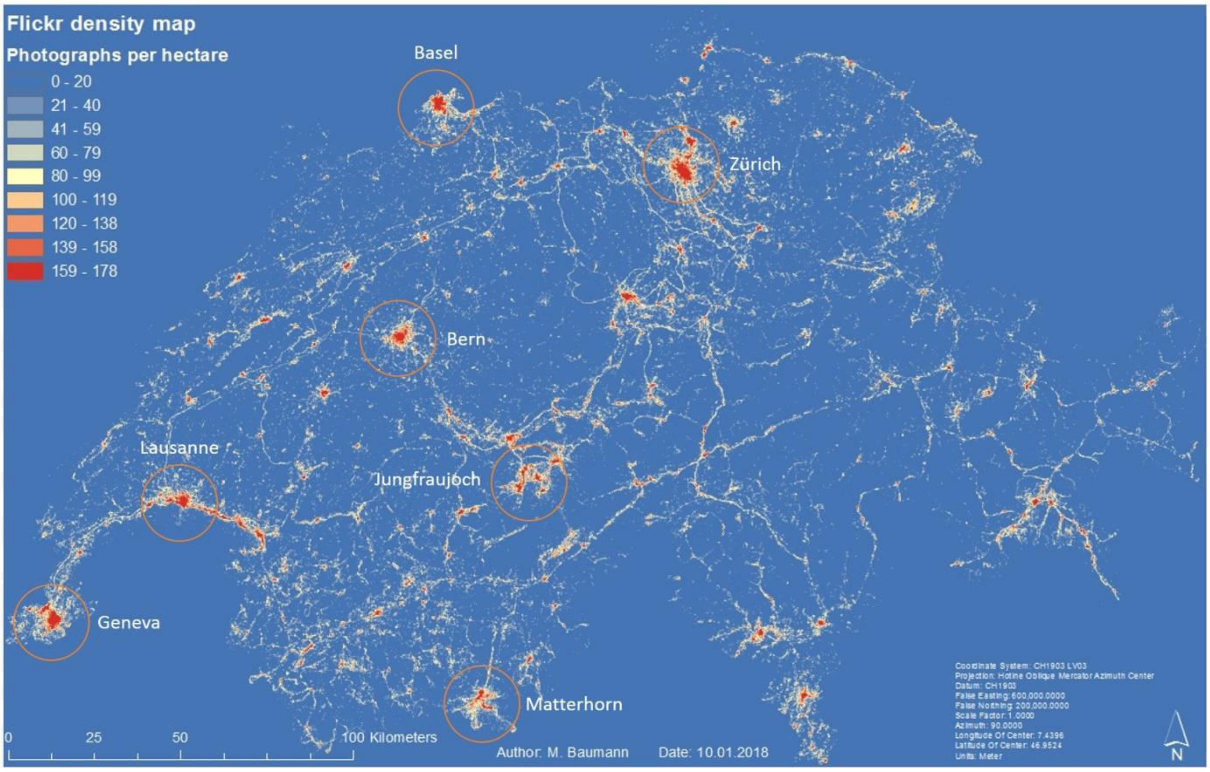


Figure 5.1 – Flickr density map (circles symbolize largest hotspots). Own creation

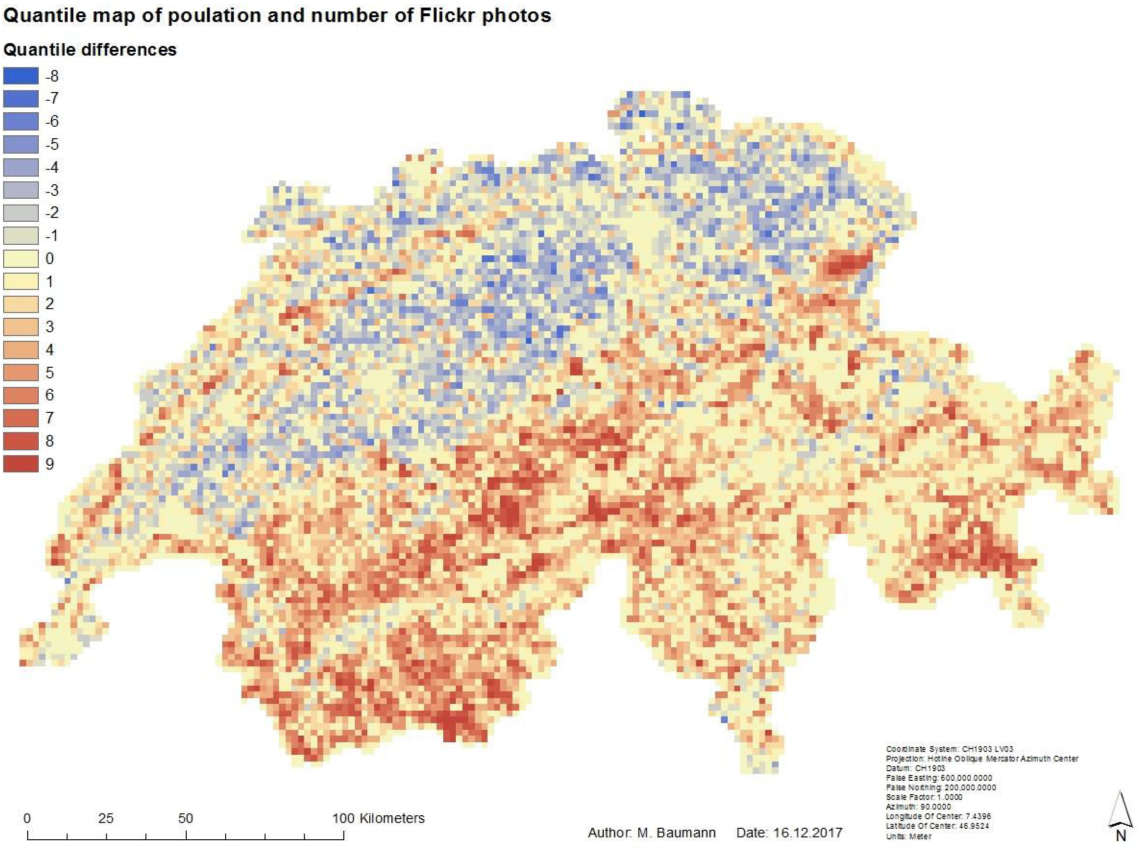

Quantile map of poulation and number of Flickr photos
Quantile differences
-8
-7
-6
-5
-4
-3
-2
-1
0
1
2
3
4
5
6
7
8
9
0
25
50
100 Kilometers
Author: M. Baumann
Date: 16.12.2017
Coordinate System: CH1903 LV03
Projection: Hotine Oblique Mercator Azimuth Center
Datum: CH1903
False Easting: 600,000.0000
False Northing: 200,000.0000
Scale Factor: 1.0000
Azimuth: 90.0000
Longitude Of Center: 7.4396
Latitude Of Center: 46.9524
Units: Meter
N

Figure 5.2 – Quantile map comparing Flickr photo density to population density.
Created by O. Chesnokova, C. Derungs and R. S. Purves

5.1.2 Participation inequality and prolific users

A list where users are ranked according to their number of uploaded photographs has been retrieved from the base Flickr dataset. The participation of Flickr users is expected to represent a bimodal curve, according to other experiences working with Flickr data. Due to large differences in participation, the graph illustrating the conditions of Flickr participation inequality has been simplified (Figure 5.3). Out of 52’313 users, only 420 generated more than thousand photographs. On the one hand, the most active thirteen users generated 10% of all Flickr photos of the main dataset. On the other hand, approximately 62% of all users generated not more than ten photographs. An extreme bimodal curve would be the real case but an abstracted version served better to show further decisions to the prolific user evaluation.

Figure 5.3 visualizes symbolically which part of the data would be excluded if the prolific user analysis would reveal negative influences on further evaluation outputs. The dashed red lines symbolize the upper and the lower threshold according to the definition in section 4.3.3. The blue areas in the graph symbolize the data generated of all users defined as prolific. Since the bimodal curve has been abstracted, the real area would be much smaller. But before carrying about the results of prolific user evaluation, some relevant statistics about the combination of the GIS-model and the Flickr dataset need to be explained.

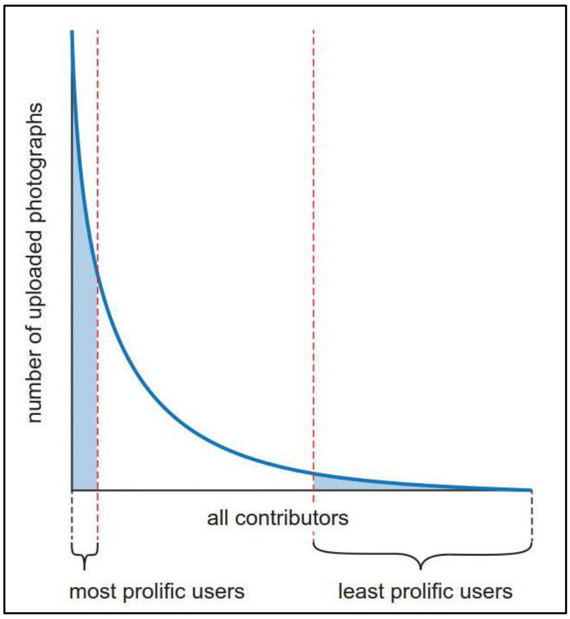


Figure 5.3 – Bimodal activity curve representing participation inequality of Flickr contributors. Symbolizing the relative part of data generated by contributors classified as prolific users. Own creation

The twenty wilderness categories of the wilderness quality map are not equally distributed neither are the number of cells per category equalized. In order to get a better understanding of the distribution Figure 5.4 illustrates the distribution of the map cells across all categories. Three characteristics are relevant at this point. First, the categories 1, 2 and 3 contain nearly no cells, as already mentioned in the data description (section 3.2). Second, especially the category 6 and other high numbers of cells with lower wilderness quality point out nicely the superior presence of regions with low wilderness quality classified by the GIS-mdoel. Third, the three categories with highest wilderness quality values show only a low presence in the GIS-model, which must be taken into consideration when interpreting results considering these categories.

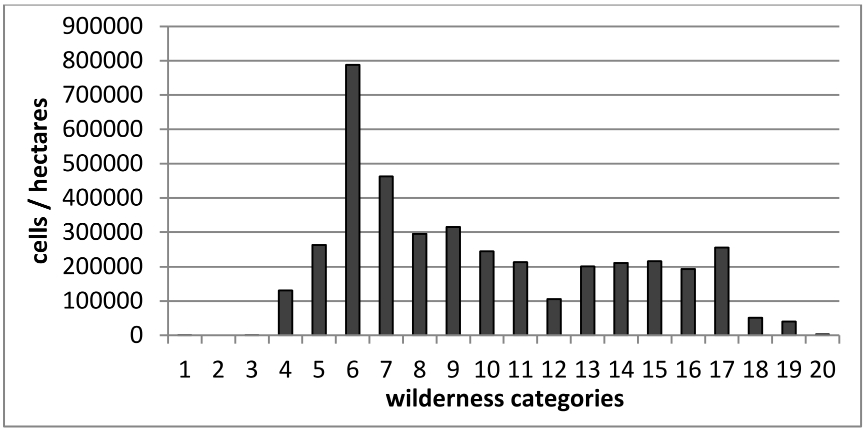


Figure 5.4 – Distribution of map cells or hectares per wilderness category

The number of cells per category gives insights into how Swiss landscapes have been classified by the GIS-model. The combination of the model and the Flickr data has been processed, so that each photograph has been assigned to the corresponding wilderness category. Figure 5.5 illustrate the distribution of all photographs across all wilderness categories, normalized by the number of category cells mentioned before. Despite the normalization, the majority of tags keep being in lower wilderness categories. These insight need to be respected while interpreting and discussing the results.

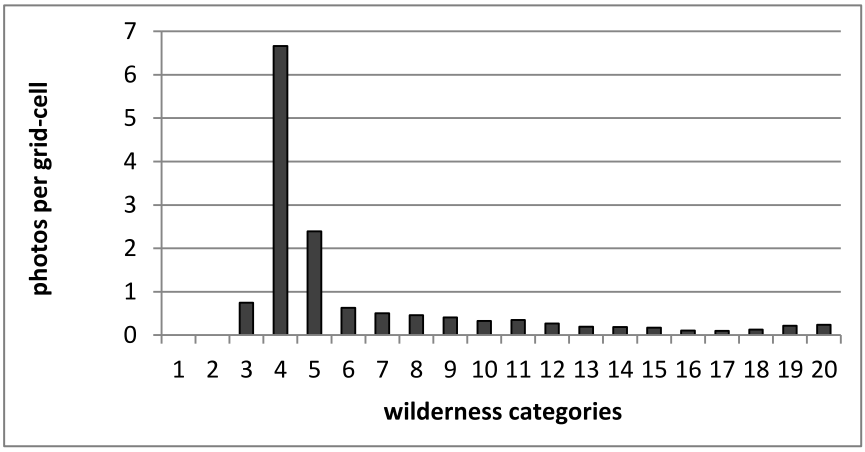


Figure 5.5 – Distribution of Flickr photographs per wilderness category, normalized by the number of map cells per category

In order to evaluate whether the exclusion of prolific user data is necessary or not, the differences between the dataset either containing or not containing data of prolific users have been analysed. But not only has the total number of photographs a decisive role in order to answer the prolific user question. Since the major part of this work concentrates on analyses with tags, the influence of prolific users on tags seems to be more adequate than only the number of photographs. Therefore, the difference in number of tags per wilderness category when excluding the prolific users has been evaluated. Additionally, the same process has been done to illustrate the effects to unique tags to consolidate the evaluations. Unique tags in this context mean tags which have only been generated by one single user. A simple statistical evaluation shows the loss of tags when excluding the prolific users across all categories (Figure 5.6). All graphs show the wilderness categories on the horizontal axis. While the graphs a) and b) show the number of all excluded tags, graphs c) and d) illustrate only the loss of unique tags.

While a) and c) are not normalized but show the bare numbers, b) and d) are normalized to the number of cells per category. The evaluation shows that all high peaks are restricted to the lowest wilderness categories in all diagrams. One could think that considering the number of cells per category would at least partially equalize these extreme distributions but the normalization even increases it. The only counter-trend can be observed in wilderness categories 17 to 20 for the unique tags. But this trend is rather small and the wilderness category 20 still has to be interpreted with care since the number of assigned hectares is much smaller compared to the other categories.

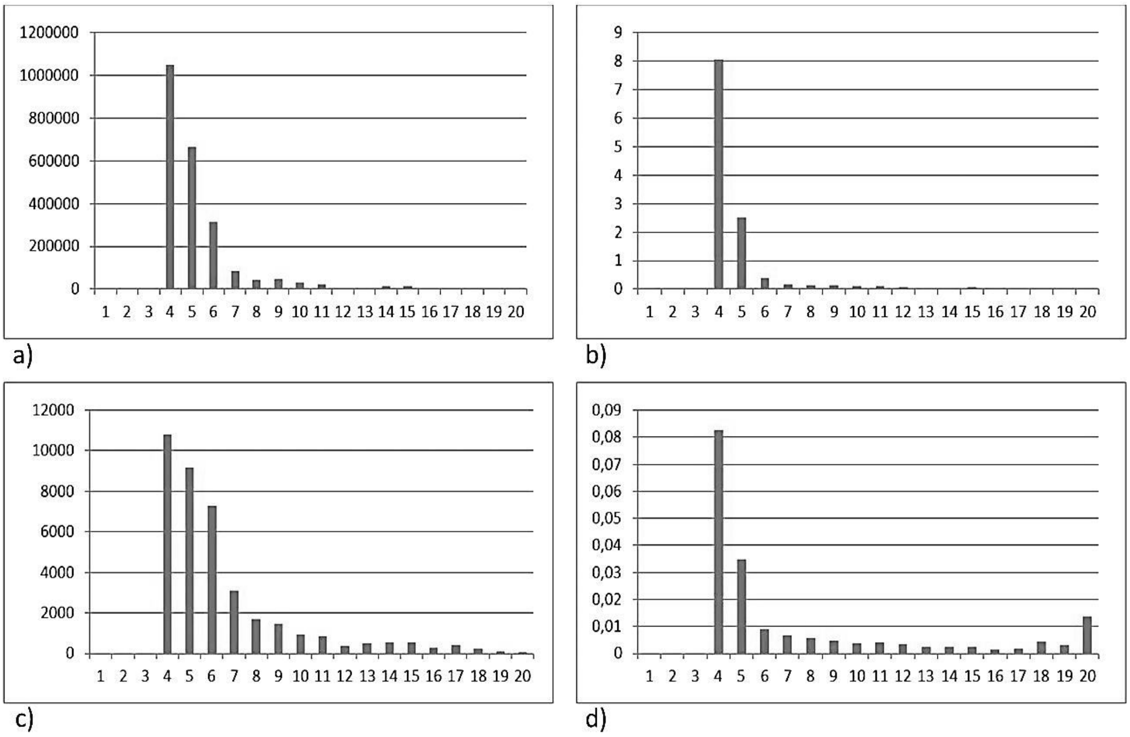


Figure 5.6 – Differences in excluded data generated by prolific users of all tags (a, b) and unique tags (c, d).

5.1.3 Bulk uploads

The third part of user behaviour analysis that attempts to reveal information about the aptitude of Flickr data to wilderness evaluation deals about bulk uploads. As described in section 4.2.3, literature cannot be consulted to get answers to methodological questions about how to identify bulk uploads or which to exclude. The developed evaluation assigns to each potential user a bulk index according to which they are classified according their uploading behaviour.

UTC [%]	MFTCa [%]	bi	bulk upload tendency
0.1	90	4.91	high
5	50	2.70	
14.84	**11.55**	**0.95**	**medium**
50	5	-0.30	low
90	0.2	-3.35	

Table 5.1 – Examples of parameter values and their effect to the bulk index.

Table 5.1 illustrates some examples of these parameters and how the bulk index has been classified. A high bulk index means the user tends to be a bulk uploader, illustrated by the red highlighted cells. For the grey cells, the mean UTC and mean MFTCa across all 885 users have been used to simulate a mean bulk index where no tendency can be interpreted. The green cells indicate examples of parameters of users which have applied a high percentage of unique tags and have applied their most frequent tag-combination only to a few of all uploaded photos. For the UTC a standard deviation of about 21.87 has been calculated while the MFTCa has a standard deviation of 15.4. A requirement for testing the output of bi to normal distribution is stated by the scientific context. In order to have normal distribution, two requirements are set. On the one hand need 2/3 of the whole examined data be within one standard deviation from the mean. On the other hand, 95% need to be within two standard deviations from mean. With a mean of 1.35 and a standard deviation of 2.03, the data of the analysed 885 users is narrow to normal distribution, since 62.26% of the data is within one standard deviation. Thus, the data distribution cannot be counted as a normal distribution.

To illustrate the total range of all calculated values, the result has been plotted in Figure 5.7 illustrating the bulk index on the vertical, and the users with descending ranks of activity on the horizontal axis. In order to evaluate if bulk uploading users mainly correlate with very active users, a trend line in white colour has been included. The corresponding coefficient of determination, the R-squared value, reached a value of 0.02 and therefore leaves no argument for an optional relation between the user activity and the bulk uploads. These results will be further discussed in section 6.1.3. The red line symbolises the achieved mean across all 885 users. Although the value range reaches from -5 to 5, the mean of all bulk indices (bi = 1.35) is higher than the mean of the value range.

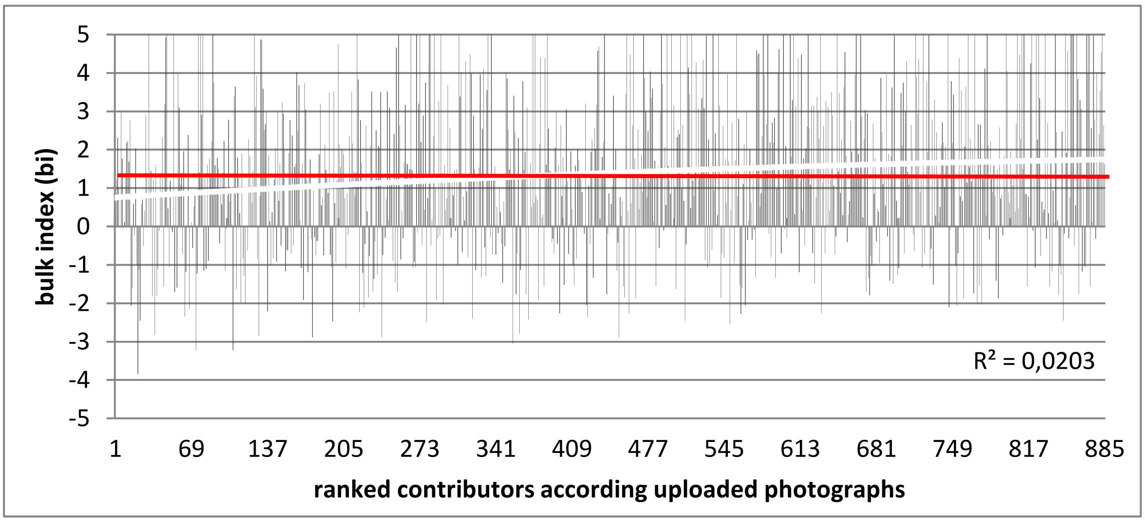


Figure 5.7 – Classified bulk index per Flickr contributor which are ranked with decreasing uploading activity from left to right.

No clear structure or peaks can be determined when plotting them according to the activity of the users. The highest bulk indices have been generated by two sorts of users. First, 68 users or 7.7% have applied no single tag to all of their uploaded photographs. Second, two users have applied only one single tag-combination to all of their uploaded photos. Since all users have uploaded at least 500 photos and all empty tags have been removed, these two users are the ones most tending to bulk uploading. The lowest bulk index of -3.85 has been achieved by a user with high activity but also high generated data quality. This user has applied 72.6% of all his

9’870 uploaded photographs with unique tag-combinations and has assigned his most frequently applied tag-combination to only 0.1% of his data. Such values most likely can only be achieved by much temporal effort spent for uploading and tagging the data. Such individual interpretations of behaviour could be managed for each user and reveal new insights into user-specific uploading behaviour.

5.2 Wilderness in Flickr tag data

5.2.1 Specific tags representing wilderness

Tags hold valuable semantic information about the content a user assigns to his uploaded photographs as highlighted by Purves and Mackaness (2016). This information can be used to analyse semantic variations across all wilderness categories of the GIS-model. Ranking the tags according to the representativeness within a certain wilderness category allows interpreting the variations across all categories and also gives information about the relation between Flickr data and the wilderness model. Detecting wilderness representative tags in Flickr data requires the definition of a collection of tags which could alternatively be used instead of the term *wilderness*. The collection of terms which most represent wilderness has been built in two separate steps. First, wilderness literature review (section 4.4.2) was applied to find potential terms and second, by analysis of tags highly co-occurring with wilderness (section 4.4.3).

Literature review

In scientific literature not much information can be found about semantically representative alternatives to the term wilderness according to the knowledge of the author. Synonyms do not exist in formal English or German language which is why other terms have to be taken into account. Some information can be found by having a look at and comparing different definitions which will be discussed in the next paragraph. Subsequently, further literature will be consulted to complete the collection of terms generated by literature review. Only those definitions will be explicitly cited here which have not already been mentioned in the theoretical part (section 2.2).

Wild Europe[9] defines wilderness as follow: *“A wilderness is an area governed by natural processes. It is composed of native habitats and species, and large enough for the effective ecological functioning of natural processes. It is unmodified or only slightly modified and without intrusive or extractive human activity, settlements, infrastructure or visual disturbance.”* (Wild Europe, 2012).

Many definitions, such as the one of Wild Europe[9] address the natural character of wilderness. The fact that not all nature is wild but wilderness zones are defined as being part of natural environment makes the term wilderness to a sub-category of nature. Thus, wilderness might not be perfectly represented by the term ***nature*** but depending on the context it can serve as a valuable indicator and is therefore added to the collection of representative tags.

Returning to the argument of the European Environment Agency (EEA) (2010) most of the areas with high wilderness classification lay in mountain regions, the correlation between wilderness and the tag ***mountain*** might be very high. This argument fits well to the results of the initial GIS-map where most wild zones lay in either the Alps or in the foothills of them. At this point it has to be reflected, that mountains can be seen from far away and people taking a picture in the flat midlands having some mountains in the background also tag that photo with

[9] https://www.wildeurope.org/

mountain which can cause significant bias. Anyway, following the definition of EEA (2010) this term can be added to the collection.

Until now the collection contains only nouns but what is about other word-types like adjectives or verbs? According to the findings of Purves et al. (2011) take verbs a minor part in general Flickr tags whereas adjectives are quite common. The closest adjective describing wilderness obviously seems to be the term *wild*. Unfortunately, this term can be used not only to mean wild in form of distinct, natural or beautiful but depending on the describing noun the meaning changes. The term can be used for instance to describe a wild party or a wild dog which both have not necessarily much to do with wilderness. Thus, this term will not be added to the collection due to this ambiguity reason. Taking the findings of Wilson (1979) in Kliskey and Kearsley (1993), 90% of New Zealand's population describe their wilderness with the adjectives *natural* and *beautiful* and 80 percent with *unspoilt*, *free*, *wild* and *valuable* amongst others. This confirms the aforementioned selection of the term nature to the collection of representative words on the one hand, but on the other can none of the other terms be added to the collection since they don't represent the term wilderness but only describe it and are not specific enough.

With a collection currently containing three terms evaluated by reviewing definitions it is appropriate at this stage to consult further literature.

The terms *wild* or *Wildnis* are according to Stremlow and Sidler (2002) a Christian European concept and in German literature the term's roots are closely related to *Wald.* This cognation can be reconnoitred back to the ninth century where large-scale closed woods were still unexplored and counted as dangerous in Europe. But *Wildnis* was not only describing forests but also deserts and unknown mountains. Thus, the semantic of the term does not implicitly mean woods but since Switzerland has no deserts the terms ***wood*** or *forest* can be used to represent the term *wilderness*. The term w*ood* is therefore added to the collection.

Many scientific articles and papers describe wilderness as being a certain extent in ***landscape*** (Carver & Fritz, 1995; Lupp, Höchtl, & Wende, 2011; ÖBF und WWF - Österreichische Bundesforste and World Wide Found For Nature, 2012; Stremlow & Sidler, 2002). The best way to take a photograph of wilderness is not by photographing a single tree or a wild bird but by capturing a landscape photograph. Beautiful views and landscapes inspire people to take a photo and tagging photos with the tag *landscape* is quite common according to Purves et al. (2011). Although landscape can be spatially referenced everywhere, photographs tagged with *landscape* might represent higher wilderness values according to the base data GIS-model. Thus, *landscape* is appended to the collection as well.

According to the aforementioned findings the collection currently contains the following terms:

Nature, ***mountain***, ***wood***, and ***landscape***

Co-occurrence

To amend the collection of tags evaluated by literature review, the co-occurrence of tags with all photographs tagged with *wilderness* has been analysed. Such co-occurrence analyses can be well illustrated by word clouds which has been realized in Figure 5.8. The size of a tag hereby represents its co-occurrence coefficient with *wilderness.* Thus, a larger size means a more representative tag. In order to represent the whole language spectrum not only the English versions of the terms have been used for this analysis but also translated versions in German and French. Since mentioned in previous sections are toponyms the most common form of Flickr tags. To avoid the strong presence of toponyms in the word cloud they have been removed manually.

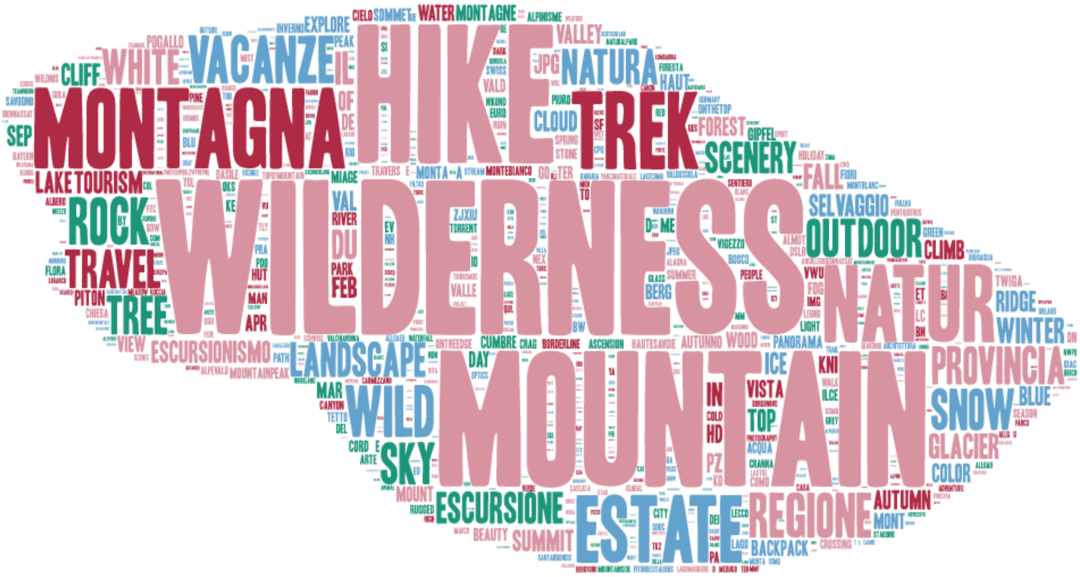


Figure 5.8 – Word cloud of the tag *wilderness,* generated online on WordArt[10]

The largest represented terms in the word cloud shows that the terms defined by literature review (section 4.4.2) are mostly also present in the word cloud which confirms the selection. *Mountain, montagna* and *nature* are highly co-occurring whereas *landscape* is less co-occurring but still present. The only appreciable representative which is close to *forest* or *wood* is the tag *tree*. But in contrast, other tags which could be taken into account to append to the tag collection like ***snow*** or ***rock*** are as highly co-occurring as the tag *tree*. As wilderness mostly occurs in mountain regions the two tags *hike* and *snow* seem to be able to represent the wilderness model as well. Therefore, they have been added to the collection.

In total, the collection counted six optional tags at this point which had to be extended by different synonyms and translations. The final collection of tags representing wilderness is illustrated in Figure 5.9. The translated versions of tags have been evaluated by manual observation considering the presence of these tags. Especially Italian tags showed a smaller presence so that only the Italian tag *montagna* has been added to the collection. In order to evaluate if the collected tags represent wilderness according to the GIS-model, their spatial distribution across all twenty wilderness categories has been evaluated. Since each photograph has assigned a wilderness quality index, described in section 4.2.2, each tag can be allocated to a wilderness category. Thus, for each tag and their translated synonyms of the collection, a graph has been evaluated showing their spatial distribution across the wilderness categories.

[10] https://wordart.com/

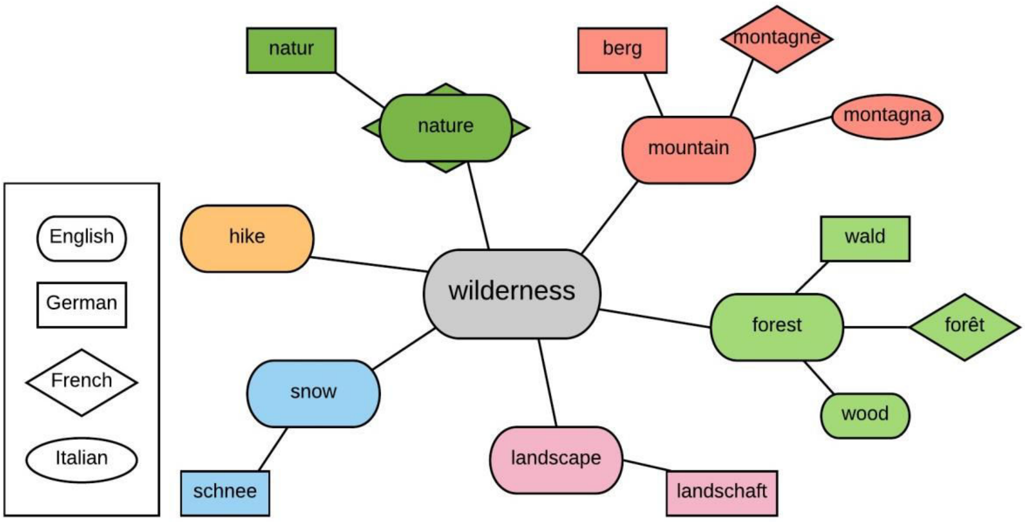


Figure 5.9 – Collection of tags selected as representatives for the term *wilderness.*

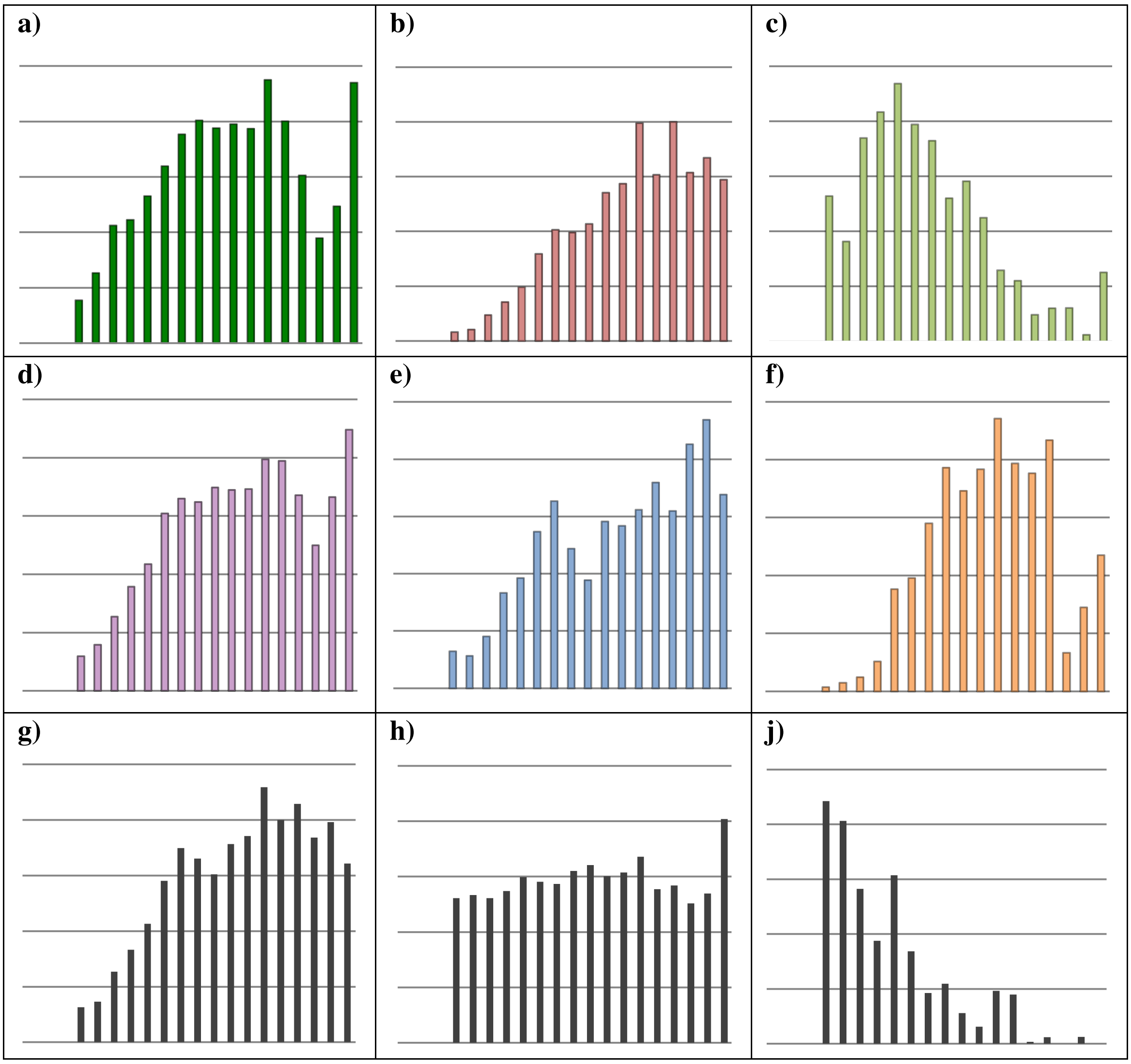


Table 5.2 – Spatial distribution of representative tags across all wilderness categories. Normalized number of tags (vertical axis) per wilderness category (horizontal axis).

Before arguing about the spatial distribution of tags across the wilderness categories it has to be mentioned, that the range of wilderness categories reaches from 1 to 20 whereas the model defines higher category values with higher wilderness quality. The lowest appreciable wilderness category is category four which is one of the most present categories in the GIS model. It has to be noticed that the GIS-model has assigned the categories 1 to 3 only nine hectares / cells, which is why they won't be taken into consideration within all tag-based analyses. Additional statement to these three categories can be found in the discussion (section 6.4.1).

Table 5.2 visualizes the spatial distribution of the representative tags of the collection according to the wilderness model. The horizontal axes represent the wilderness categories, with increasing wilderness quality indices from left to right. The vertical axes of the graphs represent the relative number of photos per category cell. The exact values of the vertical axes are not relevant since the interpretations are only interested about the relative distribution of the tags across the wilderness categories. Thus, the graphs are not compared by their exact values on vertical axes but more by the distribution on the horizontal axes. All graphs are normalized by the number of map cells /hectares per category applied by the GIS-model, which makes them comparable. But strong differences in the number of these cells per wilderness category may cause bias to the values. Most affected by this bias may be category 4 and 20 which contain by far the most, respectively the least number of cells.

The graphs a) to f) are coloured according to tags they represent in Figure 5.9. In addition to the collected tags in the previous steps, Table 5.2 contains three more graphs. Graph g) shows the distribution of all photos tagged with tags of the collection but without the tag wood and its translated terms. Photos tagged with multiple tags appearing in the collection have only been counted once in this graph. The graph h) of the tag *switzerland* serves as an example of a tag occurring in most of the categories. The last graph (j) illustrates the spatial distribution of the tag *architecture*, which serves as an example for a counter trend to the other evaluated tags.

The graphs b), d), e) and g) show an increasing number of tags with increasing wilderness quality, though the tendencies are not linear but have some breaks. Also graph a) and f) show that tendency till the highest wilderness categories where a strong break can be observed. Graphs c) and j) show decreasing number of photos per category with increasing wilderness quality, which is a counter tendency to the other graphs. The only graph which shows no clear tendency but equal number of photos per cell across all categories is graph h).

5.2.2 Most representative tags according to tf-idf

The second approach to detect the relations between the wilderness model and Flickr data was to apply the tf-idf equation which is frequently used in GIR research. This approach attempts to find the most representative tags for each wilderness category. The following paragraph describes the output of the first evaluation according to the standard formula described in section 2.3.5. Subsequently, the adaptions to this equation will be explained, then the outputs of the adapted version of the standard equation will be described and finally, the results will be interpreted.

For each wilderness category a ranked list of all tags applied to photographs lying on a hectare with the corresponding wilderness quality value has been developed as further explained in section 4.4.4. The top positioned tags have been classified as the most representative for that specific wilderness category. The first observation which clearly stated over all categories was the strong presence of toponyms. Different kinds of toponyms could be identified reflecting their spatial scale they describe. Most present were county-scale and city-scale tags but also more specific tags referring to mountain names or other place names were observable. A short overview about their spatial distribution according to the GIS-model led to valuable information about the standard tf-idf method. Even if the presence of toponyms is not equally distributed across all categories tags representing the country such as *switzerland*, *suisse* or *schweiz* are present in the top ranks in all categories. Other toponyms like *zurich*, *geneva*, *basel* or *lausanne* were only in the top ranks in wilderness categories lower than 10. The toponym *alps* stands in large contrast to the aforementioned city toponyms because it appeared in all categories with wilderness quality values larger than 8. Beside the high presence of toponyms also many tags have been observed within the top tf-idf ranks which most likely have been generated by a few contributors because they seemed so uncommon or specific. Examples are tags such as *bergtouraletsch*, *bikinitest2300* or *fliegerschiessenaxalp2012*. Tagging photographs is a very individual process which is why generally most of the tags generated are unique. This observation cannot be ignored since most representative tags should not be generated by only few or in worst case only one single user. Additionally, it is known from previous evaluations that prolific users and bulk uploading users can cause significant bias with their large quantities of data which is why this phenomenon has to be verified in more detail before describing the results.

In order to verify the impact of this observation for all wilderness categories the ten tags with highest tf-idf values have been analysed on their characteristic by how many users they have been generated. The results have shown that more than 46% of all tags were produced by not more than 10% of all users and 35% have even been generated by only a single user, thus were very user-specific. On the other hand 36% of all tags have been generated by at least 90% of all users which represents for example the toponyms generated by many users, as mentioned before. This evaluation has shown that the normal tf-idf equation, applied to this Flickr dataset, favours specifically tags generated either by very few users or by most of the users. Accordingly, the requirement for an adaptation of the standard version of the tf-idf arose. The applied adaptions, the exclusion of toponyms and the consideration of the user frequency led to a new equation called tf-idf-uf, as already described in section 4.4.4. But before discussing the output of this newly initialized equation some statements have to be mentioned about the adaption steps.

Tag exclusion

Excluding the toponyms and other biasing tags from the lists generated new insights into the actual evaluation and the relation between the Flickr photos and the wilderness model. The manual exclusion process was executed chronologically, beginning with the lowest wilderness category. Until category 9, most excluded toponyms could be classified to country scale like *suisse*, to regional scale like *tessin* or to city scale like *geneva*. Starting at category 9, the higher the categories the more frequently local scale place names like e.g. *schatzalp* or *breithorn* have been excluded. Most of the local scale terms were either referring to mountain peaks, mountain alps, specific alpine huts or ski resorts. The mentioned distributions of toponyms across the wilderness categories already show simple tendencies which are important to mention insofar as it shows that the exclusion of toponyms does not mean that no valuable information could be extracted by these tags. But since this exclusion has been processed no further interpretation and discussion will concern toponyms.

Adaption of tf-idf formula by considering user frequency

Applying the user frequency to the standard version of tf-idf changed the ranks of the tags in a strong degree. While the lists of the standard version contain many tags generated by very few users, all these tags disappeared from the top ranks when applying the tf-idf-uf equation. The effects of considering user frequency and the related changes of the output lists come up to the attempted consequences of isolating tags generated by very few users. The disappearance of tags like *99ersporthalle* or *derekflett* from the highest ranks of the lists also describes the most remarkable changes between the outputs of the standard and the adapted version of the equation. No tag has been generated by only one user until the rank 24 in adapted version whereas in the standard version more than 35% of the top-ten ranked tags were produced by single users. This fact illustrates well how effective the uf-parameter is to the whole evaluation. Another observation is that in general, most of the terms in the top ten ranks are nouns while the tags skiing and hiking are the only two exceptions. The output of the standard equation resists of such a simple classification.

Before describing the output tables of the evaluation it is necessary at this point to recall that the categories one to three will not be part of the evaluation due to the GIS model. The exact reason for that is explained in section 3.2. Accordingly, all annotations to the lowest wilderness category refer to the category four. Additionally, it has to be mentioned that the number of grid cells assigned to each wilderness categories are not equally distributed which has to be considered while interpreting the results.

Description of the results of tf-idf-uf evaluation

Some of the highly ranked terms within the output lists of tf-idf-uf (Table 5.3) have already been selected and added to a tag collection as representative tags in the last evaluation (section 5.2.2). In order to verify if the tags selected for that collection can be confirmed by the tf-idf-uf approach these tags will stand in focus of the first descriptive part of the tf-idf-uf evaluation. Further terms and tendencies will be described following on that paragraph.

Most noticeable are the tags *mountain* and *mountains* which are still placed in the top three ranks across all wilderness categories higher than seven. In other words, the most representative tags for all regions with higher wilderness quality indices than seven are the tags *mountain* and *mountains*. These observations are in line with the European definition of wilderness which defines wilderness to be related to mountain regions (European Environment Agency (EEA), 2010). In the lowest categories these tags decrease in ranks continuously. The same observation can be made with the tags *montagne* or *montagna* just that they appear in lower ranks due to lower frequency they have been tagged. The tag *berg* never reaches higher ranks than 18 and also has its highest ranks in the categories 15 to 17. In contrast, the tag *berge*, which has not been respected in the collection of representative tags, reaches the rank seven in category 20 and keeps high ranks of around nine until it strongly decreases beginning with category ten. The tag *nature* and *landscape* show a similar behaviour like the tag mountain as their ranks fall out of the top ten in lowest categories. In higher categories than seven they keep ranks between four and seven. The tags wood and forest and their German and French translations cannot be found in the top ranks as well. The highest rank achieves the tag forest on rank 21 in the category seven. The other tags related to wood are spread in lower ranks but particularly across the categories six to nine they reach their maximal ranks. At this point another tag seems important to be mentioned as well.

	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	street	lake	zoo	snow	mountain	snow	mountain	mountain	mountain	mountain	mountain	mountain	mountain	mountain	mountain	mountain	mountain
2	architectu	street	snow	mountain	snow	mountain	mountain	mountain	mountain	mountain	mountain	mountain	mountain	mountain	mountain	snow	snow
3	lake	concert	nature	winter	mountain	mountain	snow	snow	landscape	snow	snow	snow	glacier	snow	snow	mountain	mountain
4	train	train	alps	landscape	landscape	landscape	landscape	landscape	snow	landscape	landscape	hiking	snow	glacier	glacier	glacier	landscape
5	concert	architectu	winter	mountain	river	winter	winter	nature	nature	hiking	hiking	landscape	landscape	hiking	montagne	landscape	glacier
6	city	zoo	landscape	nature	winter	nature	nature	hiking	hiking	nature	glacier	glacier	hiking	landscape	hiking	ice	nature
7	night	festival	lake	airport	nature	river	hiking	lake	lake	montagne	nature	nature	nature	montagne	landscape	nature	berge
8	snow	music	mountain	motorsho	lake	lake	train	river	winter	winter	montagne	montagne	montagne	berge	ice	travel	skiing
9	winter	snow	mountain	train	bridge	train	montagne	winter	berge	berge	berge	berge	berge	nature	ski	berge	sky
10	art	nature	airport	salon	train	water	berge	water	panorama	lake	winter	sky	winter	winter	clouds	clouds	montagn
11	church	landscape	castle	palexpo	water	clouds	water	berge	montagne	glacier	lake	panorama	ice	clouds	nature	sky	rock
12	music	winter	train	motor	clouds	sky	clouds	montagne	clouds	clouds	clouds	winter	gletscher	ice	berge	montagne	clouds
13	sbb	night	festival	auto	travel	berge	river	clouds	sky	ski	panorama	lake	clouds	panorama	gletscher	winter	carrel
14	fasnacht	city	sky	car	sky	hiking	lake	panorama	2015	sky	sky	clouds	panorama	montagna	climbing	mountain	nordwan
15	bw	castle	architectu	show	night	panorama	schnee	waterfall	train	panorama	montagna	ice	sky	sky	panorama	climbing	climbing
16	travel	mountain	tree	lake	church	montagne	sky	wasserfall	water	travel	schnee	montagna	lake	gletscher	winter	ski	natur
17	water	water	concert	tree	panorama	travel	panorama	sky	travel	schnee	travel	montagne	ski	climbing	mountain	hiking	ice
18	ville	church	water	sky	city	schnee	ski	train	schnee	water	ice	gletscher	berg	ski	berg	topofeuro	alpine
19	graffiti	sbb	clouds	2011	architectu	ski	waterfall	schnee	landschaf	berg	berg	alpi	montagna	montagne	sky	train	hiking
20	festival	travel	green	flughafen	forest	bridge	sun	travel	glacier	montagna	gletscher	schnee	schnee	lake	montagne	panorama	jungfraub
21	tram	car	street	forest	schnee	sun	travel	fluss	river	summer	montagne	summer	wandern	summer	summer	blue	winter
22	landscape	lac	salon	sbb	hiking	summer	summer	natur	berg	skiing	summer	ski	skiing	randonné	randonné	rock	derperfek
23	bahnhof	art	flughafen	clouds	green	waterfall	berg	glacier	summer	montagne	pass	randonné	rock	mountain	travel	summit	snowboar
24	zug	bw	night	automobil	fluss	green	green	landschaf	montagna	wandern	hike	paysage	summer	wandern	rock	gletscher	liz
25	sky	sky	museum	green	berge	blue	trees	green	montagne	ice	water	travel	landschaf	rock	ghiacciaio	schnee	berg
26	nature	auto	sunset	schnee	ski	landschaf	pass	ski	wandern	hike	wandern	wandern	randonné	landschaf	montagna	white	fels
27	cff	live	travel	autosalon	sunset	forest	landschaf	berg	herbst	paysage	ski	hike	montagne	hike	derekflett	hochtour	2008
28	town	musique	car	travel	tree	tree	neige	summer	paysage	autumn	sun	blue	wanderun	schnee	schnee	skiing	3571m
29	urban	museum	forest	water	autumn	night	herbst	wasser	autumn	randonné	paysage	water	water	berg	alpinismo	nieve	panoram
30	people	portrait	portrait	airbus	trees	trees	autumn	wandern	pass	blue	randonné	landschaf	hike	travel	wandern	sun	greatrail
31	museum	mountain	auto	sun	castle	autumn	forest	pass	lac	sun	lac	pass	travel	wanderun	monte	peak	landschaf
32	portrait	river	2011	eos	summer	neige	blue	hike	see	landschaf	wanderun	2010	sun	blue	neige	berg	glaciercav
33	lac	white	church	berge	blue	bw	glacier	montagna	lago	train	skiing	wanderun	climbing	adventure	outdoor	cloud	cablecar
34	black	2012	lac	trees	montagne	fluss	montagna	lac	railway	rock	landschaf	sun	blue	paysage	hautemon	paysage	family
35	white	zug	animal	2012	art	see	tree	bridge	hike	river	blue	skiing	paysage	white	matterhor	bw	best
36	mountain	2011	show	castle	bw	berg	montagne	autumn	sun	lago	flowers	rock	trekking	outdoor	sun	neige	excursion
37	light	tree	trees	zug	railway	railway	railway	see	natur	wanderun	train	climbing	ghiacciaio	trekking	landschaf	alpinism	4221m
38	car	black	bw	old	sun	natur	skiing	blue	blue	herbst	climbing	lago	lac	trift	topofeuro	4000er	piz
39	medieval	oldtimer	blue	wef	herbst	lac	paysage	neige	wasser	gletscher	autumn	people	fog	neve	4000m	montaña	summit
40	carnival	graffiti	art	ski	brücke	pass	wandern	alpi	green	lac	rock	neige	neige	lago	skirando	holiday	travel

Table 5.3 – Ranked tf-idf-uf list of tags according to their representativeness per wilderness category. Tf-idf-uf ranks on vertical axis and wilderness categories on the horizontal axis.

The tag *tree* scores higher ranks than the selected tags like *forest* and *wood* but shows a similar distribution within the same categories. This is important insofar as the terms selected in the last approach could have been improved by the tag *tree* to be more adequate. The tag *snow* scores constantly very high ranks between one and four while the only exceptions are the categories 4 and 5 where they fall back to ranks eight resp. nine. It is the tag with the most constant ranks across all categories within the top-ranks of tf-idf-uf. The German term *schnee* is ranked much lower in contrast and only reaches the top-twenty ranks between category 10 and 16. Its rank fall constantly in lower and fall slightly in higher categories. The remaining tag *hike* never reaches the top-twenty ranks but keeps top-thirty ranks in the categories 13 to 17. In lower categories than 13 the tag constantly decreases in rank and in higher categories than 17 it does not occur within the top hundred. Also for this term a related tag has to be mentioned which scored better ranks. The tag *hiking* remains within the top ranks four to seven across all categories 10 to 18. In lower categories a strong decrease in ranks can be observed as well as a slight decrease in the categories 19 and 20.

The previous paragraph referred to the collection of tags evaluated in section 5.2.2. But not only have these terms shown valuable information about wilderness in the Flickr tags. The following paragraph describes some other characteristics across the tf-idf lists which are relevant to be mentioned for wilderness detection.

Switzerland is known as sustaining large water bodies, in German also called the "Wasserschloss" of Europe according to NZZ[11] while the glaciers up in the Swiss Alps symbolize these water reservoirs nourishing whole Europe with fresh water. This is only one example illustrating that not only the mountains of the Swiss Alps are well known and attract to take pictures but also the massive glaciers are relevant photo targets. Tagging a photograph captured in high wilderness quality areas with *glacier* seem to be quite common since this tag shows highest ranks in tf-idf lists between three and six in categories 14 to 20. In categories lower than 14 the term constantly decreases in rank until the tag disappears from the top hundred ranks in categories lower than 8. The German term *gletscher* shows a similar distribution just in highest ranks between categories 14 and 19. Until now only terms have been described which in some way are related to wilderness. Do tags which have less in common with wilderness also show the same distributions? Generally, such tags only occur in the top-ranks in wilderness categories lower than 8. Exceptions for that are tags like train or bridge which scored top-twenty ranks until category 12. Most of the tags referring to human-made constructions or objects, like *architecture*, *street*, *concert* or *zoo* strongly decrease in rank around the categories 7 and 8. All these observations show that a direct connection between the Flickr dataset and the GIS-model exists. These tendencies will be further discussed in the section 6.2.2 which also interprets the relevance of these results to the broader wilderness context.

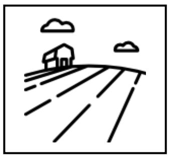
5.3 Characterization of wilderness according to tf-idf-uf evaluation

The results of the third research question could be evaluated by the same methodological output as the previously discussed tf-idf-uf analysis. The third research question attempts to further characterize the wilderness information depicted by the GIS-model. The discussion of the last evaluation (section 6.2.2) has evaluated a subdivision of the tf-idf-uf wilderness spectrum into four differentiable classes. The characterization process has considered these classes so that not each wilderness category will be characterized but rather each of these four classes.

[11] https://www.nzz.ch/wasserschloss-schweiz-1.16921466

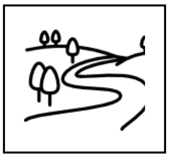
Nonetheless, the categories will be used as base for argumentations. Compared to the last evaluation not only global tendencies but more local ones are of specific interest. The following subsections will not mention all wilderness categories but only these which show interesting potential for wilderness characterization. In order to simplify understanding, all interpretations are described just directly following on the description of the observations.

5.3.1 Urban and cultivated regions



The categories 4 to 7 represent the lowest wilderness class describing urban and cultivated regions. Searching for tendencies characterizing wilderness in such regions with very low wilderness quality indices does not make much sense. Additionally, these categories are most affected by the influence of prolific users. Most of the highly ranked tags of the tf-idf-uf evaluation represent anthropogenic features which do not have much in common with wilderness. Thus, no further wilderness specific tendencies will be described in these lower categories, though a large potential for tendencies differing from wilderness would exist.

5.3.2 Flat to hilly natural regions



The first described category represents areas with wilderness quality indices of value eight. While category 7 contains many tags referring to the motor show of Geneva such as *car*, *motorshow*, *autosalon*, *motor* or *auto* are none of them observable in the higher ranks of category 8. More natural features take place on the highest ranks like *river*, *fluss* or *lake* on the contrary. Also the first appearance of the tag *hiking* can be observed within this category. The only presence of anthropogenic features in higher ranks of category 8 is reflected in tags basically referring to landmarks such as *bridge*, *brücke*, *church* or *castle* which disappear from top ranks in category 9. Some tags constantly increase in ranks within this class with higher wilderness categories. Examples for such observations are the tags *hiking*, *waterfall, glacier* or *ski*. In contrast, the tags *river*, *forest* and *tree* constantly decrease in ranks or even disappear from the top ranks. Some top-ranked tags also remain on constant ranks across the whole class like the tag *lake* or *clouds*. The appearance of the tag *pass* in top 30 ranks of category 10 and 11 is worthy to be mentioned at this point as well.

The described observations give arguments for further characterization of this wilderness class. The disappearance of anthropogenic feature tags in top ranks of category 8 illustrates the shift from urban to more natural environment, although some landmark tags like *bridge* or *castle* are still present in category 8 but disappear from top ranks in category 9. The wilderness GIS-model has therefore assigned wilderness quality indices eight to some regions which Flickr users still observed and tagged anthropogenic features. Since castles in Switzerland are due to defence strategy commonly built on elevated places and bridges are also basically built to surmount larger elevation differences in landscapes the appearance of these tags in category 8 seems to be reasonable. Also the combined appearance of the tags *bridge* and *river* seem to be comprehensible as traffic normally passes rivers by bridges. The increasing tendency of tags *hiking* and *glacier* in categories 10 and 11 shows that this wilderness class might not only refer to flat or hilly regions as they can be found in the canton Thurgau or in the hilly landscapes of the canton Appenzell but also to more mountainous regions where glaciers are present or at least can be seen. The wilderness attribute ruggedness seems to be decisive in this class. Thus, the definition of a flat natural region seems not to be absolutely adequate for all hectares this class has been applied by the GIS-model. Also the appearance of the tag *pass* in higher ranks in this class symbolizes that not only flat regions have been assigned to this class. On contrast, some larger lakes like the Rhine or the Aare have eroded their surrounding environments to flat landscapes in certain regions over time and also the tag *lake*, which constantly keeps high ranks in this class, in some cases like the Bodensee might have been tagged in flat regions. Thus, taking the term flat as a definition for all regions within this class cannot be totally approved. To conclude the characterization of this wilderness class, the regions described by this class potentially reach from flat to hilly regions according to the tf-idf-uf evaluation. The distribution of the Flickr tags and their ranks show that the Flickr users percept these regions as containing many natural features like waterfalls, rivers and lakes but also containing anthropogenic features like bridges and castles. The tags referring to water bodies most appeal to flat regions where as many more tags refer to features usually findable in landscapes with more differences in altitude. Especially the tag *glacier* provokes the interpretation for regions with higher altitude.

5.3.3 Mountain regions and glaciers

The last described class was defined by the shift between hilly to mountain regions. Since the tag *hiking* lays on the top six rank in category 11 and 12 the tag *glacier* reaches rank 20 in category 12 with increasing ranks within this class, the shift to mountainous regions seem to be initialized. This class contains all categories reaching from wilderness quality indices 12 to 18. The most important observations within this class are the appearance of tags like *skiing*, *hike, mountaineering* or *wandern*. These tags might not reach that high ranks as the tags *glacier* or *hiking* but are present in the top 40 ranks across multiple categories within this class. Others like the tags *randonnée* or *paysage* show a larger presence of French tags in the top ranks. But also the multilingual tags referring to mountains and glaciers like *montagne*, *montagna*, *montaña* or *ghiacciaio* indicate that this class is dominated by tags referring to mountain regions. And not only tagging in English language is frequent within this class but also other languages. In category 17 and 18 the tags *mountaineering*, *topofeurope* and *hautemontagne* appear in higher ranks which give additional information about the perception of the Flickr users and allow further wilderness characterization.

The users who uploaded the photographs in wilderness areas classified with the wilderness quality indices 12 to 18, according to the GIS-model, frequently applied activity tags such as *skiing*, *wandern*, *randonnée* or *hiking* in this class. Actually, hiking is in general a sport performed in natural environments and many touristic organizations especially in mountain regions promote their landscapes for being nice hiking regions. Since hiking is a certain kind of walking it has to be distinguished between promenading or having a walk, hiking, mountaineering and climbing. The differences between these terms are probably the physical effort of the movement process which reaches its maximum when climbing. This physical effort is related to the steepness of the terrain amongst others which is respected by the GIS-model applied in this work. The wilderness attribute ruggedness again comes into focus here. In higher categories within this class the tags *climbing* and *mountaineering* increase in rank which indicates also increasing steepness of the according regions. Compared to the two classes described before, this class contains more verbs expressing leisure activities in the top ranks. This shows that the Flickr community uploads many pictures during their free time in regions counting to this class. Therefore, an additional character of this wilderness class is that people like practicing sport activities in these regions. While *hiking* and *skiing* can be found in all categories on high ranks, *mountaineering* and *climbing* are more practiced in regions assigned to higher categories around 17 and 18. The tag *topofeurope* is a touristic annotation to the

Jungfraujoch which is promoted as being the top of Europe[12]. Since the Jungfraujoch is a hotspot of Flickr activity in Switzerland according to the evaluations described in section 5.1.1 it is not surprising to find this tag in top ranks in these categories with high wilderness quality indices. But one could argue that the region around this remarkable touristic attraction should be classified with lower wilderness quality due to high presence of men and also infrastructure. Figure 5.10 shows that the distribution of high activity around the Jungfraujoch is not spatially concentrated on a few map grid cells around the mountain's peak but rather shows large scale distribution with a range of several kilometres. So the tag *topofeurope* has been assigned to photos taken kilometres away from the actual summit as Figure 5.11 demonstrates. The majority of tags concentrate on an area of about one square kilometre around the touristic hotspot. Since the GIS-model has respected alpine huts and infrastructure, the wilderness quality indices around the mountain observatory building of Jungfraujoch are reduced to values 12 or 13, represented by the brighter hectares. But this has only been applied according to a certain radius but not to the actual areas where tourists take photographs as illustrated by Figure 5.11. In order to increase the quality of the GIS-model, Flickr data hotspots would therefore be an adequate alternative data source. However, this shows that the GIS-model does not necessarily respect all local effects within a smaller region to their classification, although the resolution is very high. Another interesting observation in Figure 5.10 is the curved line of photographs heading to north-east from the Jungraujoch. These photographs most likely have been taken within the tunnel of the Jungfraubahn heading through the Eiger. This tunnel and the train itself seem to fascinate people and persuade them to capture photographs. The Jungfraujoch is a perfect example where the beauty of wilderness and landscape aesthetics is used for touristic reasons which affect the wilderness quality of the environment.

The wilderness of this third class has been characterized as attracting people to upload photographs during sportive leisure activities like skiing or mountaineering. In higher categories within this class also climbing becomes increasingly important. Furthermore indicate tags like *lake* or *water* the fascination of Flickr users for water bodies in higher mountain regions such as alpine glacier lakes.

5.3.4 High mountain regions with steep slopes

This last wilderness class represented by the wilderness categories 19 and 20 stands for high mountain regions having steep slopes. But before this last class can be interpreted, it has to be reconsidered that the base dataset is affected by data generated in form of bulk uploads. The biasing effect of such data has neither been evaluated in this work nor in any comparable approach, according to the knowledge of the author. But it seems to be logic that the effect increases with decreasing numbers of total users and decreasing data volume generated by all users in the examined area. Since the user frequency has been applied to the tf-idf-uf equation, tags generated by only a few users have decreased in ranks while tags generated by many users increased. The majority of tags in high ranks most likely have been generated by many users and therefore the influence of user-specific bulk uploads are relativized. This does not mean that the bulk uploads have no influence, but the influence might be reduced due to the consideration of the user frequency. However, most vulnerable to the influence of bulk uploads are therefore the highest wilderness categories and thus, this last wilderness class since it contains the fewest users and also the fewest data volume. This needs to be respected during the interpretation.

[12] https://www.jungfrau.ch/de-ch/jungfraujoch-top-of-europe/

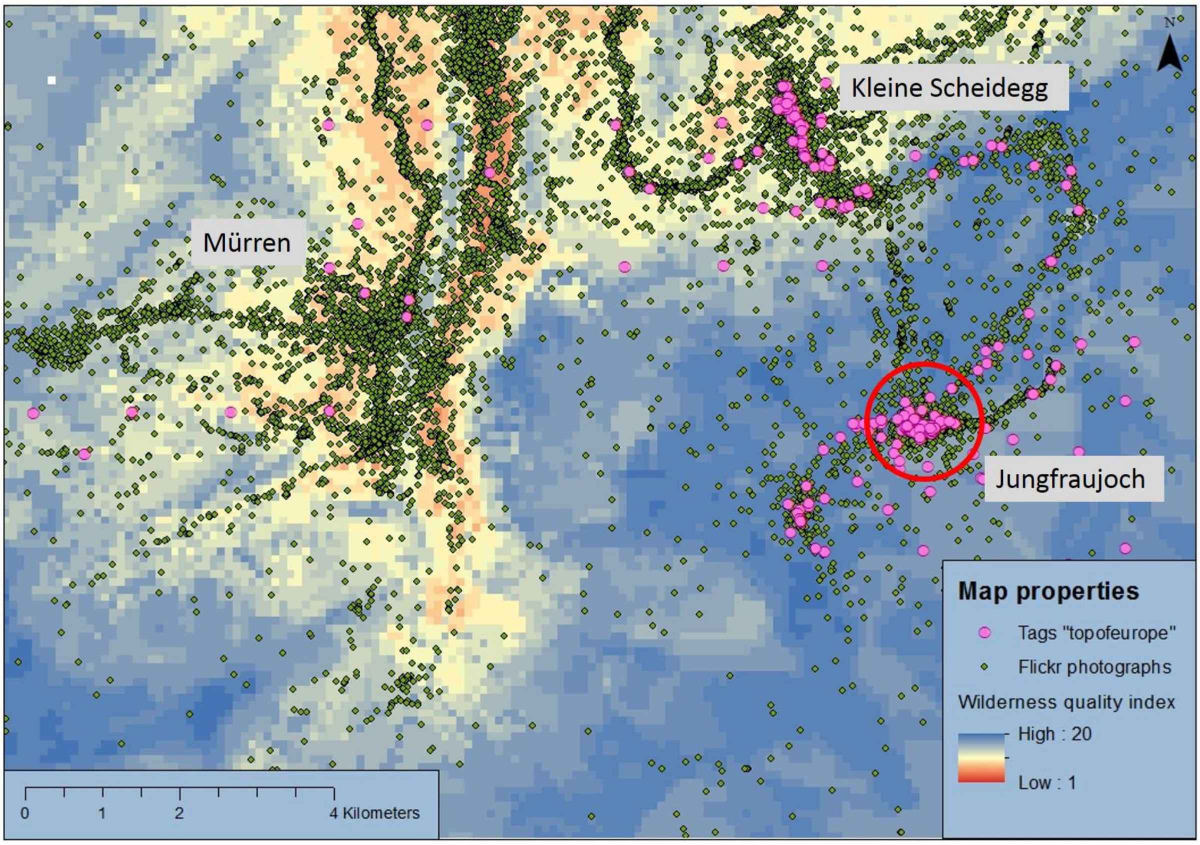
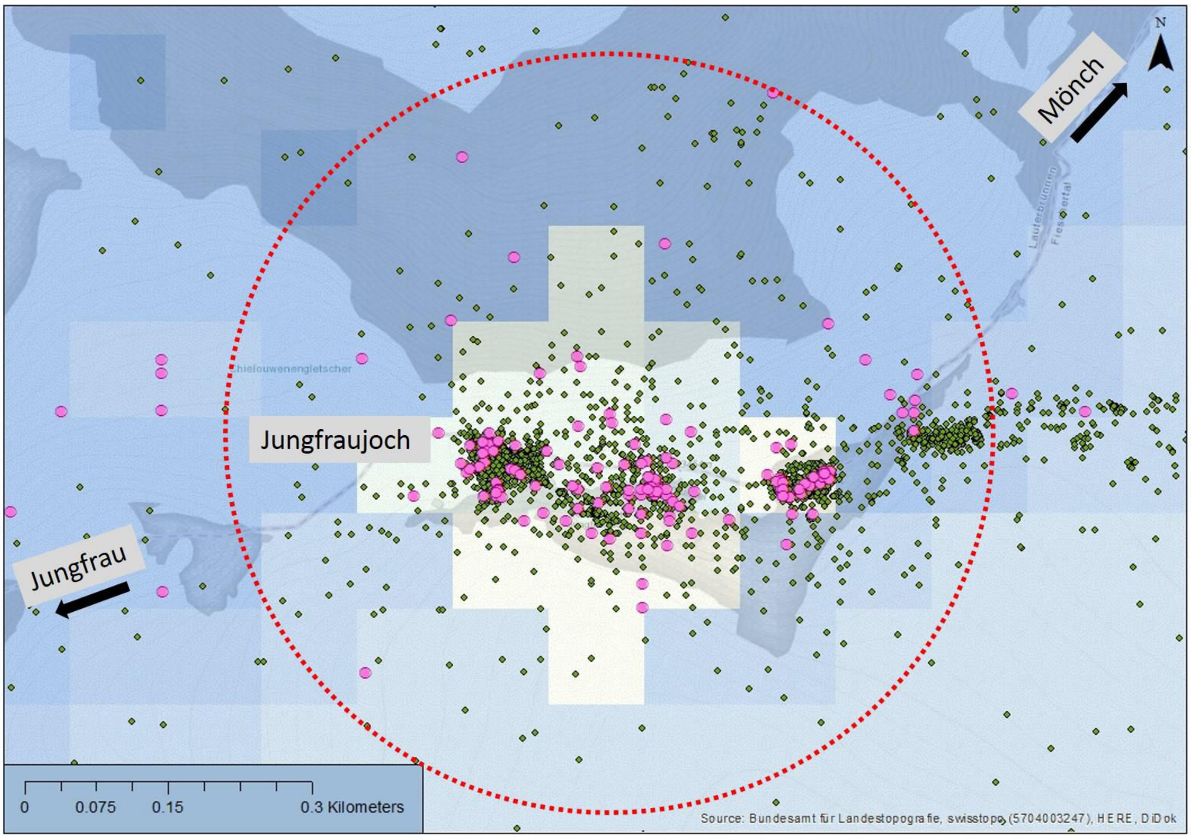


Figure 5.11 – Flickr photograph distribution in the touristic region of the Jungfraujoch and Mürren. Own creation



Source: Bundesamt für Landestopografie, swisstopo (5704003247), HERE, DiDok

Figure 5.10 –. Zoomed focus on the touristic hotspot of the Jungfraujoch. Own creation.

The most conspicuous observation within this class deals about the tag *hiking* in combination with *climbing*. While the tag *hiking* remained on constantly high ranks in the last class, it decreases in ranks in this class again. As a counter-trend, the tag climbing keeps high ranks. Especially in category 20 a higher presence of tags referring to climbing activities is detectable in high ranks like *climbingthematterhorn*, *nordwand* or *carrel*. These tags indicate very steep slopes. Many specific tags indicate highest mountains like *summit*, *peak*, *piz, cablecar* or *4000er*. Other tags in high ranks indicate that either many pictures have been taken during the winter or in regions where snow is present during the whole year like *snow*, *neve, skiing* and *snowboarding*. Especially the tags *carrel* and *nordwand* do not seem to be very popular. It is therefore possible that these tags have been generated in form of bulk uploads.

According to the tf-idf-uf evaluation are most top-ranked tags in this class either referring to winter sports, climbing or mountain features. Users therefore like doing sports in these areas but to compare to the last described class the sports seem to be more extreme since the tag *hiking* is lower ranked than *climbing*. High ranks of tags like *hochtour* or *climbingthematterhorn* confirm this observation. The tag *nordwand* refers to the well-known north wall of the mountain Eiger which is infamous for its high climbing difficulty. The tag *carrel* might either refer to the name of a well-known Italian climber with the name Jean-Antoine Carrel[13] who was one of the first climber of Matterhorn in 1865, or it refers to a specific refuge place on Matterhorn itself which was named as Carrel in memory of his death. Interestingly, the tag *carrel* exclusively appears in the top-ranks of category 2 which is also an indicator for a potential bulk upload. But since the Matterhorn is one of the main hotspots of Flickr photographs in Switzerland and is also famous as being difficult to climb, the appearance of this tag in high ranks in this class fits the classification. Thus, this wilderness class can be further characterized as being interesting for climbing and winter activities. The tag *derperfektetag* stands for a perfect day in German and even implies positive feeling of Flickr users in this class. But these results have to be interpreted carefully since the classified hectares assigned to each wilderness category are not equally distributed which especially affects these two categories 19 with 40’298 resp. 20 with only 2’462 hectares assigned. Compared to the category 6 with highest number of 787’534 assigned hectares, a certain distortion within these two categories has to be taken into consideration. Additionally, this distortion becomes visible by comparing the number of users who uploaded geotagged photos in these regions. While all photos assigned in category 4 were generated by 29’251 users, the ones in categories 19 and 20 have been generated by 1161 resp. 150 users. Considering the differences in uploading behaviours of social media users described in section 2.3.3 reduces the expressiveness of the characterization of this class. Nevertheless, characterizing this class is possible although the results have to be regarded critically.

According to the evaluated classification of the wilderness continuum a new aggregated map can be calculated which spatially visualizes and localizes each of the classes (Figure 5.12). For each of the four classes the knowledge gained from the tag-based analyses and characterizations can be applied. Since the regions coloured in red represent urban areas, no further characterization has been done in that class. But the other classes can be characterized according to the mentioned observations what illustrates that the characterization process was a successful approach to extend the GIS-based information and gain new insights into how Flickr community perceives, describes and interact with wilderness.

[13] https://de.wikipedia.org/wiki/Jean-Antoine_Carrel

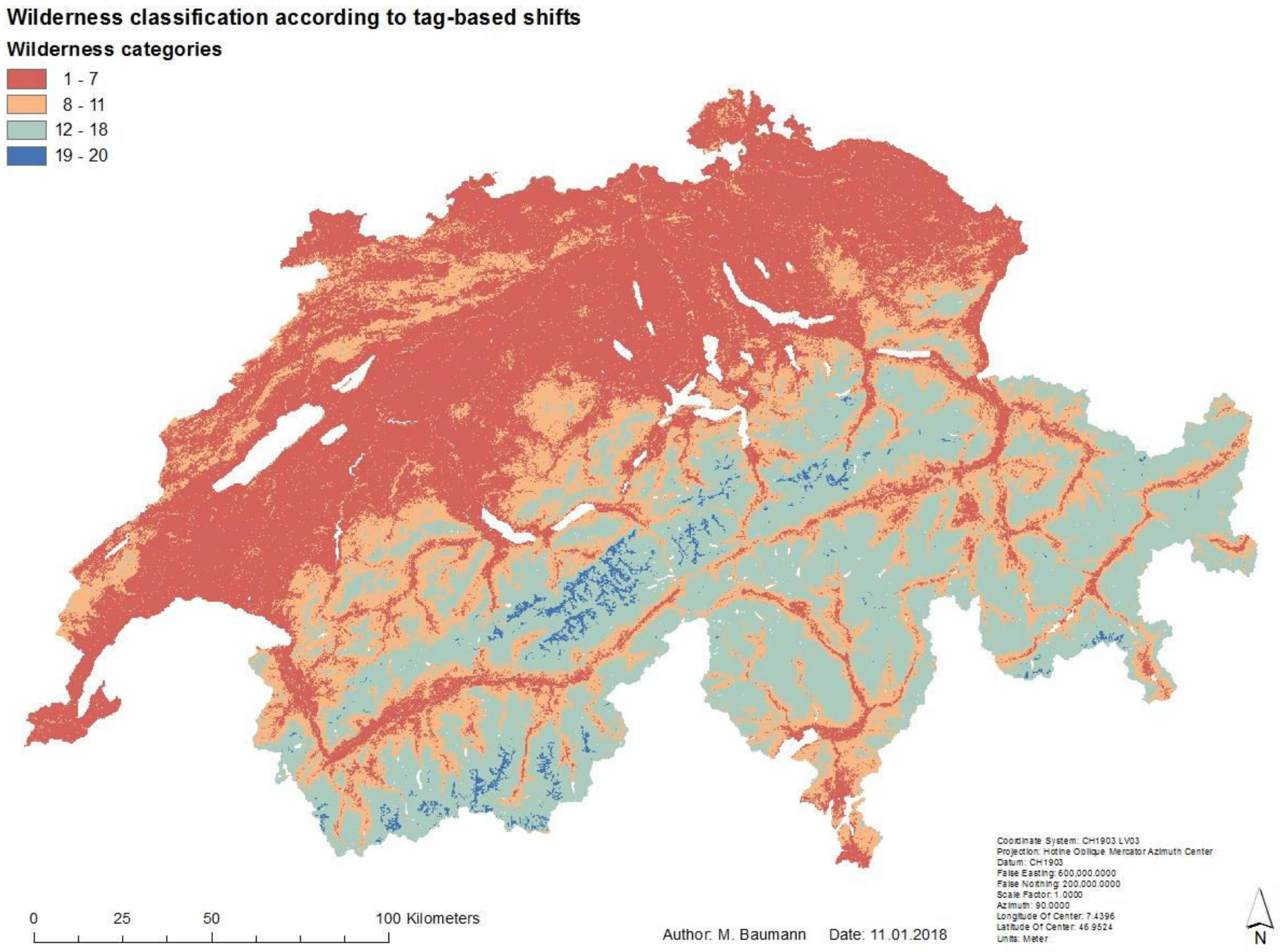

Wilderness classification according to tag-based shifts
Wilderness categories
🟥 1 - 7
🟧 8 - 11
🟩 12 - 18
🟦 19 - 20
0
25
50
100 Kilometers
Author: M. Baumann
Date: 11.01.2018
Coordinate System: CH1903 LV03
Projection: Hotine Oblique Mercator Azimuth Center
Datum: CH1903
False Easting: 600,000.0000
False Northing: 200,000.0000
Scale Factor: 1.0000
Azimuth: 90.0000
Longitude Of Center: 7.4396
Latitude Of Center: 46.9524
Units: Meter
N

Figure 5.12 – Wilderness map based on shifts in tags of the tf-idf-uf evaluation

6 Discussion

This section critically discusses methods and results. The base structure is identical to the results (section 5) so that for each research question a subsection is defined. A fourth subsection will illustrate the strengths and weaknesses of the applied data, research methods and results.

6.1 Effects of user-specific behaviour to the aptitude of user generated content in wilderness research

The wilderness debate has a high demand for new technological opportunities that aid to evaluate and visualize the necessity of wilderness preservation and protection (Tims, 2014). Geotagged social media data suit the requirement of being both, social and spatial which is required to represent the complex wilderness character. In this study the aptitude of applying this kind of data for explicit wilderness research purpose is analysed. Limitations of social media data have already been detected in previous studies (Hochmair & Zielstra, 2012; Rattenbury & Naaman, 2009). Lack of representativeness, spatial inaccuracy, temporal inaccuracy, variations in tagging behaviour and other problems have been analysed. The latter stands in focus of the first research question which concentrates on the effect which varying user behaviour can have on the whole dataset. This question is relevant insofar as if the social media data can actually be used to serve as appropriate data for wilderness research or not.

6.1.1 Spatial hotspots and granularity of Flickr photos

Interpreting the two generated maps representing the spatial distribution of geotagged Flickr photographs in Switzerland have shown specific characteristics of the Flickr community and their uploading behaviour. Although the quantile map cannot be used to interpret any correlations between the population and the number of Flickr photos per area, the observed tendencies help to understand Flickr user uploading behaviour. The quantile map illustrates that basically, the regions with highest Flickr photo density concentrate on areas with high population density. But the majority of populated places are not or only sparely covered by Flickr photographs comparing to the population. The density map confirms the expectation that Flickr spatial distribution generally concentrates on urban areas, regarding the hotspots. This finding has been made by other approaches for different areas than Switzerland (Antoniou et al., 2010; Gliozzo et al., 2016; Gschwend & Purves, 2012). A few other hotspots could be identified which do not appear at very populated areas. They rather take place at touristic hotspots such as the regions around Jungfraujoch and the Matterhorn (Figure 5.10). The blue-coloured regions in the midland of the quantile map (Figure 5.2) confirm that Flickr users most likely upload photographs not where they live but rather in regions with high touristic attraction. Although this finding is not representative for the whole Flickr community, it can be valuable for tourism regions insofar as they could consider these photographs to analyse user landscape preferences for touristic purposes. But the granularity of Flickr photographs in remote regions requires an aggregation of the dataset to two kilometres in order to visualize the map appropriately. Thus, for more local information the granularity is in general just too low. Additionally, the generated maps have shown that local hotspots, most likely generated by very few users, can have strong influence on a local scale. It can be concluded, that high granularity of Swiss Flickr photographs basically concentrate as expected on urban areas, which is not a positive argument for the selection of this data source. A higher granularity also in non-urban regions would be preferable for wilderness research but since this evaluation does not concentrate on local wilderness interpretations but is more interested on a country-scale, the granularity is dense enough. But especially in the categories assigned with very few hectares the low granularity is remarkable and makes the interpretations vulnerable for influences of bulk uploads.

6.1.2 Prolific user data exclusion required?

It is known that on social media platforms the user activity differs in a high range. Purves et al. (2011) compared Geograph and Flickr user activity and both of them show a bimodal activity curve that illustrates that very few users tend to extremely high sharing activity and the majority of users show just very small activity. This could be confirmed for the Flickr community in Switzerland as illustrated in Figure 5.3. The curve of the social media platform Geograph is even extremer than the one of Flickr, according to the findings of Purves et al. (2011). This is hardly thinkable considering the extreme participation inequality this work has detected for Flickr. However, the bimodal curve indicated the requirement of analysing the effect of prolific users to the applied dataset.

The results of the prolific user analysis have shown that the most active users are by far more active in zones with lower wilderness quality. The influence on tags is strongly concentrated on the lowest wilderness categories which is also the case for unique tags. The normalization process does not change this tendency but clarifies it even more. Since this work is generally interested in the regions with higher wilderness quality the effects of the prolific users are negligible to this research and an exclusion of 11.6% of total data volume would be a waste of valuable data in this case. In literature, most of comparable approaches do neither evaluate nor mention the awareness of prolific users (Hausmann et al., 2017; Tims, 2014; Wood et al., 2013). This shows the necessity of the toolbox Purves and Mackaness (2016) suggested for scientific work with social media data. The decision to keep the data generated by prolific users has consequences especially for the lower wilderness categories. Further analyses concentrating on the whole spectrum of wilderness categories within the GIS-model have to respect that the photos in lower wilderness categories have been generated by only few users. Thus, Flickr might not be the most adequate social media platform for wilderness research since most of the users concentrate on urban areas. But as the evaluation has shown is an exclusion of the data generated by prolific users not an absolute necessity. When concentrating on a more urban feature, it would be suggested to exclude the prolific data from the evaluation.

6.1.3 Bulk uploading users and their influence to Flickr data

Although the effects of the prolific users have been classified as not fundamental to wilderness research the aptitude of Flickr photos has to be evaluated by considering bias of bulk uploads. Even though research is aware of bias caused by bulk uploads, only very few mention or handle it such as Purves and Mackaness (2016) and Hollenstein and Purves (2010). Since no clear procedural methods are defined in research to deal with bulk uploaded data, an own approach has been initialized (section 4.3.3).

A first evaluation has shown that not only coordinates but also tags can be used to identify bulk uploads. Thus, the whole evaluation for identifying bulk uploading users has been executed on the base of tags which stands in contrast to the approach of Hollenstein and Purves (2010) which used the coordinates for that purpose. The new approach classifies users according to the bulk-index (bi) which allows distinguishing between users tending to bulk uploads or not according to their uploading behaviour. The bulk index evaluation has shown three major insights. First, that the phenomenon is not restricted to the prolificness of users, according to the coefficient of determination of the trend line. Thus, no tendency of a relation between a very active user and one with high bulk index could be detected. These findings are relevant insofar as the pre-determined threshold, limiting the analysed number of users for this evaluation to all those who have uploaded more than 500 photographs, has to be scrutinized. Second, that uploading social media data in form of bulk uploads is quite a common characteristic. So not only highly active users tend to bulk uploads but the effects of bulk uploads from active users can be stronger due to a larger quantity of affected photographs. This evaluation has not analysed the effect of bulk uploads in general nor could such evaluation be found in other research approaches. The third and most important insight is that an exclusion of data generated as bulk uploads can be suggested when working on tag-based social media evaluations. Although this has already been suggested by Purves and Mackaness (2016), the applied evaluation has approached the bulk upload problem from another perspective which has confirmed that suggestion. Therefore, a threshold would be required to distinguish between normal and bulk uploading users according to the classified bulk index. In this work, the awareness of bulk uploads and the biasing effect to tag-based evaluations came up in an advanced step of the work process when certain tag-based evaluations have already been done. The actual exclusion process required according to the evaluations could therefore not be implemented. Thus, the data of bulk uploading users has not been removed from the Flickr base dataset of this work. But the bias coming from these data has been taken into consideration while interpreting the tag-based evaluations. Thus, this evaluation has concentrated on the development of a classification method, but has not specified a clear threshold. The definition and evaluation of such a threshold could be a potential for a future master's thesis, though the theme would be rather mathematical or technical. Additionally, an improvement of the bulk index equation could be taken into account in order to evaluate its effectiveness and validity and to make the data having a normal distribution. It has to be mentioned at this point that the equation does not help to decide which data should be excluded from a dataset since it only classifies users and not their data. Removing all data of a user identified as a bulk uploader would make less sense than an exclusion based on coordinate similarity, applied by Hollenstein and Purves (2010). Because most likely not all data generated by a user tending to bulk uploads have been generated in form of bulk uploads. Thus, an exclusion of all data generated by a bulk uploader would also remove the data not providing any bias. The potential of this bulk index can be found in other directions, exemplified in the following paragraph.

Optional use cases for the bulk index equation

The potential of the bulk index lies in its classification of users according to their uploading behaviour. If for a certain analysis only users want to be taken into account which show none or just a few bulk uploads, this bulk index can be valuable in order to distinguish between the users. Another example would be if one wants to analyse specifically the characteristics of bulk uploaders the information could help to detect the wanted candidates according to their uploading behaviour. A third use case can be stated for social media platforms which could apply this index to their users in order to rank them according to their generated data quality. If a platform is especially interested in data sharing based on a certain quality, this index could encourage users to improve their tagging behaviour and to assign tags more specifically and accurately. Also science could profit from such an index, since the data quality on social media platforms applying this index could be increased accordingly. The bulk index works with relative values and could be applied to any kind and quantity of social media data containing tags and user identification numbers.

In general, bulk uploads are seen as a biasing problem. But they are not problematic in all use cases of social media data. Goodchild and Glennon (2010) illustrate the use of social media data for fire detection and observation in severe cases like large-scale wild fires. In such a use case, the data quantity available in a short time period is prior to the data quality, although good spatial accuracy is highly requested. In such a case, bulk uploaders might even be preferred to normal users.

How to improve the bulk index equation

The developed approach to bulk uploads is by far not perfect nor is it complete. Certain improvements could be detected during the appliance. First, the pre-determined threshold to limit the evaluation only to users with a certain uploading quantity has to be questioned. The evaluation has shown that no correlation between the number of uploaded photographs and the bulk index exists. Thus, for future appliance, the evaluation could either be applied to all users or the threshold could be reduced to a lower number. Second, 7.7% of all considered users have uploaded no single tags. These users should already be removed at the beginning of the evaluation in order to reduce the number of bulk index outliers, since these non-tagging users are all classified with a bulk index of 5 which refers to a highest tendency for bulk uploading. This has been applied in a supplementary step which changed the required percentage for determining the data as having a normal distribution to 64.13%. Although the percentage is closer to 66.67% than before, it could still not be classified as a normal distribution. However, excluding these users is a necessary step to improve the equation. Third, to consult temporal metadata and check for differences in uploading time to better identify bulk uploads. It is possible that not all photographs with the same tag-combination have been uploaded within the same process. This could be verified by consulting the temporal metadata and check for uploading time differences. The last improvement and also the most important one would be to implement the discussed threshold which allows the determination of which users should be excluded according to their bulk index. This would complete the bulk index evaluation.

6.1.4 Concluding findings about the aptitude of Flickr data to wilderness research

The applied evaluations have shown that the knowledge acquired in the theoretical section about the potential biases of social media data can be found in Swiss Flickr photographs. The behaviour of contributors of the Flickr platform has relevant influence on spatial evaluations like this approach. Most active users of the Swiss Flickr community are generally active in regions classified with low wilderness quality indices, which is the reason that the bias due to prolific users especially affects regions with low interest to this work. The bulk upload evaluation has illustrated that bias due to bulk uploads affect not only a few regions like the bias of prolific users, but has been observed across the whole spectre of wilderness. Although the exact effect of this bias could not be evaluated, an exclusion would have been required. As a consequence for not respecting that at the beginning of the workflow, the tag-based approaches of this work increase in risk to be affected by such bias. Nonetheless, Flickr data can be seen as an optimal source to extend the information of a GIS-model with perceptional information about regions where sparely data is available. More about the positive and negative characteristics of the Flickr dataset is discussed in the strengths and weaknesses (section 6.4.2).

6.2 Detecting wilderness variations in Flickr data

The tag-based approaches described methodologically in section 4.4 and their results described in section 5.2 have been evaluated in two separate steps which will be discussed in this section. On the one hand, a collection of tags has been evaluated according to which wilderness tendencies within the Flickr data have been determined. On the other hand, an adapted tf-idf equation has been applied in order to detect the most representative tags within each wilderness category of the GIS model. These two steps will be discussed separately in the next two subsections (section 6.2.1 and 6.2.2) and then compared in the third subsection 6.2.3.

6.2.1 Wilderness detection by tag selection

The selection applied to detect wilderness in the Flickr base dataset has been evaluated in two methodologically different steps. The literature review and analyses of definitions resulted in four tags whereas the co-occurrence analysis returned two additional terms. Since Flickr is not language-specific, the Flickr base dataset contains many photographs generated by different users applying various languages. Therefore, some of the terms evaluated by the selection had to be translated in order to cover an appropriate language volume of tags. Visualizing the distribution of these tags across all wilderness categories according to their normalized number of occurrence is a simple but effective way to detect serious differences between wilderness in social media data and the GIS-model.

The observed tendencies described in section 5.2.1 show that in general, the distributions of the tags correspond to the expectations. One specific exception is graph c) which represents the tags of *forest* (*wood*) and *wald.* Since these tags have been determined as representative for wilderness, an increasing number of tags per hectare with increasing wilderness quality has been expected. Since this is not the case, it indicates that either these tags cannot be seen as representative tags for wilderness or the GIS-model does not respect woods in their evaluations. The GIS-model effectively considers information about the degree of forestry use in their wilderness attribute naturalness so this cannot be the reason for that distribution. Reflecting that the wilderness quality index represents the whole spectre of where wilderness can be perceived in Switzerland points out that also high mountains without any vegetation can be classified as wilderness. With this argument the decreasing tendency in higher wilderness categories makes sense and most likely is the reason for that tendency.

In general, the graphs b), d) and e) show most accurate results according to the expectations. The breaks in higher wilderness categories of the graphs a) and f) can also be detected in graph d). Since all these tags show comparable breaks, the source for that is rather model-based than tag-specific. To argue that this is due to bias caused by the little number of hectares assigned to the category 17 to 20 is an optional explanation. Another reason can also be that these tags are simply not assigned so many times in regions with highest wilderness categories. The determination of the exact reason for that would go beyond the scope of this work. However, when excluding the tag with most counter trend, in this case the tag wood and its synonyms, a clear increasing tendency of number of tags with increasing wilderness quality can be observed, as graph g) illustrates. Also the distribution of a tag appearing in all wilderness categories like *switzerland* behaves as expected. Graph h) shows that this tag has equal relative numbers across all wilderness categories expect category 20. The high number of this category can also be explained by number of assigned hectares to this category, which is very small. In order to verify, that also counter-trends to all that existing correlating trends can also be detected, graph j) illustrate that on bas of the distribution of the tag *architecture*. This verifies the expressiveness of the other graphs.

The results have shown that some of these tags within the collection are better represented by the GIS-model than others. Grave tendencies that would argue against a correlation between the two datasets have not been detected. This means, that the manual selection of tags was accurate insofar as the representations of these tags correspond to the wilderness distribution of the GIS-model. However, a manual selection of tags can be quite error-prone since the handling with tags requires many precautions and considerations. The exact terms selected to the collection of tags depends on multiple features like the social media platform, on the language in which the tags have been generated and also on the term which is attempted to be represented by the collection. Flickr for example transforms the capital letter of each tag to lowercase which results in zero found tags when searching for upper case tags. Capitalization issues have thanks to Flickr not to be managed separately but are also an important research topic in GIR (Manning et al., 2008). The language is a very decisive factor for tag-based analysis in general, especially if the research area crosses multiple language borders which is the case in Switzerland as illustrated by Hollenstein and Purves (2010). Their discussion about vernacular tags basically was focused on toponyms for which a certain chance exists, that the toponym is written the same way in multiple languages. Searching for tags representing wilderness is different insofar as already within the same language multiple synonyms can exist for the same expression. The best example is the term *wood* which was required to be extended by the term *forest*. The final results of the tf-idf evaluation in section 5.2.2 have shown that the term *tree* would have matched even better than the aforementioned two terms, although it stands only for a single object and not a feature describing some kind of landscape.

Methodologically, searching representative tags in literature and definitions is a simple, but effective way for a selection, although a risk for subjectivity exists. Also the co-occurrence method is not complex, but is based on statistically evaluated representative tags what reduces subjectivity. Nonetheless, the co-occurrence part of this evaluation has to be discussed more critically. The co-occurrence analysis to the term *wilderness*, as described in section 4.4.3 has only be implemented by analysing a corpus of 598 photographs due to a lack of use of the tag *wilderness* by the Flickr community within the bounding box of Switzerland. Applying the co-occurrence approach to a tag with higher presence would be preferable but in this work only the explicit term *wilderness* or *wildnis* could have been respected. However, the advantage of this co-occurrence method lies in its small temporal and computational effort which would be more beneficial when analysing a tag that appears more frequently.

6.2.2 Wilderness detection according to tf-idf-uf evaluation

The ranked lists of tags classified by the adapted tf-idf-uf equation described in section 5.2.2 have illustrated several specific characteristics which will be discussed here in reference to the wilderness model and the whole wilderness approach. But first, two questions have to be answered before actually discussing the tendencies observed within the ranked tf-idf-uf lists. In order to verify that all applied methodological steps were relevant and advantageous to the final evaluation, the question arises, if the exclusion of tags like toponyms and other specific tags was a necessary step. Also the adaption of the standard tf-idf equation with the user frequency (uf) parameter has to be discussed critically. Therefore, the next two paragraphs take note to these critics before the interpretation of the tf-idf-uf results will be discussed.

Was the exclusion of tags a necessary methodological step?

Through literature review and own findings described in section 5.2 have evaluated beneficial information gained from toponyms, the decision has been taken to exclude all toponym tags and also some tags auto-generated by devices and applications from the main Flickr dataset. As illustrated in the subsection 5.2.2, some valuable information has been eliminated in this step. Especially the occurrence of specific toponyms like mountain or glacier names denotes important information about which kind of photographs occur in which wilderness categories. It is therefore important at this point to question the necessity of the exclusion process. The most decisive factor advocating this process was that the spatial information contained by Flickr toponym tags are irrelevant information insofar as the spatial referencing of the photographs has already been defined by classifying them to wilderness categories according to their coordinates. As evaluated in this work and also confirmed by Hollenstein and Purves (2010) most of the toponyms refer to the country or continent where the photos have been taken. Thus, the majority of the toponyms refer to the study area which is no gain of information. A second argument is that they disturb the output lists of the tf-idf-uf evaluation because toponyms like *switzerland* or similar are applied very frequently and by many users which is why most of the top ranks of the tf-idf-uf evaluation across all categories are scored by such toponyms when avoiding the exclusion process. Comparing the outputs before and after the exclusion leaves no doubt that this process was absolutely required in order to detect wilderness in the Flickr base dataset. But during the exclusion process new challenges have been faced which are important to be mentioned.

The initial exclusion of toponyms was challenging insofar as some of the tags could not be definitively identified to which type of tag they belonged. Classification difficulties due to term ambiguity has already been observed and described by others (Hollenstein & Purves, 2010; Jones et al., 2008). The tag *zug* for example has multiple senses as it might stand for the canton Zug in central Switzerland, for the capital city of the canton Zug with the same name or for the German term of a train. Especially the latter case could semantically not be excluded since some very active Flickr users showed large interest in public or private transports like trains, cars or airplanes. Particularly in the lower wilderness categories are the tags *train*, *zug*, *sbb*, *cff*, *car* or *auto* very present in the high ranks of tf-idf-uf lists, which illustrates the interest for public transports in areas assigned by low wilderness quality. Due to these ambiguity reasons, tags which could not be classified with certainty could not be excluded. In order to increase the certainty for classification according to the semantic meaning the corresponding tags would have to be compared with the other tags of that photograph. In this work, term ambiguity was only relevant for toponym exclusion.

Critical reflection to the adaption of the tf-idf evaluation

The ranked output list of the standard tf-idf method (Appendix D) and the previously gained knowledge about user behaviour of social media data has led to an adaption of the standard version of the tf-idf equation as described in section 5.2.2. But how successful was this adaption to the final table (Table 5.3) and which advantages could have been achieved by this adaption? The description of the aforementioned tables has already illuminated that the strongest difference between the standard and the adapted version, which respects the user frequency (uf) of a tag, are the tags generated by single users or very few users. By only applying the standard version, many top-ranked tags can be classified to certain events or specific interests of these few users. These event-referring tags disappear from higher ranks when applying the adapted version of tf-idf. The semantic extraction of events is a highly discussed research field and other researchers have initialized approaches based on either only temporal patterns (Rattenbury et al., 2007) or spatio-temporal patterns (Naaman et al., 2004). Because this work has only applied a simple tf-idf evaluation, some event-based information can be gathered out of the ranked lists as well. These event-based tags can also be related semantically to wilderness, although they do not necessarily refer to positive wilderness quality. Some examples illustrate that these event-tags can be used to verify if the wilderness model and wilderness represented by Flickr tags are related or not.

Goûts et terroirs, Bulle (FR)

The only category containing tags like *gout*, *gouts*, *terroirs*, *zigermeet* or *salonsuissedesgoûtsterroirs* in its top rank is category six. They might all refer to the same annual event called Goûts et Terroirs[14] located in Bulle (FR) which lies exactly in an area

[14] http://www.gouts-et-terroirs.ch/de/home/

dominated by hectares applied with wilderness quality index of six according to the wilderness model. Without having any knowledge about the exact coordinates, the potential location where these photographs could have been evaluated. In addition to that it is known now that the top tf-idf ranked tags evaluated by the standard version are basically generated due to a very specific event. And since these tags refer to an event which takes place at the same location annually and this location is close to the center of the city, the wilderness quality index would be expected to be small. This expectation has been confirmed by the GIS model which assigns this region the index six.

Fliegerschiessen, Axalp (BE)

Many tags within the top-twenty ranks of the categories 12 to 16 refer to planes or military and similar. Tags like *atterraggio*, which is the Italian word for landing, *fliegerschiessenaxalp2012*, *airshow* or *warplane* have most likely been generated at one of the international events called Fliegerschiessen[15], where military planes and aircrafts are demonstrated in a show near Brienz (BE). Visitors to that event are asked to climb up to the mountains to get a nice shot of the demonstration, which can also be detected by the higher wilderness quality values these tags have been assigned to. Compared to the last described event, which could only be detected in the top ranks of one single category, these event-based tags are distributed over at least four categories. Multiple reasons can be suggested for that characteristic. On the one hand can planes be seen from far away and they fly over a large area which increases the chance that some users uploaded photographs taken down in the valley and not up in the mountains. On the other hand the resolution of the wilderness model is so high that within a small distance, large variabilities in wilderness quality indices can occur. Additionally, the slope from Brienz up to the Axalp, where most visitors watch the demonstration is quite steep. Since the GIS-model obviously classifies mountain regions as having higher wilderness quality and the effect of ruggedness seems to be a weighty component of the classification this characteristic seems to be reproducible.

World Economic Forum (WEF), Davos (GR)

A third event represented by top-ranked tags is the annual event called World Economic Forum[16] (WEF) taking place in Davos (GR). Davos is a touristic ski resort in winter and though located in the mountains higher than Brienz and has more than 10’000 inhabitants. The expected wilderness quality index for the congress center, the actual location where the event takes place, would be between seven and ten according to the presence of mankind and infrastructure. Actually, the four tags *annualmeeting*, *congresscenter*, *worldeconomicforum* and *wef* are the top-ranked tags in wilderness category 12 which is also the only category these tags appear in top-ranks. Although the classification of the GIS-model does not fit the expectations of the author this event concentrates on one specific category and is highly present in their top ranks.

The mentioned three events are only some out of multiple other examples that could be mentioned. When the classified wilderness categories, these tags have been assigned to, would differ much from the expectation, the classification of the GIS-model has to be scrutinized. In the case of the events mentioned before the assigned categories make more or less sense. Thus, these event-based information can also be used to verify the wilderness of the GIS-model to the wilderness in Flickr data. On the other hand, these verifications need to be interpreted carefully,

[15] http://www.vtg.admin.ch/de/armee.detail.event.html/vtg-internet/verwaltung/2018/18-10/18-10-10_lw.html

[16] https://www.weforum.org/

since the exact reason why a photograph has been tagged by these event-based tags cannot be determined or extracted by this approach. In order to optimize the detection of event-based semantic from the Flickr photo dataset, the temporal metadata could have been taken into account to verify if the uploading approximately corresponds with the date of the event. This was not part of that work since large-scale events which have enough attention that they would be remarkable in Flickr tags, rather take place in urban areas than in wilderness. However, since these event-based tags disappear from top ranks of the adapted tf-idf-uf evaluation, they are most likely generally generated by only a few but active users.

Although some of the wilderness-relevant tags of the adapted version can already be found in the ranked lists of the standard version like *mountain*, *snow* or *hiking*, the output lists of the adapted version seem to represent wilderness much better. Detecting wilderness in the tf-idf lists of the standard version seems to be much more difficult than in the adapted one. Particularly the definition of borders between wilderness categories is challenging in the outputs of the standard tf-idf since many top-ranked tags are generated by a few users and cannot be related to wilderness like the tags *chrindi* or *wasif*. The conclusion of these findings is that the adapted version fitted much better to this approach since the outputs were more accurate than the one of the standard tf-idf equation. Respecting the number of users who generated the corresponding tags was a necessary and beneficial adaption which increased the quality of representative wilderness tags.

Discussion of the ranked output lists and their tag distribution

Having verified the methodological steps it is now time to concentrate on the interpretation of the tf-idf-uf results and their relevance for this approach. Since all wilderness annotations in this work refer to the GIS-model initialized by WSL, the expectations for wilderness tendencies are oriented at the wilderness spectre represented by the wilderness categories. Referring to the applied wilderness attributes such as human influence and naturalness, it is expected that tags referring to anthropogenic objects or concepts are observed in lower wilderness categories and more natural terms are expected to be found in the top ranks of higher wilderness categories. In order to verify these expectations global patterns of tags across all wilderness categories will be discussed on the base of varying distributions of tags. The analysed tags will be divided into those referring to anthropogenic features and in others referring to natural features whereas the latter will be discussed first.

The best examples of tags counting as referring to **natural** features are the tendencies of tags like *mountain* and *mountains* which show lower ranks in lower wilderness categories and *glacier* or also *gletscher* which continuously increase in rank in the mid categories and show top ranks in higher categories. The highest ranked tags across all categories are the tags *mountain* and *mountains*. The most obvious reason is that these tags have simply been used very frequently, as confirmed by the frequency analysis for the tf-idf evaluation. But only arguing that the tags were frequently used is not sufficient to describe the high ranks across all wilderness categories. Switzerland is well known for its mountains and many visitors come annually to see them, what implies that people are attracted to photograph them. Geotagged photos represent a place by their coordinates but not necessarily by their tags as concluded by Rattenbury and Naaman (2009). The mountains height make them visible from far away and taking a photo in Zurich for example, where the mountains are visible in the background, would describe a case where the actual wilderness category of the geo-referenced photograph would be rather low but the user would have tagged the photograph with the tags *mountain* anyway. In contrast, glaciers are large objects and remarkable too, but only few of them can be seen from far away since they are generally surrounded by higher mountain ridges which isolate them at least partially from farther views. Also this is observable in the data since the glacier-tags start to occur in the top ranks in higher wilderness categories and cannot be found in the top ranks of very low categories.

The tags *snow* and *landscape* as well as *nature* keep standing on equal ranks across all categories except in the lowest categories as Figure 6.1 illustrates. Theoretically, all of these features can be photographed in all wilderness categories since snowfall affects whole Switzerland in winter times and landscapes can even be photographed from a small hill or a high building within a city with lowest wilderness quality values. Why the tag snow still appears in the top ranks of the lower wilderness categories most likely can be explained by the fact that snow is rarer in cities but more remarkable and therefore attractive to take a photograph of it. Tagging a photograph with *nature* is even imaginable when taking a picture of a growing tomato plant on a balcony in a large city where the plant symbolizes a natural contrast to the urban environment. Thus, these distributions all make sense according to the expectations stated above.

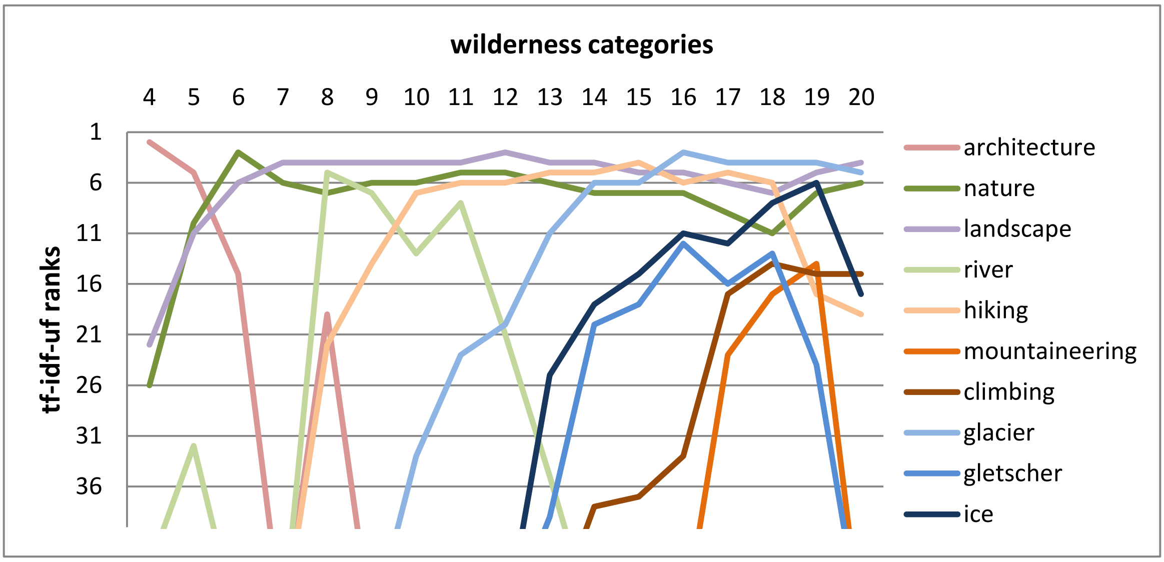


Figure 6.1 – Observed tendencies of ranked tags according to the tf-idf-uf evaluation across all wilderness categories.

Another tendency can be observed by highlighting the tag *hiking* within all wilderness categories. Increasing ranks, beginning with category 8, show that photographs tagged with *hiking* have been uploaded already in lower categories but keep being highly representative until decreasing again in category 19. The high ranks of this tag are distributed across a large range of categories which may refer to the varying regions and environments where hiking can be processed. Also the disappearance from top ranks in highest and in lower categories makes sense insofar as regions with very low wilderness quality mostly are placed in urban regions and the highest quality values can be found at very steep slopes in mountain regions. Hiking is rarely practiced in urban or very steep regions.

Other tags which refer to more **anthropogenic** features show contrasting behaviour to the natural ones but re-emphasise the expectations as well, as the examples of the tags *architecture*, *concert* or *street* illustrate. All of them disappear from the top fifty ranks until the category 9 whereas their maximal achieved ranks are in the lowest categories. This means that their representativeness is much higher in lower wilderness categories and their disappearance from top ranks in higher ones illustrates that they have much more specifically been applied in regions categorized with low wilderness quality index.

The disappearance of the tag *architecture* from top ranks in categories higher than 9 is an excellent example demonstrating that contrast distribution to natural tags. Since most architecture can be found in urban areas, the distribution of this tag across the wilderness categories is in line with the stated expectations. Also concerts or especially open air festivals are restricted to flat terrain and principally take place in urban areas. But other anthropogenic features are not as strongly restricted to urban areas as architectures or concerts. The tag *street* shows the same distribution like the tag *architecture* even if the street network in Switzerland also reaches regions applied with medium wilderness quality indices at least. Thus, Flickr users tend to upload photographs applied with the tag *street* preferably in regions with lower wilderness quality even if the streets but also architecture could be found in higher categories as well. The disappearance of anthropogenic tags from top ranks around the categories 7 and 8 reveals interesting information about the photographing behaviour of the Flickr users. In regions applied by wilderness categories above quality indices 9, users tend to take more photographs of natural features whereas in regions with lower wilderness quality than 9 tend to capture anthropogenic features.

According to the described tendencies, first variations of wilderness within the base Flickr dataset have been described. The categorization of the GIS-model respects the coarse subdivision between anthropogenic and natural features but at this point, only two different characteristics in wilderness spectrum could be detected whereas the model categorizes the whole spectrum in Switzerland into twenty different categories. But are these two characteristics the only detectable divisions of the Flickr wilderness continuum?

The theoretical evaluation (section 2.2) has shown that wilderness is a concept depending strongly on its definition. Principally, the wilderness phenomenon is not seen as having clear borders but rather as representing a continuum. This continuous nature is based on the vague characteristics of wilderness which are also reflected in the output of the tf-idf-uf evaluation. The described tendencies of tags decreasing or increasing in ranks barely showed abrupt character but rather slight tendencies across multiple categories. This illustrates that such patterns describing wilderness cannot be seen within a specific wilderness category but rather as global patterns across a wider range of different regions. Nevertheless, some patterns indicating borders with changing wilderness characters can be identified. Defining borders within the wilderness continuum has still been a challenge for research and is an ongoing debate which changes with changing definitions of wilderness. The vague character of the phenomenon resists defining strict borders but certain categories show a shift of high-ranked tags, which can be seen as indicating a change in wilderness. The most remarkable shift can be placed in wilderness categories 7 and 8 where most of the tags referring to anthropogenic features strongly decrease in ranks, while tags referring to natural features start to score higher ranks as described above. This shift most likely appears due to the fact that most Flickr users are active in cities and urban areas as described by Antoniou et al. (2010). A second shift can be observed around the wilderness categories 11 and 12 where tags like *river* or *lake* slowly decrease in rank and *glacier* and *hike* increase. A third shift sticks out when the tag *hiking* decreases and the tags *climbing* and *mountaineering* achieve their highest ranks between the categories 18 and 19. Less obvious tendencies like the increasing ranks of tag ice around the second described shift and the occurrence of tags like *summit* or *topofeurope* around the second shift support these observations. Classifying the four categories formed by these three observed shifts could result in a subdivision of wilderness as follows:

	Urban and cultivated regions	wilderness categories 4 – 7
	Flat to hilly natural regions	wilderness categories 8 – 11
	Mountain regions and glaciers	wilderness categories 12 – 18
	High mountain regions with steep slopes	wilderness categories 19 – 20

Source for icons: Flaticon[17]

This division describes the wilderness detected in Flickr photographs according to the wilderness GIS-model. It reflects the output of the tf-idf-uf evaluation and illustrates that wilderness variations within the GIS-model can also be reflected by the spatial distribution of social media data. The most remarkable factor in this tf-idf-uf evaluation was the influence of the wilderness attribute ruggedness, manifested by the tags in the output list of this evaluation. Many tags like *hiking*, *summit*, *glacier* or *climbing* refer strongly to this wilderness attribute. But also counter examples like the tags *river*, *flughafen* or *lake* can be used for interpretations referring to ruggedness. The remarkable influence of the wilderness attribute ruggedness will be further discussed in subsection 6.4.1. Also the effects of the negative weighting for infrastructure like streets can be observed in the Flickr dataset since the tag *street* decreases in rank rapidly with increasing wilderness quality index. In order to verify the influences of each wilderness attribute to the social media data further research would be required.

At this point it is required to mention that the interpretations of the different tag-tendencies within the wilderness categories have to be set carefully since a risk for distortion coming from two sources has to be taken into consideration. On the one hand, the hectares applied to each wilderness category are not equally distributed. This especially affects the interpretation of the categories with highest varying number of hectares, which would in this case be the categories 6 and 20. This consideration has already been mentioned in the last subsection (section 6.2.1), where particularly the values of wilderness category 20 have shown remarkable results. But since the number of grid cells is defined by the GIS-model and no changes to that model have been applied in this work, the first bias could not have been avoided here. Furthermore, the

[17] www.flaticon.com

unequal distribution of the photographs across the categories can cause bias as well. This bias is a result of the combination of the GIS-model and the Flickr dataset and would be a larger problem when comparing tf-idf-uf values across different wilderness categories. Since the exact values are not taken into consideration but only the ranks of the tags, this bias only needs to be respected in the interpretation of the results.

6.2.3 Concluding findings

Applying the standard tf-idf evaluation to the Flickr dataset in combination to the GIS-model has led to multiple ranked lists which have been optimized by specific tag exclusion and an adaption respecting the user frequency of each tag. The output result of the final tf-idf-uf evaluation gave valuable information about wilderness in Flickr data according to the GIS-model. The results even allowed further subdivision of wilderness into four distinguishable wilderness characteristics reflecting the wilderness continuum. The continuous nature of that phenomenon is expressed in the vague borders in the tf-idf-uf output. Although the categorization in different wilderness categories is based on many more data sources in the GIS-approach, social media data provide valuable insights into the perception of space and especially wilderness for Flickr community in Switzerland. Thus, the variations of the GIS-based wilderness approach of Radford et al. (unpublished) could also be detected in georeferenced Flickr photographs by identifying representative tags and applying the tf-idf-uf method.

6.2.4 Methodological comparison of approaches attempting to find representative tags

This subsection discusses the two evaluation techniques applied to find the most representative tags within the base Flickr dataset described in section 5.2.2.

Methodologically, especially the different effort required to process these to techniques show large difference. While the specific tag analysis can be implemented with low calculation and programming time, the tf-idf-uf method requires a multiple of that effort. But in contrast, the simpler specific tag approach requires much more knowledge about the actual term in order to be able to find representative terms, whereas the tf-idf-uf evaluation could be applied with very limited knowledge.

Taking a look to the results and comparing them obviously, the tf-idf-uf approach returns much more qualitative information and offers a broad potential for eventual interpretation and discussion. Also very specific information about the GIS-model and the Flickr data are detectable and describable whereas the specific tag analysis only allows interpretation of global matches and mismatches between the two dataset. Working with the specific tag method also implies some subjective or perceptual influences insofar as the selection of representative tags is a manual process by interpreting literature and co-occurrence evaluations. A potential risk exists to favour certain tags according to personal perception. The better the theme is known by the researcher the smaller should be the risk for such subjective influence. Also language is a challenge for the simple method since multiple tags have to be translated whereas the tf-idf-uf evaluation simply ranks tags nondependent their language. However, the described communalities between the outputs of these two methods have shown that some wilderness representative tags selected in the specific tag method could have been set better, like the tag *tree* instead of *wood* or *forest*. But the majority of tags within the selection could have also been detected by the tf-idf-uf evaluation and their detected tendencies are generally related in both approaches. This shows that the specific tag evaluation cannot be seen as an alternative approach to the tf-idf-uf evaluation but that these two methods are attractive to combine in order to verify their results.

6.3 Characterization of wilderness according to Flickr data

This section discusses the interpretations of the results mentioned in section 5.3 and takes note to the findings described in the theoretical section (section 2) and to wilderness research.

The last evaluation has shown that wilderness variations within the model can also be detected within Flickr tags, which allowed a classification of the wilderness continuum within the Flickr tags, as described in the last subsection. This classification has been used as a base for further wilderness characterization, as the third research question requires. Several different observations were helpful to detect more specific wilderness characteristics within the wilderness classes out of which the most important ones will be mentioned in the following subsections.

6.3.1 Global vs local tag distribution

While the second research question could be answered by the more global tendencies within the ranked tf-idf-uf lists (Table 3.1), tags appearing in a few or even one category are much more interesting for wilderness characterization. Rattenbury and Naaman (2009) describe spatial patterns of tags and argue that certain tags show no spatial pattern on a global scale but rather on a local scale. Referring to Tobler's first law of geography (Tobler, 1970), closer things are more related than distant things which means for local tag patterns that they tend to have approximately similar wilderness quality indices. This could be confirmed by the fact that certain tags only occurred in high tf-idf-uf ranks of one or few wilderness categories like the tags *wasserfall* or *mountaineering*. Thus, for characterization purposes the local tendencies within the tf-idf-uf output lists have more priority. Nevertheless is it questionable, if this evaluation can be used for local wilderness characterizations. Although the GIS-model has a high resolution of 100 meters, the Flickr photographs manifest some spatial inaccuracy which might cause bias to local wilderness information. However, the spatial scale of this work mainly concentrates on country level interpretations of the overall situation of wilderness. Tims (2014) has applied Flickr evaluations for solitude mapping on a local scale for a national park in Iceland which then, was combined with other factors to a wilderness map like the GIS-model applied in this work. His approach concentrates more on local conditions than this work, since it evaluates a national park and not a whole country. But to apply local characterization to Swiss national parks or also touristic regions or ski resorts, it is questionable if the granularity of Swiss Flickr photographs would be high enough to gain valuable information about wilderness situations in these regions. Further work is required in order to answer that question with certainty.

6.3.2 Verbs vs nouns

Not only local tag patterns have revealed valuable information for characterization but also the difference in distribution of verb tags and noun tags in high ranks of the tf-idf-uf evaluation. The third wilderness class referring to mountain regions and glaciers mainly could be characterized by the verb tags like *hiking*, *skiing*, *mountaineering* and *climbing* while verbs are much less present in high ranks of the other classes. This tendency could be used to interpret, that Flickr users in Switzerland like practicing leisure activities in that specific wilderness class. Thus, also differences in term types can be used to determine differences in wilderness.

6.3.3 Wilderness attribute ruggedness

One of the most decisive factors for this evaluation was the separation between tags in some kind referring to the ruggedness of the terrain they might have been taken. This was rather an interpretation than a measured classification but some separations could be processed with relative certainty. From a relative perspective, the tag *airport* refers more to a flat region than for example the tag *mountain*. The argument of such tendency has also been applied to tags like *hiking*, *mountaineering* and *climbing*, as well as for example the tags *river*, *peak* or *summit*. These relative separations suffocated to further characterize the wilderness classes and how Flickr users perceive them. Considering another thematic field than wilderness most likely this attribute could not be used in order to distinguish between different Flickr tag characteristics. Thus, the benefit for wilderness characterization of this specific wilderness attribute comes from the GIS-model and changing the thematic field would also change the effectiveness of this attribute for characterization purpose. In other words, evaluating a GIS-model representing life quality or biodiversity as examples, the attribute ruggedness most likely would not have the same effectiveness to characterize the corresponding theme. The relevance of ruggedness to this whole evaluation has been mentioned in multiple parts of this work and it will be further discussed in section 6.4.1.

The described observations were a big help for identifying how people perceive and describe these wilderness classes and to imply perceptional information to the abstract and mechanic GIS-model. The evaluation has demonstrated that tag analysis of georeferenced Flickr photographs give valuable insights into Flickr user tagging behaviour according to a specific concept like wilderness. While some attributes of the GIS-model were more salient like ruggedness, others like human influence could be augmented by applying social media data as an additional component. The variations in wilderness according to the GIS-model applied in this work have been detected in Flickr data as a first step and then be further characterized.

6.4 Strengths and weaknesses

The content of this section concludes the strengths and weaknesses of this work by referring to the two applied main data sources in the first and second subsection and states out some critical reflections to the main method tf-idf-uf in the last subsection.

6.4.1 GIS-model

The wilderness model provided by WSL has shown some characteristics which are important to be mentioned since this work was strongly oriented to this model.

Scrutinizing wilderness quality range

The general strength of such GIS-evaluations are, that large-scale mapping is realizable with relatively low temporal effort as already mentioned by Carver et al. (2012). For evaluating the specific concept of wilderness many different data sources are required, due to the complexity of the phenomenon. Combining different data sources might negatively affect the control and overview of the final map and its content. According to the MCE, the range of classification of the wilderness qualities distinguishes between 20 categories, which is not a fixed number of partition but just what Radford et al. (unpublished) have decided to be adequate in their case. More important is the fact that almost no hectares are assigned to wilderness categories 1 to 3, which raises questions about the classification. Interestingly, all of the few assigned hectares are at the administrative borders of Switzerland and the only explanation of their low wilderness quality indices is therefore that not all data sources applied to that model have had the same spatial extent. Since the quality index has been calculated within the MCE by additions and multiplications of different paragraphs for each raster cell, a lacking paragraph might have led to very low quality indices at the corresponding hectares. Thus, a weakness of the GIS-model is the spatial matching of all applied data sources which had in this case influences to the range of wilderness quality indices. This work has prevented for influences by ignoring the categories 1 - 3 for all evaluations.

Augment the wilderness attribute *human influence* by considering Flickr hotspots?

In section 5.3.3, the wilderness GIS-model has been critically analysed by observing the Flickr data hotspot around the Jungfraujoch. This critical observation revealed another weakness of the GIS-model. The evaluation has shown that the GIS-model has respected the building of the Jungfraujoch observatory and has reduced the wilderness quality indices of the surrounding hectares accordingly. But the full extent of tourism influence in that region has not been covered. As a reason for that, some hectares with high tourism frequency, represented by many uploaded photographs on Flickr, are still assigned with high wilderness quality indices by mistake. This could be augmented by applying Flickr photographs and take into consideration that hotspots in Flickr spatial distribution might serve as valuable indicator for detecting strong human influence. To examine the potential of Flickr photographs for such purpose, further research is required on that.

Is the wilderness attribute *ruggedness* over weighted?

During the tag evaluations within this work one specific characteristic of the GIS-model has been faced multiple times. As the European Environment Agency (EEA) (2010) argues, most of wilderness zones in Europe are in mountainous regions which is in line with the classification of the GIS-model applied in section 6.2.2. Nevertheless, the wilderness attribute ruggedness seems to be weighted strongly since none of the other wilderness attributes was as remarkable during the tag-based evaluations as ruggedness. This does not necessarily need to be related to the weighting within the MCE of the GIS-model but it can also simply refer to the varying topography of Swiss landscapes. Thereby ruggedness seems to be the most remarkable attribute. It has been weighted much less than the other three main attributes according to Radford et al. (unpublished). Thus, the remarkable effect of ruggedness does occur due to an overweighted parameter in the MCE, but rather because of the specific landscape of Switzerland. It would be interesting to find out how important the steepness of a terrain is in relation to the wilderness quality index. Selecting another country like France, which consists of more flat regions than Switzerland, would help to verify this relation to the analysed landscape. For future work, analysing this relation could reveal valuable insights into the relation between wilderness and the ruggedness of landscape.

The debate about the different wilderness attribute is broadly discussed in wilderness research as the theoretical section has shown. Questioning these attributes is necessary insofar as the collection of wilderness attributes defined as relevant by Carver et al. (2012) can still be improved. Considering social media data may be an opportunity to further improve these attributes or how they are weighted in a GIS-model. Thus, the challenges related to the uncertain definition of wilderness are also remarkable in current wilderness research. Although this work gives no answer to that specific problematic is it strongly related insofar as another wilderness definition would have changed the GIS-model and therefore also the output of this work.

6.4.2 Working with social media data

Most of the mentioned challenges in the theoretical section (section 2.3.3) dealing about social media data have been faced within this work. Bimodal user participation, tag ambiguity, difficulties caused by different languages, prolificness of users and bulk uploads have all been faced and considered. But not only weaknesses of social media data have to be mentioned here. The clear strength of such data sources are that they are open and free, they are at least partially structured and are accessible with a few programming knowledge. Especially the structured format Flickr offers supported a quick access to the different metadata and allowed quick analyses. As already concluded by Purves and Mackaness (2016), the challenge when working with social media data lays not in the analysis per se, but in the initial processing of the data and in the interpretation of the results. An additional benefit specifically to wilderness research is the fact that social media data represent perceptual data generated in a social context. Other methods to gather or generate such volumes of perceptual information are hardly realizable, due to large required temporal, financial and work effort. Thus, UGC is a valuable source of information to extend a wilderness GIS-model.

The above mentioned challenges have already been discussed in previous sections. But other challenges need to be mentioned at this point. Since this work basically concentrates on a spatial feature, spatial accuracy is one of these challenges.

Positional accuracy of Flickr photographs

Hollenstein and Purves (2010) describe that Flickr contributors either choose to geo-reference the photograph's location or the scene being photographed. Remarkable landmarks like the Eiffel Tower or similar, are particularly affected by that. This causes spatial bias, which has also been detected in this work. Figure 5.11 illustrates the mountain Jungfrau as a natural landmark where many tags, like the tag *topofeurope*, have not been positioned at the location of the tagged phenomenon but rather at the location where the photo has been taken. According to the findings of Hollenstein and Purves (2010) are precision and accuracy of Flickr photographs accurate enough to describe city neighborhoods. The spatial accuracy of the GIS-model of 100 meters is higher than the spatial extent of a neighborhood. Thus, either the GIS-model resolution is too high for the applied Flickr data or if such a high resolution is necessarily required in an approach, Flickr accuracy might be too low for accurate evaluations. This illustrates that applying social media data for research depends much on the required spatial accuracy. Hochmair and Zielstra (2012) examined differences of spatial accuracy between Panoramio and Flickr and evaluated a median error distance of about 58.5 meters for his Flickr dataset. This threshold cannot be directly applied to the dataset in this approach but based on the fact that this distance would be the same in our dataset would mean that many photographs would have been classified to the wrong hectare. This would have a relevant effect on the distribution of the photographs across all wilderness categories. Although a risk for such distortion exists, its effect is not radical since wilderness is a continuum and therefore, neighboring hectares are generally assigned with same or slightly different wilderness quality indices, considering the first law of geography (Tobler, 1970). Therefore, this distortion would only assign small index deviations by mistake. In order to avoid strong bias of this distortion, an aggregation of the GIS-model could be taken into account so that the grid cell size would be increased from 100 meters to 500. To increase accuracy of social media images in general, either one could only concentrate on photographs with highest accuracy metadata or as Zielstra and Hochmair (2013) suggest, only use data geotagged by automated camera positioning.

Spatial differences in granularity of Flickr photographs

The Flickr density map (Figure 5.1) illustrates how spatially concentrated the Flickr community uploads photographs. Gliozzo et al. (2016) assert that for their research area in Wales, Flickr covers 60% of the research area while in contrast Geograph covers 99%. These two examples illustrate on the one hand, that the granularity of Flickr photos varies much within space and that other platforms may have a larger coverage on the other. Thus, applying Flickr to spatial wilderness research depending on the granularity of the data might be not the best solution. But more adequate alternative sources are rare, since wilderness generally concentrates on mountain regions, where social media contributors are limited due to cell phone coverage (Boller et al.,

2010) or accessibility (Fritz & Carver, 1998) amongst others. Limited granularity also constricts the use of Flickr for wilderness interpretations on local scales as evaluated in previous section (section 6.1.1). Local interpretations of very specific regions such as parks or touristic regions stand in particular interest of organizations like Mountain Wilderness, the WSL or tourism agencies. Depending on the purpose of the evaluation and the local granularity, Flickr can be taken into consideration also for local wilderness evaluations. But in general, Flickr data is more useful for country-scale or at least regional wilderness evaluations as realized in this work. Thus, it can be concluded, that the differences in photograph granularity are a weakness of Flickr which has particular influence on local-scale wilderness interpretations.

6.4.3 Tf-idf-uf – methodological restrictions and limitations

The tf-idf is strongly connected to the number of compared documents. As the evaluations in section 5.2.2 have shown is the breakpoint within the range of tags promoted by the equation dependent on the number of documents which is in this approach defined to twenty, since the wilderness model has twenty wilderness categories. It is therefore recommended to apply the standard tf-idf equation to a number of around hundred documents, if the requirement of the approach requests for an equalized idf component. Equalized means in this case that the tf-idf values of terms occurring in less than the half of all documents are promoted and values of terms occurring in more than the half of all documents are decreased. In case of this work, the number of documents has been set to 20, which could only have been adapted by recalculating the GIS model and change the number of wilderness categories. But splitting up the spectrum of wilderness into even more than 20 categories would not be realistic, since distinguishing between these twenty classes is already challenging. At this point it can be concluded, that the equation can be applied to an approach based a number of twenty documents but the inequality has to be taken into consideration during the interpretation.

In section 5.2.2, the four characteristics according to which the tags in Table 5.3 were ranked, have been described. Especially the second listed characteristic referring to the idf-value has been described as having too low weight to the adapted tf-idf equation. This characteristic has to be discussed in more detail since two of the three research questions of this work refer to the results of this method.

The goal of the idf-parameter is to promote tags which are very specific to a document, or in this work to a wilderness category. This goal does not seem to be achieved by the tf-idf-uf version, since many of the top-ranked tags can be detected in most of the wilderness categories. An additional approach has been initialized to find out if the results of the tf-idf evaluation would be more adequate, if the weight of the idf-value would be increased. Therefore, the idf-value has been squared by 2 and a new ranked list of tags has been evaluated. The differences were not large but the only difference which could have been clearly detected was that tags which have been generated by very few users like *hauterouteimperiale* or *annualmeeting* increased in rank again, which has been antagonized by respecting the user frequency per tag (uf). Playing around with the different variables and weights of the equation barely helped to optimize its output. Therefore, increasing the weight of idf-value so that only tags occurring in a few or even one category appear in the top ranks of the tf-idf lists would operate against the effect of the uf-value. This evaluation has confirmed the fact that no model or method is perfect and that depending on the dataset still some potential for optimization exists. In order to find the optimal balance of weight for each of the function variables further research has to be done on that. In science, most of the approaches known to the author have kept the tf-idf simple and applied as standard version. This is the reason why also in this work the final version of the tf-idf function has not been further developed and additional weighting has been avoided. Thus, the tf-idf-uf equation does not necessarily return the promised output, since not the most category-specific tags are on top ranks but such that appear in all categories. This had no influences to this work directly but could have one when certain approaches search only for category-specific information.

6.5 Answering the research questions

This section attempts to answer the three stated research questions considering the discussed results from sections 6.1 to 6.3.

6.5.1 Research question 1

Is the user-specific behaviour on a social media platform relevant to the aptitude of user generated content for scientific wilderness research?

The individual user activities and behaviours of Flickr contributors are challenging characteristics for scientific work with Flickr photographs. Evaluations to analyse the influence of these different behaviours to the aptitude of Flickr photos are required when applying them to specific wilderness research. This first research question has been addressed by analysing spatial distributions, user-specific activities in form of prolific user analyses and user-specific behaviours in form of analyses to bulk uploads. The evaluations have revealed that Flickr photographs are a qualified source of information which can also be applied for wilderness research. But especially the influence of bulk uploads has impact to the evaluations and requires an exclusion of biasing data. This has been confirmed by a new initialized approach classifying bulk uploaders according to their uploading behaviour. Prolific users are less influencing, since they basically concentrate their activity to regions with low wilderness quality values, which were of low interest to this work. Thus, the user-specific behaviour has a relevant influence to the aptitude of the data and needs to be handled by excluding biasing data generated in form of bulk uploads. These evaluations have shown that the appliance of Flickr photographs for other approaches concerning the wilderness concept is justified. Especially for decision makers and stakeholders interested in detecting, protecting and preserving wilderness areas like Mountain Wilderness may profit from consulting Flickr photographs to their approaches.

6.5.2 Research question 2

Are variations in wilderness as quantified by a spatially explicit model reflected in the spatial distribution of user generated content?

The variations in wilderness classified by the wilderness GIS-model of Radford et al. (unpublished) has been analysed in Flickr data by two separate approaches. The first approach has collected specific tags representing wilderness and analysed if their spatial distribution corresponds with the classification of the GIS-model. The second approach has classified all photographs with a wilderness quality index, corresponding to their location and the GIS-model wilderness classification. By applying the tf-idf-uf equation, all tags of these photographs have been ranked according to their representativeness for each of the wilderness categories determined by the GIS-model. While the temporal effort for implementation was smaller for the first approach, the second approach returned more qualitative outputs. But both approaches have revealed that the variations of wilderness quantified by the GIS-model are also detectable in Flickr tags. This illustrates that the GIS-based approach applied in this work, which bases on the principle of Steve Carvers wilderness attributes (Carver et al., 2012), can be reflected by the perceptual data of the Flickr community. This information is valuable insofar as further evaluations on the base of Steve Carvers wilderness attributes can be extended by Flickr data to retrieve more perceptual information about the wilderness conditions in the corresponding researched area. An interesting approach would be to consider alternative social media sources and compare the results with the Flickr output. Like that, the most adequate social media data source for wilderness research could be determined and the revealed information could be improved.

6.5.3 Research question 3

Can GIS-based wilderness information be further characterized by consulting user generated content?

The characterization process of this work is based on a classification of the wilderness information evaluated during the process of analysing the variations of wilderness in Flickr tags. The classification revealed four different wilderness classes which could then be further characterized by consulting the ranked tags according to the tf-idf-uf evaluation. Although the characterization process is more interpretative than measured, clear tendencies could have been determined and prepared for argumentation of further wilderness characterization. The benefit from the outputs is an advanced insight into how Flickr community perceives and describes these classified wilderness categories. Although the Flickr community is not representative for the population of Switzerland, tourism activity and preferences can be detected on a regional scale. The observed tendencies of tags indicated differences between the wilderness classes referring to the activities the contributors performed when taking photos, referring to the landscape features they captured but also the approximate surrounding terrain when they took the photographs. Such specific information has been evaluated and gives answers to the third research question.

7 Conclusion

This section concludes the evaluated findings by referring to the discussion and the stated research questions at the beginning of this work. An outlook to future work completes this conclusion.

7.1 Summary

This work represents a first approach combining social media data to the scientific wilderness context, represented by a spatially explicit wilderness model. The general aim of this work is to demonstrate the aptitude of Flickr photograph metadata to the wilderness context and to extend the wilderness information of a GIS-based wilderness model with perceptual information. While the majority of the work has been processed by implemented Java code, also Microsoft Excel and ArcMap (GIS) have been consulted for calculation and visualization purposes. Since social media data has many specific characteristics which are potential sources of bias, the first process step has evaluated the influence of these characteristics to this work in order to prevent the outputs of the following evaluations from bias. The retrieved Flickr photo metadata consist of a variety of different data types whereas this work mainly concentrates on tags and their spatial distribution. The consulted GIS-model spatially classifies Switzerland according to a wilderness quality index, which has been accessed to analyse if comparable variations in wilderness could be detected in the Flickr data. That evaluation has initialized a classification of the Flickr data according to the wilderness concept. Considering this classification, the GIS-model has been further characterized by consulting Flickr tag semantics. Accordingly, the Flickr photo data extends the wilderness information of an already existing wilderness approach with additional perceptual information of the Flickr community, which reveal new insights into how the community perceives and describes their environment.

7.2 What has been achieved?

The output of this work has confirmed that Flickr data is a valuable source of information to combine with wilderness research. The perceptual data generated by the Flickr community is open and free accessible and can be used to address the critics of GIS-based wilderness evaluations. The evaluations in this work have revealed additional perceptual wilderness information compared to the GIS-based model on the one hand, but also specific information about the Flickr community on the other. Tag-based evaluations have indicated important shifts within the semantic of tags which allowed a reclassification of the wilderness continuum. Regions of the GIS-model classified with highest wilderness categories for instance have shown a high tendency for tagging pictures with tags referring to climbing or mountaineering activities. Such information can be used to allocate human-nature interactions and can help to determine the way they interact at these locations. Although local interpretations of the outputs need to be considered carefully due to several characteristics of the Flickr photos and other bias discussed in this work, the distribution of Flickr photos and their assigned tags give valuable information about wilderness conditions. This information can be used by stakeholders and decision makers to increase the understanding of human-nature interaction and to promote further efforts considering the general goal of wilderness research, the protection and preservation of the remaining pristine regions and the search for a balanced coexistence with the environment.

7.3 Outlook and future work

The evaluations in this work have answered some research questions but have also detected many research gaps. This evaluation is a pioneer approach insofar as no other work has directly applied Flickr photograph metadata to the concept of wilderness. Thus, many steps could be improved, such as the prior exclusion of data generated by bulk uploaders or the definition of a higher spatial accuracy restriction for the retrieval process. This work has discussed the lack of scientific determination for methods to deal with biasing social media characteristics. Especially the determination of thresholds classifying data into bulk uploads would have been a large necessity for this work. The initialized approach which addressed that research gap has potential to give answers to the bulk upload problem when further tested and applied to different data sources. Further work has to be done to improve the benefit of social media data, and also to determine restriction for using such sources in research context in general. Much research has been done with UGC but only a few mentioned the risk and distortions of bulk uploads or prolific users. In order to increase the attention of UGC to public and science and to establish social media data in wilderness research, it needs to be presented more attractively. A higher use case would also increase normalization processes so that methodological standards for dealing with social media characteristics could be initialized.

In order to verify the expressiveness of the tag-based analyses, the tf-idf-uf equation needs to be applied to different GIS-models covering different areas. Although the tf-idf equation is an established method in IR research, it has not been used for wilderness research until recently. Thus, additional knowledge and verifications are required on that. To further improve the output of this work, other parts of the Flickr metadata, such as the temporal metadata but also the photo description could be consulted to get additional opportunities for verification.

8 Appendix

Appendix A: The wilderness attributes considered by the MCE of the GIS-model

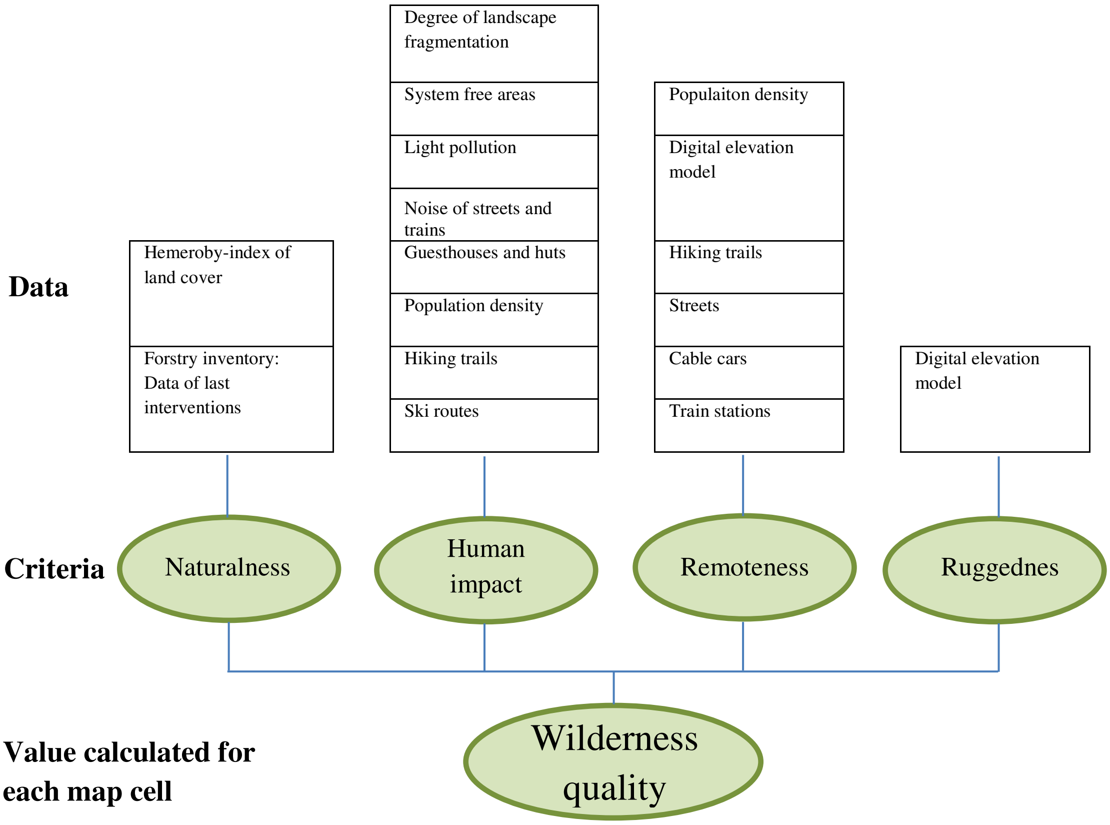


Figure 8.1 – Applied wilderness attributes and their contained data of the GIS-model of WSL. Own translation.

Appendix B: Applied Chi-squared approach

An additional way to the quantile map and to analyse the spatial distribution of Flickr photos compared to the population is to apply a Chi-square test to the two dataset. The two datasets, number of photos per raster-cell and the population per cell can be seen as surfaces while one of them is the expected and the other the observed surface (Antoniou et al., 2010). In order to compare the number of Flickr photos to the population in this work, the first is the observed and the latter the expected surface (Figure 8.2). Also for this evaluation an aggregation of one kilometre was required to avoid dividing by zero errors within the calculation process (Figure 8.3). Since this output is comparable to the quantile map but less obvious tendencies can be detected, this approach is not further discussed but builds an alternative to the quantile approach.

$$chi = \frac{(ObservedValue - ExpectedValue)}{\sqrt{ExpectedValue}}$$

Figure 8.2 – Chi-squared equation (Antoniou, Morley, & Haklay, 2010)

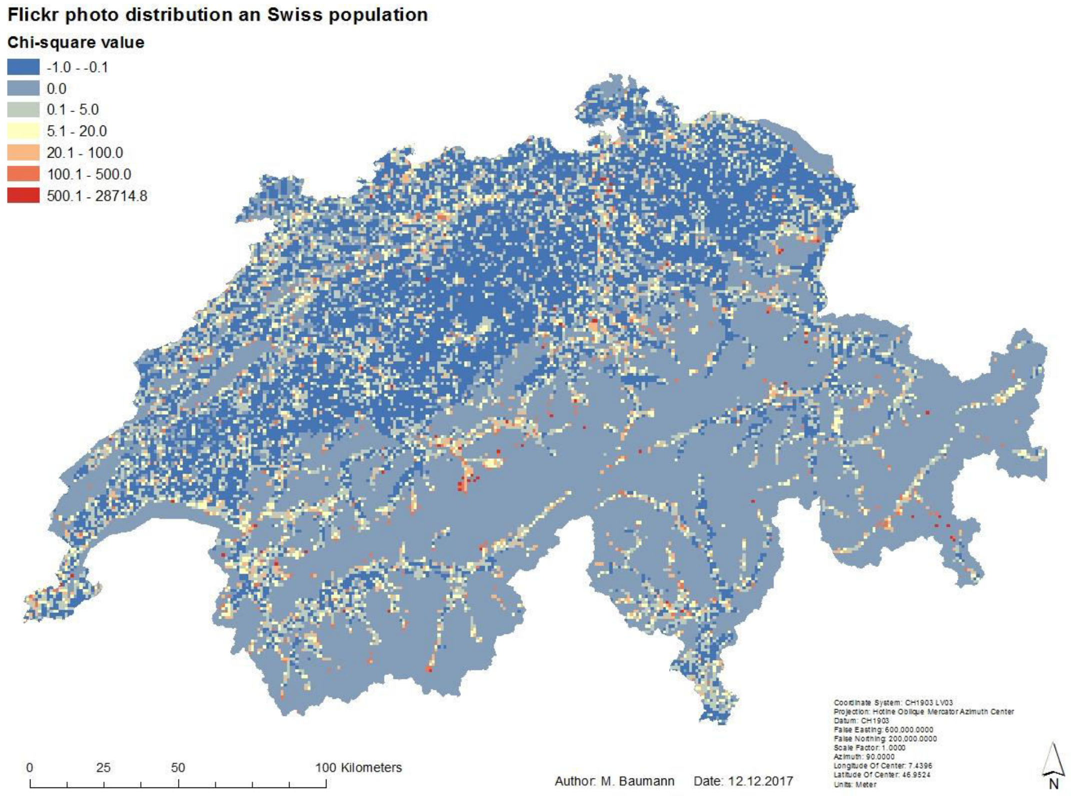

Flickr photo distribution an Swiss population
Chi-square value
-1.0 - -0.1
0.0
0.1 - 5.0
5.1 - 20.0
20.1 - 100.0
100.1 - 500.0
500.1 - 28714.8
0
25
50
100 Kilometers
Author: M. Baumann
Date: 12.12.2017
Coordinate System: CH1903 LV03
Projection: Hotine Oblique Mercator Azimuth Center
Datum: CH1903
False Easting: 600,000.0000
False Northing: 200,000.0000
Scale Factor: 1.0000
Azimuth: 90.0000
Longitude Of Center: 7.4396
Latitude Of Center: 46.9524
Units: Meter
N

Figure 8.3 – Chi-squared map of the Flickr photo density and the population.

Appendix C: Excluded tags from tag-based analyses

The following list shows all tags explicitly how they have been excluded from the tag-based evaluation of this work. As in Hollenstein & Purves are the toponyms divided into spatial scales in order to simplify and structure the order of the tags (Hollenstein & Purves, 2010).

Toponyms

Country-scale

switzerland, schweiz, suisse, swiss, svizzera, suiza, šveits, ch, italy, italien, italia, europe

Regional scale

alps, alpen, alpi, swissalps, tessin, ticino, valais, wallis, cantonduvalais, walliseralpen, graubünden, kantongraubünden, grison, grisons, waadtländerjura, vaud, jura, tessinswitzerland, berneroberland, berneseoberland, bernervoralpen, nidwalde, aargau, engadin, uri, oberland, appenzell, glarus

City-scale

zurich, zuerich, zürich, zurigo, zrh, geneva, genf, geneve, genève, ginevra, gva, basel, bâle, bern, berne, berner, bernese, schaffhausen, payerne, fribourg, lugano, luzern, liuzerna, lucerno, montreux, romont, therwil, bleienbach, oberembrach, altbüron, uster, zermatt, chamonixzermatt, heidadorf, neuhasuen, frutt, interlaken, meiringen, brienz, kronten, vevey, winterthur, neuchâtel, bellinzona, davos, lauterbrunnen, kloten, verbier, neuhausen, kandersteg

Local scale

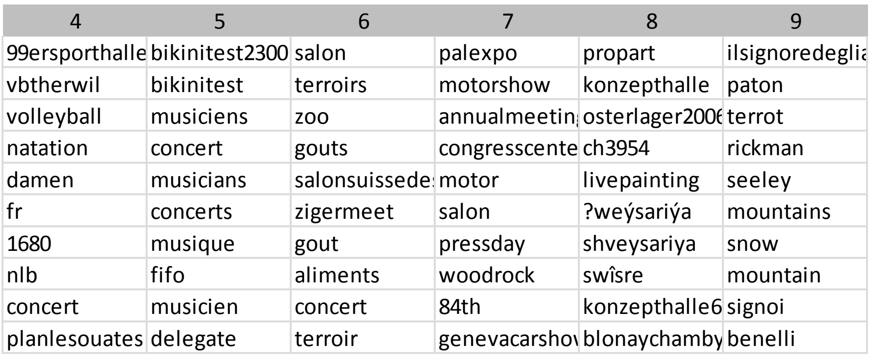
zurichairport, uetliberg, pilato, pilatus, fractusmons, chuenisberg, kruezboden, greiffensee, rheinfall, rhinefalls, furtschella, rigi, schatzalp, stelviopass, chrützlipass, fürka, jungfraujoch, jungfrau, jungfrauregion, flumserberg, axalphorn, aletschglacier, poschiavo, porshiavo, ebenfluh, rhonegletschersee, ammertenspitz, vasevay, levasevay, oberstockenalp, stockenflue, hinterstockensee, oberstockensee, jägihorn, mieschflue, bunderalp, tierbergli, rothorn, rothornli, dentdemorcles, morcles, lenzspitze, eselgrat, weissmeis, monterosa, üssersbarrhorn, martinsloch, grünbergpass, breithorn, taschhorn, grandelui, wesenalp, eiger, eigerwand, dürrenhorn, matterhorn, cervino, grindelwald, mittelegi, liongrat, brienzrothorn, zwächten, firnalpeli, schwarzseeblinnenhorn, dossenweg, grassen, zinalrothorn, bernesealsp, pizgloria, scialp, möschelenspitz, summitwetterhorn, creuxduvan, schynigeplatte, grimsel, aletsch, aletschgletscher, alpstein, sustenpass, arolla, schilthorn, canon, titlis, engelberg, brissago, isoledibrissago, morteratsch, rhonegletscher, bernina, saasfee, mönch, monch, dom, nadelhorn, dentsdumidi, kleinescheidegg, rhinefalls, rheinfall, rhine, rhein, aare, verzasca, reuss, maggia, triftbrücke,

Other tags

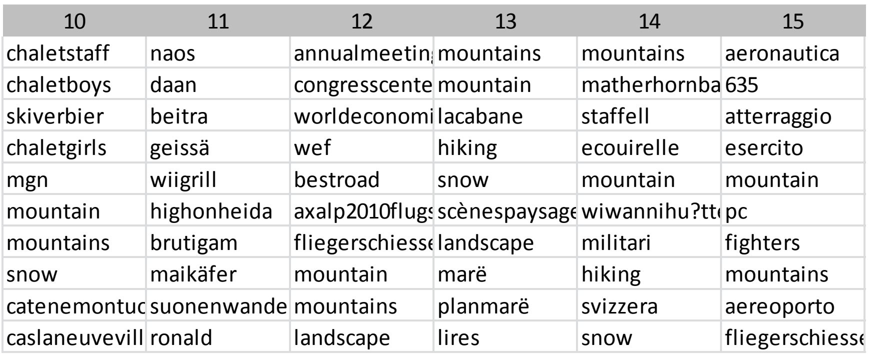
nikon, ??, -, am2013, sony, sony70400mm, ????????????????, ?????????, geotagged, approximatelygeotagged, digitalekompaktkamera, 200605, stillimage, instagramapp, uploaded:by=instagram, uploaded:by=flickrmobile, flickriosapp:filter=nofilter, iphoneography, square, squareformat, iflickr, d80, zzid24

Appendix D: Top ten ranks of standard tf-idf lists

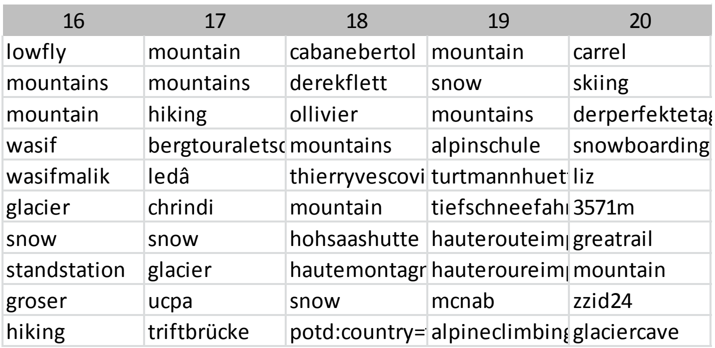
The ranked lists of the standard version of the tf-idf equation look completely different to the one of the adapted version (Table 8.1).



4	5	6	7	8	9
99ersporthalle	bikinitest2300	salon	palexpo	propart	ilsignoredeglia
vbtherwil	bikinitest	terroirs	motorshow	konzepthalle	paton
volleyball	musiciens	zoo	annualmeetin	osterlager200	terrot
natation	concert	gouts	congresscente	ch3954	rickman
damen	musicians	salonsuissede	motor	livepainting	seeley
fr	concerts	zigermeet	salon	?weýsariýa	mountains
1680	musique	gout	pressday	shveysariya	snow
nlb	fifo	aliments	woodrock	swîsre	mountain
concert	musicien	concert	84th	konzepthalle6	signoi
planlesouates	delegate	terroir	genevacarsho	blonaychamby	benelli



10	11	12	13	14	15
chaletstaff	naos	annualmeetin	mountains	mountains	aeronautica
chaletboys	daan	congresscente	mountain	matherhornba	635
skiverbier	beitra	worldeconomi	lacabane	staffell	atterraggio
chaletgirls	geissä	wef	hiking	ecouirelle	esercito
mgn	wiigrill	bestroad	snow	mountain	mountain
mountain	highonheida	axalp2010flugs	scènespaysage	wiwannihu?tt	pc
mountains	brutigam	fliegerschiesse	landscape	militari	fighters
snow	maikäfer	mountain	marë	hiking	mountains
catenemontuc	suonenwande	mountains	planmarë	svizzera	aereoporto
caslaneuvevill	ronald	landscape	lires	snow	fliegerschiesse



16	17	18	19	20
lowfly	mountain	cabanebertol	mountain	carrel
mountains	mountains	derekflett	snow	skiing
mountain	hiking	ollivier	mountains	derperfektetag
wasif	bergtouraletsc	mountains	alpinschule	snowboarding
wasifmalik	ledâ	thierryvescovi	turtmannhuet	liz
glacier	chrindi	mountain	tiefschneefahr	3571m
snow	snow	hohsaashutte	hauterouteimp	greatrail
standstation	glacier	hautemontagn	hauteroureimp	mountain
groser	ucpa	snow	mcnab	zzid24
hiking	triftbrücke	potd:country=	alpineclimbing	glaciercave

Table 8.1 – Ranked top 10 tags of the standard tf-idf evaluation

9 References

Abbasi, R., Chernov, S., Nejdl, W., Paiu, R., & Staab, S. (2009). Exploiting flickr tags and groups for finding landmark photos. *Advances in Information Retrieval*, 654-661.

Antoniou, V., Morley, J., & Haklay, M. (2010). Web 2.0 geotagged photos: Assessing the spatial dimension of the phenomenon. *Geomatica, 64*(1), 99-110.

Aplet, G., Thomson, J., & Wilbert, M. (2000). Indicators of wildness: Using attributes of the land to assess the context of wilderness. *USDA Forest Service Proceedings, 2*, 89-98.

Bauer, N. (2005). *Für und wider Wildnis: Soziale Dimensionen einer aktuellen gesellschaftlichen Debatte*: Haupt.

Bauer, N., Wallner, A., & Hunziker, M. (2009). The change of European landscapes: human-nature relationships, public attitudes towards rewilding, and the implications for landscape management in Switzerland. *Journal of environmental management, 90*(9), 2910-2920.

Boller, F., Hunziker, M., Conedera, M., Elsasser, H., & Krebs, P. (2010). Fascinating remoteness: The dilemma of hiking tourism development in peripheral mountain areas: Results of a case study in southern Switzerland. *Mountain Research and Development, 30*(4), 320-331.

Brunner, T. J., & Purves, R. S. (2008). *Spatial autocorrelation and toponym ambiguity.* Paper presented at the 2nd international workshop on Geographic Information Retrieval.

Carver, S., Comber, A., McMorran, R., & Nutter, S. (2012). A GIS model for mapping spatial patterns and distribution of wild land in Scotland. *Landscape and Urban Planning, 104*(3-4), 395-409. doi:10.1016/j.landurbplan.2011.11.016

Carver, S., Evans, A. J., & Fritz, S. (2002). Wilderness attribute mapping in the United Kingdom. *International Journal of Wilderness, 8*(1), 24-29.

Carver, S., & Fritz, S. (1995). Mapping the wilderness continuum. *Proceedings of the GIS Research UK*, 15.

Carver, S., Tricker, J., & Landres, P. (2013). Keeping it wild: mapping wilderness character in the United States. *J Environ Manage, 131*, 239-255. doi:10.1016/j.jenvman.2013.08.046

Coeterier, J. F. (1996). Dominant attributes in the perception and evaluation of the Dutch landscape. *Landscape and Urban Planning, 34*(1), 27-44.

Cordell, H., Tarrant, M. A., & Green, G. T. (2003). Is the public viewpoint of wilderness shifting. *IJW, 9*(2), 27-32.

Da Rugna, J., Chareyron, G., & Branchet, B. (2012). *Tourist behavior analysis through geotagged photographies: a method to identify the country of origin.* Paper presented at the Computational Intelligence and Informatics (CINTI), 2012 IEEE 13th International Symposium on.

Di Minin, E., Tenkanen, H., Hausmann, A., Heikinheimo, V., Järv, O., & Toivonen, T. (2016). Social media data for analysing spatio-temporal patterns and nature-based preferences of people in national parks 1-2.

European Environment Agency (EEA). (2010). *Europe's ecological backbone: recognising the true value of our mountains* (6). Retrieved from Copenhagen, Danmark: http://www.eea.europa.eu/publications/europes-ecological-backbone

Fritz, S., & Carver, S. (1998). Accessibility as an important wilderness indicator: modelling Naismith’ s rule. *GISRUK’ 98*.

Fritz, S., Carver, S., & See, L. (2000). New GIS approaches to wild land mapping in Europe. *USDA Forest Service Proceedings, 2*, 120-127.

Gliozzo, G., Pettorelli, N., & Haklay, M. (2016). Using crowdsourced imagery to detect cultural ecosystem services: a case study in South Wales, UK. *Ecology and Society, 21*(3). doi:10.5751/es-08436-210306

Gomez-Pompa, A., & Kaus, A. (1992). Taming the wilderness myth. *BioScience, 42*(4), 271-279.

Goodchild, M. F., & Glennon, J. A. (2010). Crowdsourcing geographic information for disaster response: a research frontier. *International Journal of Digital Earth, 3*(3), 231-241. doi:10.1080/17538941003759255

Gschwend, C., & Purves, R. S. (2012). Exploring geomorphometry through User Generated Content: Comparing an unsupervised geomorphometric classification with terms attached to georeferenced images in Great Britain. *Transactions in GIS, 16*(4), 499-522.

Habron, A. D. (1998a). *Defining wild land in Scotland through GIS based wilderness perception mapping.* (Doctor of philosophy), University of Stirling,

Habron, A. D. (1998b). Visual perception of wild land in Scotland. *Landscape and Urban Planning, 42*(1), 45-56. doi:10.1016/S0169-2046(98)00069-3

Harris, P. G. (2006). Environmental perspectives and behavior in China: Synopsis and bibliography. *Environment and Behavior, 38*(1), 5-21.

Hausmann, A., Toivonen, T., Slotow, R., Tenkanen, H., Moilanen, A., Heikinheimo, V., & Di Minin, E. (2017). Social media data can be used to understand tourists’ preferences for nature-based experiences in protected areas. *Conservation Letters*. doi:10.1111/conl.12343

Hochmair, H. H., & Zielstra, D. (2012). *Positional accuracy of Flickr and Panoramio images in Europe.* Paper presented at the Proceedings of the Geoinformatics Forum, Salzburg, Austria.

Hollenstein, L., & Purves, R. S. (2010). Exploring place through user-generated content: Using Flickr tags to describe city cores. *Journal of Spatial Information Science, 2010*(1), 21-48.

Huang, H. (2016). Context-aware location recommendation using geotagged photos in social media. *ISPRS International Journal of Geo-Information, 5*(12). doi:10.3390/ijgi5110195

Humayun, M. I., & Schwering, A. (2012). *Representing vague places: Determining a suitable method.* Paper presented at the Proceedings of the international workshop on place-related knowledge acquisition research (P-KAR 2012), Monastery Seeon, Germany.

Jones, C. B., Purves, R. S., Clough, P. D., & Joho, H. (2008). Modelling vague places with knowledge from the Web. *International Journal of Geographical Information Science, 22*(10), 1045-1065.

Kliskey, A. D., & Kearsley, G. W. (1993). Mapping multiple perceptions of wilderness in southern New Zealand. *Applied Geography, 13*(3), 203-223.

Lesslie, R. G., Mackey, B. G., & Preece, K. M. (1988). A computer-based method of wilderness evaluation. *Environmental Conservation, 15*(3), 225-232. doi:10.1017/s0376892900029362

Lesslie, R. G., & Taylor, S. G. (1985). The wilderness continuum concept and its implications for Australian wilderness preservation policy. *Biological Conservation, 32*(4), 309-333.

Lindemann-Matthies, P., Keller, D., Li, X., & Schmid, B. (2014). Attitudes toward forest diversity and forest ecosystem services–a cross-cultural comparison between China and Switzerland. *Journal of Plant Ecology, 7*(1), 1-9. doi:10.1093/jpe/rtt015

Lupp, G., Höchtl, F., & Wende, W. (2011). “Wilderness” – A designation for central European landscapes? *Land Use Policy, 28*(3), 594-603. doi:10.1016/j.landusepol.2010.11.008

Manning, D. C., Raghavan, P., & Schütze, H. (2008). *Introduction to Information retrieval*: Cambridge University Press.

Marshall, R. (1930). The problem of the wilderness. *Scientific Monthly, 30*(2), 141-148.

Montello, D. R., Goodchild, M. F., Gottsegen, J., & Fohl, P. (2003). Where's downtown?: Behavioral methods for determining referents of vague spatial queries. *Spatial Cognition & Computation, 3*(2-3), 185-204.

Muir, J. (1898). The wild parks and forest reservations of the West. *Atlantic monthly, 81*(483), 15-28.

Naaman, M., Song, Y. J., Paepcke, A., & Garcia-Molina, H. (2004). *Automatic organization for digital photographs with geographic coordinates.* Paper presented at the Digital Libraries, 2004. Proceedings of the 2004 Joint ACM/IEEE Conference.

Nash, R. F. (2014). *Wilderness and the American mind*: Yale University Press.

Nielsen, J. (2006). Participation inequality: Encouraging more users to contribute. Retrieved from https://www.nngroup.com/articles/participation-inequality/

ÖBF und WWF – Österreichische Bundesforste and World Wide Found For Nature. (2012). *Wildnis in Österreich?: Herausforderungen für Gesellschaft, Naturschutz und Naturraummanagement in Zeiten des Klimawandels*. Purkensdorf: Österreichische Bundesforste AG (ÖBf AG).

Purves, R. S., & Derungs, C. (2015). From space to place: Place-based explorations of text. *International Journal of Humanities and Arts Computing, 9*(1), 74-94.

Purves, R. S., Edwardes, A., & Wood, J. (2011). Describing place through user generated content. *First Monday, 16*(9).

Purves, R. S., & Mackaness, W. A. (2016). A methodological toolbox for exploring collections of textually annotated georeferenced photographs. *European Handbook of Crowdsourced Geographic Information*, 145.

Radford, S., Senn, J., & Kienast, F. Assessing wilderness quality in highly modified, structured landscapes: a quantitative assessment in Switzerland. *(unpublished)*.

Rattenbury, T., Good, N., & Naaman, M. (2007). *Towards automatic extraction of event and place semantics from flickr tags.* Paper presented at the Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval.

Rattenbury, T., & Naaman, M. (2009). Methods for extracting place semantics from Flickr tags. *ACM Transactions on the Web (TWEB), 3*(1), 1.

Schmitz, P. (2006). *Inducing ontology from flickr tags.* Paper presented at the Collaborative Web Tagging Workshop at WWW2006, Edinburgh, Scotland.

Smith, D., & Crane, G. (2001). Disambiguating geographic names in a historical digital library. *Research and Advanced Technology for Digital Libraries*, 127-136.

Stremlow, M., & Sidler, C. (2002). *Schreibzüge durch die Wildnis: Wildnisvorstellungen in Literatur und Printmedien der Schweiz* (Vol. 8): Haupt.

Swanwick, C. (2009). Society's attitudes to and preferences for land and landscape. *Land Use Policy, 26*, 62-75. doi:10.1016/j.landusepol.2009.08.025

Tenerelli, P., Demšar, U., & Luque, S. (2016). Crowdsourcing indicators for cultural ecosystem services: A geographically weighted approach for mountain landscapes. *Ecological Indicators, 64*, 237-248. doi:10.1016/j.ecolind.2015.12.042

Tims, W. (2014). *New approaches for wilderness perception mapping: A case study from Vatnajökull National Park, Iceland.* (Magister Scientiarum degree in Geo-information Science and Earth Observation for Environmental Modelling and Management), University of Iceland, Reykjavik.

Tobler, W. R. (1970). A computer movie simulating urban growth in the Detroit region. *Economic geography, 46*(1), 234-240.

van Zanten, B. T., Van Berkel, D. B., Meentemeyer, R. K., Smith, J. W., Tieskens, K. F., & Verburg, P. H. (2016a). Continental-scale quantification of landscape values using social media data. *Proc Natl Acad Sci U S A, 113*(46), 12974-12979. doi:10.1073/pnas.1614158113

van Zanten, B. T., Zasada, I., Koetse, M. J., Ungaro, F., Häfner, K., & Verburg, P. H. (2016b). A comparative approach to assess the contribution of landscape features to aesthetic and recreational values in agricultural landscapes. *Ecosystem Services, 17*, 87-98. doi:10.1016/j.ecoser.2015.11.011

Wild Europe. (2012). A working definition of European wilderness and wild areas. *Wild Europe Initiative*, 18.

Wood, S. A., Guerry, A. D., Silver, J. M., & Lacayo, M. (2013). Using social media to quantify nature-based tourism and recreation. *Sci Rep, 3*, 2976. doi:10.1038/srep02976

Zielstra, D., & Hochmair, H. H. (2013). Positional accuracy analysis of Flickr and Panoramio images for selected world regions. *Journal of Spatial Science, 58*(2), 251-273.

Personal declaration

I hereby declare that the submitted thesis is the result of my own, independent work.
All external sources are explicitly acknowledged in the thesis.

January 25, 2018

Markus Baumann