



UNIWERSYTET IM. ADAMA MICKIEWICZA  
WYDZIAŁ NAUK GEOGRAFICZNYCH I  
GEOLOGICZNYCH  
INSTYTU GEOEKOLOGII I GEOINFORMACJI

Kierunek: Geografia  
Specjalność : Geoinformacja

**Przemysław Dusza**

***Analysis of visitors behavior patterns based on GPS  
tracks from Müstair Valley, Switzerland***

***Analiza ścieżek obrazujących sposoby  
zachowania się turystów w oparciu o dane GPS  
z szwajcarskiej doliny Müstair***

Praca magisterska napisana w Zakładzie  
Geoekologii pod kierunkiem naukowym  
prof. UAM dr hab. Zbigniewa Zwolińskiego

**Poznań 2011**

*Dziękuję Panu Profesorowi Zbigniewowi  
Zwolińskiemu za naukową opiekę i pomoc  
w realizowaniu pracy.*

*I would like thank Professor Reto Rupf  
for giving me a chance to work in an interesting  
project and for many hours spent on Skype.*

**RODZICOM**

## Table of contents

1.	INTRODUCTION .....	6
1.1.	Motivation.....	6
1.2.	Thesis organization.....	7
2.	OVERVIEW OF PREVIOUS WORKS ON VISITORS MONITORING AND MANAGEMENT .....	9
2.1.	Background of visitors monitoring and management.....	9
2.2.	Specification of monitoring and management project.....	9
2.2.1	Monitoring techniques .....	11
2.2.2.	Computer simulation as a new tool for monitoring and management.....	14
2.3.	Project Mafreina – Management-Toolkit und Freizeit Natur .....	16
2.3.1.	Introduction.....	16
2.3.2.	Research area .....	17
2.3.3.	Project goals.....	18
2.3.4.	Materials and methods .....	18
2.3.5.	Summary of previous research .....	20
3.	PLACEMENT IN THE RESEARCH PROJECT.....	21
3.1.	Geographic information system.....	21
3.2.	Time and spatial analyses of the GPS data.....	22
3.2.1.	Spatial data mining .....	23
3.2.2.	Map matching .....	24
3.2.3.	Fuzzy logic and MCE .....	28
3.2.4.	Geovisualisation .....	29
3.3.	Overview of technical details .....	29
3.3.1.	ArcGIS .....	29
3.3.2.	Python .....	30
4.	DATA PREPARATION.....	32
4.1.	Data description .....	32
4.1.1.	Number of satellites.....	34
4.1.2.	HDOP .....	35
4.1.3.	Distance from previous point.....	37
4.1.4.	Speed .....	37
4.1.5.	Metadata .....	37
4.1.6.	Landcover .....	38
4.1.7.	Trails 25 .....	39
4.1.8.	Trails .....	39

4.2.	Data preparation and selection.....	40
4.3.	Summary of data preparation and selection.....	50
5.	MODELING AND ANALYSIS OF GPS DATA.....	52
5.1.	Visual analysis of the data .....	52
5.2.	Spatial data mining .....	56
5.3.	Multi-Criteria Evaluation – Boolean and Weighted Linear Combination approach.....	63
5.4.	Fuzzy logic analysis.....	65
5.4.1.	Fuzzy distance .....	66
5.4.2.	Fuzzy speed .....	68
5.4.3.	Fuzzy HDOP.....	70
5.5.	Weighted Linear Combination .....	72
5.6.	Modeling of various scenarios.....	73
6.	RESULTS .....	76
6.1.	Statistical summary of the results .....	76
6.2.	Movement patterns and trends.....	79
6.3.	Points of interest .....	82
6.4.	Summary of visitors behaviours patterns in Müstair Valley .....	84
7.	DISCUSSION.....	87
8.	SUMMARY AND OUTLOOK.....	92
9.	REFERENCES .....	93

# 1. Introduction

## 1.1. Motivation

In today's world the interactions between humans and nature are increasing very rapidly. The number of people travelling all around the world in search of the most remote and wild places is growing every day. This process will probably last for decades and it will be even more visible. The need to discover and explore is very strong in each of us but how can those needs of humans be reconcile with the needs of the nature. Is there a way to better understand and manage the interactions between humans and nature?

In recent years numerous approaches have been presented how to monitor and manage visitors in various environments. All of those approaches are trying to give an answer how and why do the people interact with the nature. An answer to those questions can help to better understand if the people have a big influence on the nature. It may also give information how managers can effectively and confidentially manage tourism so that it eventually has only slight impact on the environment.

Fast development of technology plays a key role in helping to increase the quality of visitors monitoring techniques. Global Positioning System (GPS) connected with modern devices, creates tools that allow to even profounder understand tourists behavior. In the last decade numerous visitors monitoring and management projects, tried to cooperate with the tourists in order to analyze and create very precise visitors behavior patterns.

In the Swiss Müstair Valley a new project Mafreina was brought to life to present and better understand how the visitors behave in protected areas. The biggest challenge for the creators of this project is to create on the basis of GPS tracks, clear and comprehensible methods of aggregating and analyzing data. The main difference between this project and numerous others is that the visitors in the Müstair Valley are allowed to move all around this region. They are not forced to stay on trails, which mean that they can do the things that they want, only relying on their needs.

This factor creates new problems that need to be overcome. Due to the high complexity of the visitors behavior patterns precise algorithms need to be created. They need to eventually show how do the tourists behave in the whole region, where do they leave the trails and where do they come back. They also need to give information what could possibly attract the tourists to leave the trail and can some clear trends in the

tourists behavior be determined. The creation of those algorithms and further visualization of the results on the maps will be a solid foundation in drawing the ultimate conclusions of the project Mafreina.

## **1.2. Thesis organization**

The second chapter of master's thesis describes earlier visitors monitoring and management projects with an emphasis on new monitoring techniques. The description focuses on advantages and disadvantages of those projects and how can project Mafreina benefit from their results. This chapter also presents the specification, methods and materials used in the project.

The third chapter focuses on the placement of my master's thesis in the research project. Detailed description of the problems, methods from various literature is presented in order to explain the essence of the master's thesis. One of the most important parts of this chapter is "Map Matching" methods, which is one of the key elements of my master's thesis. This chapter also guides through the technical aspects such as software and programming language "Python".

Next chapter presents data preparation and the description of this process. This chapter helps to comprehend how a proper data preparation influences the speed and the quality of work. It also gives a clear overview of various data types and methods, used before implementing the data into further analysis. Additionally complex scripts for data preparation and selection and their results are presented in this chapter.

Chapter number five focuses on the data analysis. Statistical and visual analysis of the data is supposed to explain the complexity of the research problem and suggest possible solutions. This chapter describes different methods of multi criteria evaluation and eventually presents the concept of fuzzy logic. The final subsections of this chapters show different scenarios of data modeling and their results on the map.

In the chapter number six results are summarised to show different visitors behaviour patterns, unusual trends, point of interest etc. Additionally statistical results are described in order to compare them with the visual results. Eventually the drawbacks of the results are underlined and all results are summarised.

In the last chapter's whole research project as well as each step of my master's thesis is discussed and compared with the results from other projects. Each part of the master's thesis is subjected to discussion in order to emphasize their shortcomings as well

as advantages. This discussion is an essential part of the final conclusions that will help the managers of project Mafreina, to answer the most important questions formulated at the very beginning of the project.



## **2. Overview of previous works on visitors monitoring and management**

### **2.1. Background of visitors monitoring and management**

Monitoring of vegetation and wildlife in recreational and protected areas has a long tradition. In particular national parks and recreation areas the scientific interest in creating inventories and in observing the development of environment has often been a driving force for the establishment of monitoring schemes.

According to Cessford and Muhar (2003) the visitors monitoring does not have a long and well-established research tradition. However the managers tend to realize that only complex information about visitors will help the park management agencies to comprehensively understand the visitors impacts.

In many countries systematic long-term research programs are seen as a part of duty of a national park services. Opposed to that, a continuous monitoring of recreational uses and visitor flows is rarely carried out. This is particularly true for the situation in most European countries, where visitor monitoring, if at all done, is usually organized on an ad-hoc basis without complex planning. Very often, results from improvised one-day countings are being extrapolated and used for management decision without consideration of the significance of the results (Arnberger et al. 2002)

This way of monitoring and management will never bring satisfying results. The managers however begin to understand that only continuous monitoring can help them to answer many questions, which are presented in the next chapter.

### **2.2. Specification of monitoring and management project**

Before a continuous monitoring of recreational use and visitors flows is carried out, numerous issues need to be discussed and prepared. Only well prepared visitors monitoring and management projects can bring satisfying results and draw appropriate conclusions. According to Muhar et al. (2002) at the very beginning of each project following questions need to be answered:

- Why should be monitored?
- What should be monitored?

- Who should be monitored?
- Where should be monitored?
- When should be monitored?

These questions are supposed to create a solid frame-work which will guide through the whole process.

First question, why should it be monitored is probably the most important one. The goal of monitoring process has to be clearly defined. There can be numerous goals like minimizing the conflicts between nature and humans, specifying problems in some protected areas, collecting comprehensible data for planning decision such as allocation of infrastructure and services. Every such goal needs a different monitoring scheme in order to understand the basics of the problem.

According to Cessford and Muhar (2003) there are five main reasons why a monitoring process should be carried out:

- for operational auditing of performance measures and budgets,
- to find out the condition of specific natural, historic and cultural heritage features and processes of conservation priority and related sustainability issues,
- to know the visitors numbers, their behavior patterns and their characteristics,
- physical impacts – visitor effects on natural, historic and cultural heritage features and processes,
- social impacts – visitor conflicts and satisfaction with the quality of their recreation experiences,

Next question is what should be monitored. Basing on the definition of the monitoring scheme goals expected measurements can be defined:

- number of visitors
- visitors load
- visitors flow (e.g. persons/hour/direction)
- visitors density (e.g. length/units of trails)
- visitors points of interest
- visitors' activities etc.

Additionally to this question some external factors should also be registered. Only the numbers of visitors will not be able to help to understand the exact situation in the research area. External factors as the weather condition, tourist infrastructure, points of interest or even holidays also have to be taken under consideration.

Question regarding who should be monitored helps to determine whether some people can be identified as visitors or not. Not every person encountered in the park or recreation area is a visitor. The typical motives of a visitor are outdoor recreation or cultural appreciation (Hornback & Eagles 1999). Therefore park workers, forest workers or farmers should not be considered as visitors. They should not be included in the statistics but this kind of distinction is only possible in remote areas. In urban areas it is not possible to determine the motives of the people entering a park or recreational area.

The next question is where should be monitored. According to Cessford and Muhar (2003) very often monitoring is carried out at the entrance points of parks or visitors centers. There are also other locations which are often visited so called points of interest. Monitoring can also be carried out in places where counting devices can be easily installed. In order to estimate interactions between humans and environment, monitoring devices need to be placed all around the park especially in its core. In the European context, the most typical situation is an open trail or road network with multiple entrance points. This is particularly the case in urban forests. In such situations, numerous pre-tests are important while they need to determine the most significant nodes in the trail network for the placement of counting stations (Arnberger et al. 2002).

The last question is when should be monitored. The best solution is to carry out a long-term visitors monitoring project but that happens very rarely. The most frequent types of counting activities are single-day countings. Very often, expected peak visitation days (e.g. Sundays in early summer) are selected for counting campaigns and the results from these days are then being used to alarm the public because of excessive use-levels (Arnberger et al. 2002).

From numerous monitoring projects both in urban and in remote locations the managers draw one more significant conclusion. For a better understanding of the dynamics of recreational uses it is essential to have data, which covers all seasons and concerns many additional external factors such as weather, daytime etc. Only this kind of complex methodology will give accurate and desirable results.

### **2.2.1 Monitoring techniques**

Nowadays numerous techniques are available for the monitoring of visitors flows in recreational areas. The question which technique is the best depends on the character of the monitoring project. For example when a single-day countings need to be carried out, it

will be unnecessary to buy expensive monitoring equipment. It is also important whether a large area or a small city park needs to be monitored. The choice of monitoring technique is strongly connected with the five questions discussed in the previous chapter.

According to Arnberger et al. (2002) monitoring techniques can be divided into three groups: interviews, direct observations and indirect observations. However Cessford and Muhar (2005) say that monitoring techniques fall into four groups: direct observations, on-site counters, visit registrations and infrared counts. These classifications are slightly different but the exact methods are nearly the same in both examples. In each classification oral and written interviews are still an integral part of visitor monitoring concepts. Their main advantage is that they provide mainly qualitative data which combined with quantitative data can provide interesting results. Another advantage of interviews is that they inform what were the needs and motivation of visitors, their activities and routes within the research area.

Direct observations can be divided into two groups: roaming observers and fixed counting stations. Very often in national parks rangers count the visitors that they meet. This is a much desired data especially when they concern remote areas. Although this information cannot be the main source from which conclusions can be drawn, it can serve as an important additional information. Fixed counting stations are mainly created for short-term monitoring concepts but souvenir shops or information booths can also be integrated into a long-term project.

The last group proposed by Arnberger et al. (2002) is indirect observation which is the biggest of the three groups. Indirect observations can use automatic cameras or time-lapse videos. Most of those devices are located at popular trails and can take a picture every 5 seconds. Advantage of this kind of recording is that not only quantitative but also qualitative data is gathered. From the photos or videos not only the number of visitors but also their gender, mode of transport, direction of movement etc. can be seen. The main problem connected with this kind of observation is infrastructure that needs to be built. Wireless connection, energy supply and cost of maintenance make this an expensive choice. Aerial and satellite imagery are also good ways to gather information. They are mainly used for the detection of visitors in open areas such as beaches, lakes, grassland or roads. When it comes to tracking visitors following trails, aerial imagery is not the best solution. Another method of indirect observation is counting access permits or tickets. This method is however only possible when a visitor needs to buy a ticket while entering a park or recreational area.

The development of technique provides numerous new ways of indirect observations. Turnstiles, photoelectric counters, pressure sensitive devices or inductive loop sensors are becoming very popular as the need of knowing the visitors behaviour increases. For example photoelectric counters like light barriers or active or passive infrared sensors, linked with data loggers can provide very useful data. The main challenge for all counting devices is the calibration of the counting station which is site-specific. Sometimes wild animals, big groups of visitors can be wrongly recorded. A very big disadvantage of those devices is that most of them do not show the visitors direction, yet in some regions like in the Swiss National Park this problem has been overcome. Special acoustic slabs sensors which consist of two pressure sensitive slabs register not only the number of visitors but also their direction.

The rest of the methods are based on the self-registration e.g. trail registers, summit books, hut or campground registers. There are also methods which map of traces of use like garbage, trail deterioration, damage to vegetation or footprints and sandbeds. It is clear that the probability that a big group of tourists will leave more garbage is higher rather than a small group. Still this kind of correlation will definitely not help to draw expected conclusions.

Cessford and Muhar (2003) created a table in which they divided the most popular monitoring techniques and defined what data they can normally store. Knowing these techniques and devices a good manager should adjust them to his monitoring scheme. It is advisable to mix various methods as most of them have disadvantages which can be only compensated by other methods. This way a wide range of data can be assembled and they can be used to crosscheck each other

**Table 1: Coverage capacities of the different monitoring methods. A tick “✓” is a direct yes, this method can collect that data, “?” is an indirect yes and “-“ means no, source: Cessford & Muhar, 2003**

Count methods	Visitor no's	Date & time	Travel direction	Route taken	Spatial distribution	Group size	Visitor features	Visitor behaviour
Observations								
- Roaming observers	?	✓	✓	?	?	✓	✓	✓
- Fixed observers	✓	✓	✓	?	?	✓	✓	✓
- Video recordings	✓	✓	✓	-	?	✓	✓	✓
- Time-lapse photo/video	✓	✓	✓	-	?	✓	?	?
- Aerial/satellite imagery	?	✓	-	-	?	?	-	-
On-site count devices								
- Mechanical	✓	?	?	-	?	?	-	-
- Pressure	✓	?	?	-	?	?	-	-
- Seismic/vibration	✓	?	?	-	?	?	-	-
- Active light beam	✓	?	?	-	?	?	-	-
- Passive IR sensor	✓	?	?	-	?	?	-	-
- Magnetic field	✓	?	?	-	?	?	-	-
- Microwave beam	✓	?	?	-	?	?	-	-
Visit registrations								
- Voluntary registers	?	?	?	?	?	?	?	-
- Compulsory registers	✓	✓	?	?	?	?	?	-
- Permits/bookings/fees	✓	?	?	?	?	?	?	-
Inferred counts								
- Indicative counts	?	?	?	?	?	?	?	?
- Interview counts	?	✓	✓	✓	?	✓	✓	✓

## 2.2.2. Computer simulation as a new tool for monitoring and management

Traditional monitoring techniques as well as those more up-to-date deliver a lot of information which is very helpful to managers. Thus they are able to create new adequate management and monitoring systems in order to implement sustainable tourism in the protected areas. The challenge is to protect the natural and cultural resources of these areas and the quality of visitors' experiences in the face of increasing use. Unfortunately, the traditional monitoring techniques do not seem to keep the pace with the rapid changes in the tourism. A good solution to this problem, suggested by many researchers, is computer-based simulation modelling which can facilitate the planning and management of the nature-based tourism.

The following characteristics of computer-based simulation modelling represented by (Daniel and Gimblett 2000; Gimblett et al. 2000; Lawson and Mannin 2003a; Lawson et al. 2003; Wang and Manning 1999) show high potential of this technique.

Firstly, simulation modelling can be used to describe visitor use levels and behaviour patterns. By providing the managers with relevant information about where and when the visitors are concentrating in the park or protected area, it can help to identify e.g. “hot spots”, which are always a big concern. Additional questions like are those places located near fragile ecological or cultural resources can be answered.

Secondly simulation modelling can be used to monitor the condition of indicator variables that are inherently difficult to measure through direct observation (Lawson et al. 2006; Wang and Manning 1999). For example how the number of people at popular attraction changes during a day, week or even a month.

Thirdly simulation can be used to maintain appropriate carrying capacity of parks and protected areas. Indicator values can be used to check if the minimum acceptable conditions are maintained. Computer simulation modeling provides a tool to “proactively” manage carrying capacity by providing estimates of the number of people that can visit an outdoor recreation area without violating standards for crowding related indicators (Lawson et al. 2003a; Hallo et al. 2005). This can be a very important addition while preparing complex strategies for parks and protected areas for example such as permits systems.

Fourthly simulation modelling can be used to test the effectiveness of alternative management practices in a manner that is more comprehensive, less costly and less politically risky than on-the-ground trial and error (Lawson & Manning, 2003a). How creation of new bike trails would affect the bike traffic in the park? How a new hut would influence the attractiveness of specific region?

Fifthly simulation modelling can be used to guide the design of more realistic research on public attitudes concerning the management of visitor use in protected natural areas (Lawson and Manning 2003b; Lawson et al. 2003).

The first generation of simulation modelling applications to outdoor recreation was introduced in the 1970s and continued through the mid-1980s. The modelling approaches used during that time, referred to as the Wilderness Travel Simulation Model (WTSM), were designed to represent a protected natural area’s entire travel network, including entry points, trails, campsites and the attraction sites (Van Wagtendonk 2003). With the improvement of computer-based simulations capabilities a new generation of simulation modelling has been created. Main two related approaches were Recreation Behaviour Simulation (RBSim2) and Extend Simulation. The first approach combines computer simulation modelling with artificial intelligence technologies and geographic information systems (GIS) to simulate the visitors use in protected natural areas (Gimblett et al. 2000). Second approach uses Extend software, developed by Image That, Incorporated to create probabilistic, discrete-simulations similar to the WTSM.

These new methods gained very fast popularity among managers around the world. Agent based models, decision support systems, discrete choice models and GIS combined with GPS data enabled the managers to create a very sophisticated simulation models. With their help they will better understand and manage the visitors flows and therefore better protect the nature.

## **2.3. Project Mafreina – Management-Toolkit und Freizeit Natur**

### **2.3.1. Introduction**

In the last years pressure on nature in alpine regions has increased and this trend seems continue. One cause is the various winter and summer outdoor sports activities: snowshoeing, backcountry skiing, freeriding, hiking and mountain biking (Lamprecht et al. 2008). More users lead to diverse conflicts. Mountain biking, backcountry skiing and snowshoeing are seen as most important activities regarding sports-nature conflicts. Other land use concerns are the establishment of new wildlife sanctuaries and the sitting of new mountaineering cabins (BAFU 2009). Increasing income for local people with visitor enjoyment and the protection of nature are the main management concerns. Therefore it is important to approach new development initiatives in a pro-active manner with suitable planning tools, avoiding possible conflicts.

This new initiative is represented by the project Mafreina, which goal is to create new methods and tools to better understand and control the interactions between humans and nature. The project took place in the Müstair Valley in Switzerland where a group of visitors agreed to track their behavior patterns using GPS-Loggers. Data collection process was divided into two years with additional distinction to summer and winter periods. The very unique thing regarding this project is the fact that the visitors were allowed to move freely all around the research area. This way the managers of the project were be able to analyze all behavior patterns even in the most remote areas. Methods like Discrete Choice Models, GPS-Monitoring, GIS, Decision Support System and Agent Based Model will be used to help the researchers to create on the basis of collected data complex results. Within those methods GPS-Monitoring and GIS analysis will be one of the most important.



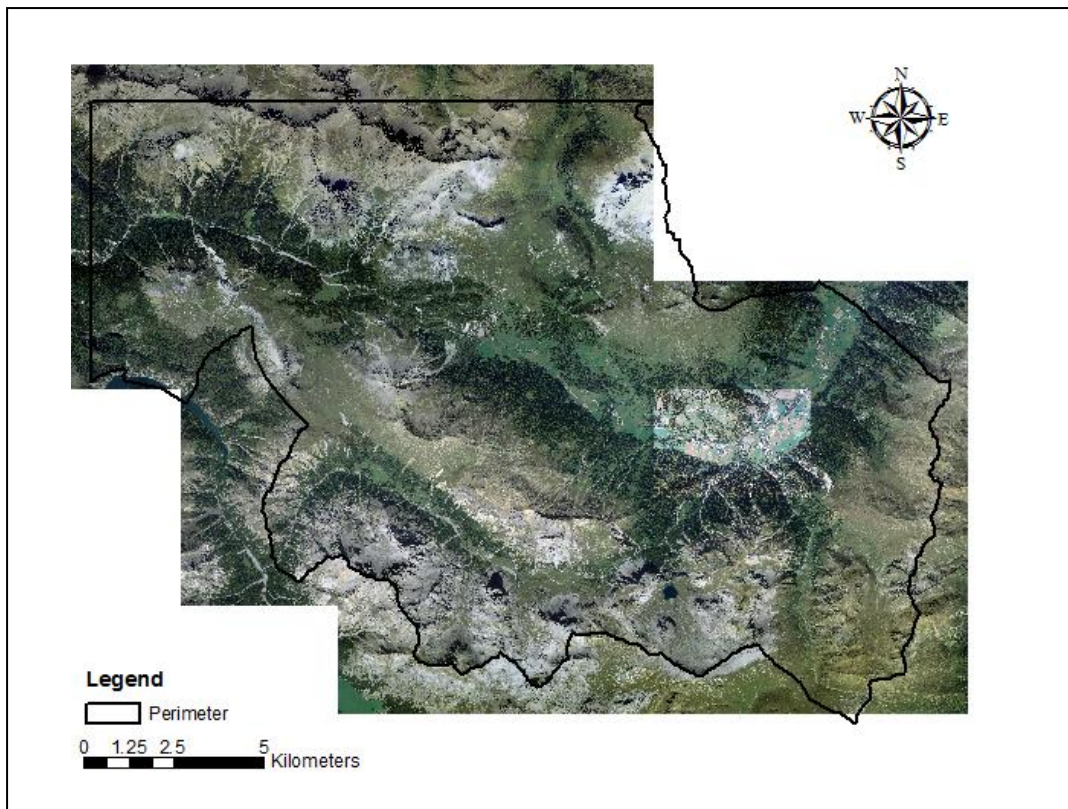
### 2.3.2. Research area

The Müstair Valley is located in the south-eastern parts of the Swiss canton Graubünden and a small part of it lies within the Italian region Trentino-Alto Adige/Südtirol. The region is only accessible from the remainder of Switzerland via the Ofen Pass. The whole region is remote from the rest of Switzerland and was always a secluded place which encouraged the growth of fauna and flora. The valley is characterized by traditional village's capes, an inviting terrain and thanks to the southern aspect a pleasant, mild sunny climate.

The Müstair Valley is an Unesco Biosphere Reserve adjacent to the Swiss National Park, which makes it very attractive to the tourists from all over the region as well as from Switzerland, Italy and Austria. The research area has 271, 14 km<sup>2</sup> and lies entirely in Switzerland. The main villages are Tschier, Fuldera, Lü, Valchava, Santa Maria Val Müstair and Müstair. Those villages and the Unesco Biosphere Reserve are main tourist attractions in the region and they will be the most interesting areas for the researchers of the project.



Figure 1: Red rectangle indicating Müstair Valley in Switzerland, source: [www.maps.google.com](http://www.maps.google.com)



**Figure 2: Research area Müstair Valley**

### **2.3.3. Project goals**

The main research goals defined by the managers for the project Mafreina are concerning:

- documentation of the existing spatial and temporal outdoor uses in Müstair Valley,
- documentation of the outdoor recreationist's requirements,
- research visitor preferences for planned projects,
- development of a predictive environmental planning tool to simulate results of management decisions on the recreation-wildlife-system,

### **2.3.4. Materials and methods**

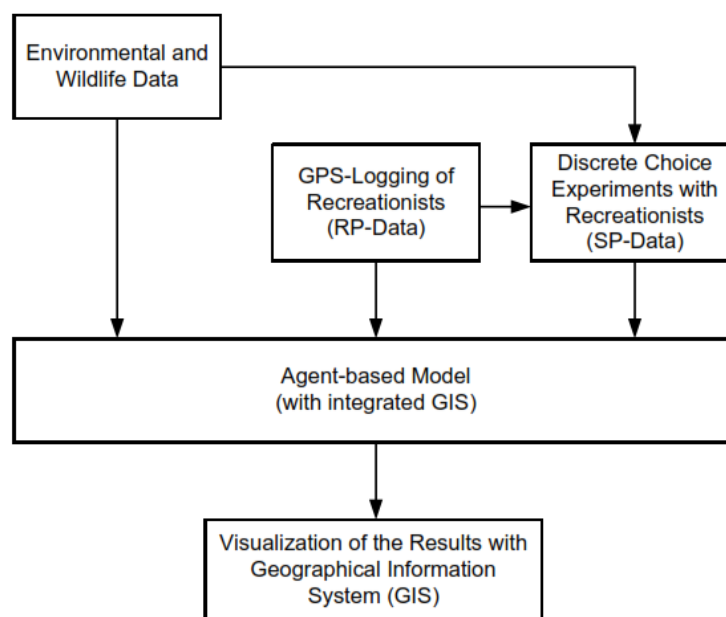
Until now no relevant decision tools exist for recreation and wildlife management. Agent-based models (ABM) are said to fill this gap (Lawson 2006). Methods to obtain rules for agents based models vary depending on author. GPS-monitoring is a method with a high potential to record real spatial and temporal movements as revealed preference data. A disadvantage of such revealed preference data is that they only deliver

information of already existing situations and not about planned alternatives or anticipated scenarios (Taczanowska et al. 2008).

In the Mafreina toolkit the geographical information system (GIS) plays a key in three various tasks. First it serves as a database of all environmental and GPS-Logging data. Second the GIS is an integrated tool for the ABM and the base of the virtual area in which the different scenarios will be computed and third results are visualized with GIS (Rupf et al. 2010).

The GPS-Loggers prepared for this project have a capacity to record data over 50 hours of activity over a period of 14 days. The time interval between each recorded point is 5 seconds which allows to create a very consist and precise behavior patterns of the tourists. The next important part of the project is the Decision Support System. DSS is supposed to help the mangers to make appropriate decision on the basis of the system results. For example the mangers may find out how an increase of 50% of bikers visiting Müstair Valley can affect the nature. It may also answer the question what might be the economic consequences of increase or decrease of the tourist's number.

The last vital part of the project is the Discrete Choice Model. DCM is supposed to provide information what are the tourist choices in to everyday situations. The tourists will need to answer various question represented on photos. This information as well as the data from GPS-Loggers will be assembled and analyzed to create various (ABM and DSS) simulations and prediction.



**Figure 3: Project Mafreina – methodological system, source: (Rupf et al. 2010)**

### **2.3.5. Summary of previous research**

Data aggregation was divided into two periods, summer and winter. In the year 2009 and 2010 data aggregation process took place and some data have been already analyzed. During the test of GPS-Logging from February to April 2009, the movement of 111 persons was recorded. These persons made over 300 daytrips (5% hiking, 15% downhill skiing, 25% snowshoeing and 55% backcountry skiing).

The collected GPS data allows diverse analyses to detect rules for the agent-based model, e.g. the frequency and location of starting points or the trip duration: the average backcountry skiing trip starts around 08:30 in the morning and lasts about 4 hours 10 minutes. The duration is distributed unimodal contrary to snowshoers.

Similar analyses will be made for other periods but they need to concern that different types of activities will occur depending on time of the year. This may cause that the behaviour patterns will vary in many parts of the region and therefore they will need to be comprehensively analysed.

### **3. Placement in the research project**

#### **3.1. Geographic information system**

Geographic information systems are widely used around the world in many disciplines to store, edit, analyse, visualise and manage data according to their geographic localisation. They allow managers to make accurate and fast decision basing on various data types. In fields like geography, geology, archaeology, remote sensing or natural resource management, GIS plays a key role. Fast development of spatial data infrastructure combined with GIS creates new opportunities to better and more comprehensively understand the surrounding environment.

Applying GIS to the visitors monitoring systems gives a new perspective to visualise many aspects connected with the visitors flows via maps which helps to easier and faster interpret the data. GIS also allows analysing the data not only in the qualitative but also in the quantitative manner. This way managers can compare their results among different areas and draw appropriate conclusions.

When it comes to visitors behaviour patterns the most useful tools are network analysis tools. They offer the possibility to analyse spatially referenced data describing traffic flows. This type of analysis is mainly carried out in the cities or urban areas, whereas in parks or recreational areas they are slowly gaining popularity. Due to fast the increase of tourist's number, the potential of the GIS and its analytical advantages was spotted and nowadays many managers use it very willingly.

In the project Mafreina GIS plays one of the key roles. The combination of GPS data, various topographic, geologic dataset and aerial photos with the advanced functions of GIS is supposed to describe and analyse the visitors behaviour patterns. Basing on the GPS data and the maps, managers will be able to see how the tourists move around the research area, where and when do they interact with the nature etc. These results will be further implemented with the results from DCM into the ABM to stimulate the possible visitors' flows scenarios.

### 3.2. Time and spatial analyses of the GPS data

Visitors monitoring and management projects vary depending on many details one of which is the method of data aggregation. Most popular methods were described in previous chapters but there is simple classification which represents static and dynamic data aggregation. First method shows how many tourists entered or left the park, how many tourists visited popular sight or how many slept in a hut. Depending on the method this data can be enriched with qualitative information such as age, gender, nationality etc. In order to even profounder analyse the visitors flows it is recommended to use GPS receivers which can continuously and very exactly save information regarding tourist's position in the research area. This is the second method which is especially usefully, when the research area is very big and tourists do not need to follow the designated trails.

The managers of the project Mafreina decided to chose the second method as the Müstair Valley represents exactly that kind of area. During the project they have asked random tourist if they liked to carry with them GPS-Loggers which would track their moves. The GPS-Loggers were able to store data over 50 hour's hours of activity over a period of 14 days. According to the producers of the GPS-Loggers the spatial position accuracy is 2.5m CEP<sup>1</sup> and 5m SEP<sup>2</sup>. The main problem with those devices was the fact that they were constructed for the localization of trucks, cars and containers. Therefore they were too heavy and too big for a monitoring project. Next problem was that the battery did not last too long and due to that, insufficient number of data was gathered.

The solution to this problem was a new generation of GPS-Loggers with the (U-blox-module) which were perpetrated for the project with the help of Federal Department of Defence, Civil Protection and Sport of Switzerland and a company "Art of Technology". These devices were adjusted to the needs of the project and equipped with appropriate software. According to the producers U-blox they support following operating modes Continuous Tracking Mode (CTM) and Power Saving Modes. According to the producer in the first mode, the Autonomous Power Management (APM) automatically optimizes power consumption. It powers off parts of the receiver when they are not used. Also, the CPU speed is reduced when the CPU workload is low. The second mode is configurable power saving mode where the GPS is put into sleep mode and

---

<sup>1</sup> CEP = Circular Error Probability: The radius of a horizontal circle, centered at the antenna's true position, containing 50% of the fixes.

<sup>2</sup> SEP = Spherical Error Probability. The radius of the sphere, centered at the true position, contains 50% of the fixes

activated up on a selectable time interval or upon external request. This mode was ideally suited for the project Mafreina where battery power savings were such an urgent issue.

All tourists who agreed to participate in the project received those GPS-Loggers which recorded every 5 seconds their exact position. GPS-Loggers data allows performing a very precise analysis not only showing where but also when the tourists were. This way the managers can adjust their management plans to the present situation knowing exactly where and how do they need to act.

During the time and spatial analyses of the GPS data researchers need to remember that its positional accuracy relies on many factors like the number of satellites, satellites constellation, GPS-Logger, weather condition as well as some obstacle which might affect the satellite signal. This factors and their role will be explained in the next chapters, where exact data preparation and analysis will take place.

### **3.2.1. Spatial data mining**

In the project Mafreina during the aggregation process, about 5 million GPS points from summer 2009 and 2010 were assembled. These points were additional enriched with other important qualitative information. This amount of data of data requires from researchers a very precise study of data structure the so called spatial data mining process.

According to Wang et al. (1997) spatial data mining is the process of discovering interesting and previously un- known, but potentially useful patterns from large spatial datasets. Extracting interesting and useful patterns from spatial datasets is more difficult than extracting the corresponding patterns from traditional numeric data due to the complexity of spatial data types and their autocorrelation. Specific features of geographical data that preclude the use of general purpose data mining algorithms are:

- rich data types(e.g., extended spatial objects),
- implicit spatial relationships among the variables,
- observations that are not independent,
- spatial autocorrelation among the features,

Data from GPS-Loggers is not only very large but also very complex. It contains many variables, such as horizontal dilution of precision, number of satellites, distance from the last point, movement direction etc., which will be very essential in further

analysis. Spatial data mining is supposed to answer many questions from which the most important are:

- Do the variables from the GPS data correlate with each other?
- Do the GPS data correlate with the variables from other datasets e.g. landcover type?
- Do those correlations influence the spatial location of the GPS points?
- Do the correlations change over time and place?

In order to perform all the analysis, the patterns in the spatial datasets need to be profoundly investigated. To find even more interesting patterns in the datasets it is recommended to enrich them with new variables but not necessarily spatial. This may outline new hidden patterns which were previously not visible. The combination of spatial and non-spatial information will be implemented into the scripts which will play a key role in the further analysis.

### **3.2.2. Map matching**

The main source of information in the project Mafreina is GPS data. The GPS data has its advantages and disadvantages, one of them is that the tourists do not need to answer many question regarding their daily trips. It is always a time-consuming process but know the managers do not need ask the visitors where they exactly were. However the spared time needs to be spent on a very extensive and complex postprocessing data analysis. One of the main analyses connected with the GPS data is the map matching process. It is defined as the process of correlating two sets of geographical positional information (e.g., GPS records of object positioning versus digital road networks or trails).

First map matching algorithms focused more on accuracy and consistency of the routes, since the survey samples have been still rather small. Accordingly, most reviews of map matching algorithms, (White et al. 2000; Quddus et al. 2007) describe accuracy as percentage of correctly identified links when they talk about the performance of the algorithm. However the increase in use of GPS devices in large-scale transport studies, caused that the need for computational speed grows. This need is further amplified by the vast use of high-resolution navigation networks, which are essential for an accurate identification of the chosen routes. However only a few authors Nielsen et al. (2004) and Marchal et al. (2005) have subjected under discussion the issue of performance in the sense of computational efficiency.



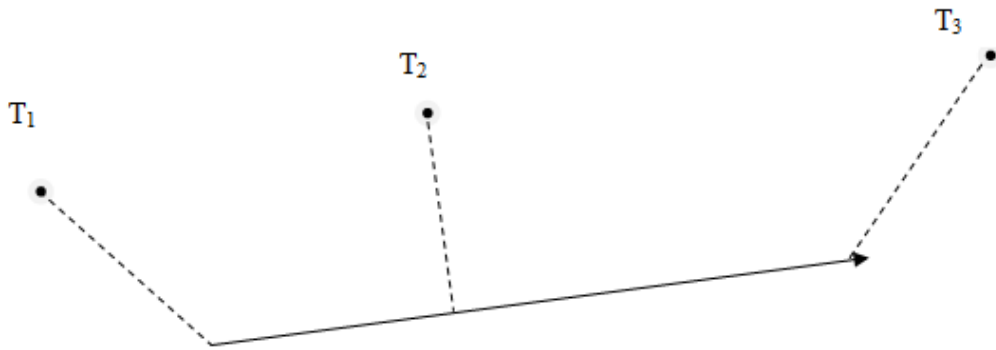
According to Schuessler and Axhausen (2009) map matching procedures can be classified into three categories:

- geometric procedures,
- topographic procedures,
- advanced procedures.

The geometric approach is the most basic one, because it only measures the distance from the GPS point to the adequate network element. The most popular example according to White et al. (2000) is the search of the nearest node or the nearest link. This search is based on the direction of GPS point and the heading of a link or a trail. The main disadvantage of the geometric procedures is that they neglect the sequence of the GPS points over time as well as the connectivity of the network links. According to White et al. (2000) they are also very dependent on the correct network coding and are rather sensitive to outliers.

The topographic procedures offer a more advanced ways of associating the point to the links. They base not only on the distance between GPS point and the nearest trail, but also on the history or sequence of GPS points and the connectivity of the network elements. Schuessler and Axhausen (2009) underline that most procedures work in two steps. First the initial node or link is found using geometric approaches. Then the route is created by choosing a link out of the set of candidate links. This set consists of the last matched link and the links succeeding that link however some authors extend it for all links preceding the last matched link Chung and Shalaby (2005) or for the links succeeding the succeeding links.

A common issue regarding this method is the choice of the link out of the set of candidate links. The best solution to this problem is the perpendicular distance between the GPS point and the link. The perpendicular distance equals the minimum Euclidean distance to the start node, minimum Euclidean distance to the end node and the minimum Euclidean distance between the GPS point and its orthogonal projection on the link (Schuessler and Axhausen 2009). Out of those three values the minimum one is chosen and it is called the relevant perpendicular distance.



**Figure 4: Relevant perpendicular distance for three example points.**

The relevant perpendicular distance for the point  $T_1$  is the start node of the link, for  $T_2$  it is its orthogonal projection on the link and for  $T_3$  is the end node of the link. The topographic approach offers other ways to match the GPS point to the nodes or links. Heading of the GPS point can be compared to the heading of the correspondent link ( Chung and Shalaby 2005; Velaga et al. 2009). According to Quddus et al. (2003) heading is based on the angle between the link and the line between the start node of the link and the GPS point. Those criteria can be merged and weighted which is especially useful when we have numerous variables. The topological approach is far more sophisticated than the geometric approach but still there are some disadvantages. In a situation when the initial node determination failed or there are parallel roads close to each other data can be misinterpreted and results can be significantly different from those expected.

In recent years many new approaches to overcome those problems have been presented. They not only take under consideration the whole sequence of GPS points and the network topology, but also the fact that due to errors in the GPS measurement as well as the network coding, the nearest link or node is not necessarily the right one. Some of those approaches are more adequate than others for the goals of project Mafreina, but a small review of them will be represented.

A very good method to explain the GPS measurement errors is the construction of error or confidence regions around the GPS points (Doherty et al. 2001; Ochieng et al. 2004; Velaga et al. 2009). The size of the errors should be calculated on the basis of the error variances. Then all links of trails within this error regions should be evaluated based on factors like heading, distance or even speed. Sometimes the concept of error regions can be approached by fuzzy logic inference systems. The fuzzy rules consider different criteria such as distance, heading, speed, HDOP value, link connectivity and the position

of the GPS point relative to the candidate link. Therefore, various rules can be applied for the initial link search and the subsequent path development.

Different approach without the use of error regions was presented by Nielsen et al. (2004). This approach is very similar to the Dijkstra algorithm for the single-source shortest path problem. The start node is determined in the preprocessing process and starting from there the route is created by adding the end nodes of all outgoing links of the current node to the set of nodes to be evaluated. The next node to be calculated is then the node that could be reached in the shortest amount of time beginning from the last node of the route so far. Nielsen et al. (2004) indicates that the score of each node is calculated based on the perpendicular distance between the GPS points and the links they are associated with and the distance between the GPS points and the start node of the link they are assigned to. Additional problem with this algorithm is that it cannot guaranty to find the optimal solution because the route development criterion differs from the scoring function.

There are many other map matching methods which could be presented in this chapter but majority of them is based on similar concepts. From all of the map matching methods none is fully suitable for the goals of the project Mafreina. The main reason for this lies in the definition. Map matching is the process of correlating two sets of geographical positional information, however in the Münstair Valley research area not all point can and should be correlated with trails. One of the project goals is to determine where the tourists leave the trails and mainly due to this reason normal map matching procedure cannot be carried out. GPS points representing the tourists who intentionally left the trail cannot be attached to the trail again. This kind algorithm's behavior is desired by the managers of the project. Therefore a new kind of map matching algorithm needs to be prepared. This algorithm needs to determine which GPS points can be attached to the trails and which should be left to mark the places where the tourist left the trails. This kind of algorithm will have to base on some basic map matching procedures and the concept of fuzzy logic presented by Ochieng et al.(2004).

### 3.2.3. Fuzzy logic and MCE

Fuzzy logic is a multi-valued logic, which allows intermediate values to be defined between conventional evaluations like true/false, yes/no, high/low, etc. In fuzzy logic values range in between 0 and 1 which is opposite to binary two-valued logic. Fuzzy logic allows determining the degree of truth, not only on the basis that something is completely true or falls. This kind of logic is used for computer software to mimic more closely human reasoning. It is especially useful when a decision is based on an incomplete or uncertain data. The concept of fuzzy logic is based on the work of Polish mathematician Jan Lukasiewicz (1878-1956) and then developed by the Azerbaijani-Iranian computer scientist Dr. Lotfi A. Zadeh who created the term fuzzy logic.

Apart from fuzzy logic which is one of the key elements in the analysis, multi-criteria evaluation will also be a significant part. Multi-criteria evaluation is a common method for assessing and aggregating many criteria/factors. It helps to profoundly analyze a problem basing on multiple criteria which needs to be evaluated. According to Proctor and Quershi (2004) it is being increasingly used in the assessment of natural resource management options which involves complex ecological, economic and social outcomes and interactions.

Currently there are several MCE techniques a simple Boolean aggregation method and more flexible and sophisticated aggregation methods like Weighted Linear Combination (WLC) and Ordered Weighted Averaging (OWA). In the decision making process some criteria can be more important than others or they may be only marginal importance. In the WLC or OWA method weights assigned to factors govern their importance or a degree to which they can compensate for another factor. This method in general not only allows retaining the variability from factors, it also gives the ability to have the factors trade off with each other. In the project Mafreina factors like HDOP, speed, distance from trail and many others can be weighted and used to evaluated whether a visitor was on the trail or not.

Merging MCE techniques with fuzzy logic gives the researches very useful tools which deliver sophisticated, flexible and realistic analysis. The researches do not have to answer questions using the binary code 1/0, they may use many different answers between true or false. This ensures that they results are more reliable and easier to interpret.

### **3.2.4. Geovisualisation**

Geovisualisation is the last step preceding results description and conclusions drawing. Visualisation of spatial datasets is very important element of the decision process. People can much easier interpret the visual results of the analysis than the computers. On the basis of the visualised results they can compare results from different algorithms or scripts. Without visualisation it would be nearly impossible to interpret the results of the spatial analysis. Advanced visualisation techniques are especially usefully when there is a big amount of the data to interpret. For this purpose traditional maps are not eligible for complex analysis and various data simulation. GIS tools enable the managers to create very complex maps, diverse scenarios and combine many different data types. GIS tools not only help the managers to interpret those results on daily basis but also to draw complex summaries of the studies.

Nowadays many programs offer numerous tools which help to create complex maps. Those maps can be displayed not only in dedicated software but also in many web browsers which makes them even more accessible and interactive. Programs which will be used in my master's thesis will be described in the subsection.

## **3.3. Overview of technical details**

As it was mentioned in previous chapters the choice of appropriate software is a very important element of each research project. The software needs to allow performing many kinds of analysis, visualizing and modifying data and adapting the software to the needs of the project. On the market there are many different companies selling GIS software but for the project Mafreina the managers have decided to buy the products of company ESRI.

### **3.3.1. ArcGIS**

Company ESRI created one of the most import GIS software ArcGIS. Depending on the user license it allows to perform many GIS analysis which can be used in many different fields like transport, geography, archeology or education. For the needs of the project Mafreina a user's license ArcInfo was bought. It consists of many components from which two are the most important:

- ArcMap is used primarily to view, edit, create, and analyze geospatial data in vector and raster formats

- ArcCatalog is geodatabase administration application

In the newest version of ArcGIS 10 those two components have been integrated to improve and accelerate the pace of work. ArcGIS has built-in geoprocessing toolbox which offers a vast number of interesting analytical tools. The most popular are Spatial Analyst, 3D Analyst, Linear Referencing Tool and Data Management Tools.

In order to make the software even more user friendly ESRI created additional possibilities to enhance the quality of work:

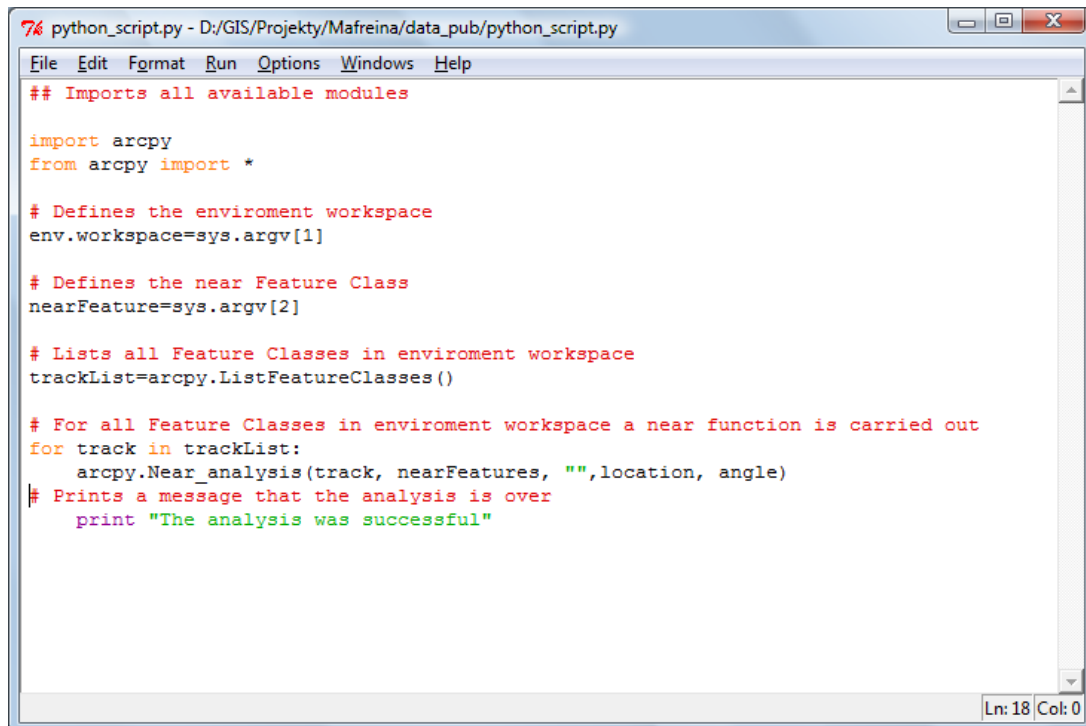
- **Model Builder** is an application in which user can create, edit, and manage models. It allows user creating their own tools on the basis of existing tools. A simple drag and drop mechanism allows user to integrate various data types and tools.
- **Batch Processing** allows the users to automate analysis of vast amounts of data
- **Scripting languages** adopted by ESRI in order to help the user to write their own scripts. They can be written for a specific task, project or just to enhance everyday work.

ArcGIS supports many programming and scripting languages like JavaScript, VBScript, Visual Basic, C++, Perl or Python. Python is programming language which was officially chosen by ESRI and now it's recommended to all users of ArcGIS. ArcGIS is clearly an undisputed market leader in GIS software therefore it was chosen for the project Mafreina.

### 3.3.2. Python

In recent years ESRI realized that many of its users do not want to be programmers but still would like to create tools to help them solve different tasks. These tools should include clear, consistent GUIs, scriptable objects and the nuts-and-bolts programming tools necessary for customization. To fulfill users needs, ESRI supports a variety of scripting languages using ArcObjects—starting with the geoprocessing framework. Python is one of those supported languages. It is an Open Source, interpreted, dynamically typed, object-oriented scripting language. Python is included with ArcGIS 9 and newer releases and is installed along with the other components of a typical installation.

Python provides many ways for integration within GIS computing systems. Cross-platform capabilities and ease of integration with other languages (C, C++, FORTRAN, and Java) mean that Python is most successful in gluing systems together.



```
python_script.py - D:/GIS/Projekty/Mafreina/data_pub/python_script.py
File Edit Format Run Options Windows Help
## Imports all available modules

import arcpy
from arcpy import *

# Defines the environment workspace
env.workspace=sys.argv[1]

# Defines the near Feature Class
nearFeature=sys.argv[2]

# Lists all Feature Classes in environment workspace
trackList=arcpy.ListFeatureClasses()

# For all Feature Classes in environment workspace a near function is carried out
for track in trackList:
    arcpy.Near_analysis(track, nearFeatures, "",location, angle)
# Prints a message that the analysis is over
print "The analysis was successful"

Ln: 18 Col: 0
```

**Figure 5: Python script for “Near” function**

Python allows using all tools from the ArcGIS Toolbox which makes it the most useful GIS scripting language. From the figure 5 it is clear that the syntax is very intuitive and it does not require advance programming knowledge.

Python scripting language is specifically useful when a big amount of complex data needs to be analysed. For the needs of spatial data mining and data preparation only Python scripts will be used in my master’s thesis. They ensure high quality of analysis and good calculation speed. For future automation of analysis it is recommended to use those scripts while they can be easily implemented into the ArcMap or ArcCatalog framework.

## 4. Data preparation

In every research project it is very important to aggregate appropriate data and profoundly examine them. This ensures that during further parts of research project it will be easier to explain what lead to particular results and how individual factors or their attributes influenced the overall results.

Solid data preparation is especially important when there are numerous data which tend to correlate. This way a researcher can draw appropriate conclusions and overcome some problems which may occur in later parts of the project.

### 4.1. Data description

In the project Mafreina there many different data types which can be mainly divided into two groups, raster and vector data. Raster data is a form a matrix of cells (pixels) organized into rows and columns. Every pixel contains specific information, such as temperature, height above sea level or slope gradient. Rasters can be digital aerial photographs, imagery from satellites, digital pictures or scanned maps. Vector data is based on very simple geometrical figures like points, lines and polygons. It can contain much different information like polygons representing landcover types or topographic elements, lines showing different roads or trail classes and points representing simple infrastructure objects.

All data which will be used in the master's thesis was delivered by Zurich University of Applied Science. It has been converted into Swiss coordinate system CH1903 which will be gradually replaced by CH1903+.

#### **Vector data:**

- **Perimeter** – Polygon feature class covering the research area.
- **Landscape** – Polygon shapefile describing various landscape types. It describes twelve various landscapes in the research area.
- **Landcover** – Polygon feature class which shows various landcover type, such as forest, meadows, lakes, boulder or marsh. It was derived from a topographic map 1:25 000 by Swisstopo<sup>3</sup>. Positional accuracy of this data is according to Swisstopo is 3-8m.

---

<sup>3</sup> Swisstopo is the official name for the Swiss Federal Office of Topography



- **Trails25** – Line shapefile representing streets and trails in the research area. It describes eleven different classes with additional classification to hiking types. Positional accuracy of this data is according to Swisstopo is 3-8m. This data was also derived from topographic map 1:25000
- **Trails25new\_class** – Line feature class with the identical information as Trails25 with exception of attribute trail construction. This information will be useful for the examination of bikers' behavior patterns.
- **Bike25\_trails** – Line feature class representing trails prepared for bikers. It was derived from a topographic map 1:25 000 by Swisstopo. Positional accuracy according to Swisstopo is 3-8m
- **Trails** – Line feature class representing the same object as Trails25. However those objects have been updated and in some regions new lines have been created. New lines have been created on the basis of orthophotos which indicate a higher positional accuracy.
- **Important\_natural\_objects** – Point feature class representing objects of high importance to nature protection. It can be a habitat of a rare snake or some flora object.
- **Tracks\_2009** – Feature dataset containing information regarding tourist trips in the year 2009. Trips are represented as GPS tracks where every trip is divided into single day trip. Each GPS point contains attributes such as X and Y coordinates, hdop, number of satellites etc.
- **Tracks\_2010**– Feature datasets containing information regarding tourist trips in the year 2010. Trips are represented as GPS tracks where every trip is divided into single day trip. Each GPS point contains attributes such as X and Y coordinates, hdop, number of satellites etc.

**Raster data:**

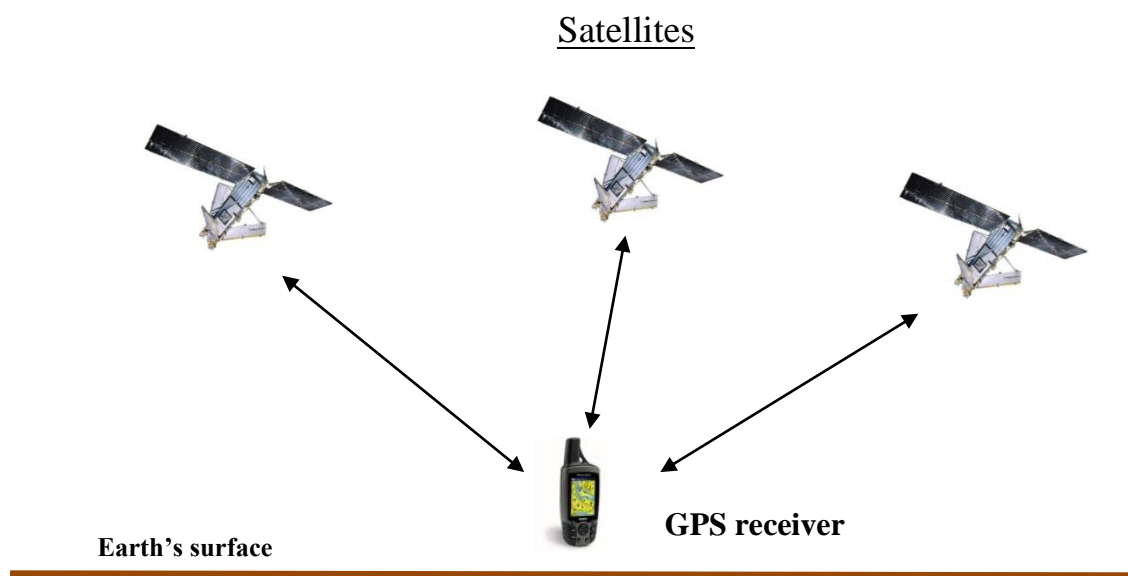
- **Orthophotos** - Aerial photographs showing the research area. The pixel size corresponds to 0.5 m in the field. This factor ensures high positional accuracy.
- **Topo25** – Topographic map 1:25 000 created by Swisstopo.

From all feature classes and rasters some need a more detailed description while they contain many important information for the project. Those feature classes and rasters are landcover, trails, tracks\_2009 and tracks\_2010.

#### 4.1.1. Number of satellites

Feature classes “tracks\_2009” and “tracks\_2010” represent points recorded by the GPS sensors which were distributed to the tourists. Their main advantage is that they contain much useful information regarding each recorded position. One of the most important information is the number of satellites which the GPS-Logger registered. On the basis of this information it can be initially specified what is the position accuracy for each point.

The Global Positioning System is a satellite-based system which provides time and location information anywhere on the Earth. To know the location on the Earth surface a special receiver is require and at least three satellites from which it can capture the signal. GPS receivers on the basis of this signal can triangulate data and pinpoint the exact position.



**Figure 6: Concept of GPS**

There are exactly 30 GPS satellites orbiting around Earth at this moment. Already three satellite can determine the location but more satellites can determine the location more precisely. The GPS-Loggers used in the project Mafreina received signal from maximum 12 and minimum 3 satellites. There are many factors which may affect the number of satellites detected by the GPS-Loggers. Popular factors are all kind of physical obstacles like buildings, trees or landforms like mountains. They may reflect the signal causing lower number of detected satellites. Time of the day is an important factor because during the day depending on the location, number of visible satellites may vary significantly. Condition of earth's troposphere, known as weather, may also affect the

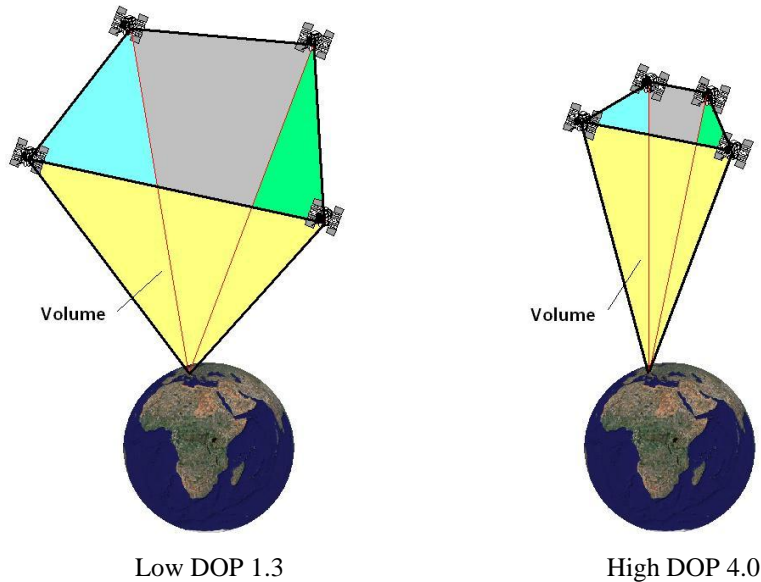
number of detected satellites. Type of the GPS receiver also influences number of detected satellites, but nowadays GPS receiver are very advanced technologically, which ensures the highest quality of measurements.

#### **4.1.2. HDOP**

Feature classes “tracks\_2009” and “tracks\_2010” contain another important information which is the horizontal dilution of precision. Apart from the number of satellites, their geometric configuration is the second most important factor influencing the positional accuracy. This configuration is expressed in terms of scalar value, which is referred in the literature as DOP (Dilution of Precision). The DOP value describes the weakening of precision and is therefore a factor or measure of the constellation dependent imprecision. When the satellites are regular distributed on the sky the DOP is low, but when they are close to each other the DOP becomes higher.

According to producers of GPS devices the most important DOP terms are:

- GDOP (Geometric-DOP): Describes the influence of satellite geometry on the position in 3D space and time measurement
- PDOP ( Positional - DOP): Describes the influence of satellite geometry on the position in 3D space
- HDOP (Horizontal-DOP): Describes the influence of satellite geometry on the position along upon a plane (2D)
- VDOP (Vertical-DOP): Describes the influence of satellite geometry on height (1D).
- TDOP (Time DOP): Describes the influence of satellite geometry on time measurement



**Figure 7: Concept of measuring the DOP values, source: [www.u-blox.com](http://www.u-blox.com)**

In mountain areas, forests and urban areas HDOP values need to be analysed very consciously. During the day the configuration of the satellites can be very unfavourable due to different number of visible satellites and obstructions affecting this number. GPS receiver calculates the DOP value from four visible satellites. When there are more than four visible satellites GPS receiver uses only those, which create the best constellation on the sky, to calculate the position.

**Table 2: Classification of DOP values, source: Langley, 2008**

DOP Value	Rating	Description
1	Ideal	The highest possible confidence indicating the highest precision at all times.
1-2	Excellent	This confidence level ensures that positional measurements are considered accurate enough to be applied in most applications.
2-5	Good	Represents a level that marks for which positional measurements could be used to make reliable in-route navigation suggestions to the user.
5-10	Moderate	Positional measurements could be used for calculations, but the fix quality could still be improved.
10-20	Fair	Represents a low confidence level. Positional measurements should not be taken under consideration or used only to indicate a very rough estimate of the current location.
>20	Poor	At this level, measurements are inaccurate by as much as 20 meters with a 4 meter accurate device.

#### **4.1.3. Distance from previous point**

GPS receiver calculates the location every 5 seconds but there are some irregularities in the data. The points are placed farther than within 5 seconds walking or driving distance. Often points are located 5-60 seconds or even longer away from the last point. This situation causes that the distance from the last point can vary significantly. Due to that fact an assumption that the distance is reliable information source cannot be made. Distance as filtration factor can be used only when the time information is provided so that it is sure that the tourists during their hiking trip made 100m not in 5 seconds but in 60 seconds.

#### **4.1.4. Speed**

This value is calculated by dividing the distance from the last point and the time difference between those points. This means that it is insensitive to the changes in the data recording. It is certain that the results are reliable because whenever the time value is higher than 5 seconds also the distance becomes appropriately longer. Speed values are represented in the feature classes in two columns one showing the speed in m/s and other km/h.

#### **4.1.5. Metadata**

Feature classes “Tracks\_2009” and “Tracks\_2010” additionally contain information regarding altitude, land aspect, slope, latitude and longitude of each point in the Swiss coordinate system CH1903 and WGS84. Movement direction, time from the last pause and time of the pause are also saved in the feature classes. Additionally information regarding location in the research area, day, track number and unique id can be found in the tables.

This feature classes contain very complex information which is also extended with extra metadata. The most important information gathered in the metadata file:

- Time and place of the disposal of the GPS receiver
- Time and place of the return of the GPS receiver
- Type of activity
- Season of the year
- Number of people in the group
- Type of transportation
- Type of visit

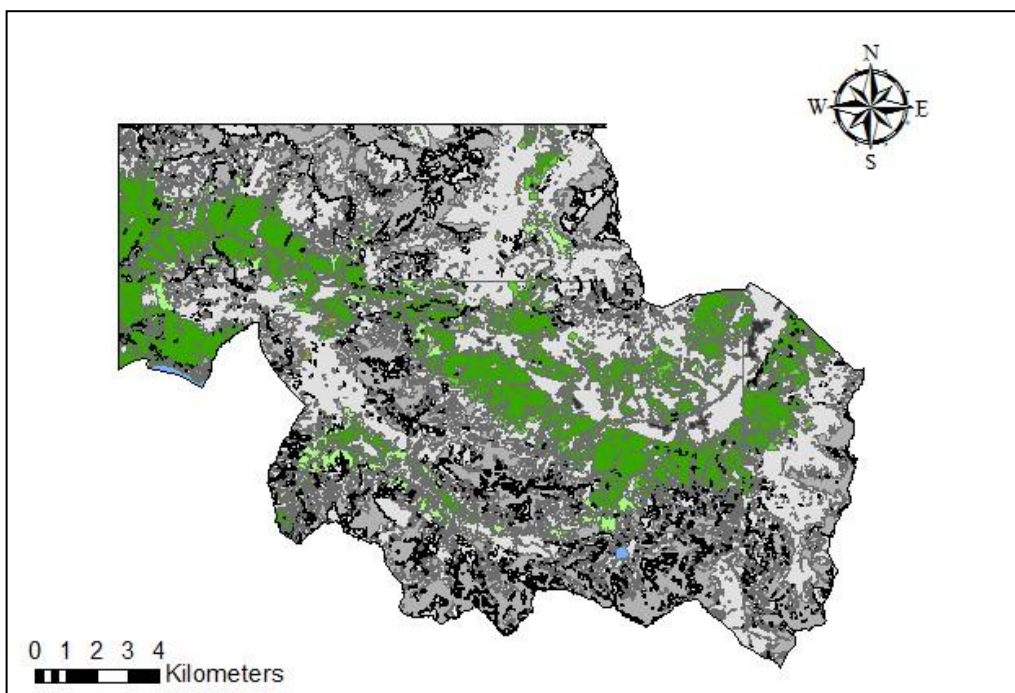
- Gender of visitor
- Age of visitor

This information will be an essential addition to data saved in feature classes, while they will allow performing more qualitative than quantitative analysis.

#### 4.1.6. Landcover

Landcover feature class contains information regarding different landcover types classified into 18 groups. Summaries made on the basis of this feature class indicate that the biggest area is covered by following groups:

- Not classified 104, 7 km<sup>2</sup>
- Boulder 71,8km<sup>2</sup>
- Forest 51, 7 km<sup>2</sup>
- Rock\cliff 24,7km<sup>2</sup>
- Coppice 11,2km<sup>2</sup>
- Open forest 4,2km<sup>2</sup>
- Settlement 1, 1 km<sup>2</sup>



**Figure 8: Landcover types in the research area.**

All other groups cover an area smaller than 1km<sup>2</sup> which makes them less relevant than those 7 groups. Characteristics of those groups may have an influence on the behaviour of the tourists as well as the positional accuracy. These landcover types can

significantly influence the number of visible satellites and therefore dilution of precision. They will be examined during data analysis part to check whether there are some significant correlations between them and data from “Tracks\_2009” and “Tracks\_2010”.

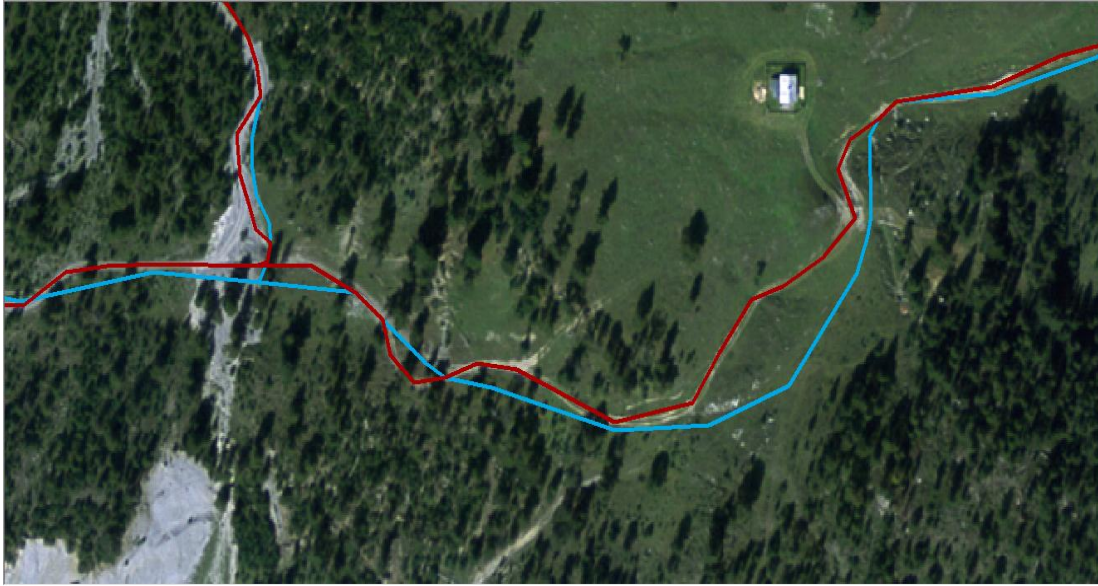
#### **4.1.7. Trails 25**

Feature class “Trails25” represent a road network in the research area. Lines creating the whole network have been derived from a topographic map 1:25 000. This means that accuracy of the position according to map varies from 3-8m. The whole network represents nine different types of roads from first to sixth class, side roads or footbridges. In this feature class there is also information concerning hiking type. Some of the roads are classified as mountain hiking trails or normal trails.

#### **4.1.8. Trails**

Feature class “Trails25” was derived from the topography map 1:2500 which ensures that its positional accuracy varies from 3-8 meters. This accuracy might influence the results of the analysis, therefore the trails had to be revised. On the basis of the orthophotos, old trails had been inspected and corrected. Old topography maps did depict new trails as well as the changes in the old trails. Old trails did not match the trails represented on the orthophotos and while the orthophotos are more precise than the topography maps (1 pixel equals 0.5 meters), the managers of project decided that they need to be rearranged.

The way the new trails have been prepared has been accepted by the managers of the project. They agreed on the fact, that the trails will vary from the official trails prepared by the Swisstopo but for the needs of the project this change had to be made as the quality of the trails is so important for the analysis.



**Figure 9:** “Trails” red lines and “Trails25” blue lines. Aerial photography showing difference in positional accuracy of both feature classes.



**Figure 10:** Trails” red lines and “Trails25” blue lines. Red lines present new trails and roads created on the basis of the orthophotos.

## **4.2. Data preparation and selection**

Data preparation and selection are an important pre-processing parts of every research project. In the chapter data description many different data types have been presented but not all of them will be used in the analysis. In this chapter Python scripts

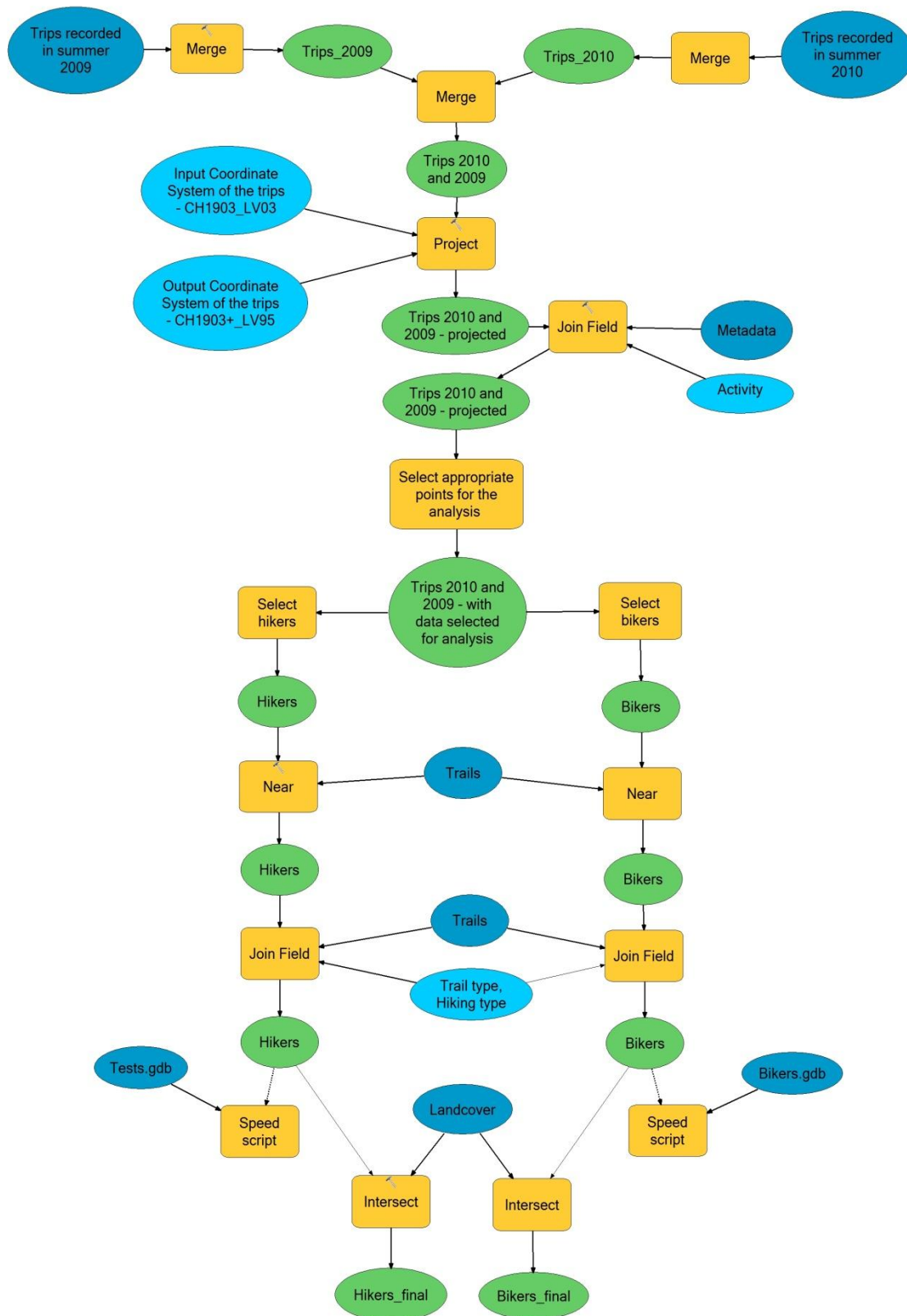


will be introduced which prepare and select specific information which will eventually be used in further analysis.

The main goal of my master's thesis is to analyse the behaviour patterns of the people visiting Müstair Valley. To ensure better understanding of the situation in the research area, many different variables need to be taken under consideration. In the analysis a key role play feature classes "Tracks\_2009" and "Tracks\_2010", however not all data from those feature classes can be used. Additionally this data needs to be enriched with some extra information.

- AKTIVITEATBEZEICHNUNG
- NEAR\_FID
- NEAR\_DIST
- MEAN\_SPEED
- OBJECTVAL\_LAN
- TRAIL\_TYPE
- HIKING\_TYPE

Those attributes need to be prepared in order to select only desired data. They will be described during the data preparation and selection process.



**Figure 11: Data preparation and selection process.**

First step of the data preparation and selection process was to merge all trips into two groups, one representing trips from the year 2010 and second from the year 2009.

Then all merged trips had to be converted into one feature class “trips\_2010\_2009.shp” where all data was aggregated. Then this feature class had to be transformed from old Swiss coordinate system CH1903 into the new coordinate system CH1903+.

Apart from those feature classes managers of project Mafreina delivered metadata about tourists who rented the GPS devices. This metadata files contain many useful information from which activity type of the tourists is a one of the most significant. Activity type tells whether tourists were driving by bike or by car, hiking, climbing, jogging or even berrying. There are 21 different categories which help to categories all trips and focus only on a specific group. This attributes were added to the feature classes and saved as “AKTIVITEATBEZEICHNUNG”.

```
import arcpy
from arcpy import *

env.workspace=sys.argv[1]
trackList=arcpy.ListFeatureClasses()
for track in trackList:
    cur=arcpy.UpdateCursor(track)
    row=cur.next()
    while row:
        if (row.aktivitaetbezeichnung == "Wandern" or row.aktivitaetbezeichnung == "Mountainbiking")
        and row.hdop <= 10 and row.imperimeter == "t" and row.numsat > 4 and row.pause == "f":
            row.FID_Identifiier = 1
            cur.updateRow(row)
        else:
            row.FID_Identifiier = 0
            cur.updateRow(row)
        row=cur.next()
    del cur, row
```

**Figure 12: Python script syntax – data selection process.**

After the data was merged and enlarged with attributes from metadata, selection process could be initialized. To enhance the selection process a python script was written, which indicates points that are meant for further analysis and those which can be omitted.

**Table 3: Mean HDOP values for different number of satellites.**

Number of satellites	3	4	5	6	7	8	9	10	11	12
Mean HDOP	5.56	4.50	2.74	1.84	1.72	1.65	1.51	1.36	1.14	0.93
Std.dev HDOP	6.06	3.82	1.91	0.85	0.83	0.84	0.76	0.70	0.51	0.28

From all points gathered in the feature class “trips\_2010\_2009.shp” only specific points were eligible for the analysis. Therefore only the points within the research area, representing hikers or mountain bikers with low hdop and high number satellites were selected. Points indicating only 3 or 4 satellites were excluded from further calculations due to high mean HDOP and standard deviation. Those points also could not represent

pauses because they always represented a low level of positional accuracy. Those points had a mean HDOP at the level of 10.2 and standard deviation 19.46. Due to so high values these types of points were not included in the selection process. For all points which met the requirements of the analysis a value 1 was assigned in the “FID\_Identifier” column and for the rest value 0. This initial selection helped to focus only on the data which was supposed to be the basis for the creation of visitors’ behavior patterns.

The next step in the data preparation and selection process was to divide the whole feature class into two groups representing hikers and bikers. This classification had to be made, because the hikers and bikers represent different behavior patterns. Hikers tend to move freely around the research, mostly due to the lack of limitation which is for example a bike. Bikers on contrary tend to move much faster and concentrate on trails which are particularly prepared for them. Apart from those obvious differences the managers of the project Mafreina wanted to create this classification in order to analyze how different visitors groups influence the surrounding nature.

After this classification was made other important steps of the process could be initialized. At this moment every group contained only selected points representing just one activity type. The next analysis had to find for each point in the feature class the nearest trail. The analysis created for each point attributes NEAR\_FID and NEAR\_DIST. First attribute represent the ID of the nearest trail and the second the distance from this point to the trail. This attributes were created with the help of ArcGIS tool ‘Near’.

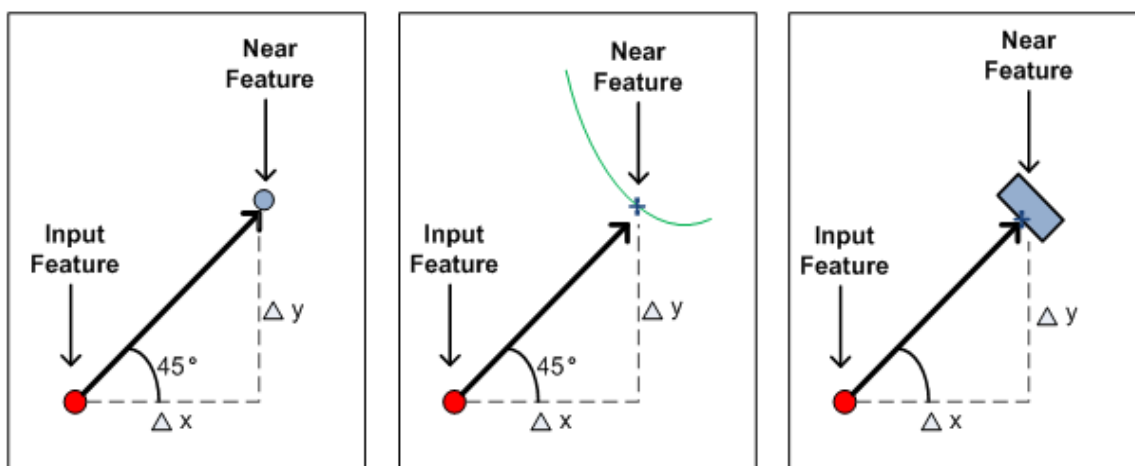
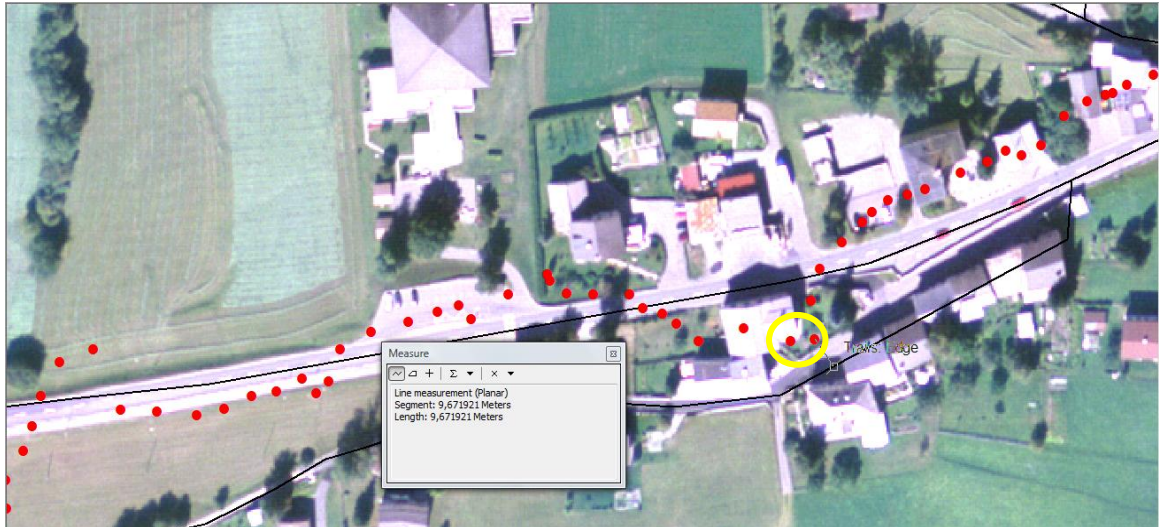
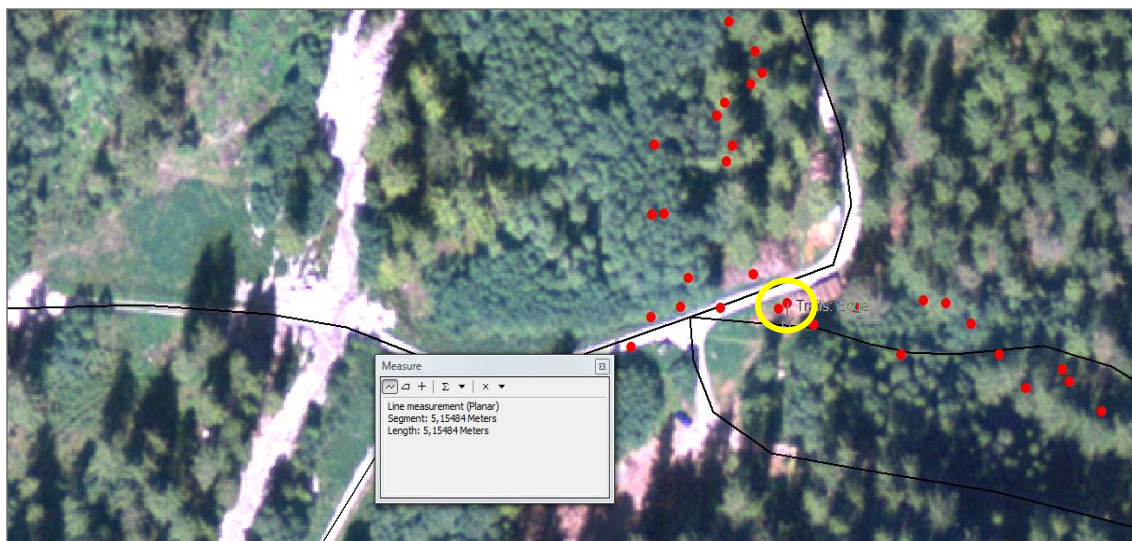


Figure 13: Concept of the “Near” function in ArcGIS, source: [www.esri.com](http://www.esri.com)

The main problem with this function is that it calculates the distance to nearest but not always to the proper trail. It may happen that the point according to the tool was supposed to lie on the trail number 1 but in the reality it should be on the trail number 2. This situation occurs at crossroads or when two parallel roads lay close to each other.



**Figure 14: Yellow circle indicating problem with the miscalculation of the nearest trail when two roads are parallel (black line). Point in the yellow frame should be closer to the upper road but according to the “near” function lower road is the one closer to the point.**

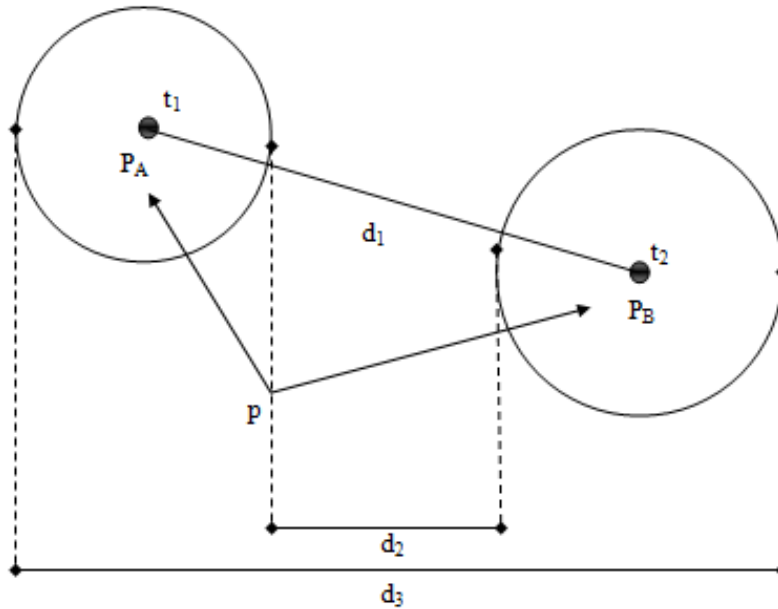


**Figure 15: Yellow circle indicating problem with the miscalculation of the nearest trail (black line) at a crossroads. Point in the yellow frame should be closer to the lower road but according to the “near” function upper road is the one closer to the point**

Ways to solve the problem with finding appropriate trail were introduced by many researchers (Chung and Shalaby 2005; Greenfeld 2002) which used in the calculations the information from previous points (sequence or history of GPS points), movement direction and ID of the trail to which points were assigned. Those scripts had a one thing in common, they all assumed that the people or cars were on the trail. In the project

Mafreina an assumption that the visitors always followed the trails cannot be made. Therefore the information from previous points cannot be used as an indicator while finding the nearest trail for the next point. This problem is known to the managers of the project and they agreed to include it in the data. Fortunately, due to large amount of data, this problem is not expected to have a meaningful influence on the further results of the analysis

After the distances have been calculated next step of data preparation and selection was the issue of speed values. For the managers of the project the speed values were very significant, especially in terms of further analysis. Yet the existing values have been calculated on the basis of distance and time values, which from the assumption were wrongly calculated. The speed values were a result of a division of a false distance between point A and B and the time difference when the points A and B were recorded. Then the calculated values were assigned to the consecutive point but the calculations omitted the issue of point's positional accuracy. Positional accuracy can vary for each point, factors like number of visible satellite and their configuration can have a big influence on the this value. During the day depending on the conditions, when and where the point was recorded, this value can change significantly. In addition to this issue, there was a problem with outliers which were spotted in all regions of the research area. Outliers are points which due to some errors in the GPS device, low number of satellites or their bad configuration have been recorded far from their original position. To ensure that this type of data will not influence the results a Python script had to be written.



**Figure 16: Comparison of the distances between point A and B.**

$t_x$  – time when the point x was recorded

$p$  – positional accuracy buffer

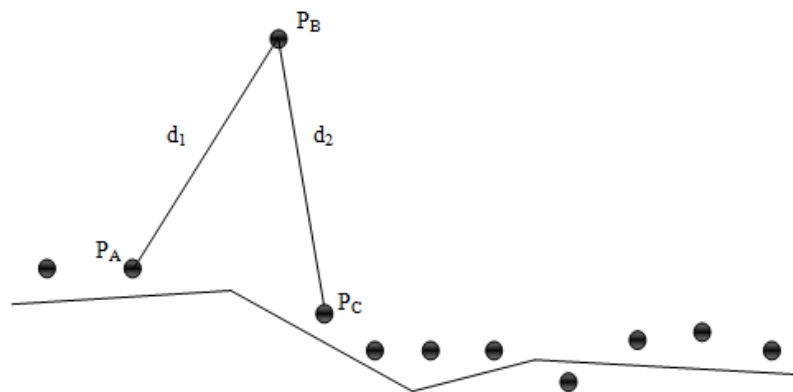
$d_1$  – distance calculated on the basis of coordinates of point A and B

$d_2$  – minimum distance between point A and B selected using positional accuracy buffer

$d_3$  – maximum distance between point A and B selected using positional accuracy buffer

$F(v_1)$  – original speed

Problem with the positional accuracy was not the only one which the managers of project Mafreina had to overcome. Second problem was connected with the so called outliners.



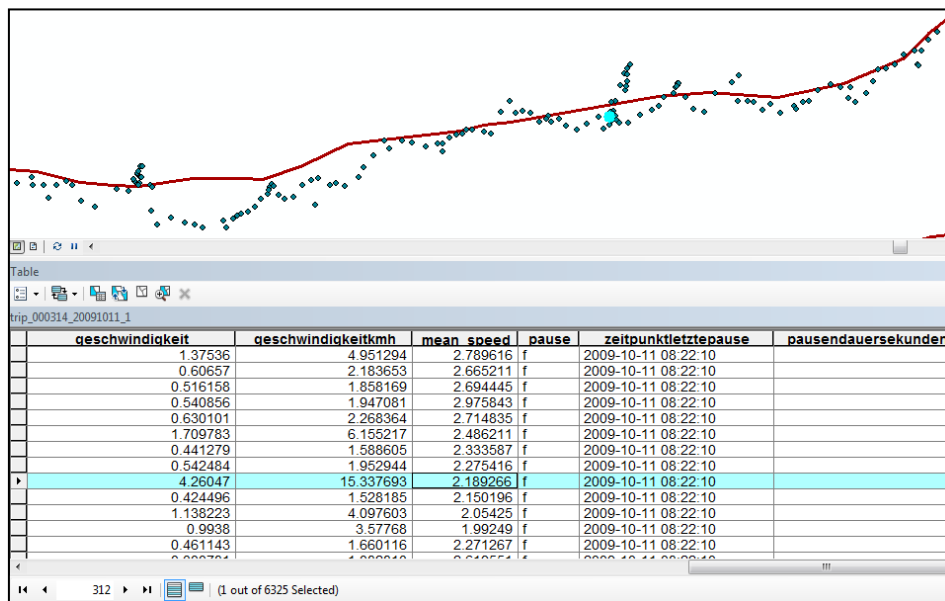
**Figure 17: Example of outliners.**

$P_B$  – outliner.

$d_1$  – distance between point A and B

$d_2$  – distance between point B and C

The figures 16 and 17 emphasized two basic problems which had to be resolved with the script. In order to minimize the problem with the positional accuracy of the points for each point five consecutive and five previous points were selected. On the basis of the speed values from those 11 points a mean value had been calculated and assigned to the selected point in the new column MEAN\_SPEED. During the calculation of the mean value two maximum speed values were found and omitted in order to eliminate the problem of outliers.



**Figure 18: Results of Python script.**

Gathering speed values from eleven points aimed to limit the problem of positional accuracy as there is no better described way to recalculate the speed values. The script was based on an assumption that the mean value from a group of nine points can be more reliable than a value from one point, which is more affected by errors in positional accuracy. Additionally the values for outliers were recalculated in order to use them again in the analysis.



**Table 4: Statistical summary of the most important attributes from feature classes “Bikers\_final” and “Hikers\_final”.**

	Hikers	Bikers
Number of points	2747615	540001
Mean HDOP	1.74	1.74
Std.dev HDOP	0.84	0.79
Mean number of satellites	7.34	7.13
Std.dev number of satellites	1.25	1.20
Mean speed [km/h]	5.63	8.95
Std.dev speed [km/h]	13.3	11.42
Mean recalculated speed [km/h]	4.95	8.03
Std.dev recalculated speed [km/h]	9.61	9.53
Mean distance to trail [m]	15.15	7.13
Std.dev distance to trail [m]	71.19	12.89

The number of GPS points selected for the analysis is five times higher for hikers than for bikers. The values for mean HDOP and number of satellites are nearly equal for bikers and hikers. A correlation between a high number of satellites and low HDOP, discussed in previous chapters, can be seen in the table. The difference between mean speed values for bikers and hikers is only 3.32 km/h and for the speed values calculated using the python’s script it is 3.07 km/h. It was expected that the values will be much different, but the difference between them is not that noticeable. This situation is influenced by the fact that hikers during their daily trips also travelled by car or by bus. This is acknowledged by the high value of standard deviation. It indicates that the speed values for hikers are not regularly disturbed and that there are values much higher than the mean speed. This fact was not taken under consideration during the preparation of metadata file. Therefore some trips registered as hiking trips were in fact also car or bus trips. Speed values recorded from those trips have definitely influenced the mean speed value in feature class “Hikers\_final”. To ensure that those values will not influence real hiking values, data needs to be precisely analysed. On the contrary the standard deviation for bikers underlines that the bikers drove with very different speeds. Those values can be sometimes much higher than the mean speed value e.g. 50km/h when they are riding down a steep road. However in the data analysis process it needs to be defined whether hikers riding on roads with high speed are an import element of the final results.

The last values in table indicate the mean distance to the trail. The values for hikers are twice higher than for bikers and standard deviation is five times higher. This can mean that the hikers tend to leave the trails more often than the bikers. On the contrary this can mean that the data for hikers is less precise than for bikers. Both of those assumptions need to be analysed in the next chapter.

The final step of data preparation and selection was to add new fields “OBJECTVAL\_LAN”, “TRAIL\_TYPE” and “HIKING\_TYPE”. “OBJECTVAL\_LAN” indicates a landcover type where each point was recorded. Using the “Intersect” function in ArcGIS information about the landcover type can be added to all points. This function selects from the input features those which overlap each other and saves them with chosen attributes. This way landcover data will be used in further steps of the analysis to check how it can influence the visitors behavior.

“TRAIL\_TYPE” is an attribute which was created on the basis of trail ID. Every trail segment has its own unique ID and also information about the trail class. Using “Join Field” function in ArcGIS new field was added to each point with the type of the nearest trail. Also using unique ID “HIKING\_TYPE” was added to the data.

### **4.3. Summary of data preparation and selection**

The final result of the data preparation and selection process was the creation of two feature classes representing bikers and hikers. First feature class contained 540001 GPS points representing various biking trips. Second feature class contained 2747615 GPS points representing different hiking trips. There are exactly 3287616 points from both feature classes which means that during data preparation and selection process 1635086 points have been excluded from further analysis. Additionally in the data analysis process points representing hikers and bikers need to be analysed again, which means that next points can be excluded.

Figure 19 and 20 show the Müstair Valley and the distribution of the GPS points for both feature classes. They cover nearly the whole research area and represent trips running through many different landcover and trail types.

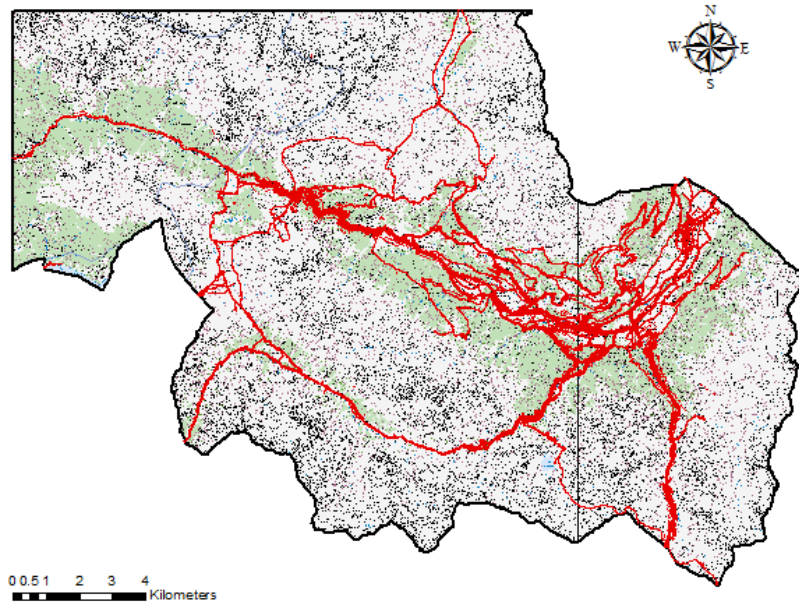


Figure 19 Müstair Valley covered with GPS tracks from the feature class representing bikers.

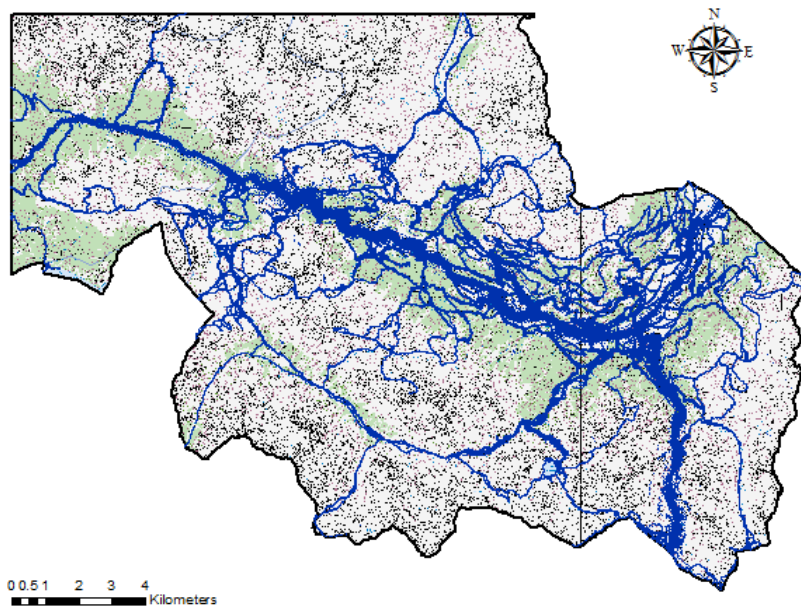


Figure 20: Müstair Valley covered with GPS tracks from the feature class representing hikers.

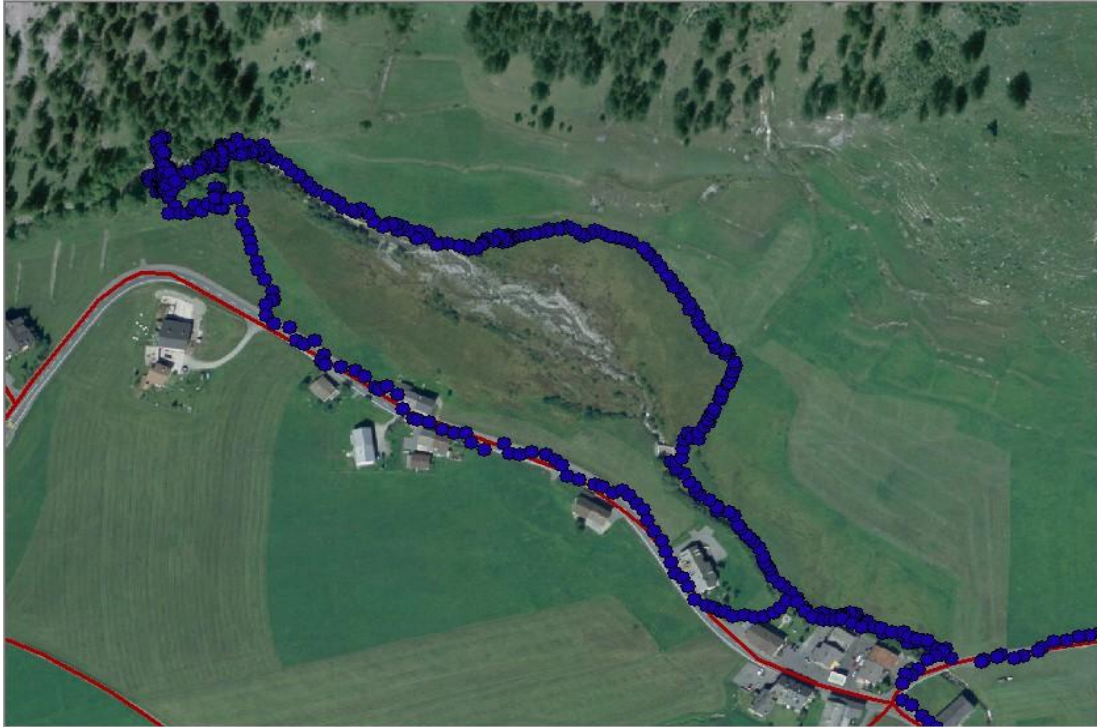
## **5. Modeling and analyses of the GPS data**

After the data preparation and selection processes were finished an answer to question, how to analyse so big amount of data, had to be found. As it was emphasized in previous chapters managing so big amount of data is a complicated task. There are many different situations that need to be analysed in order to create adequate analytical tools. Conclusion about how the data needs to be analysed cannot be drawn only basing on two or three examples. The more situations are analysed the more advanced tools can be created. Perfect analytical tools cannot be created as there will always be situations where no adequate solution can be found.

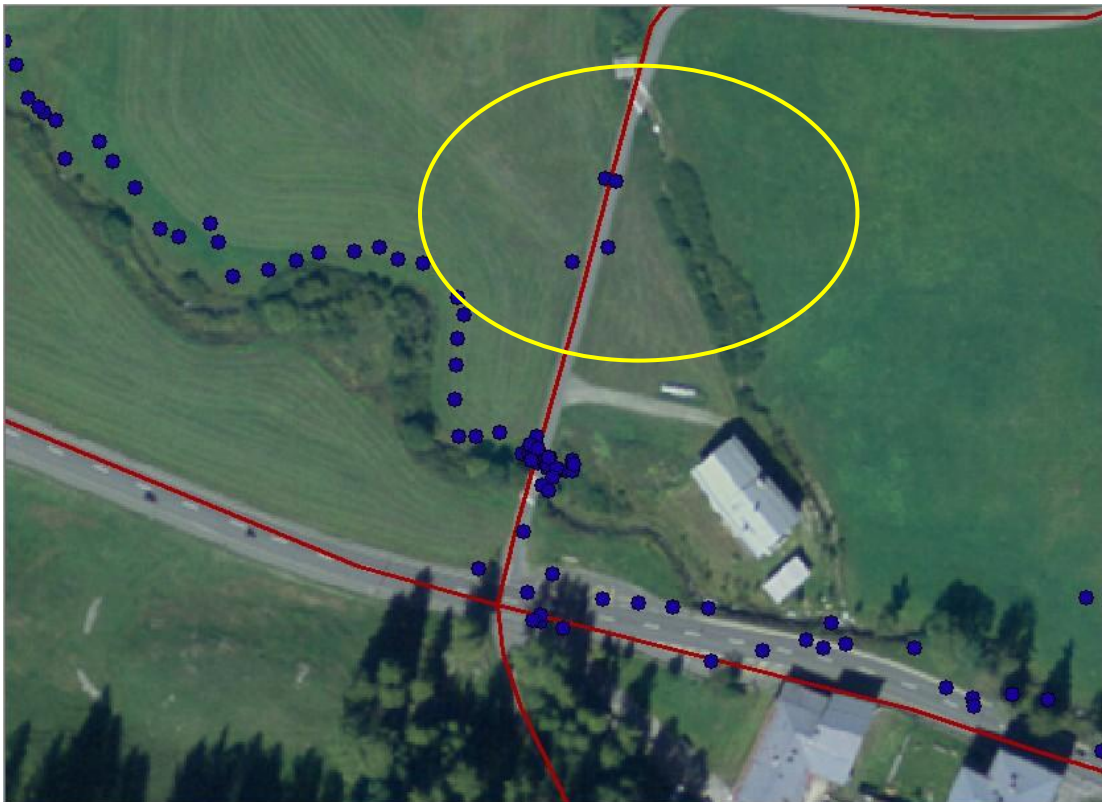
### **5.1. Visual analysis of the data**

First step of the data analysis process was the visual analysis. The visual analysis had to be complemented with statistical analysis in order examine a vaster range of data. This allowed choosing a method, which would examine the data in order to answer a question regarding visitors behaviour patterns.

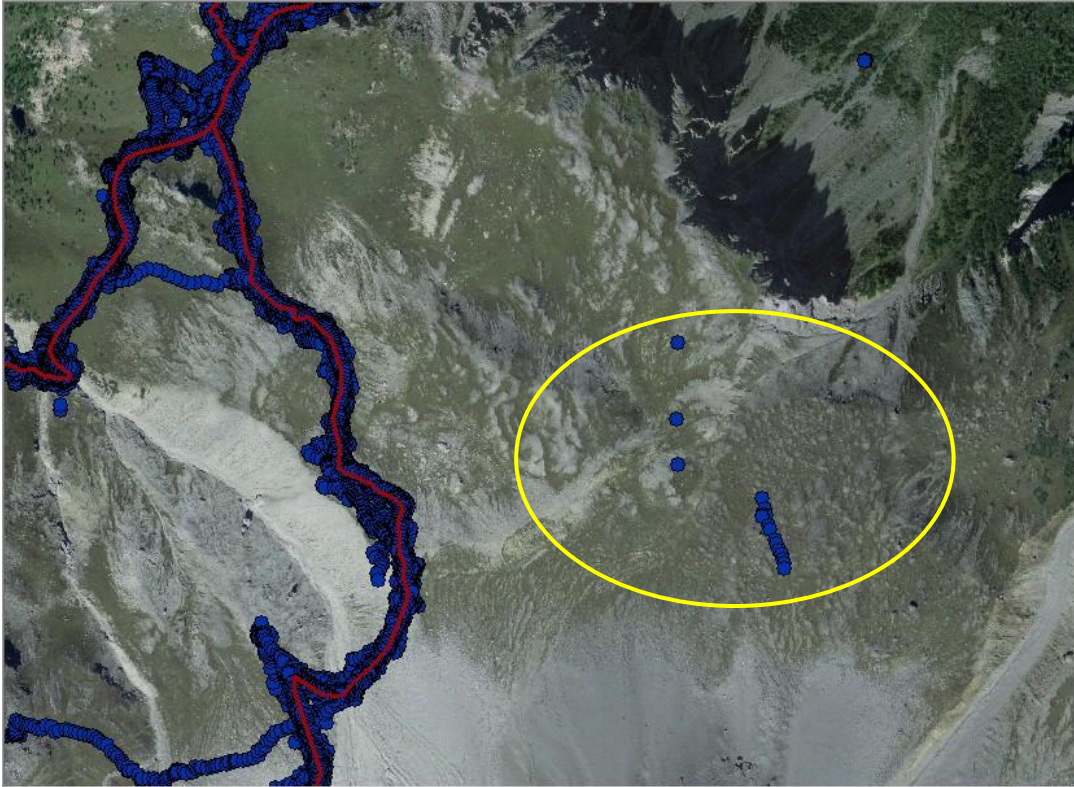
Data in the feature classes “Bikers\_final” and “Hikers\_final” represented many different tracks which the visitors followed. Sometimes the way that the visitors moved was unique which means that they did not followed the existing network of trails and roads, instead they moved freely around the research area. On the contrary many visitors did not leave the existing network and therefore helped to indicate potential movement trends. Figure below show examples of various situations which have been spotted in the research area. From the selected regions points had been identified and analyzed for better understanding how different attributes influence the location of each point.



**Figure 21:** Blue points representing a hiking trip. Points in the upper left corner represents hikers who initially left the trail



**Figure 22:** Blue points representing a hiking trip. Yellow circle indicating potential deviation from the original track.



**Figure 23: Blue points representing a hiking trip. Yellow circle indicating potential deviation from the original track.**

Figure 21 shows an area where the hikers left the road and decided to walk around a marsh or meadow. All the points representing this path have a HDOP value below 2.5, speed value below 4km/h and high number of satellites ranging from 5 to 9. Points which are located on the roads have similar values to the other points. The attribute which helps to distinguish the first group of points from the second is the distance to the trail. On this basis it can be determined which points are on the trails and which are not. However in this example basing only on the distance attributes is enough, next figures will show that the problem is more complex.

In the figure 22 the yellow circle indicates a new group of points representing outliers. The problem with outliers is that they very often lie close to a trail, not always the proper trail and their speed values are close to 10km/h. Attributes which can help with their selection are HDOP which is in the range from 3.5 to 6 and the number of satellites which is equal 5 or 6. In this example is it hard to decide whether the visitors really changed their movement direction so rapidly and then returned to the main path or are does points really outliers. In some places of the research area it easier to determine if the point is an outliers or not. Those point lie far from the trail and in most cases these are single points.

The figure 23 shows a group of points in the yellow circle which obviously are outliers. They are located far from the line, there are no other points which could suggest that this is a part of a larger path and additionally they have HDOP higher than 5 and number of satellites equal 5. Those points leave no doubts to which group they need to be assigned. However there are many other points which lie very close to the trail and have similar values with the expect of “NEAR\_DIST” value. Furthermore there are points indicating similar values and distances to nearest trail but they are a part of path representing visitors who initially left the trail.

On the basis of those three figures which represent very common examples from the whole research area first conclusion can be drawn. The points can be classified into four groups:

- points on the trails with attributes indicating high positional accuracy
- points on the trails with attributes indicating low positional accuracy
- points not on the trails with attributes indicating high positional accuracy
- points not on the trails attributes indicating low positional accuracy (outliners)

Attribute “MEAN\_SPEED” which was created using the python script, also helps to determine whether a point can be marked as on the trail or not. According to managers of the project from the speed values it can be assumed that the hikers or bikers were on the trails. When a hiker starts to move faster than 10 km\h it is nearly sure that he started driving a car or took a bus, which means that he is on the road. When a biker begins to ride faster it can be also assumed that it safer for him to follow the trail and therefore more probable that he is actually on the trail.

The visual analysis indicate that some attributes can be used to measure whether the points are on the trail or not and other attributes should be only used to verify positional accuracy. Those attributes are HDOP and number of satellites. Attributes which help to determine if the point was on the trail are “NEAR\_DISTANCE” and “MEAN\_SPEED”. They should only be used in the calculations as they directly inform about the position of the point in reference to the trail or suggest the possible position.

## 5.2. Spatial data mining

According to the subsection 3.2.1. spatial data mining this process which helps to discover interesting and previously unknown but potentially useful patterns in large spatial datasets. After the visual analysis were done it was necessary to perform statistical analysis. The values for attributes “MEAN\_SPEED”, “HDOP”, “NUMSAT” and “NEAR\_DISTANCE” were analysed in order to indicate potential trends, correlation, situations where the correlations or trends are more noticeable and where they do not occur. Crucial issue was the question how those attributes influence the location of each point and if they can help to determine which points can be marked as those on trails and which not.

First analyzed feature class was “Hikers\_final”. During the data preparation and selection process unusually high value of speed standard deviation value has been spotted. According to the table 4 standard deviation equaled 9.61 km/h and the mean speed 4.95 km/h. For hikers which normally walk with a speed around 2-6km/h so high standard deviation could only mean that the hikers also used different means of transport. In order examine this assumption a histogram had to be created.

Histogram representing hikers speed of movement was based on 51 intervals each indicating speed of 2 km/h. Data represented in the histogram had been prepared with the help of Python script.

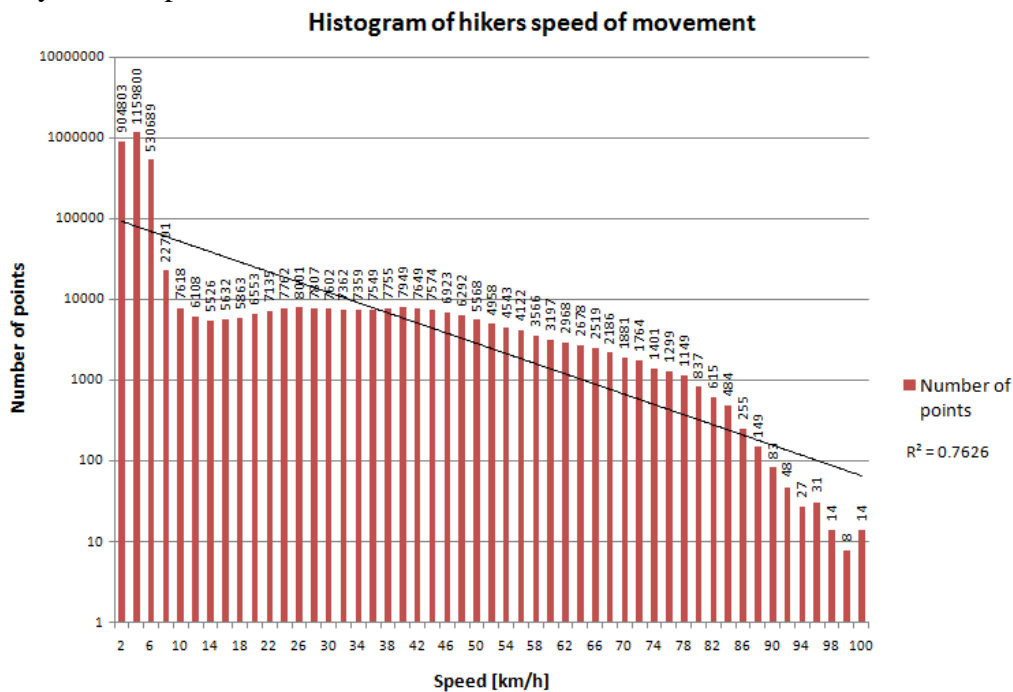


Figure 24: Distribution of hikers speed of movement.



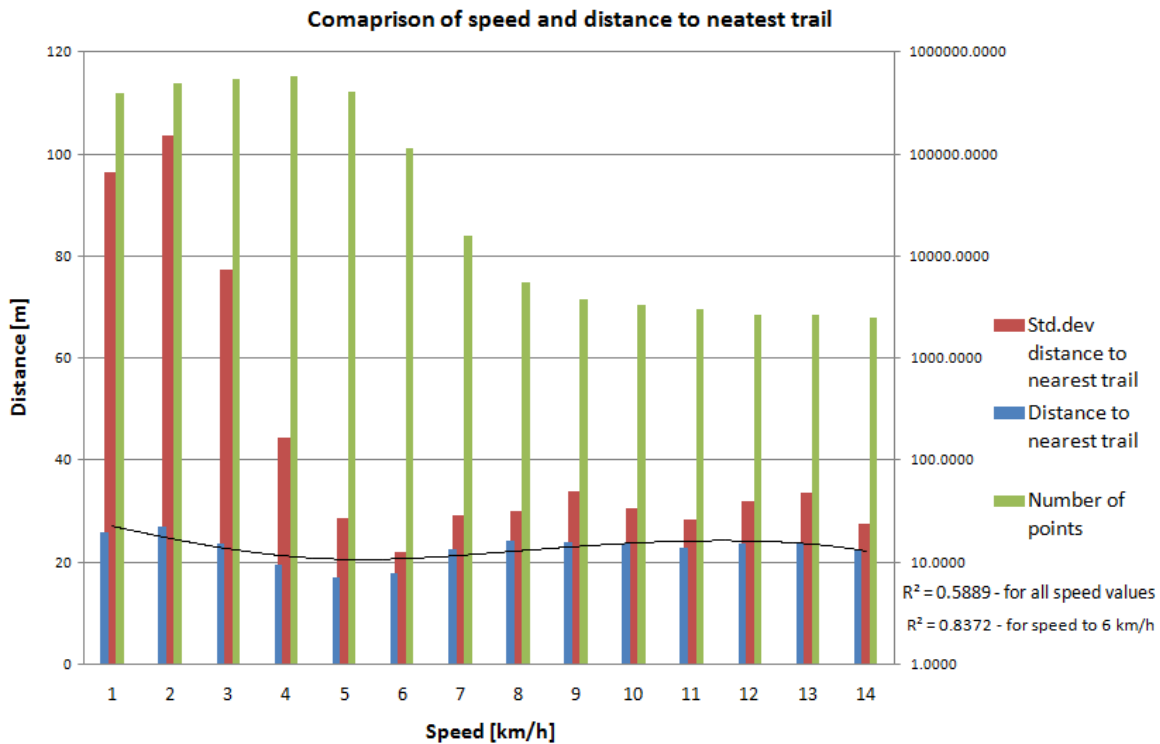
First histogram indicates that the distribution is not symmetrical and it is clearly skewed to left. Below the speed of 8km/h lie 93.2% of all points. This number was expected as the majority of points indicate the speed with which the hikers normally walk. The highest result of coefficient of determination was indicated for exponential trend line. It clearly demonstrates that with the increase of speed the number of points is rapidly decreasing. Further analysis of histogram indicates that a local minimum in the number of points can be noticed around the speed of 14km/h. Before the histogram was analyzed an assumption had been made that exactly around 10-14km/h a local minimum in the data should be observed. It was expected that the hikers around those speeds should start driving by car or by bus. As it can be seen in the histogram the number of points increase from 14km/h and decreases with the speed of 40km/h. This situation can be interpreted as an example of using public or private transport.

The whole histogram is in fact a combination of two types of data. Hikers who move on foot or travel by bus or by car. This division is very rational as the hikers sometimes need to take a bus to travel to or from some remote areas. The same situation concerns those hikers which travel by car. Additional question is how should be explained the distribution of speed values from 8km/h to 14km/h. This question can be answered with the help of figure 17. It demonstrates the concept of positional accuracy buffers and the uncertainty connected with the exact location of each point. If a point is located just a couple of meters from its true location than this difference can have a noticeable influence on the speed value. So when two points represent a medium level of positional accuracy the speed even after recalculating it with the script can still be false. Therefore those points should not be associated with hikers travelling by bus or by car. An extra explanation can be the fact that some visitors jogged during their hiking trips which might also have an influence on the results.

The goal of the analysis is to examine the behavior patterns of hikers and bikers. Therefore 162095 points which represent hikers moving faster than 14km/h have been excluded from further analysis. A new mean speed value for hikers is 2.78 km/h and standard deviation 1.58 km/h. Those values are more reliable because they correspond with actual hiker's speed of movement.

**Table 5: Statistics for the speed values**

Speed	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Percent of total [%]	15.2	18.9	21.2	22.7	15.9	4.3	0.6	0.2	0.1	0.1	0.1	0.1	0.1	0.1
Mean distance	19.3	22.2	15.4	9.5	7.1	7.7	13.4	16.3	15.8	15.3	13.7	15.3	15.1	13.4
Standard deviation	96.3	103.7	77.3	44.2	28.6	22.1	29.3	29.9	33.7	30.6	28.3	32.0	33.5	27.5

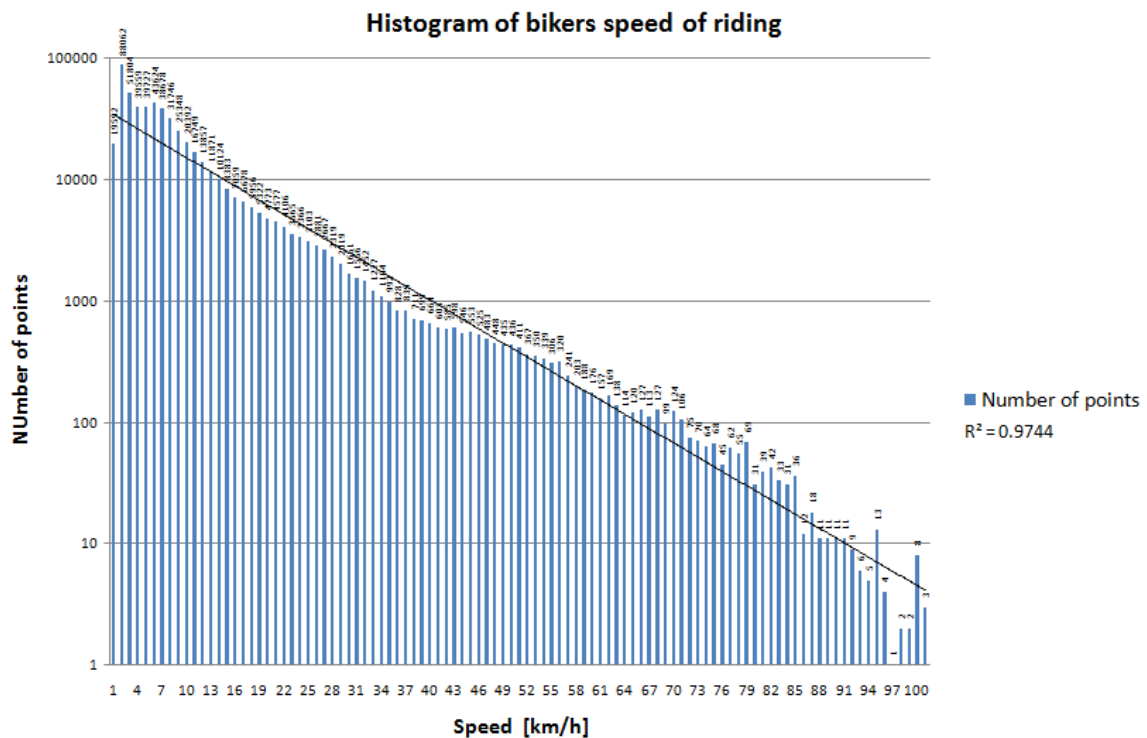


**Figure 25: Relation between different speed values and the mean distance to the nearest trail.**

Chart presented on the figure 25 show the relation between speed values and the mean distance to nearest trail. The highest mean distances 22.20 m to nearest trail is indicated for the speed below 2km/h and lowest 7.11 for 5km/h. The highest standard deviations are recorded for the speed values to 4km/h which mean that those values represent visitors on the trails as well as those who did not follow the trails. Therefore they should not be used in estimating visitors location. Additionally value higher than 6 km/h also should not be taken under consideration as they only represent 1.53 percent of all points. An Order 3 polynomial trend line indicates coefficient of determination for all speed values at the level of 0.5889 but for speeds lower than 6 km/h the exponential trend line represents coefficient of determination equal 0.8372. This is a very high value which underlines that with the increase of the speed the distance to nearest trail decreases. Apart from mean distance also standard deviation is decreasing. Second coefficient of

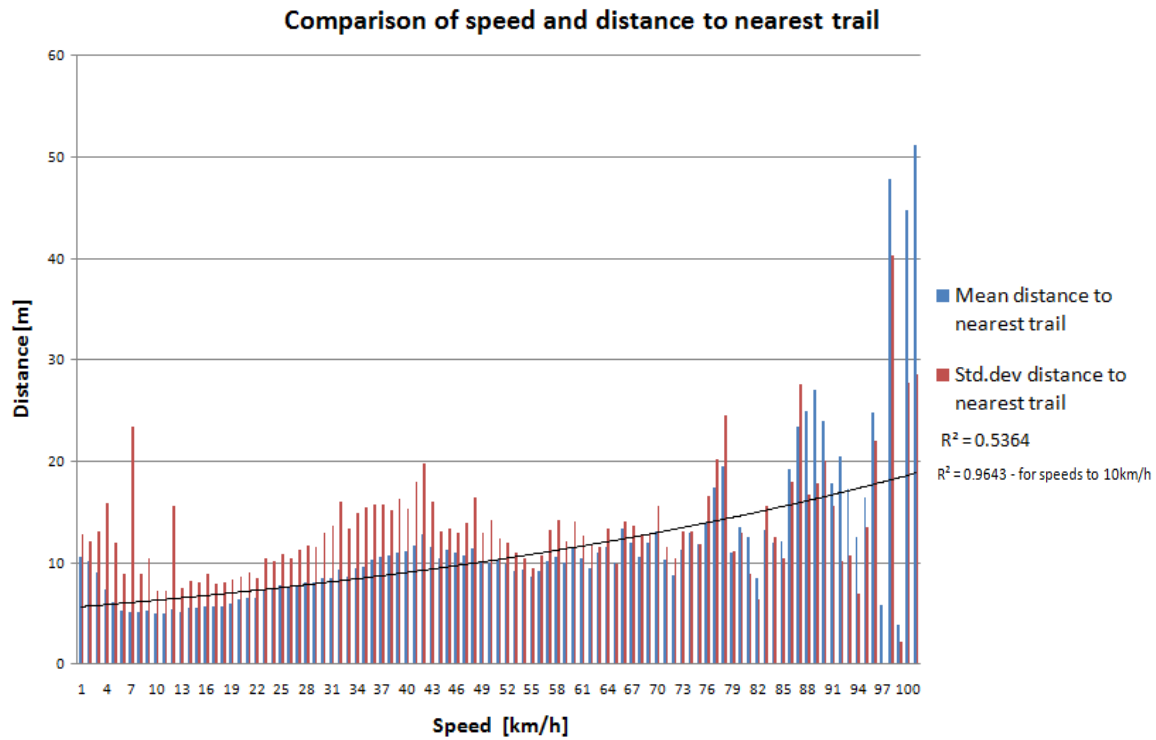
determination in comparison to the first one can interpreted as a reliable value because it is based on nearly 94% of data. Each speed is represented by 15 to 22 percent of all points which leads to a conclusion that hikers speed of movement from a specific value can help to decide whether a visitor was on the trail or not. Points indicating speed lower than 6 km/h should not be taken under consideration due to high standard deviation but for remaining points this assumption can be made.

Second analyzed feature class was “Bikers\_final”. Mean speed value for bikers was 8.03 km/h and the standard deviation 9.53 km/h. Knowing the way that the bikers ride it was expected that those values will be higher than for hikers. In order to analyze the distribution of speeds a histogram had to be created. In the histogram bikers riding speed was visualised on the basis of 51 intervals each indicating speed of 2km/h. Data represented in the histogram had been prepared with the help of Python script



**Figure 26: Distribution of bikers riding speed**

Histogram representing distribution of bikers riding speed is noticeably skewed to left. It indicates that 85% of all points lie under 15km/h. This means that the bikers do not ride so fast in the research area. 15% of data indicates that the bikers move faster than 15km/h sometimes up to 100km/h. It is possible especially when they are riding down a steep hill or road. Nevertheless only 1.91% of all points represent speed values higher than 40km/h.

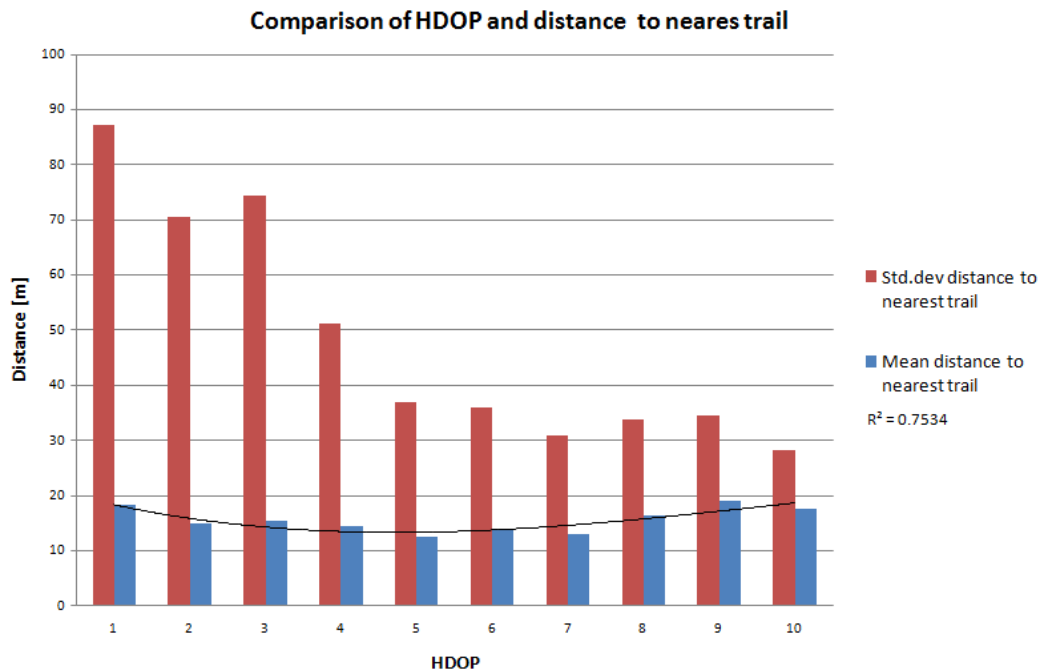


**Figure 27: Relation between different speed values and the mean distance to the nearest trail**

Chart presented on the figure 27 describes the relation between different speed values and the mean distance to nearest trail. Values to 20km/h represent 91.51% of all points but only values lower than 10km/h can be used for reliable analysis as they represent 77% of all points. As it can be seen from the chart values higher than 70km/h indicate very diverse mean distances and standard deviations. Due to low number of points representing those values and high uncertainty concerning their positional accuracy they will be omitted in further steps of the analysis. Exponential function indicates coefficient of determination equal 0.5364 but it cannot be recognized as reliable source of information. Points indicating speeds higher than 40 km/h are only 1.91% of all points and as it can be seen from the chart they have an influence on the final result of coefficient of determination. However for speeds lower than 10km/h, which have a reliable amount of data, an Order 2 polynomial trend line represents coefficient of determination equal 0.964. It means that with the increase of speed the distance to trail decreases. It causes that the bikers follow the trails or roads whenever they start to ride faster. According to the figure 28 it has been stated that speed values higher than 40km/h will be omitted from further analysis. Firstly they represent unreliable values due to low number of points and secondly so high speeds can be only reached onroad, which do not account for the area of research.

**Table 6: Statistics for HDOP values**

HDOP	1	2	3	4	5	6	7	8	9	10
Percent of total [%]	11.99	69.69	13.82	2.28	0.96	0.54	0.38	0.16	0.10	0.07
Mean distance	18.43	14.92	15.50	14.39	12.47	13.68	13.04	16.37	19.07	17.60
Standard deviation	87.17	70.46	74.32	51.26	36.95	35.93	31.01	33.84	34.52	28.18

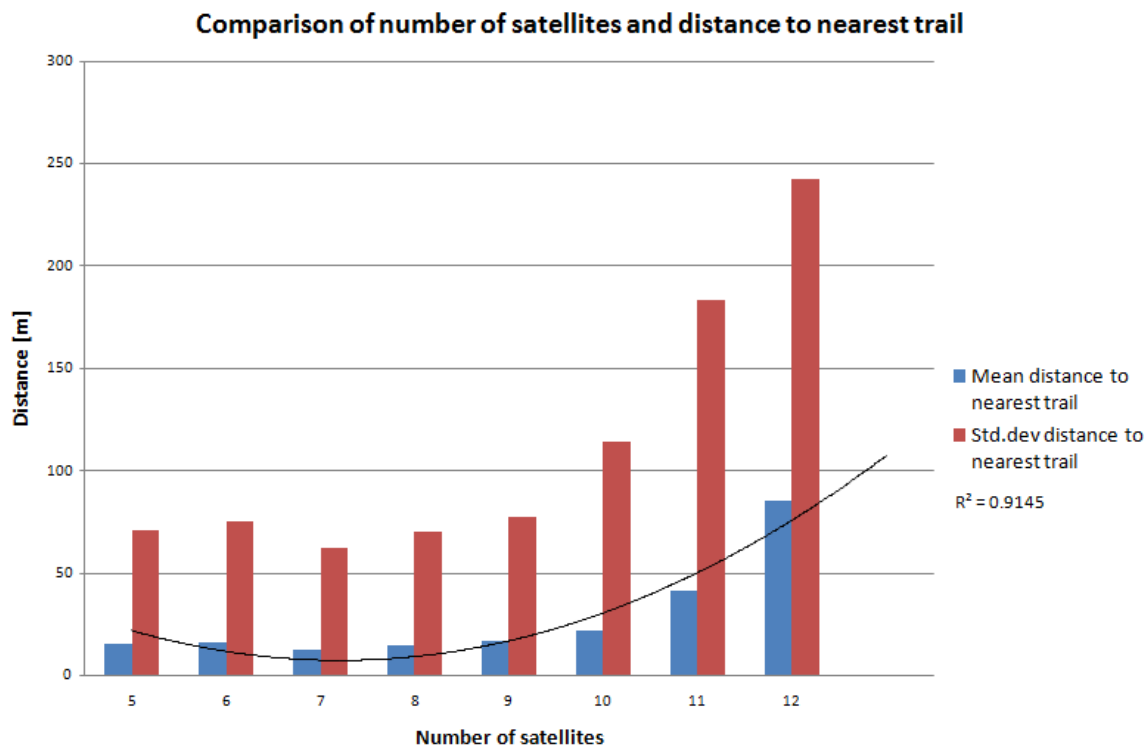


**Figure 28: Comparison of different HDOP values and the mean distance to the nearest trail.**

Chart from the figure 28 shows the relation between different HDOP values and the mean distance to nearest trail. The lowest mean distance value 12.47 m has the HDOP value 5 and the highest 18.43 m HDOP value 1. Mean distances for other HDOP values are very similar which indicates that HDOP cannot be used as factor measuring if the tourist were on the trail or not. An Order 3 polynomial trend line indicates coefficient of determination is 0.7534 for the mean distance to the nearest trail. This is a good value however it cannot be used to determine the distance to the nearest trail. HDOP values higher than 5 should not be taken under consideration as the number of those points is lower than 2.22 % of total. Additionally standard deviations for the HDOP values till 4 inform that the distances are very diverse, which mean that those values concern visitors on the trails as well as those who left the trails.

**Table 7: Statistics for the number of satellites**

Number of satellites	5	6	7	8	9	10	11	12
Percent of total [%]	7.00	17.76	29.14	27.24	15.12	3.27	0.43	0.04
Mean distance	15.47	15.99	12.98	14.67	17.19	21.70	41.31	85.47
Standard deviation	70.72	75.48	62.73	69.98	77.77	114.56	183.38	242.74



**Figure 29: Comparison of different number of satellites and the mean distance to the nearest trail.**

Chart from the figure 29 represents the relation between different number of satellites and the mean distance to nearest trail. Mean distances for the number of satellites higher than 10 and lower than 5 should not be taken under consideration as there are too few points that could create reliable values. Percent of the points which indicate 12 satellites is only 0.04 and for 11 satellites only 0.43. The lowest mean distance to the nearest trail 12.98 m was recorded for 7 satellites, the highest 85.47 for 12 satellites. Values of the standard deviation range from 62.73 m to 242.74 m, which confirms the assumptions that the number of satellites also cannot be used as factor indicating whether people were on the trail or not. An Order 2 polynomial trend line represents coefficient of determination equal 0.9145 however due to insufficient number of data for some values, it should not be used in drawing final conclusions.

The analysis of the influence of HDOP and number of satellites on the distance to the nearest trail for bikers were not performed. According to the results for hikers those two values do not correlate with the distance to the nearest trail. The analysis for hikers can be recognized as reliable because they concern similar values and they are based on reliable data. However the speed values need to be compared with the distances to nearest trail.

The visual and statistical analyses emphasize the complexity of the research problem. There are many situations in the research area, where it can be clearly stated that the point is on the trail or not. Attributes like speed, HDOP, number of satellites and distance to the nearest trail help to evaluate each point but the question is how the results of this decision process should be presented. Should the final answer be Boolean, in this case, yes the visitors were on the trail or not. Maybe it should be non-boolean and just expressing the probability that they were on the trail or not. This question needs to be profoundly analyzed to make the results easy to interpret especially that in some regions number and distribution of points will require advanced analytical and visualization techniques.

### **5.3. Multi-Criteria Evaluation – Boolean and Weighted Linear Combination approach**

According to Janssen and Rietveld (1990), Jankowski (1995) fast development of GIS led to noticeable improvements in its capability for decision making process, especially in environmental management. Multi-Criteria Evaluation (MCE) is considered as one of the most important procedures. In the context of GIS two approaches of MCE are very common. First is based on the concept of Boolean algebra where all criteria are assessed by using various thresholds. Those thresholds are used to indicate whether some values are right or wrong, if they meet the criteria or not. Those values are then combined using logical operators like intersection (AND) or union (OR). Second approach is the Weighted Linear Combination. In this concept continuous criteria are standardized and combined by weighted averaging. In both approaches there are two kinds of criteria, factors and constraints. According to Eastman et al. (1993) factors signify a continuous degree of fuzzy membership in the range 0-1 and constraints that are mostly used to limit the alternatives together e.g. fuzzy membership is either 0 or 1.

Jiang and Eastman (2000) suggest that both methods may bring very different results and there are some fundamental problems associated with their usage. Those problems are connected with different aggregation methods, particularly with tradeoff. In the Boolean method aggregation relies on the intersection (AND) or union (OR) concept. In the first example, the results had to meet all the criteria to be considered as reliable. In the second example only one criterion had to be met, which increased the number of potential solutions. On the contrary to Boolean approach Weighted Linear Combination allows to compensate low scores on one criterion by a high score on another. This is the feature known as tradeoff or suitability.

Second problem with the usage of Weighted Linear Combination concerns the standardization of factors. In some cases standardization has to take linear or sometimes non-linear form. The way that the factors will be standardized depends on the type of data and some assumptions that were made by e.g. the managers of a project. Standardized factors express the measure of suitability and therefore the higher the score the more suitable is the factor. A typical type of continuous factor is distance, which should be standardized before implementing it in the Weighted Linear Combination. For example points which lie close to the trail should receive high scores and those which are far away from the trail low scores. Those scores should be based on numerical range e.g. from 0 to 1 or from 0 to 255. However whether point is located 100m or 150m from the trail does not make any difference and so non-linear standardization should be chosen. Similar approaches should be made for other factors used in further analysis with consideration of their data type.

In most decision making processes, multiple criteria are considered to assess the degree of suitability each location bears to be allocation under consideration. Thus suitability is commonly not Boolean in character, but expresses varying degrees of set membership i.e. fuzzy set. (Jiang and Eastman 2000).

According to Hall et al. (1992 ) there a many reasons why fuzzy set membership should be applied into the criteria standardization. Firstly they provide strong logic for the whole process of standardization, which can be interpreted as recasting values into a statement of set membership. Secondly standardization based on fuzzy set membership presents a strong relation between criterion and decision set. Thirdly fuzzy set does not require clear thresholds as it in the Boolean approach. Monotonically increasing or decreasing sigmoidal functions or J-shaped function can be used for the data standardization which ensures that no concrete threshold values must be specified.



For the next steps of the data analysis process a Weighted Linear Combination combined with fuzzy sets has been chosen. The complexity and amount of the data require a more advanced data evaluation than offers the Boolean method. Different attributes need to be standardized using a common numeric range 0 to 1. The standardization needs to be done differently for each attribute as they represent different factors. Those factors will have to be weighted according to their importance and modeled to check how different weights influence the overall results. Despite the level of complexity and in some cases uncertainty the final results should be more logical and easier to interpret than those created by using Boolean method.

#### **5.4. Fuzzy logic analysis**

According to the assumptions made in the subsection 5.2, attributes can be classified as those which indicate the level of positional accuracy and those indicating whether visitors were on the trail or not. Attributes describing the level of positional accuracy are HDOP and number of satellites. However only HDOP can be used for further analysis. The HDOP value is strongly correlated with the number of satellites. With the increase of number of satellites the HDOP value decreases and the positional accuracy is getting higher. However HDOP value not only relies on the number of satellites but also on their geometric configuration. This fact causes that HDOP is a more reliable source of information concerning positional accuracy. The second group of attributes includes distance to nearest trail which directly informs about the position of a visitor. Second attribute is speed which in some cases helps to determine possible location of the visitor.

In order to use those attributes in the Weighted Linear Combination they need to be standardized as they correspond to criteria type called factor. Those factors signify a continuous degree of fuzzy membership. Each factor needs to be standardized depending on the type of data and also type of visitors which it concerns. For example results of standardization process of speed values will be different for hikers and bikers. This method was partly based on the work of Ochieng et al. (2004) who suggested usage of regions of confidence but in general it can be seen as similar method.

After the factors will be standardized they need to be combined by an appropriate equation. For each factor a weight needs to be assigned according to its importance in the

decision process. Weights can be eventually changed in order to examine how they influence overall results.

#### **5.4.1. Fuzzy distance**

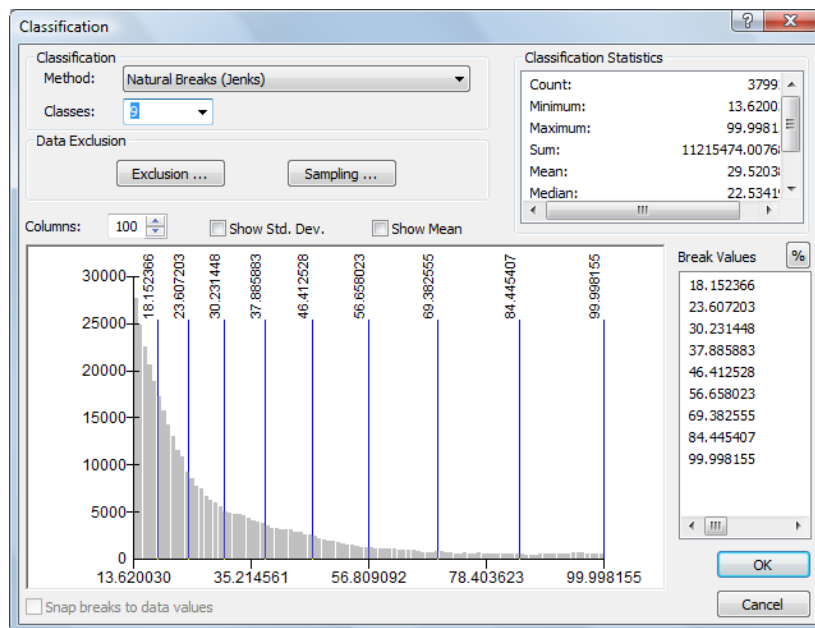
First factor subjected to standardization was distance to the nearest trail. Distance is a typical example of continuous data. Distance standardization needs to be performed separately for hikers and bikers because those groups present different movement behavior.

For the needs of the standardization the minimum and maximum values need to be determined in an advanced way not only basing on real maximum and minimum values. Threshold indicating the minimum value needs to result from the data accuracy. Data accuracy is based on two feature classes, one representing trails and second representing GPS points. Determination of trail positional accuracy is easy as this feature class was created on the basis of orthophotos delivered from Swisstopo. On the official website of Swisstopo it can be found that for orthophotos ground resolution is 0.5m and the standard deviation of precision in position varies from 3-5m in hilly terrain. On the basis of this information it can be determined that positional accuracy for vector data representing trails is 4.5m. The problem arises when the data accuracy needs to be analyzed for the GPS points. There is no fixed network of trails which the visitors need to follow which means they can move freely around the whole research area. Without a solid reference dataset it is impossible to analyze the positional accuracy of the points.

In the Swiss National Park adjacent to Münstair Valley similar problem had to be solved. The same GPS devices have been used to trace the way the visitors moved around the park. The project was held on a much smaller scale but its results are very beneficial to the needs of project Mafreina. The managers of the Swiss National Park were able to calculate the positional accuracy of the recorded GPS points due to two important factors. Firstly because the Swiss National Park has a fixed network of trails which secondly the visitors cannot leave. On this basis it could be assumed that the visitors were always on trails which position was precisely known to the managers. The method used for the calculation of positional accuracy was taken from Goodchild and Hunter (1997) who described simple positional accuracy measure for linear features. This method requires a representation of higher accuracy with which a low accuracy representation can be compared. Then the percentage of total points lying within a specified distance to the high

accuracy representation is calculated. For example if 95% of all recorded GPS points are lying in the 7m buffer around the trail, this means that with 95% probability the positional accuracy of this dataset is 7m. Goodchild and Hunter (1997) indicate that this method has three basic advantages; it is statistically based, relatively insensitive to extreme outliers and does not require matching of points between presentations. The results of this method indicated that with 95% probability the positional accuracy of GPS dataset in the Swiss National Park is 9.12m. Eventually this value could be added to positional accuracy of the digitized vector feature class representing trails and set as the minimum natural break value to the nearest trail. The distance of 100m was chosen as the maximum natural break value.

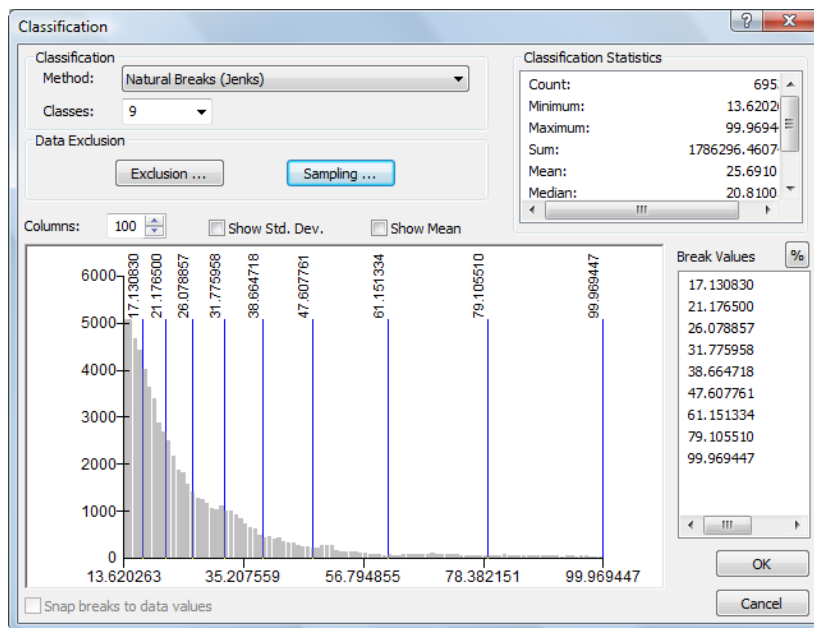
Points lying closer than 13.62m to the nearest trail were assigned to a fuzzy set 1 as they indicate the highest 100% probability that they were on the trail. On the contrary according to the decision of the managers of the project Mafreina points lying further than 100m had to be assigned to a fuzzy set 0. In order to assign other distances to remaining fuzzy sets a natural break classification method has been chosen. This method groups values within classes of similar values separated by breakpoints. Those breakpoints are chosen in a manner which maximizes the differences between classes. The creation of classes is base on minimization of the standard deviation from the class mean and maximization of each class deviation from means of other groups.



**Figure 30: Natural breaks data classification for the distances to nearest trail.**

The figure 30 represents 9 natural breaks which correspond to 9 fuzzy sets. The 10th fuzzy set corresponds to values below 13.62 m but it was excluded from the classification, as it was manually defined. Thus distances longer than 13.62 m but shorter than 18.15 m inform that with 90% probability the points were on the trail on the contrary distances longer than 84.44 m but shorter than 100m indicate only 10% probability that they were on the trail. Additionally figure shows that the width of classes is increasing with the increase of distance. This is caused by the decreasing number of points and noticeable differences in distances.

After the preparation of classes for hikers was finished the same procedure could be started for bikers. In classification process the maximum and minimum values have also been manually assigned. The minimum distance was also 13.62m and the maximum distance was 100m. The remaining distances had been analyzed with the help of natural breaks classification and assigned to appropriate fuzzy set.



**Figure 31: Natural breaks data classification for the distances to nearest trail.**

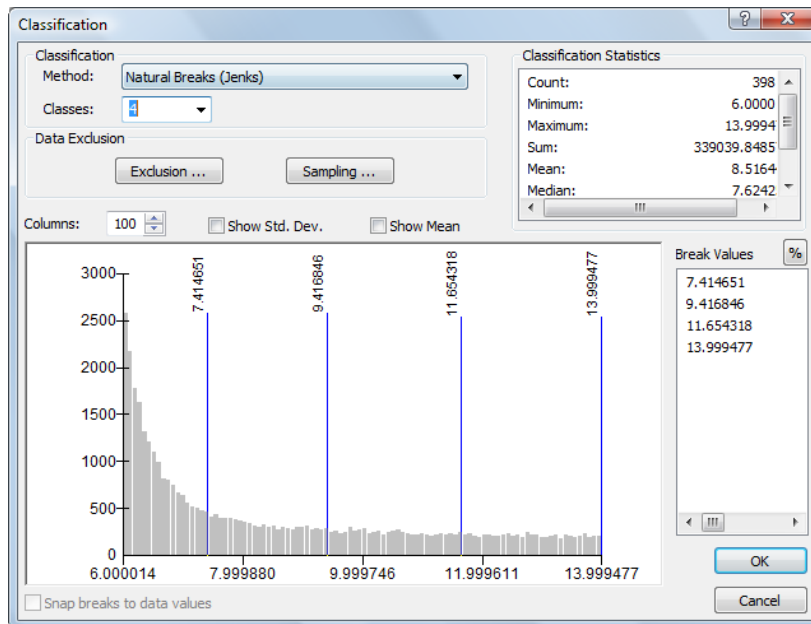
#### 5.4.2. Fuzzy speed

Second factor subjected to standardization was speed. The concept of standardization varies from the one suggested for hikers. Firstly both for hikers and bikers not all speed values have been taken under consideration. Feature class “Hikers\_final” represents only speeds lower than 14km/h and “Bikers\_final” speeds lower than 40km/h. Those values will be used as the highest natural breaks because they represent visitors

driving by car or bus and for bikers points which due to high speed are assumed to be on trail or road.

Secondly according to figures 25 and 27 the coefficient of determination indicates for hikers and bikers that with the increase of speed the distance to nearest trail decreases. However due to high standard deviation especially for low speeds they need to be omitted in the standardization process. This assumption is especially important because some visitors moving with the speed of 2km/h could be fallibly assigned to a wrong fuzzy set.

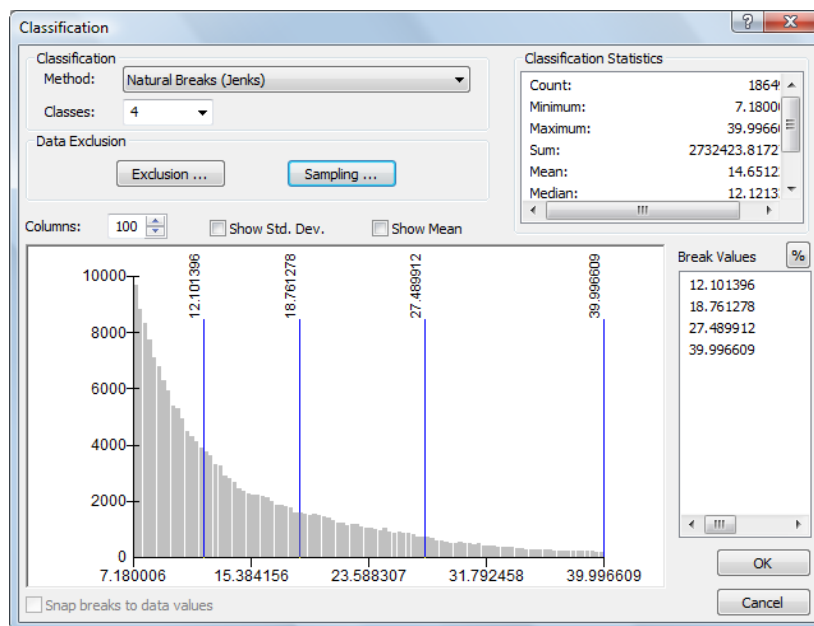
Basing on the figure 25 representing the relation between speed values and distance to nearest trail it is noticeable that speeds lower than 6km/h need to be omitted in the standardization. Points below this value can represent hikers on trail as well as those moving slowly far away from the trail. The speed of 6 km/h was chosen as it indicates the smallest standard deviation from mean distances to nearest trail. Data in this group is mostly coherent from the whole data set and additionally it is represented by a satisfying number of points. Therefore speeds below 6 km/h inform that probability that visitor was or was not on the trail is 50% and they need to be omitted in the analysis. The remaining speeds had been classified with the natural breaks method and assigned to appropriate fuzzy sets.



**Figure 32: Natural breaks data classification for the hiker's speed of movement.**

Remaining speeds have been classified into 4 groups because they represent the probability ranging from 60% to 100%. 50% probability concerned values below 6 km/h and therefore the next fuzzy set had to represent 60% probability.

Classification of speeds with which the bikers were riding was very similar to the classification made for hikers. Values below the mean speed 7.18 km/h indicate that the possibility that the bikers were on the trail or not is 50%. The figure 27 shows that especially for speeds close to 6 km/h the standard deviation is very high. On the contrary the coefficient of determination underlines that with the increase of speed the distance to the nearest trail decreases. Speed representing the highest natural break is according to earlier assumptions 40 km/h. On the basis of those conclusion speed values for bikers have also been classified with the help of natural break method.



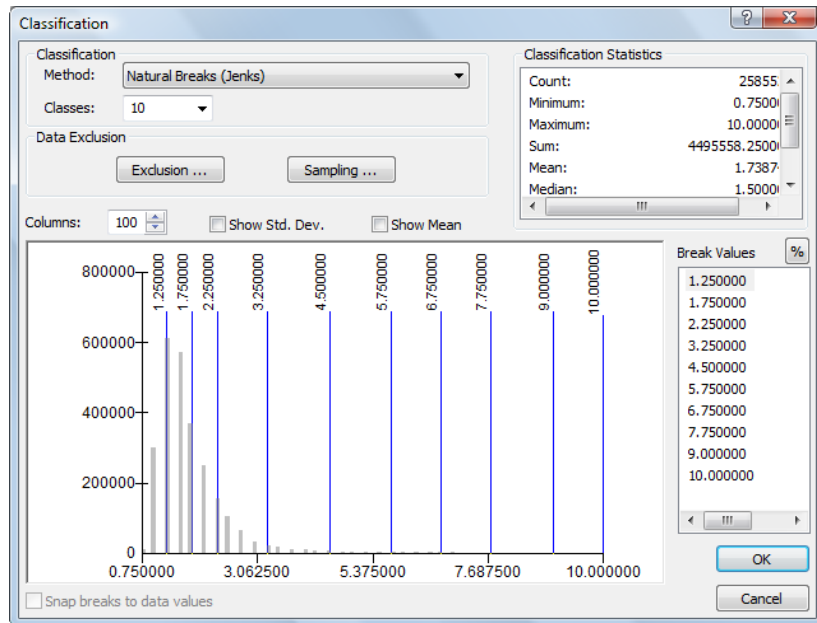
**Figure 33: Natural breaks data classification for the bikers riding speed.**

Eventually two new columns had to be created for each feature class, one “DIST\_FUZZY” representing distance fuzzy sets and second “SPEED\_FUZZY” representing speed fuzzy sets. In order to automate the process of data standardization two Python scripts have been written which assign each value to appropriate fuzzy set.

### 5.4.3. Fuzzy HDOP

The last factor subjected to standardization was HDOP. According to information from previous chapters this factor does not help with indicating if the visitor was on the trail or not. It is only responsible for informing how precise is positional accuracy for each point in other words how the recorded localization reflects the true position of visitor.

Standardization process for HDOP values was based on all points. No extra minimum and maximum thresholds have been added. According to natural breaks standardization and help of managers of the project HDOP values have been assigned to appropriate fuzzy sets.



**Figure 34: Standardization of HDOP values**

HDOP values lower than 1 are equal with 100% probability that recorded localization of a visitor matches the true localization. With the increase of HDOP the probability decreases but not in linear manner. The highest HDOP values ranging from 9 to 10 represent only 10% probability the position was appropriately set. The HDOP values are the same for hikers and bikers therefore standardization has been the same for those two groups. Similarly with standardization of speed and distance values, also for HDOP a python script has been prepared which creates a new column “HDOP\_FUZZY” where all information regarding each fuzzy set are saved.

## 5.5. Weighted Linear Combination

Weighted Linear Combination method is one of the two main approaches of MCE. This method is responsible for aggregating and weighting factors where the weights indicate the degree to which factors trade off. For the purposes of project Mafreina following equation for Weighted Linear Combination have been proposed.

$$\text{Trail} = [(X * \text{DIST\_FUZZY}) + (Y * \text{SPEED\_FUZZY})] * \text{HDOP\_FUZZY}$$

$$\text{Trail}^* = \text{DIST\_FUZZY} * \text{HDOP\_FUZZY}$$

Trail - Probability that the visitors was on the trail (1- He was definitely on the trail, 0 – He was definitely not on the trail)

Trail\* - Probability only based on the distance to nearest trail and HDOP, because speeds indicate fuzzy set equal to 0.5

X - Fuzzy distance weight

Y - Fuzzy speed weight

This equation is composed from two parts. The first part is where the “DIST\_FUZZY” and “SPEED\_FUZZY” are multiplied by the weights and next added to each other. In the second part they are multiplied by the “HDOP\_FUZZY” in order to confirm the data accuracy. Thus the equation proves whether the visitor was on the trail or not and then checks the data accuracy to provide the highest possible probability with which it can be said how visitor behaved. The equation has also second form which is used only when a specific condition is met. Whenever the “SPEED\_FUZZY” indicates values equal to 0.5 it means that those speeds need to be excluded from the calculation. The level of uncertainty is too high and therefore only “DIST\_FUZZY” and “HDOP\_FUZZY” need to be used in the equation. Then the “DIST\_FUZZY” factor is not weighted as it cannot be compensated by another factor.

The way the factors are aggregated and weighted is unique in its manner. No similar approach has been proposed yet but the research problem is also unique in its manner. There were many different map matching methods which helped to analyze so many different movement patterns but none of them concerned so complex dataset and so



many different research scenarios. This equation can be interpreted as a combination of a map matching procedure and Weighted Linear Combination method. Therefore it is supposed to create reliable and logical results which will be easy to interpret and model.

## 5.6. Modeling of various scenarios

In the WLC method weights represent the suitability of the factor for a specific purpose. Those weights can be modeled to prove how the factors trade off with each other. However the sum of all weights needs to be equal 1. For example if the weight for “DIST\_FUZZY” is 0.45 then the weight for “SPEED\_FUZZY” should be 0.55. Those weights mean that the relative importance of “SPEED\_FUZZY” is higher for the purposes of the research problem than the importance of “DIST\_FUZZY”.

For the purposes of project Mafreina different weights will be proven in order to choose the best combination. However already at the beginning it needs to be stated that according to the data type and the experience of the managers of the project “DIST\_FUZZY” will be more important than the “SPEED\_FUZZY”.

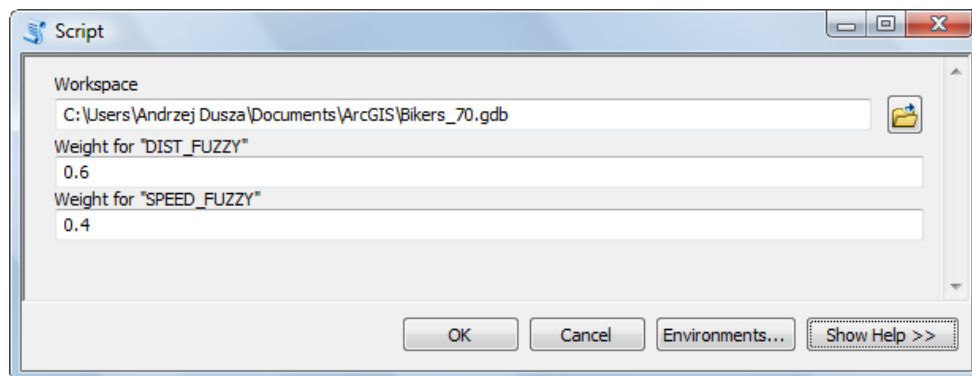


Figure 35: Weighted Linear Combination Python script.

Table 8: Comparison of the influence of different weights on mean probability that bikers or hikers followed the trails

Weights	Mean value [0.1;0.9]	Mean value [0.2;0.8]	Mean value [0.3;0.7]	Mean value [0.4;0.6]	Mean value [0.5;0.5]	Mean value [0.6;0.4]	Mean value [0.7;0.3]	Mean value [0.8;0.2]	Mean value [0.9;0.1]
Hikers_DIST_FUZZY	0.093	0.18	0.28	0.36	0.46	0.55	0.65	0.74	0.84
Hikers_SPEED_FUZZY	0.63	0.56	0.49	0.42	0.35	0.28	0.21	0.14	0.07
Bikers_DIST_FUZZY	0.09	0.19	0.29	0.38	0.48	0.57	0.67	0.76	0.86
Bikers_SPEED_FUZZY	0.61	0.54	0.47	0.4	0.33	0.27	0.2	0.13	0.06

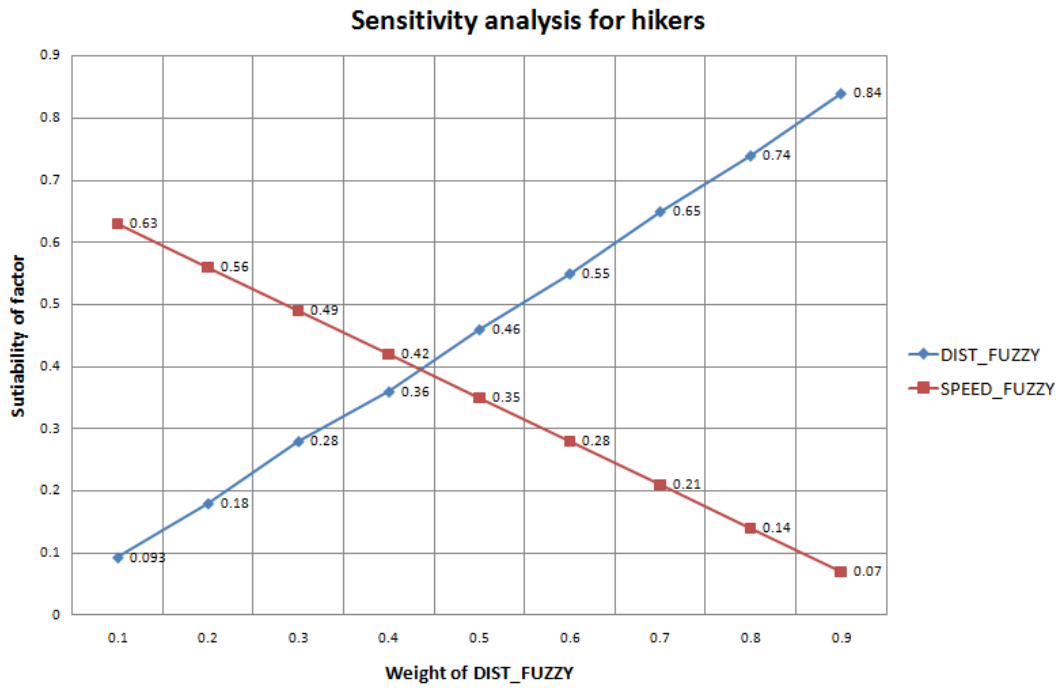


Figure 36: Sensitivity analysis indicating influence of various weight on factors.

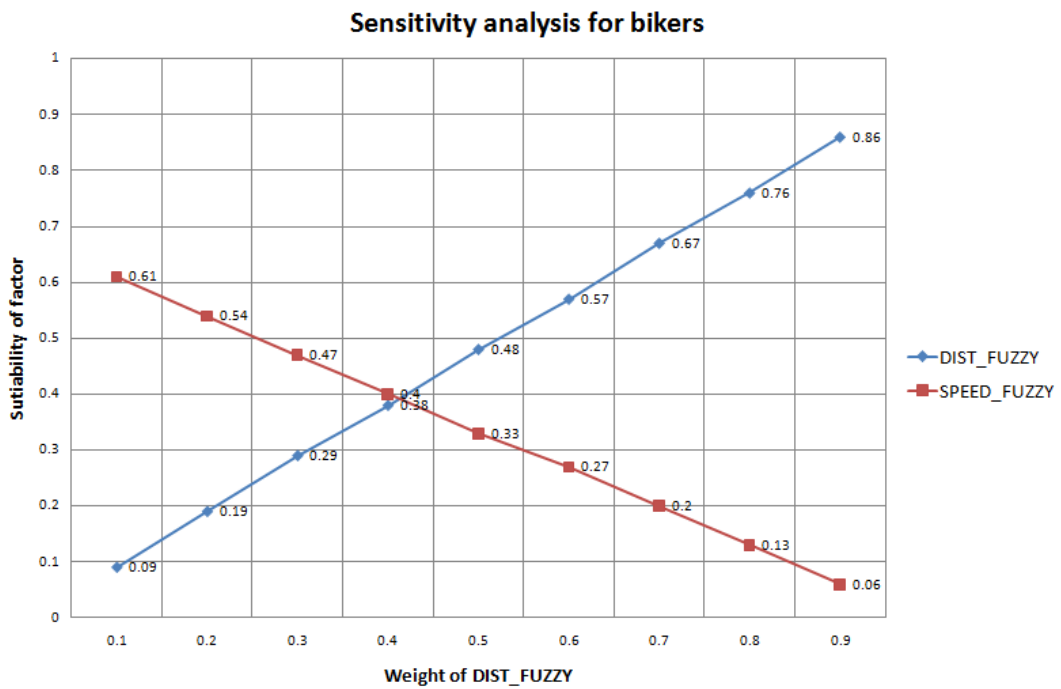


Figure 37: Sensitivity analysis indicating influence of various weight on factors.

Table 8 represents a comparison of the influence of different weights on factors for the hikers and bikers. Sensitivity analysis presented on figures 36 and 37 indicates which weights should be chosen for both factors depending on feature class. On the basis of figure 36 factor “DIST\_FUZZY” for hikers should receive weight 0.48 and

“SPEED\_FUZZY” 0.52. For bikers factor “DIST\_FUZZY” needs to be weighted with the value 0.46 and factor “SPEED\_FUZZY” 0.54. After assigning the weights to the factors python scripts have calculated the values for the probability that hikers or bikers were on the trails. The mean value of probability that hikers were on trail equaled 0.83 and for bikers 0.8. Standard deviation values were 0.2 for hikers and 0.16 for bikers, which means that the probability that the visitors stayed on the trails is rather high.

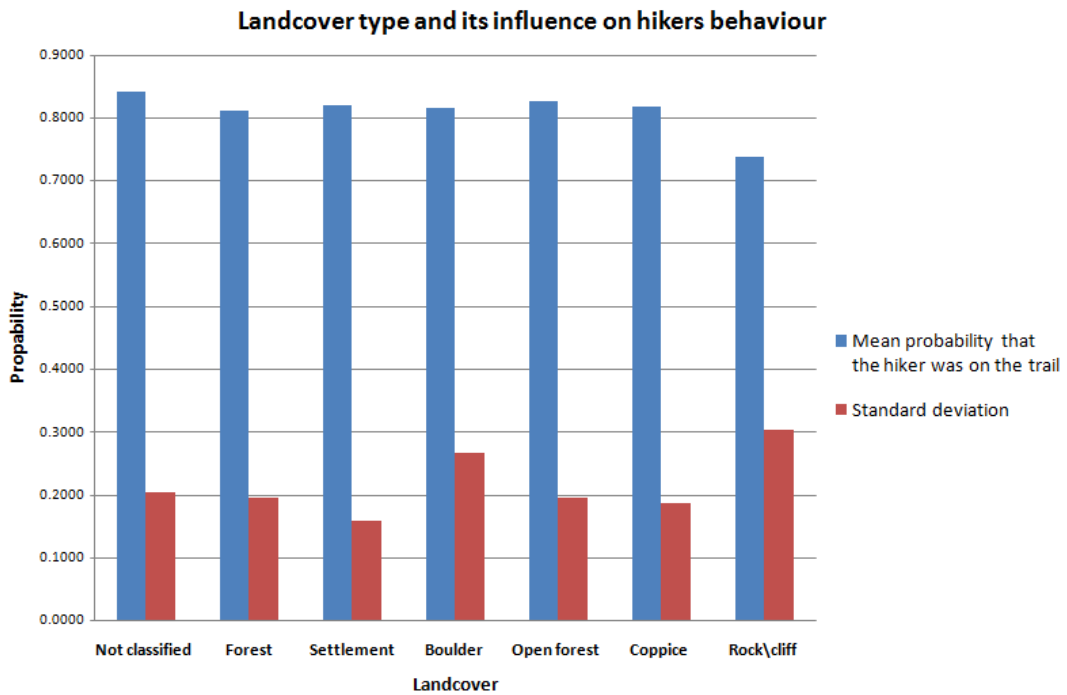
## **6. Results**

The last part of working with the data is the analysis of the results. The results have to be described in a quantitative and qualitative manner. Additionally the results need to be visualised and described concerning the statistical results.

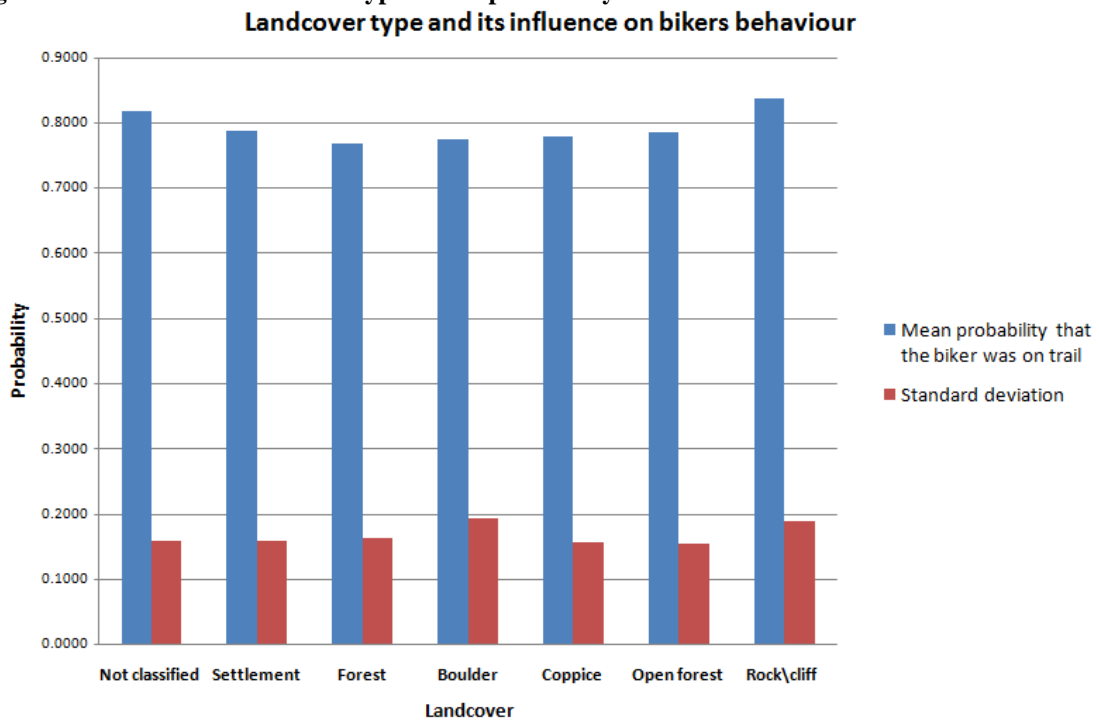
### **6.1. Statistical summary of the results**

Qualitative and quantitative analysis of the results are supposed to answer one of the following questions where do the visitors leave the trails, who leaves the trails do the hikers or bikers leave the trails more often? Question regarding age and gender of visitors can also be helpful in describing the results. In order to better interpret the statistical results the feature classes “Bikers\_final” and “Hikers\_final” have to be visually analysed. Before they will be analysed they have to be converted to raster format. This conversion will be performed using “Point statistics” tool in ArcGIS. This tool can calculate on the basis of specified attribute its mean value, sum or standard deviation in a fixed area. Feature classes will be analysed on the basis of “Trail” attribute where a mean value from all points within a square, 5 pixels wide and high, will be assigned to the central pixel. Each pixel represented 2m which means that the area is equal to 100m<sup>2</sup>. Additionally for similar area sum of points needs to be calculated.

In order to understand where the tourist deviated from the trail and are there places where they tend to leave the trails frequently, landcover data had to be analyzed. According to subsection 4.1.6 there are 6 significant landcover types in research area and therefore only for those the analysis will be performed.



**Figure 38: Influence of landcover type on the probability that hikers were on the trail**



**Figure 39: Influence of landcover type on the probability that bikers were on the trail**

Figures 38 and 39 show the influence of landcover type on the probability that hikers and bikers stayed on it. The mean values for hikers indicate that the probability is high irrespectively from landcover type. However for hikers on cliff or rocks the mean probability increases and the standard deviation decrease. This can mean that in high hill regions the hikers leave the trails in order to climb a hill on which there is no trail. Boulder which can be found also in high altitudes represents the second highest standard

deviation which mean that also there hikers tend to leave the trails more often. For all types of landcover standard deviation is higher than 0.15 which means that everywhere in the research area the hikers occasionally leave the trails.

Figure 39 informs that boulder has the lowest probability that the bikers followed it. Also forest and coppice indicate lower values than other landcover types. On contrary to other landcover types and the results for hikers the probability that the bikers stayed on the trail is very high rock and cliff. It can mean that bikers are more willing leaving the trails in lower altitudes in regions where the slopes are that noticeable rather than in high hill regions. In comparison to hikers standard deviation values are lower which means that in general the bikers are riding according to the fixed network of trails and roads.

Analysis of age is also important as it can underline how each age group behaved. It can be possible the young people tend to leave the trails more often than the elderly.

**Table 9: Influence of age on the probability that biker stayed on trail.**

Age	20-24	25-34	35-44	44-54	45-64	64-100
Mean probability that the biker was on trail	0.80	0.79	0.81	0.81	0.80	0.79
Standard deviation	0.17	0.17	0.16	0.16	0.16	0.21

Table 9 shows that none age groups represents higher probability on staying on trail than other. However, basing on the standard deviation it can be stated that elderly tend to leave the trail slightly more often than other groups. Additionally for bikers the influence of gender has been analyzed. According to the results nearly 80% of all bikers are men and only 20% are women. Results also indicate that that women with 79% stay on trails and men 80%. Standard deviation for women is 17% and men 16% which means that the gender does not have an influence on the way that bikers behave.

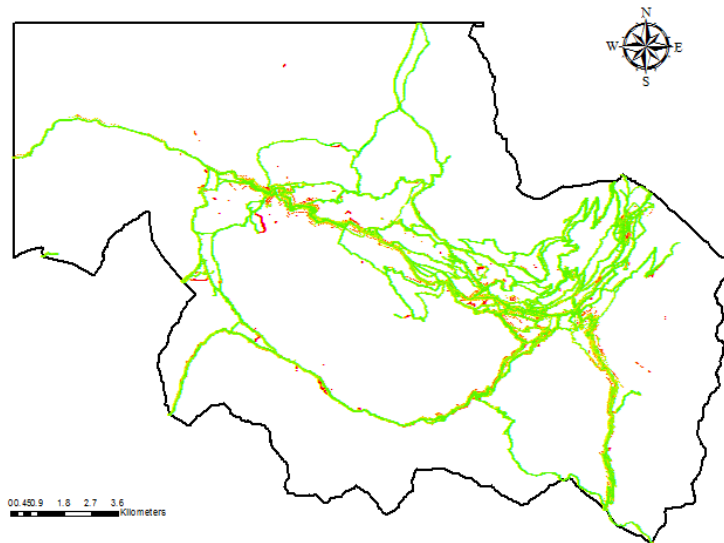
**Table 10: Influence of age on the probability that hiker stayed on trail.**

Age	20-24	25-34	35-44	44-54	45-64	64-100
Mean probability that the hiker was on trail	0.80	0.82	0.84	0.83	0.83	0.83
Standard deviation	0.24	0.22	0.21	0.21	0.20	0.20

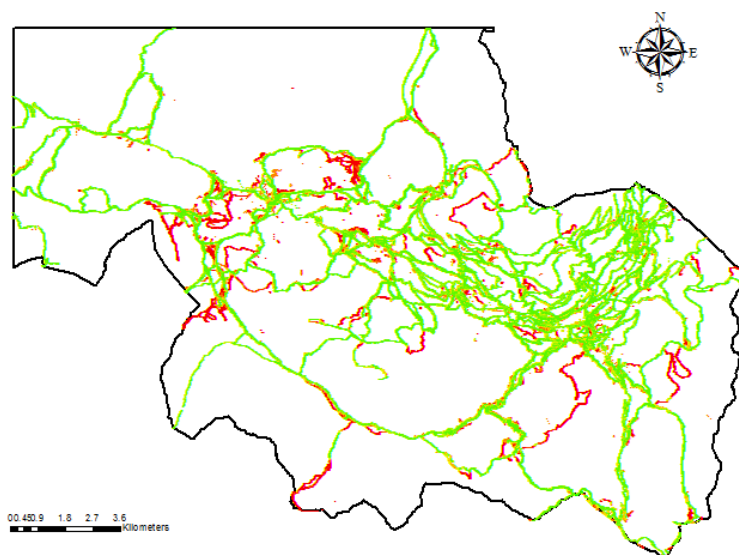
Table 10 shows that with increase of age the probability that hiker stayed on trail is also increasing. Values of standard deviation are highest for age group 20-24. With the increase of age those values decrease which mean that the people tend to more often follow the trails.

## 6.2. Movement patterns and trends

Visual analyses play a very important role in the assessment of final results. It is easier for the researcher to interpret the data as they can compare it with other datasets and also refer it to their experience. For example a researcher from a national park will find it much easier to interpret the results comparing them to well known sites which he can visualise in dedicated software



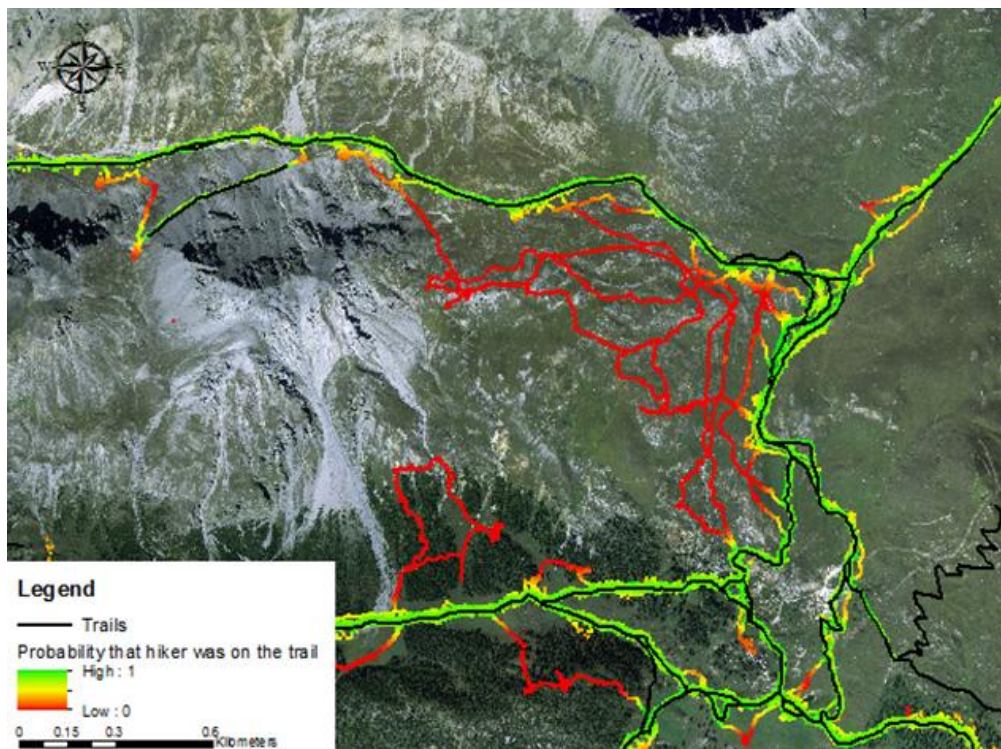
**Figure 40: Feature class „Hikers\_final” representing probability that the biker was on the trail. Green color means he was and the red that he was not.**



**Figure 41: Feature class „Bikers\_final” representing probability that the hiker was on the trail. Green color means he was and the red**

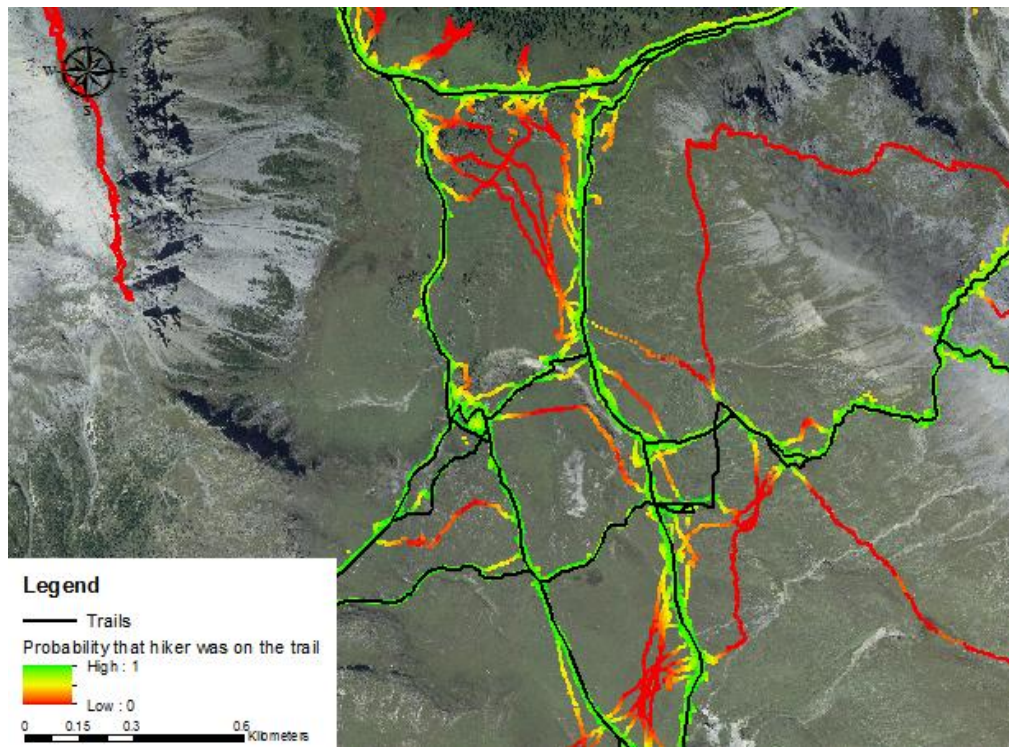
Figure 40 and 41 represent feature classes “Bikers\_final” and “Hikers\_final” classified according to the probability that the visitor was on the trail or not. Colours from red to green indicate the level of probability thus red colour equals 0% probability and green 100% that the visitor was on the trail. According to the figures hikers tend to leave the trails more often than the bikers which correlates with the results from the tables.

Identification of the trends in the movement aims to indicate areas where the visitors leave the trails frequently. On the basis of figure 41 three areas in the western and middle parts of the research area have been identified. The number of tourist leaving the trails there is significantly higher than in other regions. Those tracks clearly demonstrate trends in visitors’ behaviour.

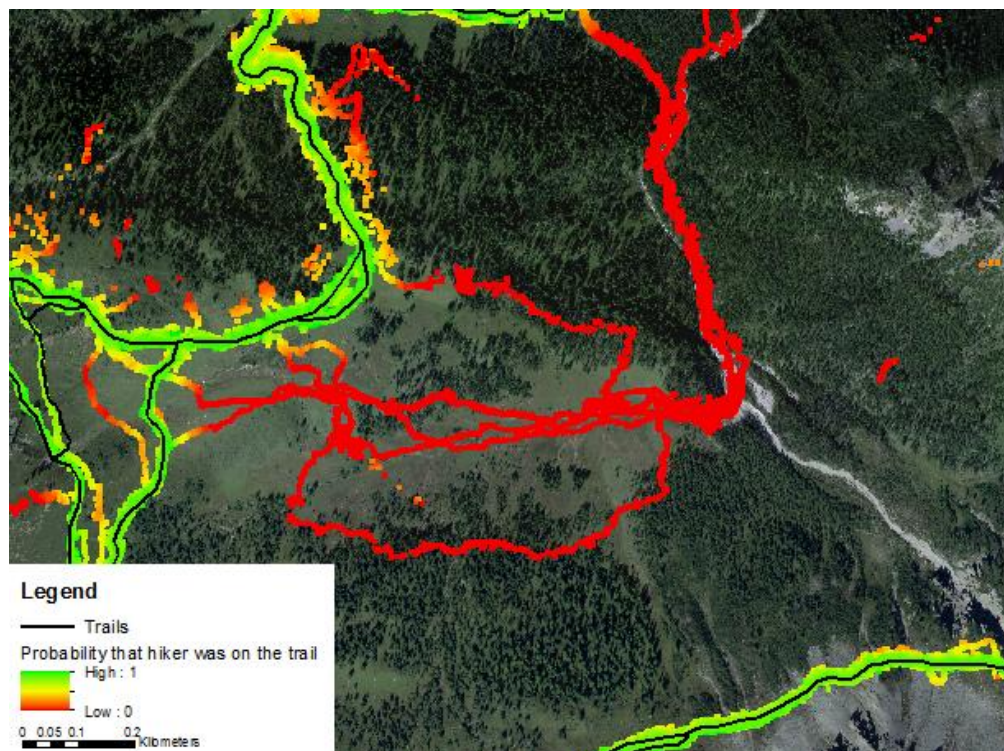


**Figure 42: North western part of the research area.**





**Figure 43: Western part of the research area.**



**Figure 44: Central part of the research area.**

Figures 42 to 44 represent the probability with which it can be said that hikers were on the trail. In those regions number of hikers who left the trails is very noticeable. Hikers leave the trails to climb a mountain or some nearby hill or they leave the trail to follow a gill as it can be seen on figure 44. The hikers very often leave the trail to take a shortcut to other trail. Number of patterns indicating that a big amount of tourist left the

trail is more noticeable in higher altitudes, which accords with the results from the tables. For the remaining tracks none significant trends have been spotted. Most of the hikers seem to follow the trail however there are individuals who leave the trails.

In comparison to hikers bikers do not create noticeable patterns indicating places where they left the trails. Most of them follow the trails and only some individual tracks have been recorded away from fixed trails.

### 6.3. Points of interest

Points of interest are locations which are frequently chosen by the visitors. This can be hut, camping, shelter, bus stop or place where some interesting natural monuments can be found. There can be many points of interest depending on the location and region. Those points are supposed to have a big influence on the way visitor move in research area. For example lakes, shelter, vast meadows may cause that the visitor will leave the trail. Tourist may also leave the trails while they are near a shelter or some interesting monument.



**Figure 45: Red circle indicating location of UNESCO World Heritage Site The Benedictine Convent of Saint John**

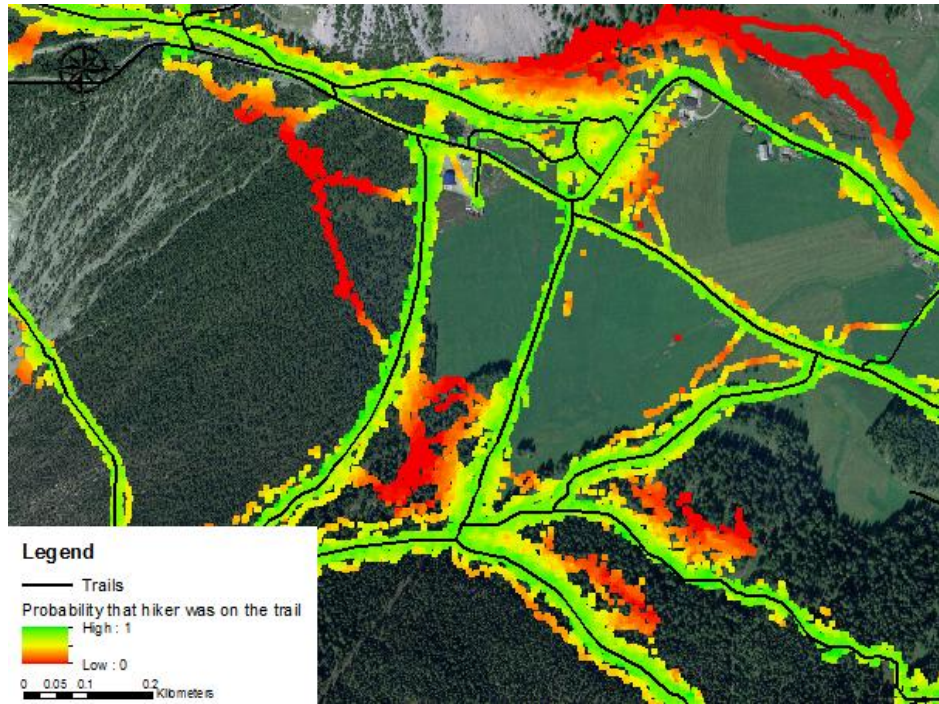


Figure 46: In the upper right corner village Tshierv and in the middle crossroad of trails.

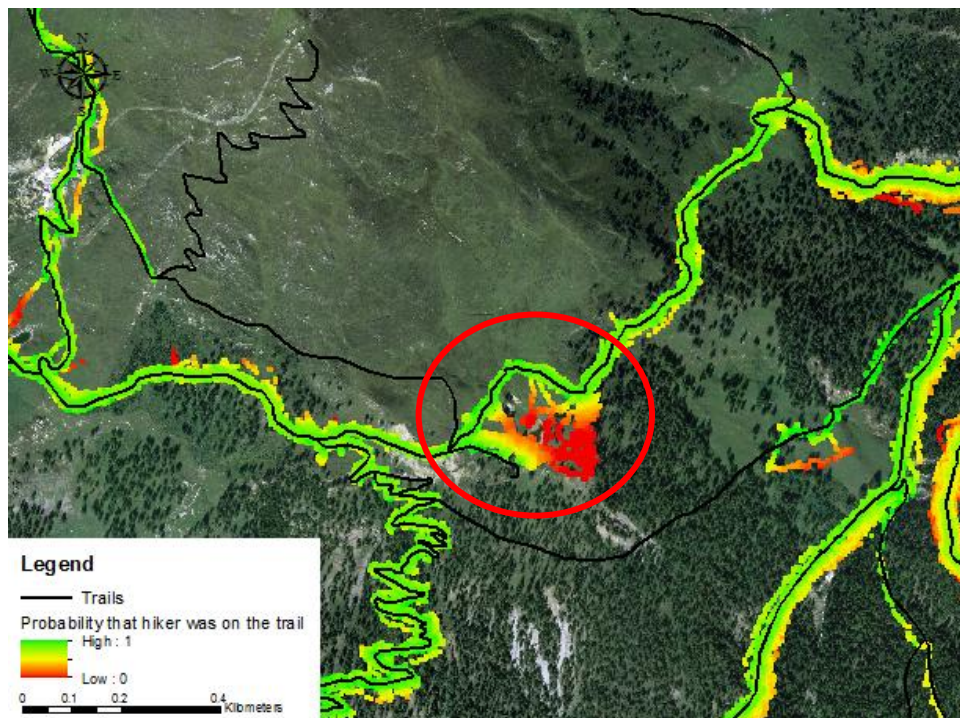
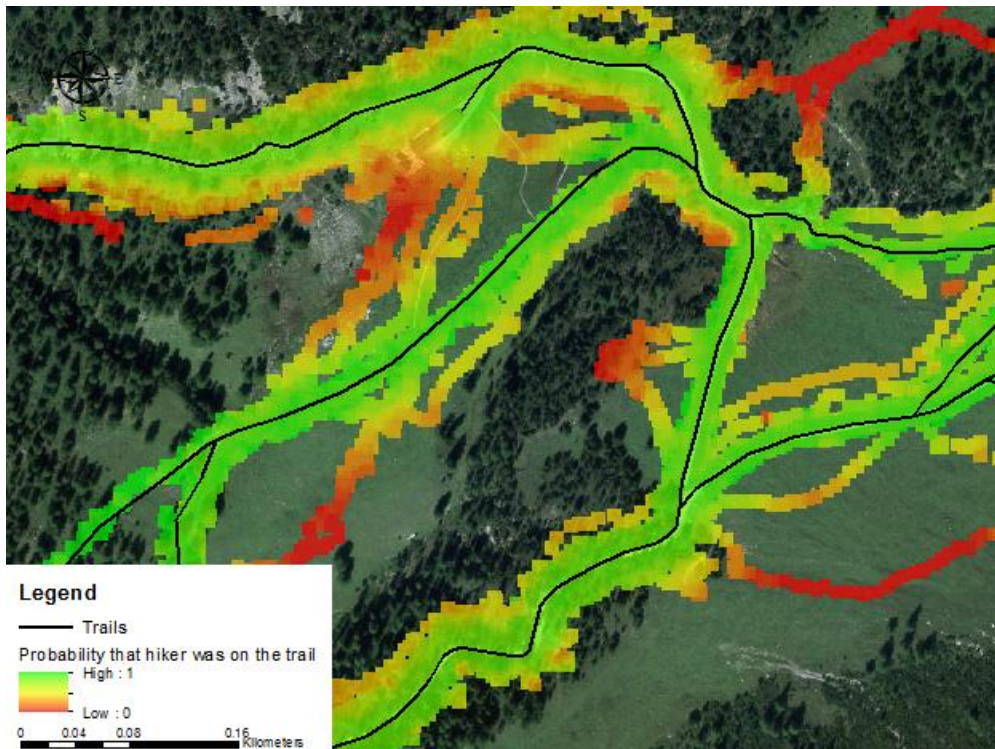


Figure 47: Red circle indicating mountain lake



**Figure 48: In the upper corner a mountain shelter**

Analysis of points of interest indicate that the tourist leave the trails on many occasions. For example they deviate from trails when they see a lake, meadow or other interesting natural monuments. These are places where they can eat, play with children or swim in the lake as it is not forbidden in Müstair Valley. Tourists also tend to leave the trails at the crossroads of trails or roads. It may be caused by their unawareness of the appropriate trail or they are taking a shortcut to get to new trail. Figure 45 shows that visitors also leave the trails near urban areas. In place like village Müstair the tourists visit the famous Convent of St.John. This leads to significantly higher density of tourists but also causes that many move freely around the whole nearby area. Figure 48 represent a shelter which is very often visited by the tourists. They very often take a shortcut to get faster to the shelter which can cause deterioration of nearby environment.

#### **6.4. Summary of visitors behaviours patterns in Müstair Valley**

The results of Weighted Linear Combination describe very precisely the probability that the visitor followed the trail or not. On the basis of those results it can be easily interpreted which tracks are or the trail and therefore should be seen as normal behaviour. On the contrary the results clearly indicate where a tourist deviate from track

and with what probability this happened in reality. Trends in visitors' behaviour patterns as well as point of interest can be easily identified on the map. Apart from those positive results there are some drawbacks which need to be improved.

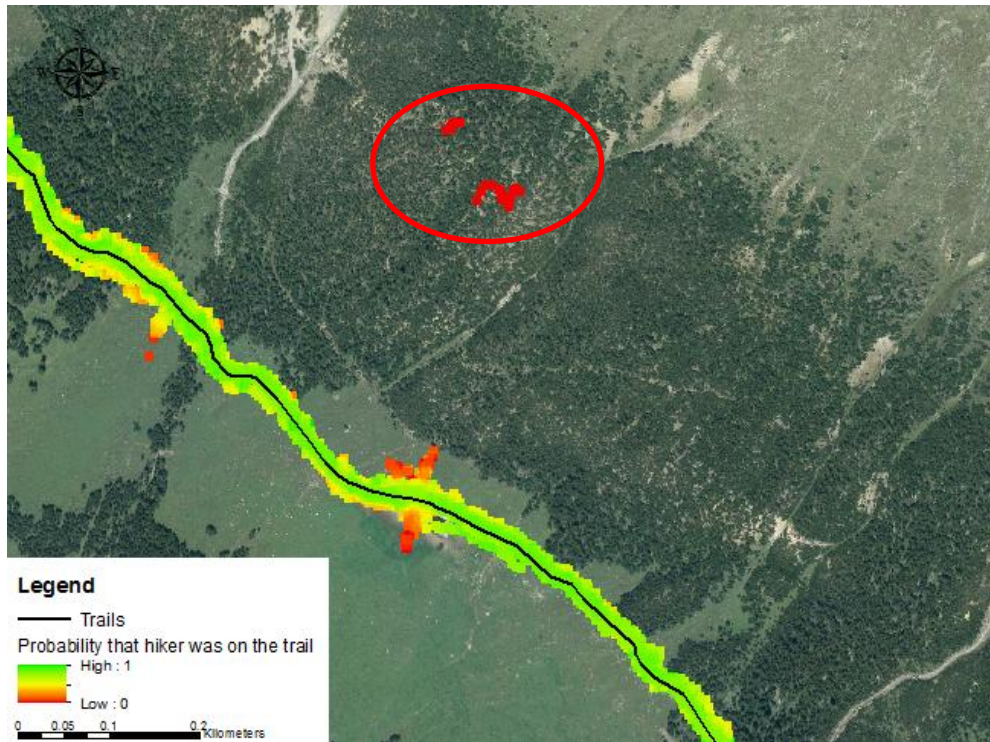


Figure 49: Red circle indicating badly interpreter points.

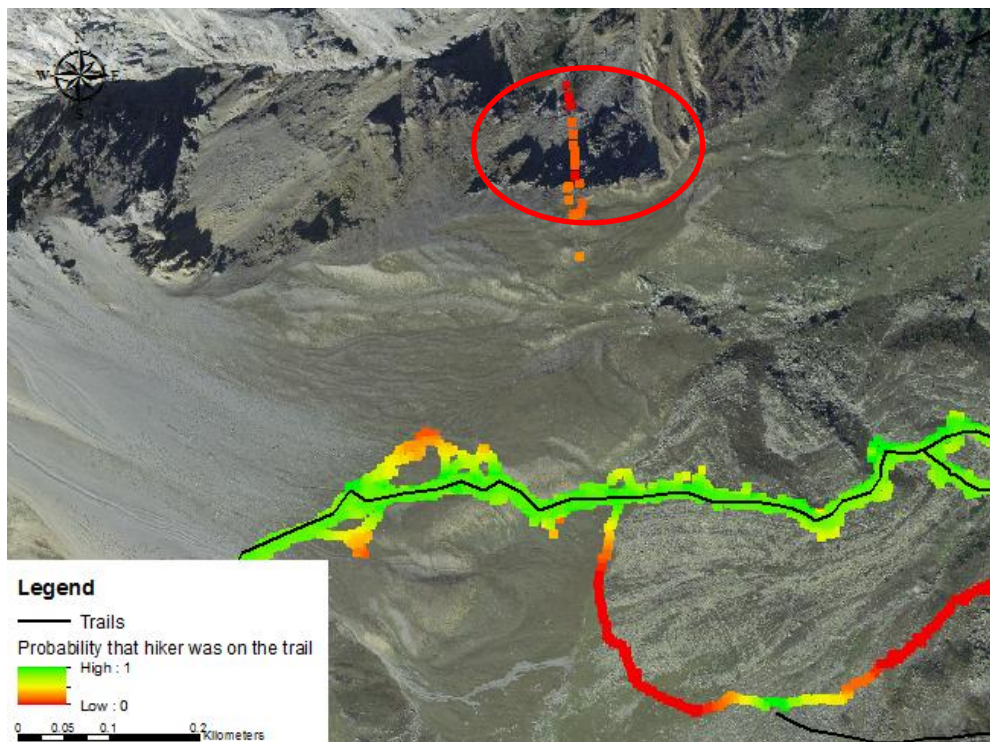


Figure 50: Red circle indicating badly interpreter points.

Figure 49 and 50 represent points which cannot be used for any interpretation as they do not show connection with the existing movement patterns. Those points could be excluded neither in data selection or data analysis part because they indicate very good results for the attributes which have been used in the selection and analysis process. Normally those points would be omitted as they lie far away from the trails but according to the assumptions of the project those could be tracks of visitors who deviated from the trail. There other points which show similarities with those from figure 49 and 50 but it is now the responsibility of managers of the project to omit them in further work. Those points could be excluded from the final results by manually selecting each of them however their number is so insignificant that they do not noticeable influence the overall results.

## 7. Discussion

Increasing pressure on the environment in alpine regions lead in recent years to conflicts between nature and humans. In order to analyze those conflict, better understand their origins and create tools two which would help to better manage them a project Mafreina was launched in Müstair Valley in Switzerland. The goals of the project were analysis of existing spatial and temporal outdoor uses in the valley, documentation of outdoor recreationist's requirements and research visitor preferences for planned projects. Additionally it was supposed to develop predicative environmental planning tools to stimulate results of various management decisions.

According to Muhar et al. (2002) every research project concerning visitors monitoring and management should define answers to five following questions:

- Why should be monitored?
- What should be monitored?
- Who should be monitored?
- Where should be monitored?
- When should be monitored?

In project Mafreina it was clearly stated that the visitors monitoring and management project aims to eliminate existing conflicts between nature and humans. Interactions between those two groups need to be monitored especially that the people visiting Müstair Valley are allowed to leave the network of roads and trails. Particularly interesting for managers of the project were two groups of visitors, hikers and bikers. They are the biggest concerns while with regards to social conflicts, mountains bikers are considered as a major conflict causing group (Freuler 2008). Monitoring was performed in the whole area of Müstair Valley because the goal was to analyze the conflicts in every place even in the most remote one. Project lasted two years where each year was divided into two periods summer and winter. This distinction was supposed to underline different behavior patterns and impacts of people visiting Müstair Valley in winter and in summer. Project Mafreina was prepared according to the requirements for monitoring and management project presented by Muhar et al. (2002).

Research goals proposed by the mangers of the project required appropriate toolkit. According to Skov-Petersen (2008) it is expected that agent based models will be gaining popularity in the recreational planning and management projects. In the Mafreina toolkit ABM was supposed be one of the key elements. However ABM to provide desired

results, needs scenarios or in other words rules for agents. According to Taczanowska et al. (2008) and Skov-Petersen (2006) GPS-monitoring is a new method to trace spatial and temporal movement of visitors. The main disadvantage of GPS-monitoring is that it delivers only information about existing situation. Data from GPS-Monitoring is not able to inform about planned alternatives or anticipated scenarios. Therefore data from GPS-Monitoring had to be combined with discrete choice model (DCE) to detect agent rules for future non-existent scenarios (Hunt et al. 2007, Haider 2007). The most important part of the whole project was the combination of all elements using geographic information system (GIS). Until now Mafreina toolkit worked according to this scenario. First test including data from winter have been performed and their results are very satisfying. This leads to a conclusion that the number of projects willing to use similar toolkit will increase. According to Lawson (2006) computer simulations similar to those performed in project Mafreina will have significant potential to assist and inform planning and management of visitor use in protected natural areas.

First important part of my master's thesis was to analyze existing works on visitors monitoring and management especially works which concerned GPS data and GIS. In my master's thesis I had to analyze GPS data which represented traces of visitors movement. The most important task was to distinguish from the data various visitors movement patterns. Numerous attributes created for each recorded point had to be used as factors for data assessment. The biggest concern was that the literature describing similar works assumed that the visitors had to follow the trails whereas in Müstair Valley they must not. Numerous projects which handled tracking of movement traces concerned problems with public transport or some logistic issues. However they were similar to my master's thesis they could not be directly implemented to my research problem.

According to literature ways of monitoring and managing tourism in recreational areas are numerous and some of them vary significantly. Most those methods are based on direct observations and only some of them use indirect observations. However methods proposed in Swiss National Park or Danube Floodplains National Park suggest that GIS and GPS-Monitoring are beginning to gain popularity. Most works on network analysis of visitors flows in recreational areas suggested that the best method for data preparation is map matching. This method was supposed to help assigning each traced point to appropriate network element which could be line or node.

Map matching methods can be classified into three categories, geometric procedures, topological procedures and advanced procedure (Schuessler & Axhausen



2009). All of those methods match GPS traces on high-resolution networks. Some of them like geometric procedures rely on distance to closest trail, topological take into account perpendicular distance and the sequence or history of GPS points. Only advanced procedures offered solutions which could be implemented into my master's thesis. According to Ochieng et al. (2004) regions of confidence or error should be created for each GPS point. Creation of errors of confidence should be based on fuzzy logic inference systems. Those fuzzy rules should consider different criteria like speed, HDOP or the position of GPS point to candidate link.

However all of those procedures were very advanced I could not use them directly due to the fact that I had to analyze places where the visitors leave the trails. Thus using this method I would have to match each point to appropriate line. In Müstair Valley visitors are allowed to move freely therefore most of the points should not be assigned to none trail.

According to Goodchild and Hunter (1997) when no matching of points is possible it necessary to base on different metric of separation between the tested and reference source. Therefore they suggested measuring the distance from each point along tested source to the closest point on reference source. On the basis of those results a histogram indicating distributions of distances could be created and analyzed. This method could help to determine how the points should be prepared, which of them should be excluded from further analysis and which labeled as on the trail. However this method is subjected to one problem, we don't know how the shortest distance correlates with distortion of individual points, which can be measure only when points can be matched (Goodchild and Hunter 1997).

Problems with choosing appropriate method caused that a new concept of data preparation and analysis had to be proposed. It has been decided that a combination of methods proposed by Goodchild and Hunter (1997) and Ochieng (2004) could be the best solution to the research project. The points had to be analyzed using fuzzy logic and regions of confidence based on different factors and histograms indicating distributions of distances from each GPS point to nearest link.

In the data preparation and selection process selection of points was based on various attributes. More than 1.5 million of points have been excluded from the analysis basing on HDOP, number of satellites, activity type and the location in reference to research area. Selection could be based on extra attributes like speed or distance from the previous point however the complexity of those attributes caused they had to be analyzed

in further parts of analysis. In the preparation process the calculation of distances from point to nearest trail was based on perpendicular distance. This caused that the distance was not always to nearest point. This problem could not be solved differently as the visitors could leave the trails. This fact could also be omitted as for millions of points it had only marginal meaning. Eventually in the data preparation process speed had to be recalculated to ensure that it would reflect real movement speed. Disadvantage of this method was that sometimes it could create false values for points representing fast speed growth e.g. when people were quickly taking a shortcut.

In the analysis process three factors speed, distance to nearest trail and HDOP have been selected for fuzzy analysis. Basing on data distribution of those factors in comparison to distance to closest trails rules for fuzzy sets have been determined. This method was partly based on the work of Ochieng et al. (2004) who suggested usage of regions of confidence but in general it can be seen as similar method. Analysis of data distribution for those factors needs to be performed in other similar projects to prove its reliability. However basing on histograms, some clear trends have been proven and on the basis of those trends and earlier assumptions, factors have been standardized. Significant aspect of standardization was the natural breaks method. It can be subjected to discussion whether standardization should be based on the knowledge of researchers or results of standardization methods. In project Mafreina it has been decided that standardization will be based on natural breaks method which means that final results are strongly dependent on the data set. The advantage of WLC approach over the Boolean does not have to be discussed. It is clear that managing so high uncertainty whether the visitors deviated from the trail or not needs a more flexible approach. Boolean approach indicating only true or false answers could cause that two points located near each other were classified as two different groups.

The final results of analysis demonstrate various behavior patterns of hikers and bikers. On the basis of different probability values it can be stated for each point whether it belongs to a trail or not. The probability results from three different variables which cause that it can be perceived as reliable source of information. According to Lawson (2006) simulation modeling can be used to describe existing visitors use conditions that are inherently difficult to observe. This advantage of simulation modeling is especially important when the protected area is large in size and has multiple points of access. This causes that the visitors behavior patterns are dispersed over a large area and can be difficult to monitor. The final results of my master's thesis clearly indicate visitors movement

patterns. On their basis it can be determined where the visitors follow the trails where are located the points of interest and are there some trends within tourists leaving the trails. Those questions and many others can be explained on the basis of the final results.

However the final results are not free from errors. Within the whole research area there are single points or areas which are hard to interpret. These are residues of data selection and analysis processes. Data selection process was not able to exclude all undesired points while they were signified by high positional accuracy and appropriate values for other attributes. Those points could not be excluded from analysis on the basis of distance to nearest trail or speed. The initial assumptions of the project stating that visitors can move freely around the whole research area caused that not all points could be properly selected. In general error points could be eliminated but this would cause that many proper points had to be eliminated. The question whether to work with smaller group of points or bigger but with some errors depends on research project. Smaller group of point can deliver results free from errors on the contrary it can omit relevant data representing interesting patterns. Nevertheless number of errors points doesn't have a major influence on the overall results. They do not cause that the final results are illegible or hard to interpret.

## **8. Summary and outlook**

The combination of GPS-monitoring and advanced GIS tools creates new possibilities for visitors monitoring and management projects. The managers get advanced analytical and statistical tools which help them to analyze unlimited number of scenarios in the research area. They may detect previously unknown conflict areas and quickly and efficiently prepare appropriate solutions. Computing results from those tools with discrete choice experiments and agent-based models can help to create new strategies on protection of diversity while taking into account growing needs of the tourists.

Results of my master's thesis suggest that the process of data preparation, selection and analysis can be very complex and differ depending on the goals of project. Methodology needs to be profoundly analyzed basing on the existing literature. However sometimes new approaches need to be presented as the monitoring of visitors movement patterns using GPS devices is still developing. Till now no relevant literature concerning problems described in Müstair Valley has been presented therefore various approaches had to be combined. The final results are very satisfying however they leave some place for further improvements.

The final conclusion of my master's thesis is that however results of GPS-Monitoring and data simulation are very appreciated only by linking them with social and environmental factors we can begin to try to better understand and manage the interactions between humans and nature.

## 9. References

1. Cessford G., Muhar A., 2003. *Monitoring options for visitors numbers in national parks and natural areas*. Journal for Nature Conservation, 11: 240-250.
2. Chung E.H., Shalaby A., 2005. *A trip bases reconstruction tool for GPS-based personal travel surveys*. Transportation Planning and Technology, 28 (5): 381–401.
3. Daniel T., Gimblett R., 2000. *Autonomous agents in the park: An introduction to the Grand Canyon river trips simulation model*. International Journal of Wilderness, 6(3): 39-40.
4. Doherty S.T., Noel C., Lee-Gosselin M. E. H., Sirois C., Ueno M., Theberge F., 2001. *Moving beyond observed outcomes: Integrating Global Positioning Systems and interactive computer-based travel behaviour surveys*. Transportation Research E-Circular C026: 449–466.
5. Eastman J.R., Kyem P.A.K., Toledano J., Jin W., 1993. *GIS and decision making. Explorations in Geographic Information System Technology*, 4 (Geneva: UNITAR) (single volume)
6. Gimblett R., Richards M., Itami R., 2000. *RBSim: Geographic simulation of wilderness recreation behavior*. Journal of Forestry, 99(4): 36-42.
7. Goodchild F.M., Hunter G., 1997. *A simple positional accuracy measure for linear feature*. International Journal of Geographical Information Science, 11(3): 299-306.
8. Hallo J., Manning R., Valliere W. 2005. *Acadia National Park scenic roads: Estimating the relationship between increasing use and potential standards of quality*. In D. N.Cole, *Computer simulation modeling of recreation use: Current status, case studies, and future direction*. USDA Forest Service General Technical Report RMRS-GTR-143, 55-57.
9. Hall G. B., Wang F., Subaryono., 1992. *Comparison of Boolean and fuzzy classification methods in land suitability analysis by using geographical information systems*. Environment and Planning A, 24: 497-516.
10. Hornback K.E., Eagles P.F.J., 1999. *Guidelines for Public Use Measurement and Reporting at Parks and Protected Areas*.
11. Jankowski P., 1995. *Integrating geographical information systems and multiple criteria decision making methods*. International Journal of Geographical Information Systems, 9:251- 273.
12. Jansen R., and Rietveld P., 1990. *Multi-criteria analysis and geographical information systems: an application to agricultural land use in The Netherlands*. The Netherlands: Kluwer Academic Publisher, 129- 139.

13. Jiang H., Eastamn J.R., 2000. *Application of fuzzy measures in multi-criteria evaluation in GIS*. International Journal of Geographical Information Science, 14(2):173-184.
14. Langley B.R., 1999. *Dilution of precision*. GPS World, 5:52-59
15. Lamprecht M., Fischer A., Stamm H.P., 2008. *Sport Schweiz 2008: Das Sportverhalten der Schweizer Bevölkerung*. Magglingen: Bundesamt für Sport BASPO.
16. Lawson S., Mayo-Kiely A., Manning R., 2003. *Integrating social science into park and wilderness management at Isle Royale National Park*. George Wright Forum, 20(3):72-82.
17. Lawson S., Manning R., Valliere W., Wang B., 2003a. *Proactive monitoring and adaptive management of social carrying capacity in Arches National Park: An application of computer simulation modeling*. Journal of Environmental Management, 68: 305-313.
18. Lawson S., Manning R., 2003a. *Research to inform management of wilderness camping at Isle Royale National Park. Part I – descriptive research*. Journal of Park and Recreation Administration, 21(3): 22-42.
19. Lawson S., Manning R., 2003b. *Research to inform management of wilderness camping at Isle Royale National Park: Part II – prescriptive research*. Journal of Park and Recreation Administration, 21(3): 43-56.
20. Lawson S., Itami B., Gimblett R., Manning R., 2006. *Benefits and challenges of computer simulation modeling of backcountry recreation use in the Inyo National Forest*. Journal of Leisure Research, 38(2):187-207.
21. Lawson S., 2006. *Computer Simulation as a Tool for Planning and Management of Visitor Use in Protected Natural Areas*. Journal of Sustainable Tourism, 14(6): 600-617.
22. Marchal F., Hackney J. K., Axhausen K. W., 2005. *Efficient map matching of large Global Positioning System data sets: Tests on speed-monitoring experiment in Zürich*, Transportation Research Record, 1935: 93–100.
23. Muhar A., Arneberger A., Brandeburg C., 2002. *Methods for visitors monitoring in recreational and protected areas: An overview*. Monitoring and Management of Visitors Flows in Recreational and Protected Areas, 1-6.
24. Nielsen O. A., Würtz C., Jorgensen R. M., 2004. *Improved map-matching algorithms for GPS-data - methodology and test on data from the AKTA road pricing experiment in Copenhagen*, 19th European Conference for ESRI Users.
25. Proctor W., Qureshi E., 2004. *Multi Criteria Evaluation Revisited*.

26. Ochieng W. Y., Quddus M. A., Noland R. B., 2004. *Map matching in complex urban road networks*. Brazilian Journal of Cartography, 55 (2): 1–18.
27. Quddus M. A., Ochieng W.Y., Noland R.B., 2007. *Current map matching algorithms for transport applications: State-of-the-art and future research directions*. 86th Annual Meeting of the Transportation Research Board.
28. Quddus M. A., Ochieng W. Y., Zhao L., Noland R. B., 2003. *A general map matching algorithm for transport telematics applications*. GPS Solutions, 7(3): 157–167.
29. Rupf R., Koechli D., Haider W., Skov-Petersen H., Proebstl U., 2010. *Framework Mafreina: Management Toolkit Recreation and Wildlife in the Swiss Alps*. 5th International Conference on Monitoring of Visitor Flows in Recreational and Protected Areas, 121-123.
30. Schuessler N., Axhausen K.W., 2009. *Map-matching of GPS traces on high-resolution navigation networks using the Multiple Hypothesis Technique (MHT)*.
31. Skov-Petersen H., 2008. *The role of agent-based simulation in recreational management and planning*.MMV4 Proceedings – Keynotes addresses.
32. Taczanowska K., Muhar A., Arnberger A., 2008. *Exploring Spatial Behavior of Individual Visitors as Background for Agent-Based Simulation*. Monitoring, Simulation and Management of Visitor Landscapes, 159-174.
33. Van Wagendonk J., 2003. *The wilderness simulation model: A historical perspective*. International Journal of Wilderness, 9(2): 9-13.
34. Velaga N. R., Quddus M. A., Bristow A. L., 2009. *Developing an enhanced weight-based topological map-matching algorithm for intelligent transport systems*. Transportation Research Part C: Emerging Technologies, 17 (6): 672–683.
35. Wang B., Manning R., 1999. *Computer simulation modeling for recreation management. A study on carriage road use in Acadia National Park, Maine, USA*. Environmental Management, 23:193-203.
36. White C. E., Bernstein D., Kornhauser A.L., 2000. *Some map matching algorithms for personal navigation assistants*. Transportation Research Part C: Emerging Technologies, 8 (1–6): 91–108.

## Internet sites:

1. ArcGIS Desktop 9.3 Help, 2010. Online: <http://webhelp.esri.com> (04.09.2011)
2. BAFU, 2009. Medienmitteilung - Sport in freier Natur: Pilotprojekt für mehr Rücksichtnahme auf Wildtiere, Online: <http://www.bafu.admin.ch/dokumentation/medieninformation/00962/index.html?lang=de&msg-id=25179> (12.06.2011)
3. Graubunden official website. Online: <http://www.graubuenden.ch/> (5.05.2011)
4. Federal Office of Topography Swisstopo. Online: <http://www.swisstopo.admin.ch> (01.09.2011)
5. Python Programming Language. Online: [www.python.org](http://www.python.org) (12.09.2011)
6. Polish Python coders group. Online: [www.pl.python.org](http://www.pl.python.org) (12.09.2011)
7. Ublox product documents and resources. Online: <http://www.u-blox.com/en/download/documents-a-resources.html> (1.09.2011)



## OŚWIADCZENIE

Ja, niżej podpisany, Przemysław Dusza, student Wydziału Nauk Geograficznych i Geologicznych Uniwersytetu im. Adama Mickiewicza w Poznaniu oświadczam, że przedkładaną pracę dyplomową pt. „Analysis of visitors behavior patterns based on GPS tracks from Müstair Valley, Switzerland” napisałem samodzielnie. Oznacza to, że przy pisaniu pracy, poza niezbędnymi konsultacjami, nie korzystałem z pomocy innych osób, a w szczególności nie zlecałem opracowania rozprawy lub jej części innym osobom, ani nie odpisywałem tej rozprawy lub jej części od innych osób.

Oświadczam również, że egzemplarz pracy dyplomowej w formie wydruku komputerowego jest zgodny z egzemplarzem pracy dyplomowej w formie elektronicznej.

Jednocześnie przyjmuję do wiadomości, że gdyby powyższe oświadczenie okazało się nieprawdziwe, decyzja o wydaniu dyplomu zostanie cofnięta.