Detection of illegal building using geodetic data gathering

Aalborg University

Master's program in Surveying and Mapping

4th Semester

Master Thesis

AALBORG UNIVERSITET

Author: Lukáš Matrka

Supervisor: Peter Cederholm

Date: 22.11.2018

Title

Detection of illegal building using geodetic data gathering

Project period
4th Semester
1.2.2018 – 22.11.2018

Semester topic
Master thesis

Supervisor
Peter Cederholm

Participants
Lukáš Matrka

Number of pages
97

Number of appendixes
3

Abstract

The goal of this project is to establish a process how can illegal buildings be detected using geodetic data acquisition. The assumption is to compare state of the Cadastre with the representation of reality, captured by surveying to identify the illegal buildings. In first part, available surveying option are researched in the means of data they produce and methods of processing. The most relevant one for this project is chosen and the experimental part takes place. Here the geodetic data along with the data from the Cadastre are processed and prepared to suitable form for their mutual comparison. This will be based on comparing polygons representing buildings in both data sets. The outliers between the two sets of data will be analyzed for the presence of illegal buildings. Evaluation of detected illegal buildings along with the conclusion for the overall process will be done afterwards.

Acknowledgement

I want to thank the supervisor Peter Cederholm for the guidance he provided to me during the project and for his open-minded approach that made this project possible. Also, to the company Helicop s.r.o for providing me with data for this work. Last, but not least I want to thank to all those who supported me on my way to the engineering degree.

Foreword

This report represents a Master Thesis in of a 4th semester in Master's program Surveying and Mapping at Aalborg University written by Bc. Lukáš Matrka. The project period started on the 1st of February and ended with hand-in report on the 22nd of November 2018.

Disclaimer: All illustrations, figures, plots, and tables found within this report and its appendixes, are produced by the author unless referenced.

Table of content

Experimental part

1. Introduction

Nowadays the urban growth can be noticed more than before as people are moving from rural areas, the cities are growing with new buildings and urban areas are expanding [14]. This brings new possibilities, advantages but also a lot of challenges. One of them is the effort to keep this growth organized by controlling the construction with construction law.

In Slovakia the law states that before a building can be lawfully constructed it has to fulfill specific rules to have a valid construction permit issued. These rules are for example supplying architecture and other documentation to the authorities, suitable location of the building, position and size within the plot and other requirements. Not meeting these criteria results in invalid construction permit and such constructed building is considered illegal. Permit is issued by Local Construction Office which oversees the process and validates once all of the requirements are met.

Even though it is against the law some people decide to build their constructions without a valid permit for various reasons, usually it's because they didn't meet all the requirement for the permit, or they want to build in forbidden areas as forests or protected environment, so they didn't even apply for it [15]. There are several problems related to this issue with an impact on different spheres such as environment, society or economy. The environmental problems can occur when the illegal buildings are constructed in nature reservations or other areas of protected nature. Very common are cottages in the mountains and forests and recreational objects near water reservoirs and lakes [16, 17]. Complications related to the social aspect such as arguments about the ownership, court orders etc. can occur when the buildings are constructed on someone else´s property. One of the problems related to the economy comes from the fact that these illegal buildings are not listed in Cadaster of real estate. Without evidence the state cannot collect taxes for these structures and is losing money that would go to the national budget.

It is clear that illegal construction is accompanied with many problems with impact on various areas and the growing number of these structures creates a serious issue in Slovakia I hope that this project focused on this issue can find a useful solution to this growing problem.

2. Initial problem statement

Nowadays buildings without a valid construction permit are a common issue in Slovakia and also other countries. Based on field investigation and news reports many buildings without a validation are already built and others are being constructed so the number is constantly increasing [15, 16, 17]. Although the exact number is not known because only a certain percentage of the illegal structures have been identified and there is no data base with evidence focusing on buildings separately. Before building can be lawfully constructed a valid permit has to be issued and after the construction it has to be surveyed so a geometric plan containing position, size, shape and other measurements can be created. Based on this plan along with the permit the building can be saved in Cadastre of real estate and drawn in Cadastral map.

It is clear that illegal buildings are a serious issue which can be tackled with geodetic methods since it´s closely linked to this part of science. Various techniques of data gathering and mapping are available today and can be useful when trying to locate buildings with certain parameters. Finding an effective way how they can be identified could bring multiple benefits to the government and society. Based on this assumption an initial problem statement can be formulated:

How can an illegal building be identified?

To find a solution to this problem statement multiple aspects have to be investigated which can be summarized by sub-questions:

Sub-question 1: What is the core idea of building detection?

Sub-question 2: Why are illegal buildings a problem?

Sub-question 3: What is Cadastre of real estate and Cadastral map?

Sub-question 4: What is the definition of illegal building?

Sub-question 5: What base data can be used for illegal building identification?

Sub-question 6: What methods can be applied to the base data to identify illegal buildings?

3. Method chapter

The purpose of this chapter is to establish a general approach to addressing the initial problem statement. To find a solution to the problem of identifying the illegal buildings many questions have to be researched and several tasks have to be fulfilled one after another. In the initial problem statement six sub-questions were asked. These together should lead to answering the initial problem statement: How can illegal buildings be identified.

The method chapter is divided into several paragraphs which correspond with the questions asked in the initial problem statement. These will be reflected and tackled in the following chapter: Pre-analysis.

The core idea of building identification

The main idea of the building identification in this project will be based on a simple assumption. If the building is constructed lawfully, it fulfills all the legal requirements for a building permit being issued and therefore it should be listed in Cadastral map as a part of Cadastre of real estate. If I can use the data regarding building from Cadastral map and compare them with the state of reality, I should be able to detect structures that are present in the reality and not present in the Cadastral map, therefore considered illegal. The state of reality will be represented by data collected by surveying method in specified area. This comparison should detect any outliers between the two data sets which will be analyzed afterwards.

Problematic of illegal buildings in Slovakia

Illegal buildings influence the country's functioning in various ways. They have impact on multiple levels of the society and also the government. These might be the societal, ecological or economical issues. It is important to evaluate these impacts and establish the overall effect and relatively to that decide upon the appropriate actions.

Cadastre of real estate and Cadastral map

The core idea of this project is to compare the legal state with the state of reality. The legal state is stored in Cadastre of real estate and the part that interests me the most is Cadastral map. It is important to have a clear idea about how is it produced, structure of it, what data does it contain etc. to proceed correctly in the experimental part. This is especially important since Cadastral map will be one of the two data sets used in this project.

Illegal buildings

Before I can start looking for a solution it is important to have a clear idea which structures can be considered illegal. In Slovakia the construction of buildings is controlled by a construction office which has the authority to issue permit based on established national laws. The constructor has to apply for the permit before the actual building process starts and in order to get it the building project is taken in examination and it has to meet certain criteria and characteristics to be validated. The construction permit is a necessity for building a structure and without it the object is considered to be an illegal building.

Another aspect necessary to investigate is which structures can be considered as buildings. This means that not all the man-made structures present in environment can be considered as buildings. For example, commercial billboards, garbage sheds, small structures are not considered as buildings by the law and different requirements are connected with them. Therefore, it will be necessary to investigate the construction law to find out which structures can be described as buildings and based on this information to decide on common parameters I will use for identifying the illegal ones.

Building that have a valid construction permit are afterwards listed in Cadastre and also drawn into Cadastral map. Therefore, in theory buildings present in reality and not listed in Cadastral map can be considered illegal. Therefore, if I can identify buildings in a set of base data and compare them with the ones in Cadastral map, I will be able to identify the missing ones.

Base data for illegal building identification

The process of illegal building identification will consist of several steps. First, I need to have a clear definition of what illegal buildings are, so the parameter that will describe them can be determined. Based on the idea that illegal buildings are not present in Cadastral map I will perform the identification by comparing the map to chosen base data. First, I will apply certain method to the chosen base data in order to locate the buildings which will be compared with Cadastral map to distinguish the illegal ones. From this process it is obvious that choosing suitable base data is an essential step to achieve good results in this task.

After defining what illegal buildings are and what parameters to look for in the identification, I can proceed to investigating the issue what base data can be used for this task. Data can be collected in various ways differing in the technology used: total station, scanners, lasers, etc., coverage: gathered from land, from air, precision, resolution, availability and many other characteristics. It is important to have a good overview of the possible data set that can be used for this task, so I can choose the most suitable option or multiple sets of data. The base data structure will mostly determine the approach and the possible methods of working with the data.

Methods for identification of illegal buildings

After investigating the base data that can be used for this task the next step is to research available methods which can be applied to each possible option. Today we live in the age of technological advancements where new solutions are emerging opening innovative ways of treating problems. It is possible that multiple approaches will be available for the base data, therefore it is important to have a good overview of the possibilities to pick the most suitable method and reach valuable results.

The desired method should be effective in the means of covering large areas and preferably automatic, so the man-power needed can be minimized. Efficiency, cost, time required, and other parameters will be considered to choose the suitable method that will be then applied to the base data. Various approaches will be examined along with the possibilities of comparing the chosen data with Cadastral map. After the methods are researched, I can start processing the data and working on the solution to the initial problem statement: How can illegal buildings be identified.

4. Pre-analysis chapter

Method chapter suggest the necessity to investigate the available possibilities before I can decide on the way of approaching this task. To successfully find an answer to the initial problem statement research the sub-questions. These together should lead to solution to the initial problem statement.

Since many different parameters can be chosen when identifying buildings and various sources of data and methods can be applied, I need to conduct a preliminary research on these topics. This chapter will be dedicated to obtaining knowledge and information necessary for further decision-making process. It will be divided into sections corresponding to the sub-questions mentioned before:

- 4.1. Problematics of illegal buildings in Slovakia
- 4.2. Cadastre of real estate and Cadastral map
- 4.3. Definition of illegal buildings
- 4.4. Base data for illegal building identification
- 4.5. Methods for identifying illegal buildings
- 4.6. Conclusion

4.1. Problematics of illegal buildings in Slovakia

As mentioned before the illegal structures are an ongoing problem in Slovakia. With the current ones remaining untacked and a number of new ones increasing it is suspected this problem has a growing tendency [17]. The construction office doesn't have database of illegal buildings based on which it could try to tackle this issue. There are multiple reasons for this situation amongst which the most significant is the insufficient method for illegal building identification and the fact that some of the cadaster offices in smaller regions still don't have complete electronic evidence of the real estates in the area. With no such evidence of buildings it is complicated to make a major progress in this problem.

The illegal buildings in Slovakia were sorted and placed based on research in three general categories considering their common characteristics regarding the missing or invalid construction permit:

- Well-known residences and constructions
- Illegal constructions build by minorities
- Constructions with missing permit on purpose

First category holds big structures that are usually constructed by major developers. These are for example Double Tree Hilton Hotel in Bratislava or private Talapkas mansion near the city of Senec. These are usually notoriously known to the public due to their appearance in TV news. They are owned by private parties as politicians, ex-politicians, big entrepreneurs, investors and so on who built such vast structures illegally. The major problem here is linked with the political connections and bureaucracy which prevent the authorities from the rightful actions. Even though these structures are publicly known, years-long court cases take place

due to flaws and interpretation of the construction law but in the end the structures are still standing. For example, the Talapkas mansion is built partially on the plot owned by the Slovak Railroads which were in court with the owner for the past 3 years and it seems that the residence will remain.

Figure 4.1 Private residence with lake and Hotel in capital city both constructed illegally [16]

The second category embraces the illegal structures that are built by marginalized communities. The biggest minority in Slovakia are the Roma people which are known to build structures without valid construction permits usually on the property not owned by them. They usually build their roughly built huts or cabins in groups creating colonies or settlements. The biggest one is in Jarovnice in eastern Slovakia holding up to 5000 inhabitants. These settlements are known to have no running water, poor hygiene standards and causing problems to the local population living close to them. The issue of this topic has been present for many years now, but no usable solution was found yet. Attempts at reallocating the inhabitants to state-provided accommodation has proved itself ineffective, equally as demolition of the settlements which only brought about diffusion of such settlements into neighboring localities. And tearing down the illegal structures is also not a long-time solution because the communities would move to a different location and build their illegal shacks there. This being a very sensitive topic in Slovakia where the human rights of the minorities have to be preserved but also the rights of the land owners and the impact on local society have to be taken in account.

Figure 4.2 Destroyed apartments and illegal shacks [16]

The last group holds buildings that were built by the constructor without a valid construction permit on purpose, what might sound strange at first. To obtain a construction permit the owner has to fulfill certain requirements that are different for every region, locality, type of building etc. If the owner knows he will not fulfill all of them for example because he wants to build in a forest, vine yard or a protected locality he might decide to build the construction anyway. In Slovakia a permit re-evaluation system is in use enabling owners to reclassify existing constructions where the already constructed building can obtain a valid permit after fulfilling the requirements and this system is many times abused. For example, an owner wants to build a house in the forest area. He might classify the house as a cottage where different requirements apply and obtain a permit for cottage since this structure might be allowed in such area. After the construction is done, he will be imposed a fine for breaching the permit definition amounting to several hundred euros to five thousand euros Afterwards he can apply for issuing a new permit on the already constructed house which he will probably obtain. In the end he might have built a family house which he classified as recreational cottage with a valid permit. This problematic is hard to demonstrate and explain because it is about people trying to find loopholes in the construction law which in many cases only apply a fee for the owner. Therefore, many owners classify their houses as a recreational objects, pay the fee and obtain a valid permit after the construction is completed.

4.2. Cadastre of real estate and Cadastral map

In this subsection I will take a look at Cadastre of real estate and its graphical part the Cadastral map. Both of these are crucial for the understanding and solving the problem this project is focused on.

4.2.1. Cadastre of real estate

Slovakia uses a system called Cadastre of real estate which is run by government and it is designed for work with real estates as plots, buildings, structures etc. Cadastre of real estate can be defined as geometric designation, an inventory and a description of real estate which includes data about the rights to these properties. Cadastre is divided into parts called Cadastral operators corresponding the cadastral territories. Usually one Cadastral operator covers the area of certain territory as city or a region. These are processed on the regional level in weekly intervals the information is updated to the central register of real estate. This register works as a database for distributing information on national level for informational, control and security functions [5].

A Cadastre operator consists of documentation necessary for the cadastre management and development of these parts:

- File of geodetic information
- Descriptive information file
- Collection of documents
- Summaries of the soil fund
- Land books and a railway book

The most interesting for me is the first part, File of geodetic information, which besides other information contains Vector Cadastral Map.

4.2.2. Cadastral map

I will focus mostly on Vector Cadastral Map (VCM) which is more useful for this project then the analog version because we will be using the VCM as a base data set for evaluation of the illegal buildings. VCM is the technical basis for registering real estate in Cadastre, it represents the vector shape of Cadastral map and it is used as a part of the geodetic information system of Cadastre of real estate. The idea is to compare the VCM with the chosen data set to estimate which buildings are constructed illegally based on assumption that structures that are present in the field (represented and located in the data set) and not found in the VCM can be considered illegal.

Figure 4.3 Vector Cadastral Map [5]

The VCM is created in the Coordinate System of the Uniform Trigonometric Cadastral Network (S-JTSK). The basis for VCM creation is both coordinates of the topographical points in the S-JTSK with required precision and complete numbering of the points and graphical overview showing the geometry of the map, the marks, the descriptions and other elements of topography [5].

VCM is produced as a result of:

- Renewal of the Cadastre operator by new mapping,
- Reconstruction of the Cadastre in a shortened procedure
- Vector processing of an existing digital Cadastral map.

The creation of VCM as a result of the redevelopment of Cadastre by a new mapping is achieved by direct interactive or batch processing of numerical and graphical data obtained by photometric or geodetic method. The objects in the field are surveyed either on site or by remote sensing methods and these data are processed into VCM [4].

The creation of VCM as a result of the reconstruction of Cadastre in the shortened procedure is carried out by redrawing the original map by processing the originally measured values for the calculation coordinates of detail points with their current interactive processing into vector shape or by redrawing the original map by a cartometric method with the current one interactive processing carto-graphically obtained data [4].

The creation of VCM as a result of vector processing of an existing digital Cadastral maps are used if numeric results exist (list of detailed coordinate points, drafting drawings), and an overview of the point numbers from the creation of a major scale map [4].

The VCM map is made up of the individual elements that are arranged in drawings in a precisely defined structure. The basic elements of VCM are:

- Points
- Lines
- Signs
- Polygons
- Texts

VCM elements are grouped into objects where each group has a defined layer in which it is located. Objects are made up of individual drawing elements and may contain text information used to identify the object or provide further information.

Cadastral parcels show plots that represent part of the earth's surface separated from adjacent parts by the boundary. The positional accuracy is expressed by the accuracy of the parcel border fracture points. The accuracy of the parcel boundary fracture points is expressed by the detail point quality code as follows: Code 1 and 2 - Basic Coordinate Median Error 0.08 m, Code 3 - Base Coordinate Error 0.14 m, Code 4 - Base Coordinate Error 0.26 m, [5].

Layers

VCM objects are grouped into layers based on the characteristics they represent for example layer holding boundaries of plots, house numbers etc. Layer approach allows manipulation with the selected category, displaying, coloring and so on. The defined layers of VCM are:

- Points – holds geodetic points
- *Boundary layer of plot usage*
- Layer of marks
- Descriptive layer
- Polygon Point Field Layer
- Boundary layer cadastral territories
- *Plot borders*
- Plot numbers

Buildings

For this project, the layers Plot borders and Boundary layer of plot usage are interesting since here are the plots of buildings marked. A major complication relating to this project is that buildings are not drawn separately and don't have specific status in the current state of the VCM meaning that there is no layer holding just buildings. Cadastral law states that buildings must have a separate plot that covers the build-up area. In this case the building boundary can be identified in the plot border layer. Or in the second case, buildings can be registered in Cadastre by the plot usage as build-up are. This is mostly the case when new structure is added to already existing building and then the building boundary can be identified in the layer of plot usage. This complication will have to be solved by either extracting the buildings or otherwise editing the layers or finding another interpretation of cadastral information that can be used for this task.

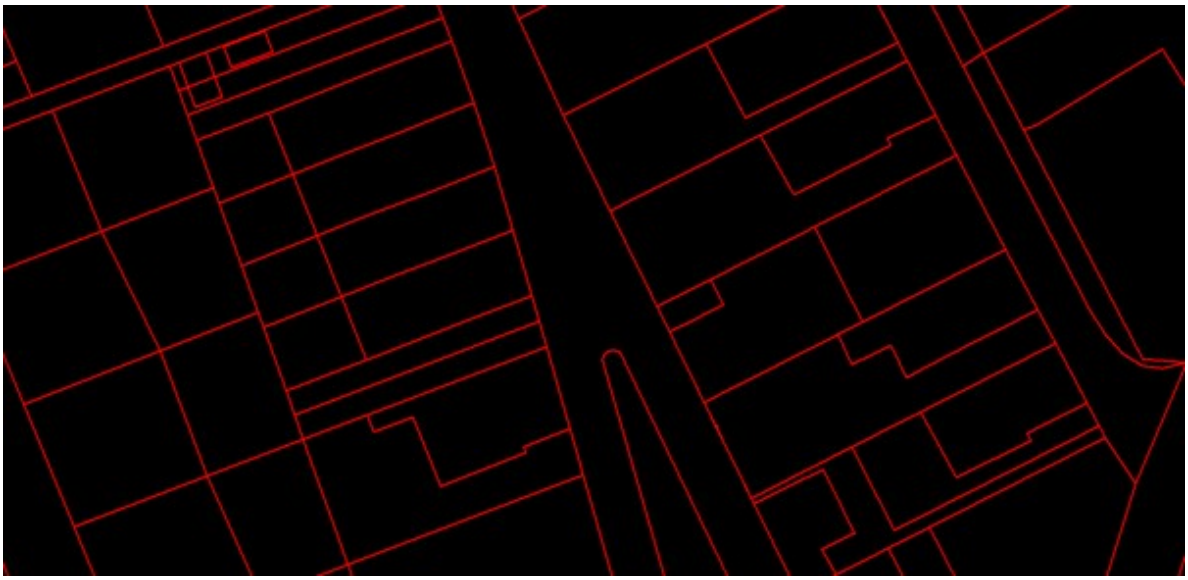


Figure 4.4 Layer with plot borders [5]

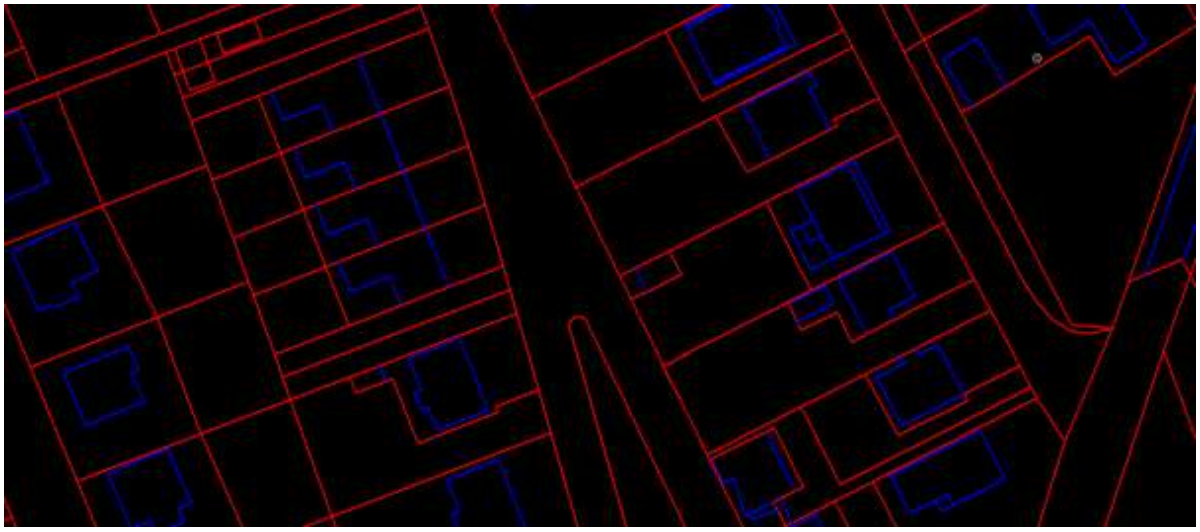


Figure 4.5 Plot borders plus the Boundary layer of plot usage (blue) [5]

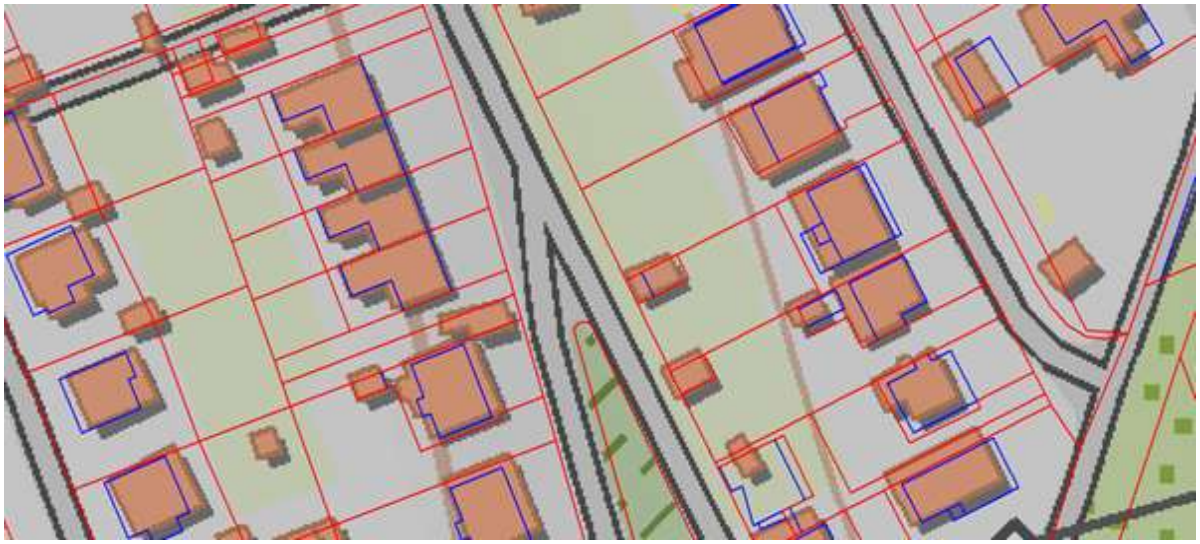


Figure 4.6 Both layers on top of raster with buildings [5]

Graphical files in the exchange format are text ASCII files of type VGI (vector graphical interface) allowing the storage and transmission of VCM data independently of the used graphic software. The exchange format of the data is in the form of comprehensible and easy-to-read text files that can be viewed or edited by any text editor.
When creating file formats in a defined format, the loss of the information content of the artwork is minimized. The format is expandable for future items if needed in the future [5].

4.3. Definition of illegal buildings

There are certain rules to be followed while constructing a building and numbers of authorities are in charge of accepting the proposal for the plan of construction. These rules are established and defined by the law. They are executed by Local Construction Department present in each city. These departments guided by the law supervise the construction work in the region allocated to them. They are responsible for issuing the construction permits as well.

Illegal building can be described as construction work (or the result of such) without a valid construction permit. Several problems influencing different spheres of human life as environment, social and economic aspects are related to these buildings. The problems are ranging from potential technical hazards on uncontrolled construction sites and in finished buildings; illegal construction activity in protected environment as nature reserves; to losses on taxation for the government.

Figure 4.7 Illegal buildings in Slovakia [17]

Construction law definitions of buildings

Buildings and structures are precisely defined and sorted by a law. In Slovakia, it is the construction law no. 50/1976 Zb. Law of spatial planning and construction order and it defines building as:

Building is a constructional structure built by construction works of construction products which is firmly connected to the ground or whose installation requires the modification of the substrate. Fixed connection to the ground is understood as:

- a solid base connection
- fastening by machine parts or by welding to a rigid base in the ground or other structure
- Anchorages by piles or ropes with anchor in the ground or other structure
- as a connection to networks and equipment of the territory
- underground location

The constructions are divided into *buildings* and *civil engineering constructions* according to their construction and purpose.

Buildings are spatially concentrated roofed buildings, including underground spaces, which are technically suitable for construction and intended to protect people, animals or things; they do not have to have walls, but they have to have a roof. By purpose, they are divided into *residential buildings* and *non-residential buildings.*

Engineering structures are highways, roads, local and purpose communications, waterfront, sidewalks and uncovered parking lots, rails, cableway and other railways, construction of power plants, gas and waste incinerators etc.

Residential buildings are buildings of which at least half of the floor area is intended for housing. The residential buildings include:

- apartment buildings,
- family houses,
- other residential buildings, such as children's homes, student homes, retirement homes and homeless shelters.

Non-residential buildings are buildings in which more than half of their usable floor space is used for non-residential purposes. Non-residential buildings are:

- hotels, motels, boarding houses and other accommodation facilities for short stays,
- buildings for administration and management, banks and post offices,
- buildings for trade and services, including car service and service stations.

The construction law includes also criteria that states: construction permit doesn't have to be issued for buildings if their built-up area does not exceed 25 m^2 and height 2,5 m but for such buildings different permission is required.

Based on the definition from the construction law I decided not to investigate all of the structures but focus on the category defined as *Buildings,* which consists of two subcategories:

- *residential buildings*
- *non-residential buildings*.

There are several reasons for this decision. Based on the investigation and news reports [15, 16, 17], the illegal buildings are in most cases the residential and non-residential buildings for example houses, cottages or other recreational buildings, hotels or industrial halls which are all included in the category defined as buildings. These are the ones with most negative impact on the society in the means of taxation, environment or social relations. Another reason is to eliminate problems with identifying small structures which can be easily confused with other object not related to the issue. To avoid such confusion, a criterion based on the restriction for small buildings mentioned above will be added and only buildings with built-up area that exceeds 25 m^2 (and height over 2,5 m) will be investigated. This way I will avoid the identification of structures as trash bin sheds, electric units and others that are not classified as buildings but as engineering constructions.

4.4. Base data for illegal building identification

Thanks to advanced technology available these days we can chose from different data sets that could be used for this task. They differ in many aspects as availability, technology, coverage, price, software requirements, etc. therefore it is important to evaluate these possibilities to pick the suitable one or a combination of multiple options.

This is a list of data sets that I found usable for this problem:

- Data from geodetic ground survey
- Satellite images
- Orthophoto images
- Data gathered by LiDAR

Each data set has its characteristics which have different pros and cons in relevance to this task. It is important to evaluate them before we can decide on a suitable data set which will be used for building identification or if possible, a combination of multiple data to obtain better and more valuable results.

4.4.1 Geodetic ground survey

This data is gathered by a surveyor performing a ground survey. Such survey is conducted using geodetic equipment usually total station or a GNSS receiver. These instruments are capable of measuring the position, angles and distances with high precision. Precision depends mostly on the instrument and method used. The total station, depending on the type, has a relative precision of measuring the position around 3 mm and a GNSS receiver using RTK method is capable of measuring points with a 3 cm deviation in the position [5].

In the ground survey, the surveyor comes to the site and by using the geodetic instruments he measures the desired parameters. The benefit is that surveyor being already present at the site he can make observations and gather additional information about the site, property, relations and other aspects that can be beneficial. In the means of illegal building identification these observations can prove to be very useful.

The main disadvantage of the ground survey is that it is time consuming and requires considerable man-power in order to be conducted. Because of this it would be very expensive to cover large areas and gather enough data to perform the building identification on bigger scale for example for a whole city. This data could be used for example after the illegal buildings are located to perform a final check by a surveyor coming to the site and conducting a measurement. The geodetic measurement could be also a part of the solution of obtaining the validation for construction permit.

Advantages:

- High precision (up to 3 cm in position)
- Surveyor present at the site (can perform inspection of the area)

Disadvantages:

- Small area coverage (one building at time)
- Time consuming (requires lot of man-power therefore also expensive)

Figure 4.8 Geodetic survey with the use of Total station [18].

4.4.2. Satellite images

These images are produced by specialized satellites equipped with technology capable of capturing images from big distances. These are Earth observation satellites specifically designed for observation of the Earth from orbit. They are intended for non-military purposes such as environmental monitoring, meteorological observations, map making etc. Majority of the Earth observation satellites are designed to be operated at relatively low altitude around 600 – 800 km, therefore they are called low orbit satellites [3].

The resolution of satellite images depends mostly on the quality of the instrument used, height of the flight, atmospheric conditions. Earth Observation Satellites can be divided in 3 groups based on the resolution of the images they produce [20]:

- High resolution satellite images: 0.3m – 2m
- Medium resolution satellite images: 2m – 20m
- Low resolution satellite images: + 20m

High resolution satellites are for example GeoEye (50cm), WorldView-3 (30cm), IKONOS (1m) satellites (satimagingcorp.com) which provide images that can be used for various purposes from mapping, monitoring certain area (in case of natural disaster for instance) to providing precise visual information. Images from this resolution group are mostly bounded by license or copyright and are sold by the provider to the customers, i.e. they are not freely available. Some of them can be accessed but usually only as examples, when the user wants to access data from a certain location, he has to buy the data first [2].

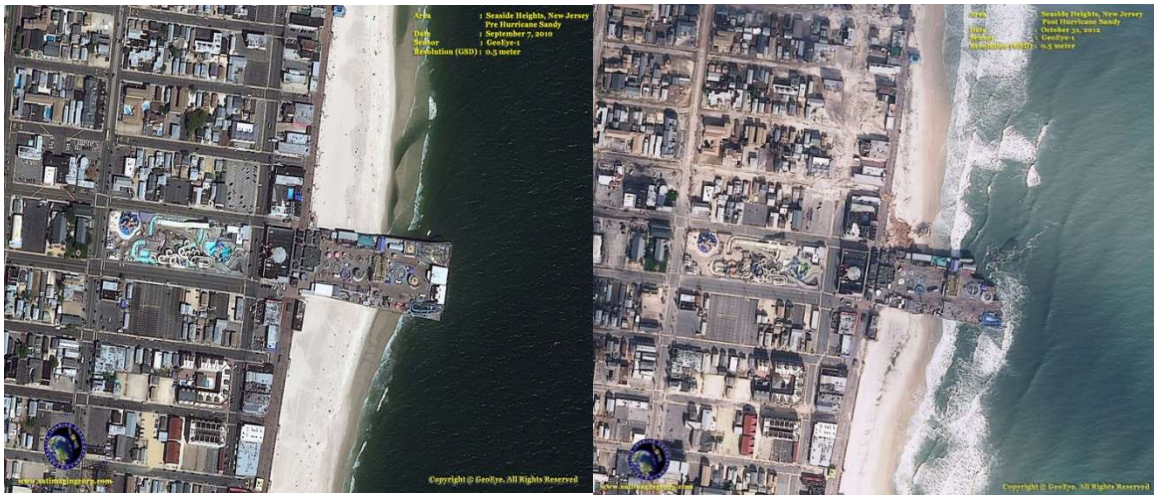


Figure 4.9 Satellite images from GeoEye (0.5m resolution), before and after hurricane [2]

Medium resolution images are used for monitoring natural phenomena or mapping the development in certain environments and producing change detection maps. Satellites in this category are for example Sentinel-2, LANDSAT 7, RapidEye or Dove. Data from these are more often free of charge. Especially the Sentinel and LANDSAT missions' data are available on various portals from different epochs of mapping making it very useful when focusing on the development in certain area over time [2].

Figure 4.10 Satellite image from ALOS with 2,5m resolution of Saint Petersburg [2]

Big advantage of satellite data is the coverage. Satellites operating in the low orbits are capable of circling the earth within a few days' period which makes them very useful in monitoring certain areas and processes. This way they can cover a large area and monitor it within a short time cycle. Another advantage of satellite image is they are available through various sources to the user. They can be obtained through national space or mapping agencies or from private sector providers. They can be accessed either in form of raw data or as processed images depending on the provider of the data [20]. Beside the raw data there are many applications that use satellite images and the most known is Google Earth.

The fact that satellite images cover large area of the Earth with a relatively high resolution brings also disadvantages. The databases to store such amount of data are rather huge and also the image processing where the raw data are transformed into useful images are demanding on technological equipment in form of hardware and software and are very time-consuming.

Another aspect that is important to consider is the resolution of the data needed to be able to successfully identify buildings from the images. Since the high resolution data are usually not freely accessible and the medium resolution do not have to be sufficient for this task the question of availability of the data might be considered as both advantage and disadvantage. Further research would have to be applied to estimate the minimal and the optimal resolution for identifying buildings from satellite images.

A disadvantage of satellite images related to this task is that these images provide only 2D projection of the reality. After they are processed, georeferenced and so on the image shows perpendicular or upright projection of the reality. This fact limits the techniques for identifying buildings to those based on spectral parameters as color, intensity or shape and because it is only in 2D coordinates the factor of position or height differences cannot be used for identification. For 3D coordinates the satellite images have to be supplied with a digital terrain model (DTM). In this case the image is positioned and laid on top of the DTM and adjusted so that the 2D position of point or object in the image corresponds with the position in DTM which holds the height coordinate [2]. The satellite images are stored in a raster format.

Advantages:

- Coverage of the most part of the Earth
- Availability of medium (low) resolution images (resolution up to 2 meters)

Disadvantages:

- Large data files (tens of Gigabytes)
- High resolution images usually licensed or not available to public
- Only spectral characteristics can be used for identification

Characteristic:

- Data structure – raster format

4.4.3 Orthophoto images

An orthophoto, orthophotograph or orthoimage is an aerial photograph or image that has been geometrically corrected. The process of correction of these images is called ortho-rectification and in the result the scale of the image should be uniform, and the image should lack distortion. Unlike an uncorrected aerial photograph, an orthophotograph can be used to measure true distances, because it is an accurate representation of the Earth's surface, having been adjusted for topographic relief, lens distortion, and camera tilt allowing their use as base maps for digital mapping and analyses in a GIS [20].

In other words, an orthophoto is an adjusted photograph taken from an infinite distance, looking straight down to nadir, where perspective must be removed, but variations in terrain should be corrected for. Multiple geometric transformations are applied to the image, depending on the perspective and terrain corrections required on a particular part of the image. Orthophotos are usually created from pairs of aerial photos by a process called ortho-rectification. By merging multiple orthophotos we can create a raster image called orthophoto mosaic. Orthophotos are commonly used in geographic information systems to create maps. Once the images have been ortho-rectified and processed they can be aligned or registered with known real-world coordinates creating an orthophoto map [22].

The advantage of orthophoto images is that they can cover large areas and since they are created from aerial images their resolution is higher than in satellite images. The resolution of orthophoto images can be up to 10 cm.

The process of ortho-rectification and creating an orthophoto map is rather complicated and demanding on both hardware and software. Orthophotos are available online in national data sets for some areas usually provided by national mapping agencies which conducted the surveys either free of charge or priced depending on the agency policy. If they are in a form of a separate orthophotos a processing software would be required to obtain a map by connecting them together. The orthophotos are frequently used by various online applications providing a visual information to the user.

As well as satellite images, the orthophotos contain only 2D coordinates unless projected on top of a DTM. This means that only spectral characteristics as color, intensity so on, can be used for identification but not the position characteristics as height differences. Also, the orthophoto images are in a raster form, therefore the same or similar methods can be applied to them as to satellite images, limiting the processing options to methods dealing with raster [2].

Figure 4.11 Orthophoto image of Bratislava [4]

Advantages:

- Resolution up to 10 centimeters
- Coverage (most of the urban areas)

Disadvantages:

- Availability (some countries have public database but in Slovakia most orthophoto data sets are licensed)
- Demanding processing and ortho-rectification (require both hardware and software)
- Only spectral characteristics can be used for identification

Characteristic:

- Data structure – raster images

4.4.4. LiDAR data

LiDAR (also called LIDAR, Lidar, and LADAR) stands for *Light Detection and Ranging,* is a remote sensing method that uses light in the form of a pulsed laser to measure ranges (variable distances) to the Earth. LiDAR uses ultraviolet, visible, or near infrared light to image objects. It measures the distance to a target by illuminating the target with pulsed laser light and measuring the reflected pulses with a sensor along with the time of flight and amplitude shift of the pulse. These light pulses, combined with data recorded by the airborne systems such as IMU, GNSS generate precise, three-dimensional information about the shape of the Earth and its surface characteristics [7].

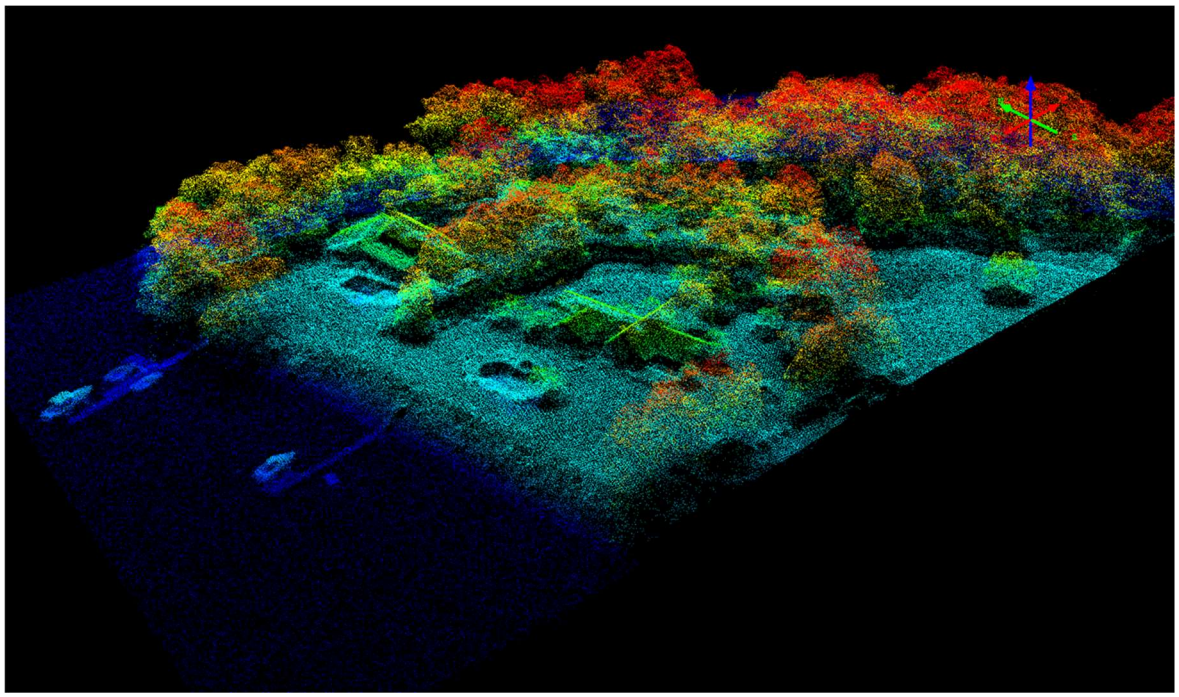


Figure 4.12 LiDAR data set [24]

LiDAR is commonly used to make high-resolution maps with applications in geodesy, archaeology, geography, geology, seismology, forestry, atmospheric physics, laser guidance, airborne laser mapping etc. The technology is also used in control and navigation for some autonomous vehicles. A narrow laser beam can map physical features with very high resolutions; for example, aircraft can map terrain at 20 cm resolution or better [8].

There are several major components to a LiDAR system:

1. Laser
2. Scanner and optics
3. Photodetector and receiver electronics
4. Position and navigation systems

When an airborne laser is pointed at a targeted area on the ground, the beam of light is reflected by the surface it encounters. A sensor records this reflected light to measure a range. When laser ranges are combined with position and orientation data generated from integrated GPS and Inertial Measurement Unit systems, scan angles, and calibration data, the result is a dense, detail-rich group of elevation points, called point cloud. Each point in the point cloud has three-dimensional spatial coordinates (latitude, longitude, and height) that correspond to a particular point on the Earth's surface from which a laser pulse was reflected [8]. The point clouds are used to generate other geospatial products, such as digital elevation models, canopy models, building models, and contours.

The airborne LiDAR (also airborne laser scanning - ALS) is a laser scanner attached to an aircraft that captures a 3-D point cloud of the landscape. This is currently the most detailed and accurate method of creating digital elevation. One major advantage in comparison with photogrammetry is the ability to filter out the data representing the vegetation from the point cloud data set and create a digital surface model which represents ground surfaces such as rivers, paths, cultural heritage sites, etc., which are concealed by trees.

There are several advantages to the LiDAR technology. The main are high point density (depending on conditions, up to tens of points per meter) and high spatial resolution is obtained when mapping the area with this instrument. The data are stored in a form of a point cloud where each point represents a measured feature from reality with three coordinates – latitude, longitude and altitude. This means that every point captured by the instrument has coordinates in 3D system what brings great advantage [28]. When processing the data, we can detect and classify the points representing the ground surface which can be used as a height reference for the other points meaning that we can compute the vertical distance of the points from the ground. This can be used along with the spectral characteristics to detect the objects as structures and building roofs.

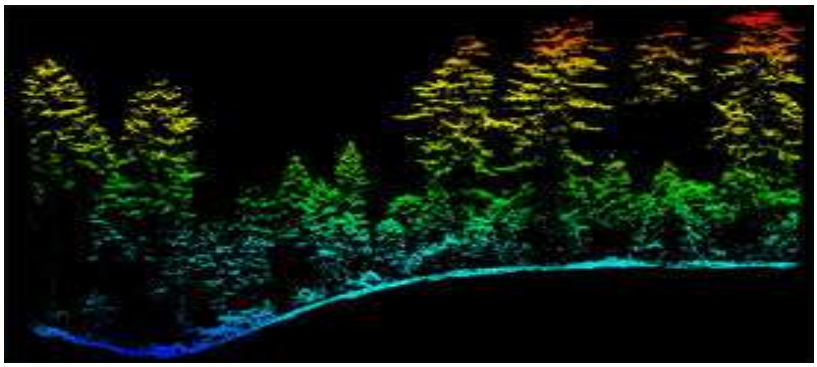


Figure 4.13 Ground points captured under the vegetation [24]

Important feature that is not achieved by any technologies mentioned before is that LiDAR is capable of mapping the ground underneath the vegetation. This allows us to filter away the vegetation data and obtaining a ground model hidden underneath. This is possible because of the working principles of LiDAR where multiple laser pulses are emitted at once and the ground points can be identified from multiple returns of the signal.

The major disadvantage of LiDAR data is that they are not publicly available. Since the technology is very expensive and the aerial mapping is usually conducted within a precise area of interest the data are owned by the executors of such survey and in some cases the access to them is available online and can be bought.

Advantages:

- Coverage (when used on plane covers large areas)
- Precision (can be up to couple of centimeters in position for each point in the cloud)
- Resolution (usually around 10 points per meter are captured depending on instrument, conditions and intensity)
- Spectral characteristics along with 3D position can be used for identification

Disadvantages:

- Availability (data sets are owned by the conductors of the survey and licensed)
- Processing (the raw data needs to be connected together, processed and filtered)

Characteristic:

- Data structure – Point cloud
- Data collected under vegetation

4.5. Methods for identifying illegal buildings

This part of the pre-analysis chapter will investigate possible methods which can be applied to the data mentioned before to reach the solution of identifying illegal buildings. First two parts of this sub-chapter will focus on two types of formats of base data and what methods can be applied to them and the third part will talk about the possibilities of comparing vectors as outcome of building identification and VCM.

Usable data set should contain valid information with characteristic as precision, resolution suitable for the task of building identification and covering area large enough to execute this survey and obtain valuable results. Coverage is crucial for the solution to be applied on large areas as town, cities etc. Because the data set from ground survey cover very limited area it doesn't fulfill the last requirement about the coverage, I decided to eliminate it from the data set possibilities.

From now the possible data sources will be limited to satellite and orthophoto imagery and LiDAR data. The main characteristic of these data sets is the format of the data, either raster or point cloud and this is also the main difference between them related to this project. Based on the data format I will divide the evaluation of the possible methods to two groups:

- raster format methods
- point cloud format methods

4.5.1. Raster methods

In its simplest form, a raster consists of a matrix of cells (pixels) organized into rows and columns (grid) where each cell contains a value representing information, such as temperature, height, count. Raster can be digital aerial photographs, imagery from satellites, digital pictures, or even scanned maps. The cell values represent the phenomenon portrayed by the raster dataset such as a category, magnitude, height, or spectral value. The category could be a land-use class such as grassland, forest, or road. A magnitude might represent gravity, noise pollution, or percent rainfall. Height (distance) could represent surface elevation above mean sea level, which can be used to derive slope, aspect, and watershed properties [24]. Spectral values are used in satellite imagery and aerial photography to represent light reflectance and color.

The dimension of the cells can be as large or as small as needed to represent the surface conveyed by the raster dataset and the features within the surface. The cell size determines how coarse or fine the patterns or features in the raster will appear. The smaller the cell size, the smoother or more detailed the raster will be. However, the greater the number of cells, the longer it will take to process, and it will increase the demand for storage space. If a cell size is too large, information may be lost, or subtle patterns may be obscured [23]. For example, if the cell size is larger than the width of a road, the road may not exist within the raster dataset.

The disadvantages of the raster format are mostly linked to the size of the cells. If the cells are too big there can be spatial inaccuracies due to the limits imposed by the cell dimensions. Raster datasets are potentially very large. Resolution increases as the size of the cell decreases; however, normally cost also increases in both disk space and processing speeds. For a given area, changing cells to one-half the current size requires as much as four times the storage space, depending on the type of data and storage techniques used.

Raster format methods

Image classification is the process of assigning land cover classes to pixels for example forest, urban, agriculture, terrain or other classes. Extracting land cover information from remotely sensed imagery can be performed through multiple methods including: parametric and nonparametric statistics, supervised or unsupervised classification, hard or soft set classification logic, per-pixel or object-oriented classification, or a combination of methods mentioned above. These are the three most widely used classification techniques for raster images [24]:

- Unsupervised image classification
- Supervised image classification
- Object-based image classification

Unsupervised and supervised image classification techniques have been the two most common approaches while object-based classification has been used more lately because it's useful for high-resolution data.

4.5.1.1. Unsupervised classification

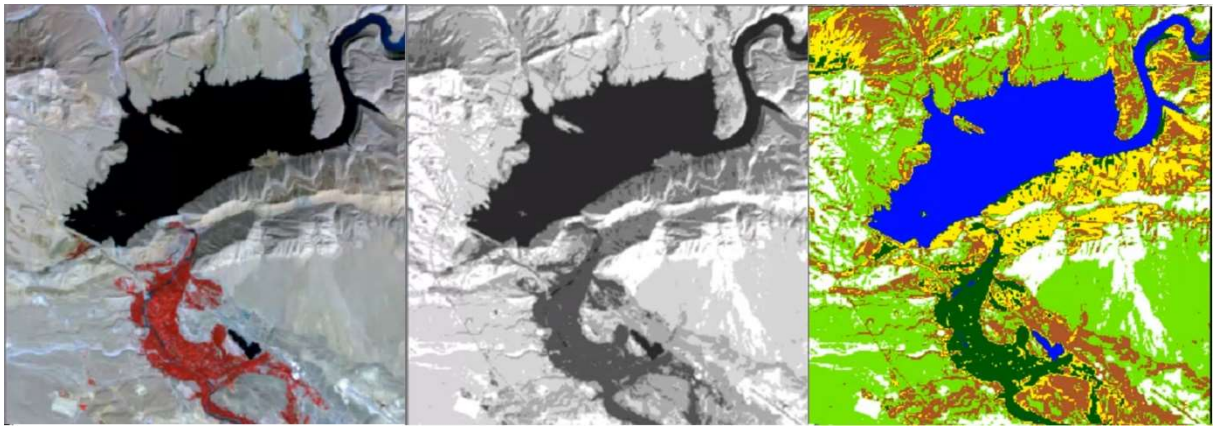


Figure 4.14 Unsupervised Classification [23]

Unsupervised classification first groups pixels into “clusters” based on their properties. In order to create “clusters” image clustering algorithms such as K-means and ISODATA can be used. After picking a clustering algorithm, you identify the number of groups you want to generate for example 10, 20 or 45 clusters. These will be unclassified clusters which we have to identify with land cover classes. For example, if we want to classify vegetation and non-vegetation, we will have to merge all the created clusters into only 2 clusters [23]. Overall, unsupervised classification is the most basic technique and an easy way to segment and understand an image.

Unsupervised classification steps:

- Generate clusters
- Assign classes

In the first step the pixels are sorted into clusters representing an unknown class. The number of different classes that will be generated is selected by the user. After the clusters are identified the user manually assign the desired land cover classes to them.

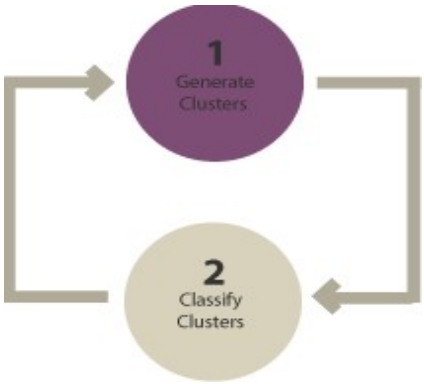


Figure 4.15 Unsupervised Classification Diagram [24]

In general, it is good to select colors for each class. For example, set water as blue for each class. After setting each one of your classes it is possible to merge the classes by using a tool for reclassification for instance if we want to join different classes of trees and plants under one major class called vegetation.

On figure 4.14 are three stages of the unsupervised classification process. First image from the left represents the original raster image before the process. Middle image displays the situation after the pixels were sorted into clusters and the image on the right shows the final stage where the land cover classes were assigned to corresponding clusters along with defining colors for each class [23].

To demonstrate the workflow and requirements on user I picked the unsupervised classification using the Iterative Self-Organizing Data Analysis Technique (ISODATA) clustering algorithm. Minimal user input is required to perform the classification, but extensive user interpretation is needed to convert the generated spectral clusters into meaningful informational classes.

ISODATA is a modification of the *k*-means clustering algorithm in that it has rules for merging clusters, based on a user defined threshold, and splitting single clusters into two. ISODATA is considered self-organizing because it requires little user input. The algorithm begins by placing arbitrary cluster means evenly throughout a 2D area based on the mean and standard deviation of each band used in the analysis. These cluster means are recalculated and shifted in feature space based on a minimum distance from mean classification rule through each iteration. Once the user defined convergence threshold has been reached, iterations cease, and the resulting spectral clusters can then be interpreted [25].

The advantages of unsupervised classification is no extensive knowledge of the study area is required. Little user input is needed to perform unsupervised classification which minimizes the likelihood of human error. However, the analyst has little control of the classes generated and often these clusters contain multiple land covers making interpretation difficult. The main difference from supervised classification is that it does not provide sample classes, but the user has to manually assign the classes to the clusters [26].

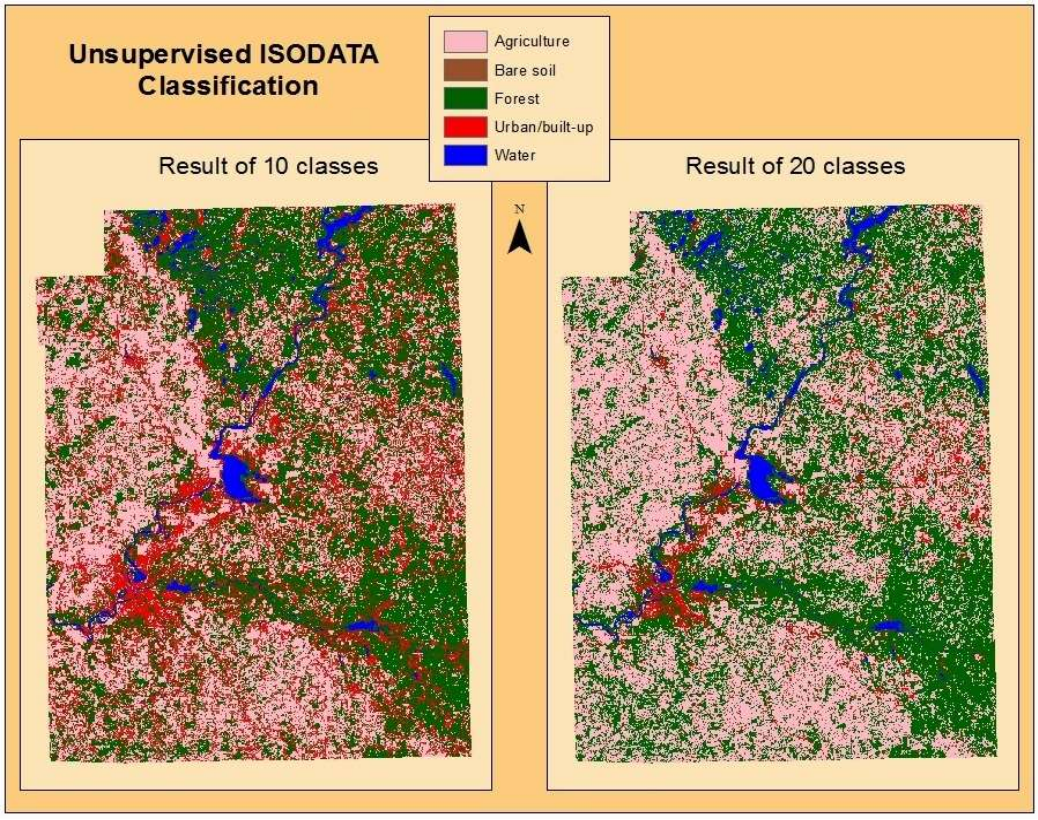


Figure 4.16 Example of unsupervised classification with different numbers of clusters [23]

In figure 4.16 we can see examples of two unsupervised classifications with 10 classes (left) and 20 classes (right). In order to display the map from classified image, the number of classes needs to be recorded from 10/20 down to the desired 5 defined classes of land cover [23]. There is a visible difference between these two images where the one with 20 classes is more detailed and accurate what is caused by the fact that more clusters were generated which allowed for more specific spectral characteristics to be singled out and classified accordingly.

4.5.1.2. Supervised classification

To perform supervised classification, an analyst will collect samples areas of known land cover, commonly referred to as training samples, from the image which will be used to train a classifier. The classifier will then classify the entire image based on the information gathered from the collected training samples. Preferably, training samples would be collected based on prior knowledge of the area using the most accurate means available (GPS, topographic survey, etc.) However, for some applications this is impractical and high resolution imagery can be used to designate areas for training samples.

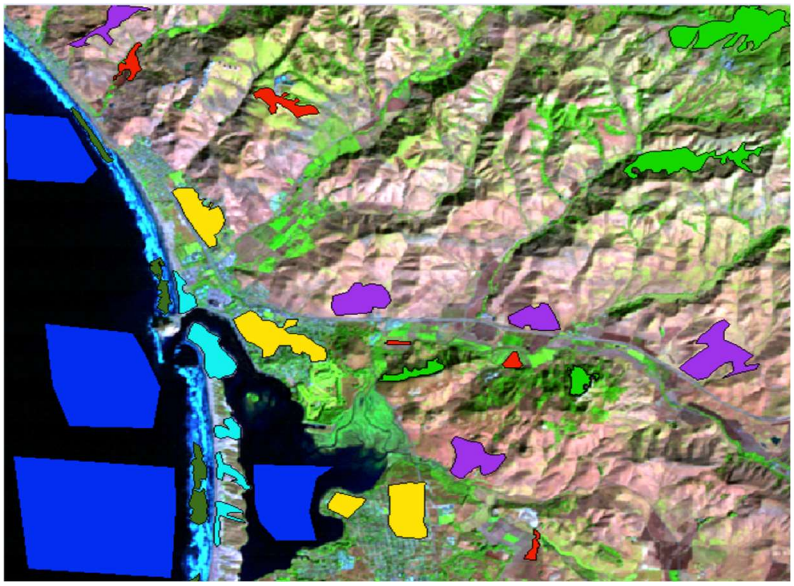


Figure 4.17 Example of training sites in Supervised Classification [20]

Important factors to consider when collecting training samples include: the number of training samples, the number of pixels, shape, location, and uniformity. The specific number of training samples for individual informational classes may vary depending on the nature of the image (spectral diversity) and project (special emphasis or resource availability). The shape of training samples will normally be a derivation of a polygon. Training samples should be distributed throughout the entire image to account for spectral variability in the informational classes and be located within a uniform and homogeneous land cover. An important concept to keep in mind is the geographic signature extension problem where differences in spectral characteristics of the same informational class result produce differences from a variety of factors like, soil type, moisture or crop species. To reduce the errors that result from the geographic signature extension problem, training samples should cover all possible variations of the desired informational classes for example for vegetation class collect a sample of forest, meadows, grass land and other types of vegetation present in the area [25].

Supervised Classification Steps:

- Select training areas
- Generate signature file
- Classify

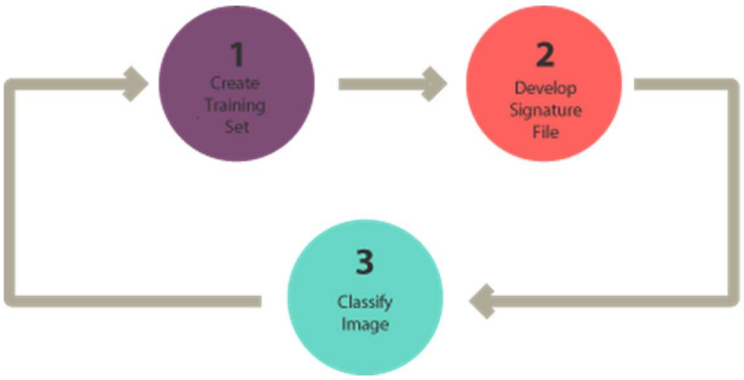


Figure 4.18 Supervised Classification Diagram [24]

After a sufficient number of samples is selected for each class, next step is to develop a signature file. Here all the samples for each class are merged together to produce one signature sample for each class. Then the whole image is classified based on the signature classes developed by user in the training set. Several options for classification were developed for example maximum likelihood, minimum distance, iso cluster, class probability and principal components [26].

Class #	>	Signature Name	Color
1	▶	Water	
2		Forest	
3		Agriculture	
4		Urban/Built-up	
5		Bare soil	

Figure 4.19 Example of signature classes from training samples [24]

Advantages of supervised classification over unsupervised is the control the analyst has over informational classes produced and not having to interpret the spectral clusters generated by unsupervised methods. Also, by analyzing the quality of training samples, the classification can be improved before its actually performed. However, collection of proper training data can be time-consuming, expensive, and may not fully represent the desired informational classes leading to errors in classification.

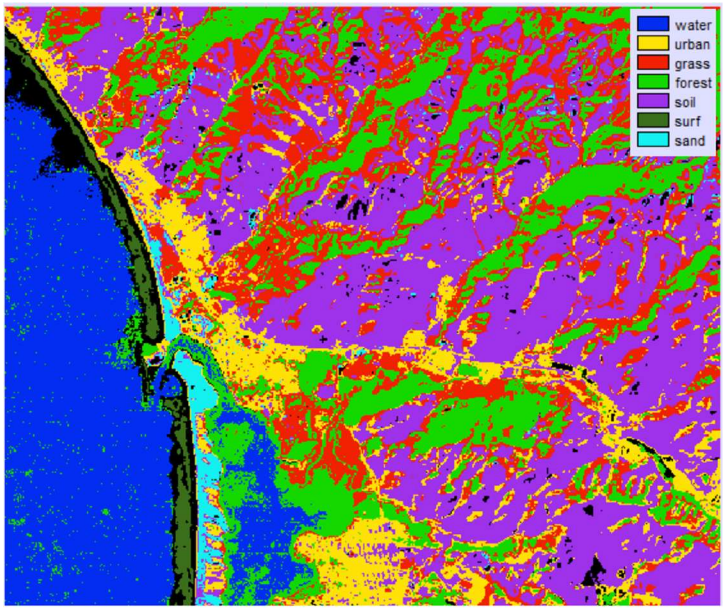


Figure 4.20 Result of Supervised Classification [20]

4.5.1.3. Object based image

Traditional pixel-based image classifiers assign a land cover class only per pixel. All pixels within the image are the same size, same shape and don't have any relations of their neighbors. By accounting for spatial properties, like distance, texture, and shape, an object-based classifier results in a more natural looking and often times more accurate classified image. Object-based classification segments an image into areas based on both spectral and spatial homogeneity criteria. An analyst can then classify specific objects and use them as training samples to classify the entire image. Object based image segments an image grouping small pixels together into vector objects. Instead of a per-pixel basis, segmentation automatically digitizes the image.

Figure 4.21 Object-Based Image Segmentation [30]

When the image is segmented, the process groups pixels to form objects. Suddenly, land cover features can be observed. After the image is segmented the objects can be classified based on spectral, geometrical and spatial properties. This is possible because each object has various statistics associated with them now. Objects can be now classified based on geometry, area, color, shape, texture, adjacency and more [30]. After the image is segmented the process is similar to supervised classification where training samples are chosen by the user from the segmented image base on which is the classification performed for the whole image.

4.5.2. Point cloud format

Point clouds are a collection of points that represent a 3D shape or feature. Each point has its own set of X, Y and Z coordinates which can be accompanied by additional attributes as intensity, return number or even class may be recorded in case of LiDAR. We can think about a point cloud as a collection of multiple points which brought together start to show qualities of the feature that they represent.

Point clouds are most often created by methods used in photogrammetry or remote sensing. Photogrammetry uses photographs to survey and measure an area or object. A combination of photographs taken at many angles can be used to create point clouds. Remote sensing is a method of collecting data of the Earth by use of satellites or aircrafts. On these aerial vehicles, LiDAR sensors can be mounted to collect information about the shape of the Earth and its features

A point cloud is a type of geometry that is useful for storing large amounts of data. The use of LiDAR allows for collection of data in large area in relatively short time period with high accuracy up to couple centimeters depending on the conditions, method and equipment used for forestry canopy measurements, landscape modeling, etc. Point cloud geometry allows for quick and efficient processing of a large collection of points in 3D space that represent the external surfaces of objects [7]. Together, these points form a model which can be transformed, and visualized. Some operations of the point cloud geometry involve thinning, splitting, and combining to produce a more useable data set. There are two widely used formats for storing point cloud data, ASCII and LAS (Long ASCII Standard) format.

Point cloud format methods

Identifying objects as buildings in a point cloud data is done in a process called filtering and classification. Before this can be done the data needs to be edited. Even though the data might be already pre-processed from the raw output from the measurement by connecting separate observation flight lines into single set, the data still needs to undergo a filtering.

4.5.2.1. Filtering

When conducting aerial mapping by LiDAR the plane flies over desired area and the laser signal emitted by the instrument interact with the first surface it hits and is captured by the sensor and points with coordinates are produced by processing these measurements and stored in form of a point cloud. The objects can be soil, vegetation, buildings, electric lines or even birds, everything that comes into the path of the emitted signal. The laser of LiDAR usually consists of four emitters so four separate pulses are emitted and captured upon return.

Filtering of the LiDAR data can be defined as a process of distinguishing and separating objects as buildings or vegetation and ground surface points. a number of algorithms have been developed to enhance the extraction process in which surface points are separated from the point clouds obtained by LiDAR. Most of these filtering algorithms classify data into at least two categories, namely object and surface, using local neighborhood operations on points [13]. Every filter makes an assumption about the structure of the bare earth points in a local neighborhood. However, they often face difficulties in practice, due to the variety and complexity of objects in urban environments and interacting with the discontinuous features of the bare earth therefore it's important to find suitable method for specific data set [7]. There are many existing filtering algorithms where the most commonly used can be in sorted in four classes:

- Morphologic filtering
- Progressive filtering
- Active shape filtering
- Segmentation / Clustering

The first class is similar to mathematical morphological filtering. The filtering method can be based for example on comparing height differences between two nearby points to determine an optimal filtering function and to preserve terrain features. In general, it deals with the object shape or shape measurement which is used to estimate the ground surface [9]. These can be based on using dilatation or erosion masks to remove non-ground candidate and usually are combined with other types of filters to achieve better results.

The second class of filters works progressively. Part of edge points are first identified and then used to construct an initial Triangulated Irregular Network (TIN). More and more terrain points are identified based on this TIN and then added to the classification from the surroundings. The principle is that identification starts with relatively small window which is gradually enlarged as more points are identified. Various approaches to progressive filtering have been invented for example Chen et al. used two set process where the algorithm first identifies most forest points using relatively small window iteration, then repeats the filtering process with larger window to further screen the area [9].

The third class of algorithms creates an estimate of initial digital terrain model (DTM) which is progressively made more dense as new points are processed in order to approximate the bare earth. The ground surface can be estimated by employing active shape models where adjustable model is used to fit the bare earth. Hierarchical approach can be applied where first a coarse DTM is generated which is then refined hierarchically [28]. The first estimate can be based for example on assumption that ground areas are usually smooth surfaces therefore points with strong curvature are defined as non-ground points.

The last class of filters is based on segments. Points are segmented by clustering analysis, region growth, or edge detection techniques, based on height, normal, curvature, slope, or gradient differences, within a small neighborhood. These segments are then classified based on their contextual information [13]. The ground features are separated into multiple categories for example prof. Filin uses three types of parameters: point position information, elevation difference to neighbors and parameter description from the point to its tangent plane. The second stage uses those segments as basic elements for a least-square linear interpolation by incorporating an adaptive weight function to minimize the weights for segments from non-ground objects [9]. This type of filtering is most suitable to be used on relatively flat surfaces while surfaces with rough terrain and less homogenous height texture can be challenging to the performance of these methods.

The main distinction between these different approaches can be seen in the strategy that they use to estimate the planimetric and height differences between object and surface points. Additionally, until now, most of these algorithms have been built for a rural environment and very few have been focused on separating urban objects from the complex urban surface. The filtering process for an urban environment is essential in order to produce an accurate DTM. In praxis usually combination of multiple methods is used to limit the weak spots of the algorithms.

There are three possible ways of filtering a neighborhood.

- Point-to-Point (1:1) - In this method, two points are compared at a time. This is based on the relative position of the two points. If the output of the function is above a certain threshold, then it will be assumed to belong to an object.
- Point-to-Points (1:m) - In this method, the neighboring points surrounding the point of interest, are used to resolve a function. The point of interest is classified based on the output of the function. Only one point is classified at a time.
- Points-to-Points (n:m) - In this method, several points are used to resolve the function. The point of interest is classified based on the output of the function. More than one point is classified at a time [7].

Filtering can be performed in various ways each having its advantages depending on the structure and morphology of the terrain, building types and point density. The point of filtering is to sort the points into segments which can be that classified into groups as buildings, terrain, vegetation etc. and exclude the point that are not belonging to any group as outliers. After the points are filtered into groups, they are classified based on chosen characteristics and the unwanted groups can be easily eliminated.

4.5.2.2. Classification

In filtering step, the points are sorted into segments or clusters. The purpose of classification is to group these parts in classes based on defining characteristics. These are for example spectral characteristics of the points, position or height relations between the points and groups, reflectance, intensity or others.

With some filter algorithms, the classification of points is done in a single step, while others classify points in multiple steps, using iteration. The advantage of a single step algorithm is computational speed, but the accuracy is less than that of the algorithm using iteration. During each iteration, more information about the neighborhood of a point is gathered and thus a much more reliable classification can be obtained.

Usually basic classes as terrain, classified, unclassified are defined and then they are divided into desired classes of objects as buildings, vegetation etc. in the next iteration.

4.6. Conclusion

Multiple options of data sets and corresponding methods were presented and discussed in this chapter. All of them having advantages and disadvantages related to the project issue. When deciding on which of these to use in the experimental part probably the most important criteria is which data set and method will provide best and most precise results when identifying the structures, we defined as buildings that are relevant to this project. This means to identify as many buildings as possible in the data set by the chosen method. High percentage of identified buildings is crucial to obtain valuable results since there might be only few buildings in the chosen area are actually illegal. If low percentage of buildings are identified the later comparison of data set with cadastral map might show small or zero number of illegal buildings. The goal here is to obtain results as close to reality as possible although evaluation of this number will be difficult because the real numbers of how many illegal buildings are in the area is not available.

Geodetic ground survey

The first data set produced by the geodetic ground survey is not considered usable for this projects purpose for one major reason which is the coverage. Although the data are the most precise and it has the advantage of the surveyor being present at the surveying site, it is not suitable for this project because the data covers only small area and are expensive due to the required man-hour to be collected.

Common characteristics in remaining options

To achieve the mentioned accuracy, I need to pick method based on advantages and disadvantages in relation to this project. From the three remaining data set: satellite images, orthophoto images and LiDAR point cloud I can conclude these facts based on the research done in this chapter:

- All these data set provide adequate area coverage for the purpose of this project.
- Resolution of all three is sufficient for the task – although the resolution varies from 10cm (orthophoto images) up to 1m (high resolution satellite images) depending on the used instrument it is still sufficient for the task of distinguishing building structure from the surroundings.
- Precision of all three methods can be considered satisfying since I only need locate and identify the buildings in the data set and it can be assumed that minor deviation of the data set from the cadastral map should not cause problems when comparing these two.

Advantages of LiDAR compared to others

The data sets have many common characteristics in relation to the project one of them brings also advantages compared to the others and it is the LiDAR. This option offers additional possibilities when it comes to identifying structures which are:

- LiDAR can capture ground points that are hidden under vegetation
- Data set contains 3D coordinates of all points, latitude, longitude and height which can be used along with spectral characteristics to distinguish buildings from surroundings

First the possibility to define the ground points from others can be used to distinguish different objects in the point cloud as vegetation, buildings etc. The seconds fact that coordinates along with height are available can be beneficial as well. When combined with the known ground points it can be used to find points that have certain height difference from the ground and define the roof of building for example.

Common disadvantages

These three options are accompanied with common disadvantages:

- Availability
- Demanding processing on both software and hardware

First major complication is the availability. There are many countries that provide these data set freely to public but the situation in Slovakia is that the Orthophoto images and LiDAR data are owned by the companies which conducted the survey and are bounded by license. Usually the surveys are done by state, university or private organizations which use the data for private purposes. Data for some areas can be free to access but this doesn't apply to the majority of the country. Another disadvantage occurs when the data are acquired in a raw form since demanding processing has to be done before the usage. The data from survey had to be connected together, georeferenced, edited, etc. before they can be used to identify structures in them. Usually professional licensed software related to the type of instrument used for the data acquisition is used for this work and since the data covers large area, they tend to be large in size and demands enough processing power.

Conclusion

From the research done in this chapter I can conclude that most suitable option is to use data acquired by LiDAR sensor in a form of a point cloud. This data will be edited and processed by filtering and classification to detect building structures with defined parameters derived from the building definition in cadastral law. After detecting the buildings in the data set, I want to outline them and export this output to a file which will be compared with the vector cadastral map and evaluate the results.

This brings me to the fact that I managed to obtain access to LiDAR data set from a surveying company based in Slovakia. This data set contains town of Stará Tura with surrounding area. It was collected in survey conducted by Technical University of Žilina in cooperation with geodetic company Helicop s.r.o and provided to me due to my previous work with this company and good relations. The data set they provided is in a form of a point cloud that was already pre-processed to the stage described in the following chapters and is ready for filtering and classification.

By this situation major problem of availability of the LiDAR data set was solved what supported my decision of using this data set with all the advantages it brings over the other ones.

5. Problem statement

In order to find an answer to the initial problem statement: How to identify illegal buildings, I conducted research described in the pre-analysis chapter with the focus on narrowing down the initial problem statement to a more precise approach to the problem. Various possibilities were introduced in the means of data sets and methods that can be applied when searching for the solution to this problematic. These differ in ways what can be reflected in advantages and disadvantages for each approach. Based on the research I came to a conclusion that LiDAR data will be used for my experiment. This conclusion is supported by the facts that LiDAR method provides enough coverage of the area, sufficient precision and resolution and besides it is also capable of mapping the ground points through the vegetation and capture 3D coordinates of the points what can be beneficial in the building identification. By this I managed to specify the initial problem statement of how to identify illegal buildings to a more precise problem statement:

How to identify illegal buildings using LiDAR technology?

6. Experimental part

This part of the project is dedicated to finding an answer to the problem statement:

How to identify illegal buildings using LiDAR technology?

The problem statement was established and narrowed down by the research conducted in Pre-analysis chapter where I found out that most suitable method for this case is to use the LiDAR data for multiple advantages compared to other methods.

In following chapters I will work on the process of comparing the LiDAR data with the Vector Cadastral Map to distinguish the buildings constructed without proper validation based on the definitions in subchapter 4.3.

Method chapter for the experimental part

For the purpose of building identification, the LiDAR technology was chosen as the most suitable option. The method chapter will describe the overall process of the building identification, finding the illegal ones and all the aspects related to the experimental part of the project. It is divided in subcategories which correspond to step taken to reach the goal.

The subcategories are:

- LiDAR overview
- Vector Cadastral Map overview
- Coordinate systems
- Workflow overview
- Processing of LiDAR
- Preparing Vector Cadastre Map
- Comparison of the data sets
- Evaluation of the experiment

LiDAR overview

To decide on correct approach to the problem I need to have a knowledge about the data, the instrument, for what purpose and how were they collected etc. Researching information about the instrument setup, factors that influence the measurement, conditions and other supporting information may help in tackling problems that might occur while working with the data set. Also, I need to decide on how to approach the problem, what methods, software and algorithms to apply, what will be based on the research conducted in Pre-analyses chapter. The LiDAR overview can be understood as a guideline upon which I will later base the processing of the LiDAR data.

Vector Cadastral Map overview

Second part of the data necessary for this project is provided by Cadastre of real estate. The overview will introduce this data and provide information of how it was produced, used structure and what information can be accessed. Since I will be using only some parts of Cadastral map its crucial for me to know the background and principles of this data set, so the correct information can be extracted.

Coordinate systems

The goal of the project is to detect illegal buildings by comparing two data sets. These are produced by different instruments and methods and as a result they are stored in different coordinate systems. To enable the joint work with them they need to be placed in same coordinates first. This will be achieved by performing transformation on one of the systems.

Workflow overview

The data set will require multiple editing and processing actions before I produce the desired outcome, outlined buildings, which will be used for the identification. Before starting the actual experimental work, I will include a workflow overview. This should provide a clear idea of what will go on, what actions and steps will be taken, for the reader to grasp the whole process. The overview should provide an understandable description of the actions and when are they executed in the time hierarchy.

Processing of LiDAR

The data introduced in LiDAR overview will need to be first prepared prior to their use. This will require multiple step that will include filtration, classification and so on that will be performed in a software environment. The outcome I want to produce is a file containing outlined boundaries of the identified building by polygons. This file will serve as a one of two bases later used for identifying the illegal buildings.

Preparing the Vector Cadastral Map

The core idea of building detection is to compare buildings contained in two different data sets, one of them being VCM which needs to be adjusted as well as LiDAR data before it can be used in the experiment. The changes may include file formatting, coordinates transformation, structure editing and others before the VCM is prepared to be compared to the output file from LiDAR data.

Comparison of data sets

When LiDAR and VCM data are processed and prepared I will compare both and hopefully identify the illegal structures in the chosen region. This idea is based on assumption presented in the first part of this project, that buildings found in the LiDAR data (representing the state of reality) and not located in the VCM (representing the legal aspect) can be considered illegal. The comparison will be performed on vector elements in both data sets in a graphical software environment. The chosen software has to fulfill criteria like compatibility, ability to perform vector analysis, preferably free of license, etc. The desired outcome is to detect outliers between the two data sets which will be analyzed for detection of illegal buildings.

Evaluation

After the comparison, as a core part of the experiment, is done I will evaluate achieved results. The results should be in the means of differences between the two compared data set, to analyze and evaluate if the detected differences can be considered as illegal buildings. I want to also take a look at the overall process of the data preparation and comparison as a part of the evaluation to see if there are areas that could be improved to achieve better results. The evaluation of the experiment as a whole, after having gain additional knowledge and skills from the actual processing can be useful for future advancements and improving the achieved results.

The parts of this method chapter correspond to the following chapters of experimental part and are proceeding in chronological order. First, the necessary information is provided on the data sets, methods that will be used, auxiliary knowledge: LiDAR overview, VCM overview and coordinate systems. Then an overview of the workflow is included for an easier and better comprehension of the overall actions and steps of the work done on the data sets. Followed by and precise description of the processing of both LiDAR and VCM data which need to be edited and addressed before they can be used further. This step will have two outputs: prepared LiDAR data and prepared VCM data that will be used in the

7. LiDAR overview

Following part will supply overview along with instrument parameters used to collect the data that I will use in the experimental part. Subchapter will provide information about:

- LiDAR data
- LiDAR instrument setup
- Methodology of LiDAR processing

7.1. LiDAR data

The LiDAR data for this project was provided by surveying company Helicop s.r.o based in town Stará Tura. They conducted a survey with the cooperation of Technical University of Žilina where they used a small aircraft mounted with the LiDAR instrument described in the section above. The purpose of this survey was to map the region of Stará Tura to obtain data for research. The research was focused on identifying objects covered by vegetation as old war bunkers in the forest based on the anomalies in the terrain, what is the main reason why they chose the LiDAR technology.

The survey covered an area of approximately 6 x 5 km with flight-lines with at least 40 percent side-to-side overlap between them. Average flight line covers the area of 5.7 km^2. The data are in a form of a point cloud stored in LAS format. The average values for the flight lines and tiles are listed in the Table 7.1.

Parameter	Average value
Flight-line	
Coverage of flight-line	5.7km2
Point density: all returns	10.33 p/m2
Point density: last returns	6.40 p/m2
Spacing btw points: all returns	0.31m
Spacing btw points: only last return	0.40m
File type	las
Coordinate system	WGS 84, UTM zone 34N
Size of single flight-line file	1.2GB
Tile	
Size of single flight-line file	1000 x 1000 m
Coverage	1km2
Number of points	16 300 000
Point density in the tile: all	14.75 p/m2
Point density in the tile: last	11.41 p/m2
Size of tile file	0.4GB

Table 7.1 Average values for LiDAR data

The acquired data was connected together to a single master file in which area of interest for this project was identified and chosen. This area represents town of Stará Tura which will be tested for identification of illegal buildings. The master file was divided into small sections, for better manipulation due to the large size of the data, by the function called tilling.

This part so far was done by the Helicop company and is closely described by Scheme 10.1. and in the section 11.1. Data preparation. I was provided with tiles containing the town of Stará Tura, where starts my work on this data set. Detailed explanation of the point cloud formats in use is included in Appendix B.

Figure 7.1 Area covered by one tile

7.2. LiDAR instrument setup

The instrument used for acquiring the data from this survey was Trimble Harrier 68i.

Sensor Head	
Beam deflection	rotating polygon
Pulse repetition rate	80 kHz - 400 kHz
Field of view	45° - 60°
Operation altitude	30m - 1600m
Vertical accuracy	< 0.15m (absolute)
Horizontal aaccuracy	< 0.25m (absolute)
Scan pattern	paralel lines

Digital Camera	
Model	Trimble AC P65+
Operating altitude	0m - 3000m
Array size	60 MP
Channels	3 (RGB)
Image pixel size	Down to 0.03m
Image scale	1:250 to 1:10 000

Computer Rack	
Log time	> 8h
Weight	43kg
Dimension (cm)	44W x 54L x 40H
Vibration isolated case	
Mouted to the aircraft floor	

Table 7.2 Instrument parameters [31]

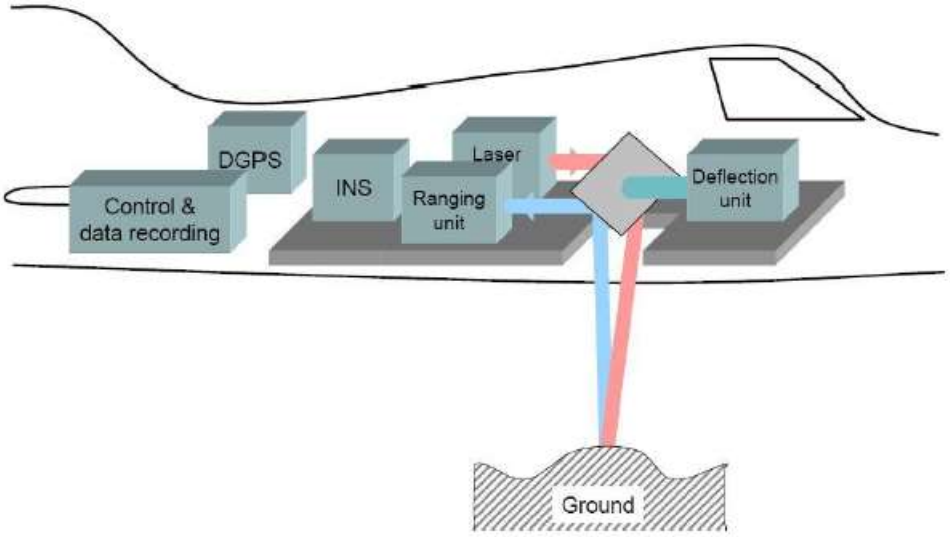


Figure 7.2 Airborne laser scanning system [24]

The instrument has a complex build that accompanies all the parts in several cases what ensures the portability and easy manipulation with it. It has an on-board computer running the mapping software it can be used and controlled by the pilot during the flight. The setup of the instrument can be seen in Figure 7.2 with main parts being marked and described [24].

Figure 7.3 Parts of the LiDAR set up [31]

A. Trimble Aerial Camera providing metric, medium format imagery.
B. Rotating polygon airborne laser scanner with high accuracy and reliability. Limited moving parts in the laser scanner and a rigid chassis design ensures stability.
C. Operator and pilot displays provide real-time monitoring of system status, imagery, and ground coverage.
D. Power supply with UPS provides the system with uninterrupted power
E. Control computer integrating and controlling all subsystems.
F. POSTrack provides GNSS-aided inertial direct georeferencing and Flight Management System
G. Data recording units for the vast amounts of imagery and laser scanning data generated by the system [31].

The outcome of the mapping are data from scanner, IMU, GPS processed by the on-board software into flight lines holding measured points with coordinates stored in point cloud preferably in a LAS format.

The errors which influence the LiDAR system measurements are closely described in Appendix A, along with the values and characteristics for the most common errors.

7.3 Methodology of LiDAR processing

When researching the problematic of working with LiDAR data I found multiple sources that were dealing with building identification from the data for example research prof. Abdullah: A methodology for processing raw lidar data [7.] or Aijazi and company: Segmentation based classification of 3D urban point clouds [12.], articles from Vosselman and Sithole regarding processing of LiDAR measurements [13]. They divided the work into set of processing steps which were similar for all of them. The main differences were in the algorithms used for classification and filtering depending on the purpose and data type. My decision, based on these publications for this workflow process is also supported by the prof. Isenburg who used similar processing steps to develop a software LAStools, designed for work with LiDAR data, which I will be using later in the experiment.

The main steps in chronological order are:

- Filtering noise
- Ground points classification
- Computing heights
- Vegetation and structure detection

First step is to remove any outliers and noise points from the data set. Afterwards the ground points are detected with the use of specific algorithm depending mostly on the structure of the terrain. From the ground level are then computed the heights of all the remaining points. When the height is known I will use it along with other parameters to classify the points representing vegetation and structures of building [10]. Classification methods were described in the section 4.5.2.

I decided to follow these steps in my project because they were used widely among the LiDAR processing and I found their chronological sequence comprehensible to understand and to work with. Also, I will be able to see how different settings in one step may influence the next one. The parameters and algorithms used will be described when tackling each of the steps.

8. Vector Cadastral Map overview

Vector Cadastral Map (VCM) is the technical basis for registering real estate in the Cadastre, it represents the vector shape of Cadastral map and it is used as a part of the geodetic information system of Cadastre of real estate. The idea is to compare the VCM with the LiDAR data set to estimate which buildings are constructed illegally based on assumption that structures that are present in the field (represented and located in the LiDAR data) and not found in the VCM can be considered illegal.

VCM is provided by the National Geodetic institution for the surveyors free of charge and can be downloaded from their web portal. The downloaded file contains the VCM in a vgi (vector graphical interface) format. Detailed description of the VCM structure is provided in section 4.2.2. For this project I obtained the VCM for the town Stará Tura which was chosen as area of interest.

Figure 8.1 Picture of VCM for Stará Tura with detail [5]

There are two layers in the VCM introduced in chapter 4.2.2. important for this project: Plot borders and Boundary layer of plot usage. These two together contain the boundaries of the buildings [5]. Major complication is that buildings are not drawn separately and don't have specific status in the current state of the VCM meaning that there is no layer holding just buildings. Instead they are present in the plot borders layer and boundary layer of plot usage what is the result of the structure and the way how are the data collected and updated to the Cadastre. This complication has to be solved prior the comparison of VCM and LiDAR data.

First idea was to find another source of cadastre information which has separate building layer. The suitable option for replacement the VCM is a data set from Ground Database of the Geographic Information System (ZBGIS). This service is run by the Department of Geodesy and Cadastre of Slovak Republic (UGKK) and can be accessed through web portal. ZBGIS is a web platform that provides Cadastral and Geodetic information to the public and it is based on the Cadastre of real estate [5]. The purpose of ZBGIS is to reduce the amount of work for the Cadastre by supplying some of its information to public via online services.

Figure 8.2 Example of the ZBGIS data [5]

From the Figure 8.2 is clear that ZBGIS should hold information about the buildings since they are separated by the other terrain and structures by different color and they are marked by boundary. In the figure above the building structures are clearly distinguished by color and most of them also by boundary. Based on this fact I concluded that ZBGIS should contain the information about buildings hopefully in a separate layer.

The major complication of the ZBGIS data is that they are only accessible through a viewer. If you want to obtain this data for use you have to buy them and the price for data covering the whole town was up to hundreds of euros. I tried to get this data by applying for a research or study purpose, but my request was not fulfilled. Therefore, this data will not be used in my project even though they might have been very useful.

There is no other source of cadastral data besides the two mentioned that's why I decided to continue the work with VCM. To make the VCM data useful and solve the problem with the layers information I will conduct editing of the VCM which will be described in chapter 12.

9. Coordinate systems

In the Slovak republic where the survey took place, are several geodetic reference systems in operation. The choice of a reference system depends mostly on the purpose which it is used for and the instrument you are using to collect the data. The geodetic reference systems are:

- ETRS89 – European Terrestrial Reference system
- S – JTSK – Coordinate system of United Trigonometrical Cadaster Network
- Bpv – Baltic Altitude system after alignment

The S-JTSK coordinate system is requested by law for every documentation submitted to the Cadastre of real estate, also the Cadastral map is saved in this realization. The LiDAR data are measured in ETRS89 coordinates in Trans Mercator projection [5.]. Precise description of the coordinate systems and the transformation is included in Appendix C.

	Cadastral Map	LiDAR data
coordinate system	JTSK	ETRS-TM34
height system	Bvp	ETRSh

Figure 9.1 Coordinates of data sets

For the transformation I used a Transformation service provided by the National Department for Geodesy and Cartography. The transformation from ETRS89-TM34 to JTSK is performed in two separate steps:

- Transformation from ETRS89-TM34 to ETRS89-Lat, Long
  This is a transformation of the measured coordinates in Easting, Northing to Latitude and Longitude, without the transformation of the height

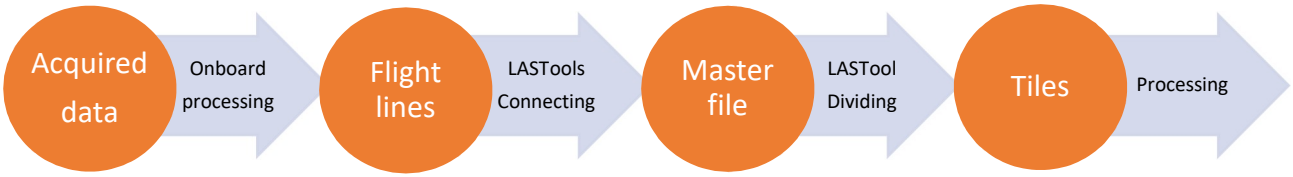
- 2. Transformation from ETRS89-LatLong + ETRS89h to JTSK + Bpv
  Here the Latitude and Longitude along with the height in ETRS89 are transformed to JTSK and height in Bpv.

As the result we have a data file in JTSK coordinates and Bpv height that can be used for further applications.

10. Workflow overview

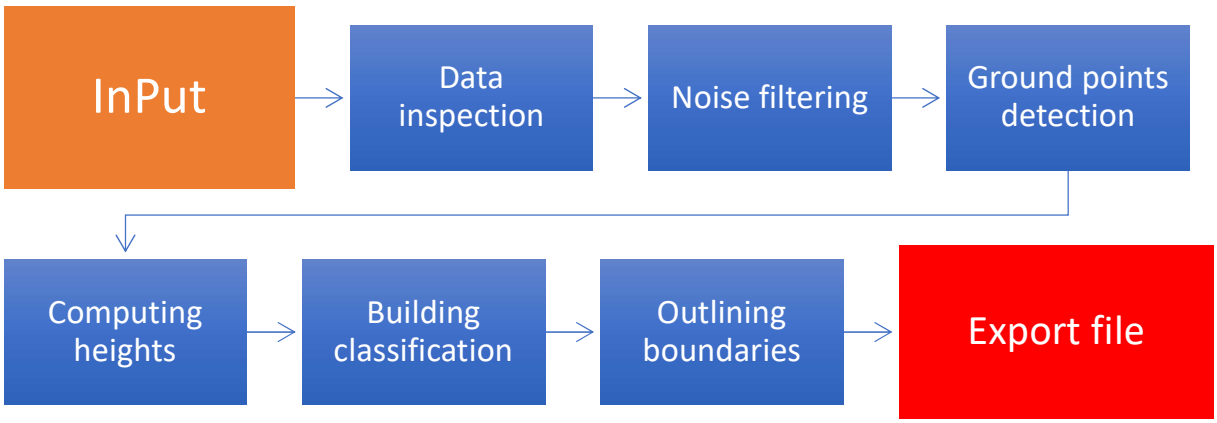
The experimental part consists of multiple actions on these separate data sets which will be compared in the end, creating a complex work process. I want to provide an overview of the overall process, to make it easier for the reader to understand the whole process and to provide possibility to orientate among the individual steps later on.

The workflow is divided into four sections each corresponding to a separate action containing several steps executed in consequential order. These actions are represented by schemes listing the main steps for them.



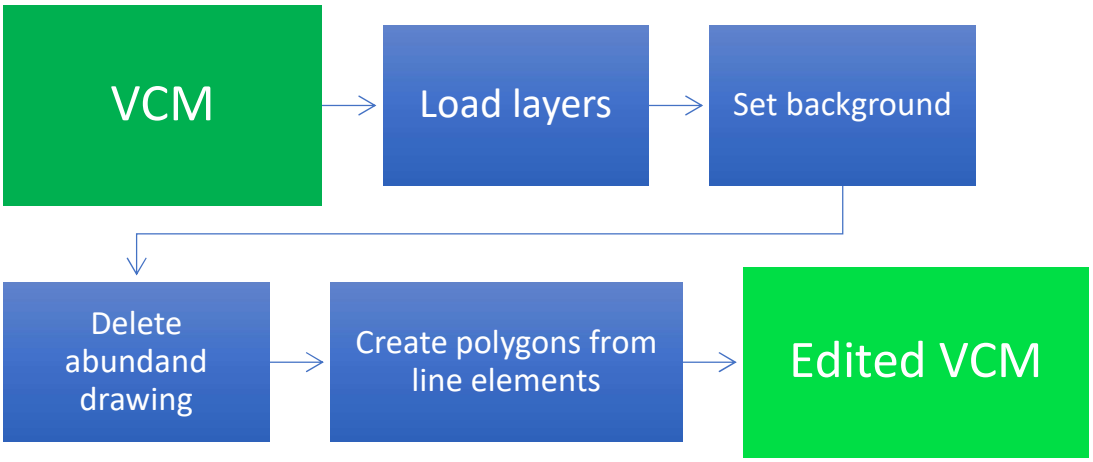
Scheme 10.1 LiDAR data preparation

Scheme 10.1 represents the work done by enterprise of processing the acquired raw data into tiles. Input for this part is a data acquired by LiDAR survey. It starts by collecting the data by the LiDAR setup created by the instrument and all auxiliary sensors. This is after the acquisition processed by an on-board software into a set of flight lines. These hold the measured terrain and objects represented by points with X,Y,Z coordinates stored in point cloud. Along the coordinates other information is saved as well for example intensity of return number. Flight-lines are connected into master file which is later divided into tiles containing the desired area of interest. This part of the work was done by enterprise Helicop s.r.o which executed the survey and provided me with the chosen tiles which are considered as an input for the next action in following scheme. Tiling is closely discussed in subchapter 11.1. Data preparation.



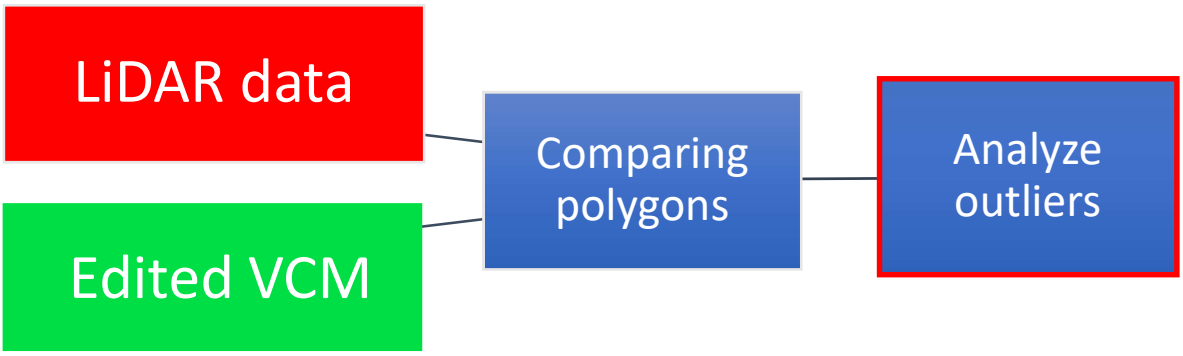
Scheme 10.2 Data processing

Scheme 10.2 describes the work steps of preparing the LiDAR data set for the comparison stage. Here the tiles holding the measured values are considered as input. They are treated with multiple processing steps like noise filtering, classification, etc. to reach the desired goal, an output file containing the boundaries of identified buildings from the LiDAR data, one of two main outputs I need to produce prior to the illegal building identification. Processing will be done in a LAStools software environment.



Scheme 10.3 VCM editing

Second half of the data used is VCM containing the boundaries of buildings. It represents the buildings listed in Cadastre of real estate and therefore considered lawfully built. Its structure was described in sections 4.2. and 8. Before the data from VCM can be used I need to prepare it with a series of editing actions to obtain the desired result: edited VCM, which will be used as the second main output for the illegal building identification.



Scheme 10.4 Detection of illegal buildings

Last step is to compare the two produced files:

- Building boundaries represented by polygons from LiDAR data
- Edited Vector Cadastral Map

In both files the buildings are represented by polygons which will be compared together. The outliers between them will be analyzed to achieve the goal and identify illegal buildings.

11. Processing of the LiDAR data

The data from LiDAR setup needs to be edited and prepared for the comparison with the VCM file. Input for the editing process are the data acquired by the survey and prepared into tiles by Helicop enterprise, this part will be explained in the first section. The desired output is a file containing polygons corresponding to outlined buildings in the data set. The work is divided to multiple actions which need to be completed in sequential order, these are:

- Data preparation
- Data inspection
- Noise filtration
- Ground points classification
- Computing heights
- Building points classification
- Outlining building boundaries

Each action will be described in separate paragraph. Data preparation corresponds to the Scheme 10.1 since it was done prior to this experiment and the remaining actions corresponds to actions in Scheme 10.2 in the Workflow overview.

The work with the LiDAR data was mostly done with the LAStools software. This software consists of multiple applications, one for every "step" in the process. These applications run in separate software windows which are controlled by the command panel. The LAStools is an open source software package developed by Martin Isenburg and it proved very useful for the work with LiDAR data [10].

11.1. Data preparation

Data for this project was supplied by Helicop s.r.o who executed the survey with cooperation of Technical university of Zilina. The workflow of this part is described by the Scheme 10.1.

First, the acquired data were processed by onboard software into flight lines, which were merged into one master file containing the whole surveyed area. Because of the large size of the master file (slowing down the software, demanding on hardware as well), it's a common practice to divide it to smaller sections in the process called tiling.

Area of interest

The survey took place in region of Nove Mesto and as area of interest I picked the town of Stará Tura which was the biggest town in the mapped area. This is a town in north-west region of Slovakia with 10 000 inhabitants and the largest town captured in the data set. The master file was split into multiple tiles, described in following section.

Figure 11.1 Survey area with town Stará Tura

Flight lines

When the survey is performed the scanner along with the sensors capture data that are processed by the on-board software into flight lines, each stored in a separate file. These lines consist of points that have 3D coordinates (x, y, z) in a chosen coordinate system along with other information as intensity or number of returns stored in a point cloud in LAS format [7].

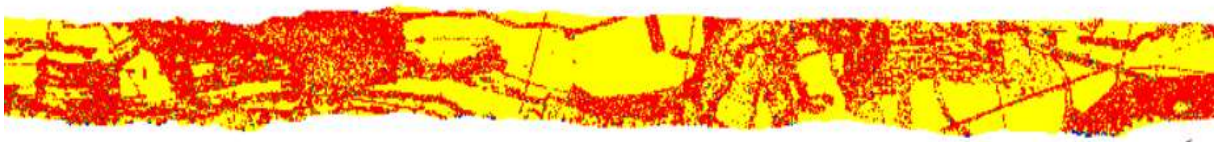


Figure 11.2 Example of a flight line (colored by return, yellow – single, red – multiple)

Tiling

Figure 11.3 Command line for creating tiles

Flight lines were joined together to produce a merged point cloud that includes whole mapped area. In average the flight lines files of this project are ranging from 1-2 Gigabytes per file and holds approximately 50-70 million points for average length flight line. Such large file is very demanding on both software and hardware what can result in lags and crashes of the software therefore it should be divided it into smaller files. This process is called tiling. Here the original file is divided based on chosen parameters into grid of tiles. Parameters are: Tile and buffer.

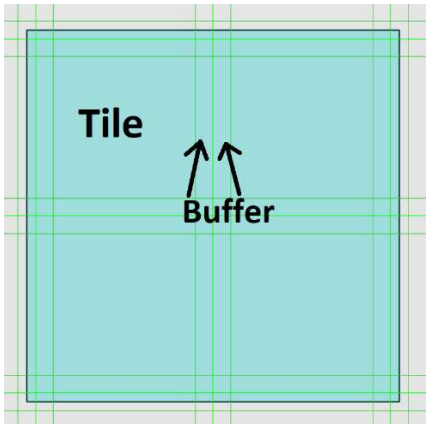


Figure 11.4 Tiles with buffer [10]

When dealing with airborne LiDAR in flight lines the grid of tiles that are typically 1000 by 1000 or, for higher density surveys 500 by 500 meters in size. Adding a buffer to the tiles significantly helps in avoiding edge artefacts when processing the LiDAR tiles afterwards. The temporary TIN of ground points is rastered at the user-specified step size onto a grid. Without a buffer in the edge areas of the tiles, software will not always be able to create a triangle to cover every pixel, caused by lack of points in the edge areas. This will cause that in the corners of the tile are often present empty pixels. Furthermore, the poorly shaped triangles along the boundary of the TIN do not interpolate the ground elevations properly. In contrast, in the tiles with buffer the TIN covers the entire area because of abundant number of points that will create better coverage by nicely shaped triangles [10].

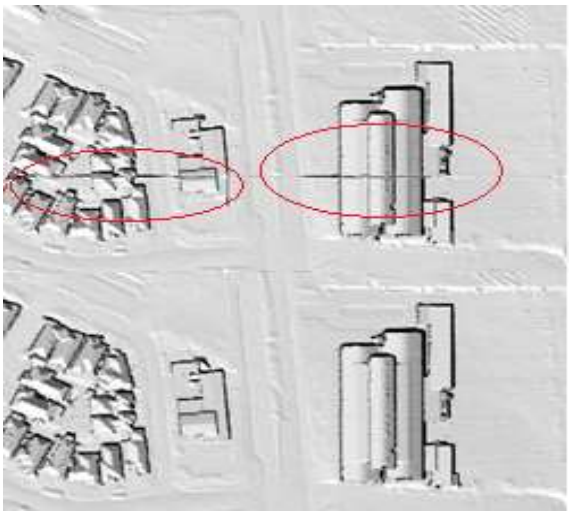


Figure 11.5 Edge artefacts, Top – without buffer, Bottom – with buffer [10.]

11.2. Data inspection

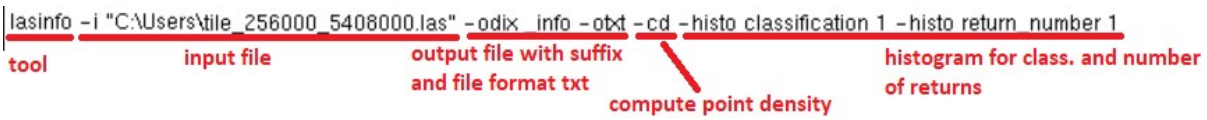


Figure 11.6 Command line for Info tool

The data prepared in previous step are used as an input for the processing. Data inspection is a first step in the processing where I analyze the data set to get a detailed overview. This inspection provides information about the loaded data set as coordinate system used, number of points etc. and let me find out information about average density, overlap of the flight lines and height differences between the lines.

Data info

```
number of point records:   16323725
number of points by return: 12625203 2392643 1007340 257319 37858
scale factor x y z:         0.001 0.001 0.001
offset x y z:               500000 5414000 0
min x y z:                  255975.000 5407975.000 267.035
max x y z:                  257024.999 5409024.999 489.074
  WGS 84 / UTM zone 34N
  gps_time 370682.962949 376865.867572
number of first returns:        12625207
number of intermediate returns: 1305812
number of last returns:         12625185
number of single returns:       10232507
covered area in square meters/kilometers: 1106700/1.11
point density: all returns 14.75 last only 11.41 (per square meter)
      spacing: all returns 0.26 last only 0.30 (in meters)
histogram of classification of points:
      16323725  unclassified (1)
             0  ground (2)
             0  medium vegetation (4)
```

Figure 11.7 Data information log

The information log displays useful data parameters:

- Total number of points in the file: 16 323 725
- Coordinate system used: WGS 84 / UTM 34N
- Area of coverage: 1.1km^2
- Point density: 14.75p/m^2
- Number of points in classes: 16 323 725 points unclassified (all points)

Visualization of the data

The data can be displayed based on different parameters as return type, intensity, class or flight-line to perform visual check for detecting gross errors, misalignment or validating the area and to see if the flight-lines were connected correctly and it gives me the first visualization of the selected area. This is performed by a view tool of the LAStools by changing the attributes that will be displayed.

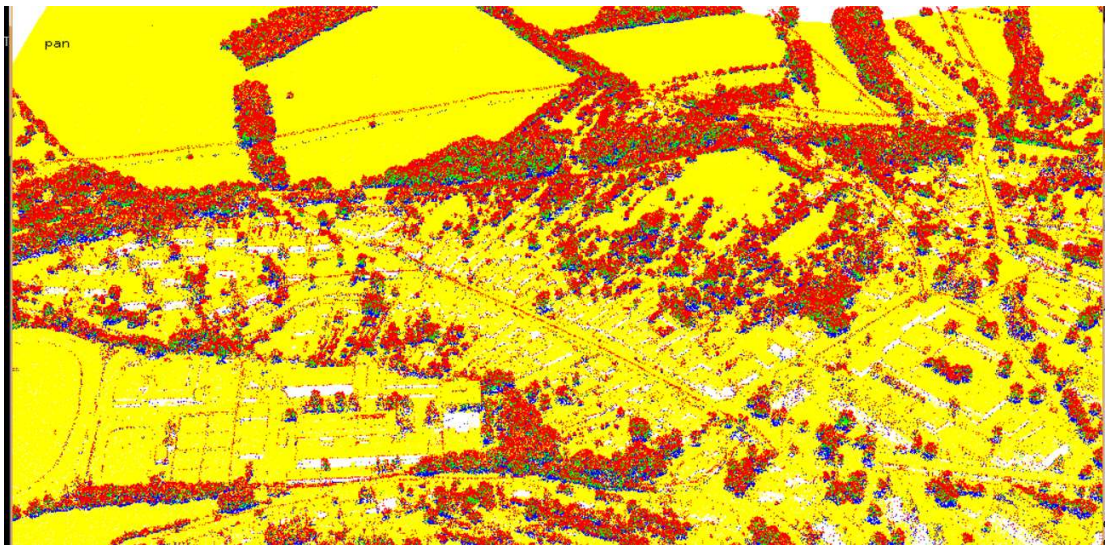


Figure Points 11.8 colored by return number (yellow-single, other-multiple)

In Figure 11.8 is a section of data colored based on the return number. Yellow points represent single returns, those are usually the ground points and building roofs. The points with multiple returns are mostly the vegetation where the laser bounces of several objects before its returned to the scanner. The Figure 11.9 shows the tile colored based on the intensity collected signal upon which I can distinguish different types of vegetation as trees or grass and structures as well.

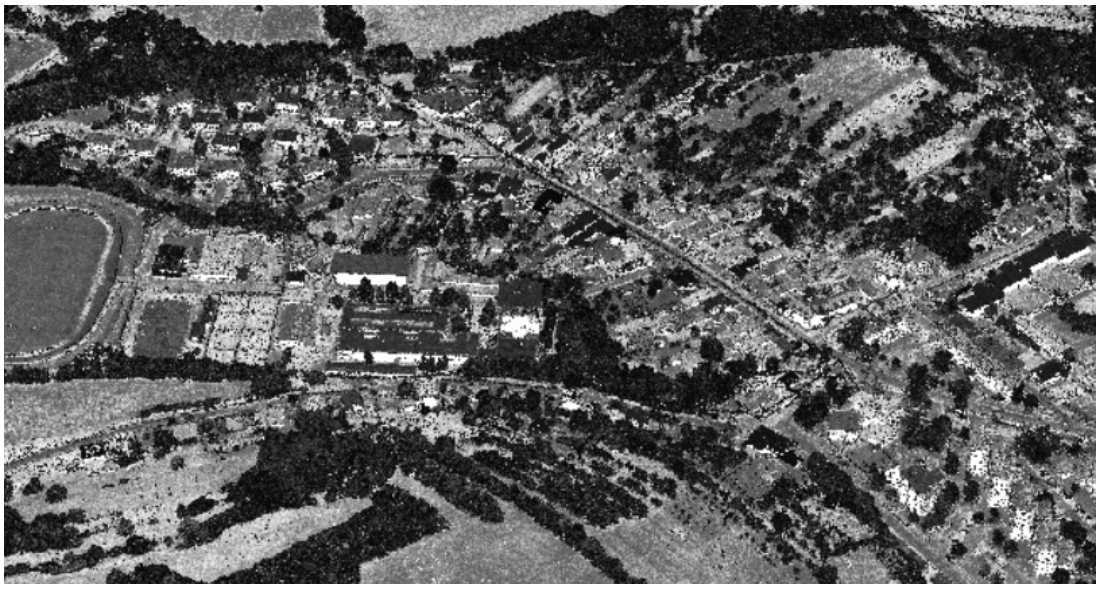


Figure 11.9 Data colored by intensity

Flight lines overlap

I can display the separate flight lines as they are connected in the selected tile, check if they are overlapping on the whole area without blind spots. Top image shows separate flight lines in different colors and the bottom image shows overlap in the area where blue indicating one flight line, turquoise indicating two, and yellow indicating three overlapping flight lines.

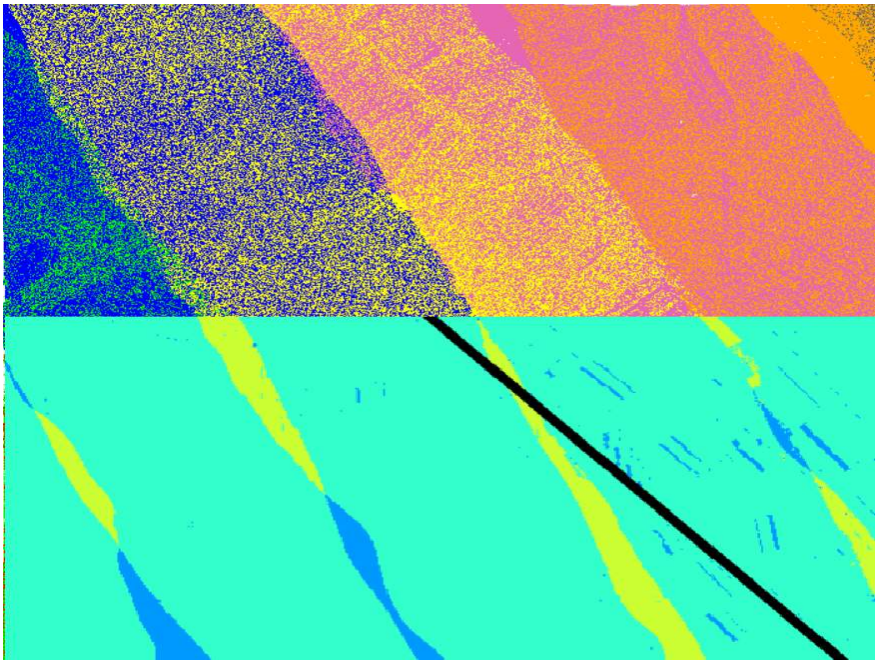


Figure 11.10 Flight lines coverage

Height difference

The Figure 11.11 shows the vertical difference of points between the flight lines, which is color coded. The white color is for differences smaller than 10cm while saturated red and blue indicate areas with more than 20 cm positive (red) or negative (blue) difference. Couple pixels wide red and blue along building edges and speckles of red and blue in vegetated areas are normal. Problem is usually displayed by large systematic errors where terrain features or flight line outlines become visible. The bottom part shows the differences with the background of satellite image of the area to help better analyze the site. No systematic or gross errors were detected in my data.

After I am satisfied with the data inspection and everything seems in order, no gross errors or misalignment was detected I can proceed to the next step as shown in the Scheme 10.2.

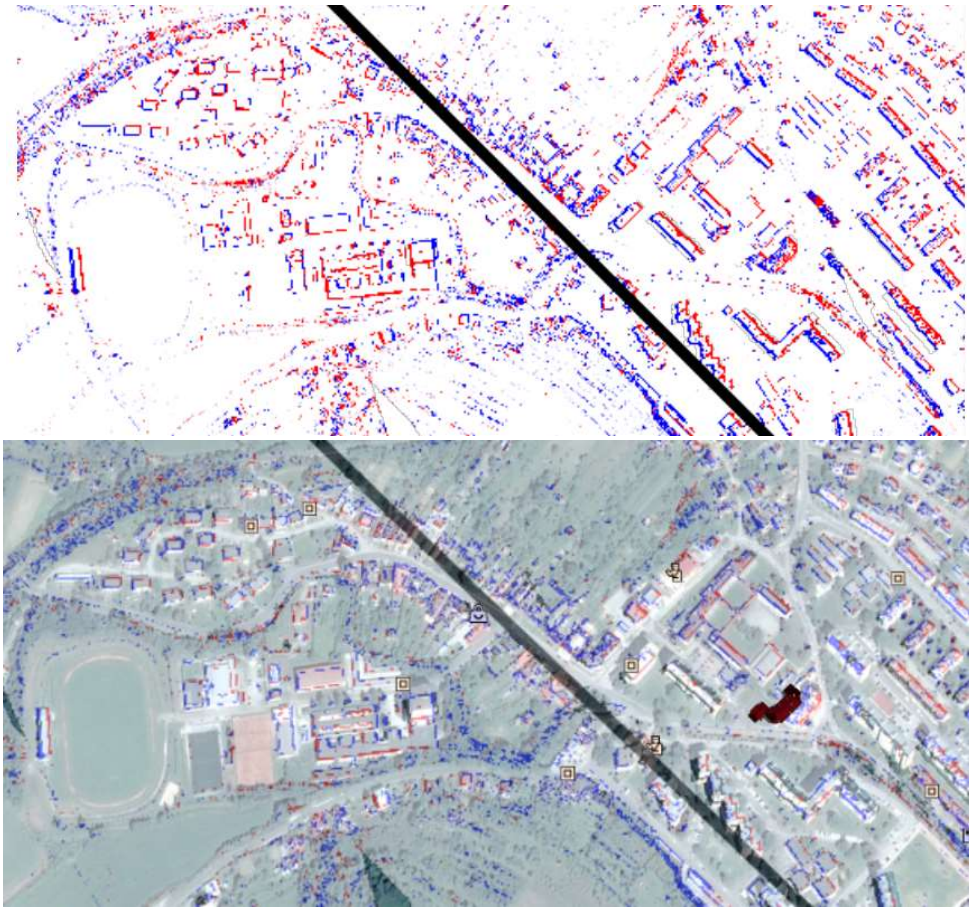


Figure 11.11 Height differences

11.3. Noise filtration

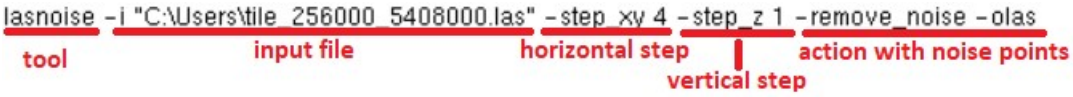


Figure 11.12 Command line for noise filtration

Important for correct data output is to classify isolated points based on their exceeding distance from most other points. These noise points are classified with a separate code what enables me to exclude them from the further processing. Especially for ground classification it is important that low noise points are excluded. While the algorithms in for ground classification are designed to withstand a few noise points below the ground, I found out based on research on prof. Isenburg paper [10.] that they will be included into the ground model if there are too many in the data set. Here the horizontal and vertical step were left to default values: horizontal 4 m and vertical 1 m. In my data only, few points were classified as noise points (app. 300 points) and these were excluded from the set.

11.4. Ground point classification

Ground filtering is a necessary step to determine which lidar returns are from the ground surface and which are from the non-ground features as buildings or vegetation. The principle is that the non-ground points are detected and removed and then ground surface is constructed.

Ground filtering algorithms often perform best when specific surface conditions are met therefore is important to choose the right approach depending on the terrain I am working with [9.]. Multiple filtering algorithms techniques were described in Pre-analysis chapter, section 4.5.2.1. The ground classification will be carried out by LAStools software, but first it is important for me to find out what kind of approach does it use to correctly decide on setting the filtering parameters.

The software doesn't provide much information about the classification algorithm it uses but I managed to find out by research that it is based on progressive morphological filtering. Mathematical morphology deals with object shape or shape measurements where these are compared and analyzed based on chosen parameters to estimate the ground surface. Study has shown that this filtering algorithm has good ability to remove non-ground objects, such as buildings and trees. It also shows that the selection of searching window size is critical to removing objects with different sizes. Progressive filtering in this case means that the window size is gradually increasing to tackle this problem [11]. The progressive morphological filter produced the least error when compared to other morphological filters for example elevation-based or maximum local slope filters and it also demonstrate a better ability to preserve the boundary of object larger that the window size [12]. Good filtering ability of morphological filters is supported by and research done by Sithole and Vosselman who tested eight different filtering algorithms on different data sets with a conclusion that this method along with line prediction method and adaptive slope filter outperformed the other tested methods in terms of average overall accuracy [27]. Supporting my choice of using morphological filtering is also the fact that it achieves better results on high-resolution data (10 points per square meter) what in my case is up to 14 points per meter. In short, the progressive morphological filter uses a progressively enlarging search window to evaluate and detect ground points based on their morphological characteristic as position, height or slope within other points in the point cloud.

Software provides options to input various filtering parameters to adjust the algorithm to the character of the terrain. The parameters I was using were:

- Step
- Standard deviation and offset
- Return

Step

The tool should produce good results for town or cities when the step size set accordingly to the terrain character, but buildings larger than the step size can be problematic. Step represents the resolution of the grid used for initial ground estimation. The default step size is 5 meters, which is good for forest or mountains. For towns or flat terrains '-town' the step size is increased to 10 meters. For cities or warehouses '-city' the step size is increased to 25 meters. For very large cities use '-metro' and the step size is increased to 50 meters [10.].

Because I am concentrating on locating buildings and the chosen area contain mostly family houses and panel houses up to 10 stories, I chose the step size to 10 meters.

Standard deviation and offset

Standard deviation describes the remoteness within which a point can be included in a planar region. The maximal standard deviation for planar regions in centimeter can be set with '-stddev __' where 10 centimeters is a default value. The maximal offset in meters up to which points above the current ground estimate get included can be set with '-offset __' where the default value is 0.05 [10.]. The offset value controls the maximum distance of points from the estimated ground surface based on which they are either included or not in the estimate.

The values for standard deviation and offset were left to default.

Return

By default, the tool only considers the last return. Earlier returns are considered non-ground. You can turn this off by requesting '-all_returns' or '-first_returns'.

Use of first or last return. There is an open discussion whether is better to use first or last returns for ground classification. The use of last returns is beneficial when dealing with vegetation where the last returns are point captured beneath the vegetation helping to improve the ground estimate in these areas, but they also tend to cause errors in classification where the pulse doesn't reach all the way to the ground. With my purpose of identifying buildings I decided to use first returns only to avoid this problem what is supported by research done by Vosselman and Sithole [13.].

Other options as adjusting the search for the initial ground points were left to default setting since they looked suitable and I found no need to adjust them at this stage of processing.

Figure 11.13 Command line for ground classification

I can check the information log described in section 10.2. to see how many points were classified as ground.

Ground: 10 219 278 points
Unclassified: 6 104 447 points

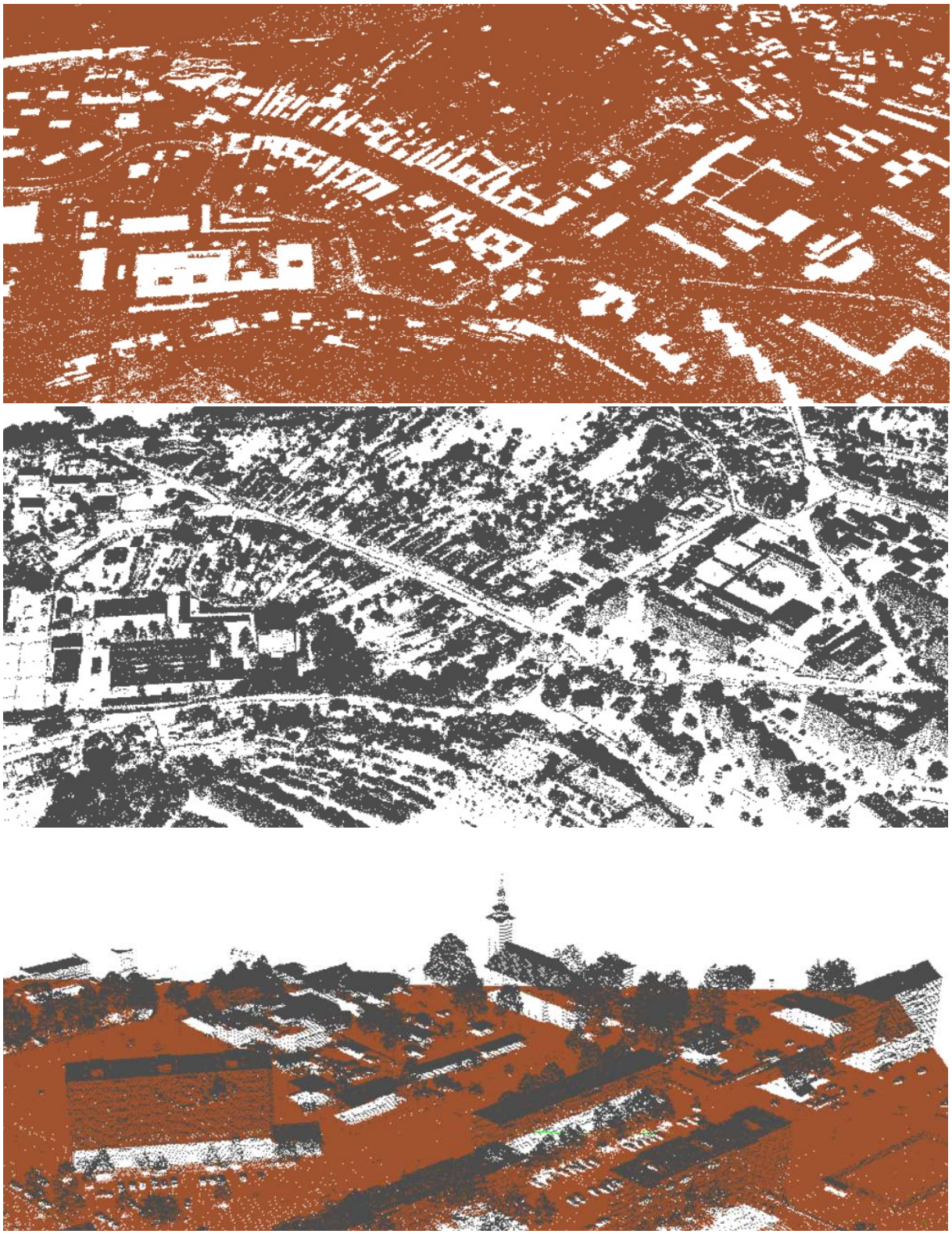


Figure 11.14 Ground points classification (ground points - brown, unclassified - grey)

11.5 Computing Heights

Figure 11.15 Computing heights

In this step the height of each point above the ground is computed. This assumes that grounds points have already been classified so they can be used to construct a ground TIN. The tool triangulates the ground points into a TIN and calculates the elevation of each point with respect to this TIN. This step is necessary for the next classification of buildings and vegetation since these are based on the height about the ground surface [10]. Output of this step is an updated file where all the points were supplied with a height value above the ground surface.

11.6 Vegetation and building classification

This tool classifies buildings and vegetation (i.e. trees, shrubs) in files. Prerequisites for this step are: bare-earth points have already been identified and computed elevation of each point above the ground was computed. The tool essentially tries to find neighboring points that are at certain height above the ground and classify them into two groups: Roofs and vegetation [10]. The classification is based on using a slope concept which is described below.

Detection of buildings from point cloud using slope concept

In this step, the morphological algorithm is improved by initially detecting and labeling all the buildings. This process is done by manipulating the slope concept based on the following assumptions:

- Terrain slopes are different from the slopes that are found between the ground and the top of a building;
- Buildings are highly elevated objects in an urban area;
- Most buildings in an urban area have smoothly sloped roofs;

Initially, points are labelled as “High” or “Low” based on their elevation compared to the ground points based on chosen height values [7]. The point cloud is transformed to raster and then, for each point in the cloud, the slope percentage is calculated for 3x3 neighborhoods around every point, Figure 11.16, left. The maximum value of the slope from these neighborhoods is considered as the slope attribute for that point, Figure 11.16, right.

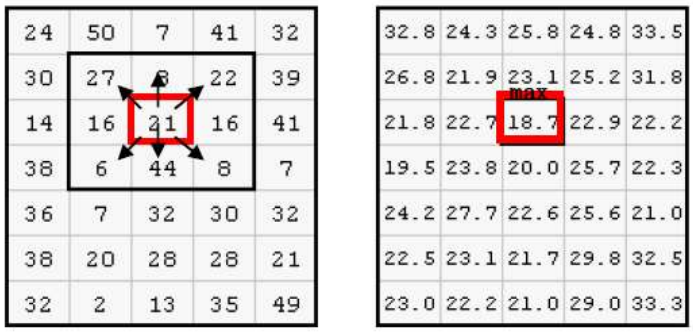


Figure 11.16 Neighborhood points slope calculation

Points are divided into two classes: “Steep” and “Slight”, based on the slope value. The initial slope value is usually defined as the slope between the smallest buildings (minimal value of height), and the surface. The members of the Steep class are those points representing *vegetation* and *walls of buildings*, while the members of the Slight class are the representatives of *building roofs* and relatively *flat areas* on the surface [7].

These points are selected and converted to a vector form as polygons. Based on the assumption that a roof should lead to the creation of a closed polygon in vector form, all created polygons are tested to see if they are closed or not. Points that do not create closed polygons are usually represented as high land and vegetation. These unclosed polygons are then removed.

For better assumption the concept of slope attribute is replaced with planarity. Planarity estimates if the neighboring points are part of the plane based on their standard deviation from the planar region they share [12].

Used morphological algorithm can be controlled by these parameters:

- *Ground offset*, minimal height from which the points can be classified as roofs
- *Planarity* correspond to a standard deviation neighboring points can have from the planar region they share.
- *Search size area*, minimal area within which the points can create plane (depends on point density)

It is important to choose correct step size which is dependent on the point density in the cloud. At least 4 pulses per square meter which is the minimum that is needed for considerably reliable roof detection. Density of points in my data set is up to 14ppm so I can afford to experiment with this value. Step I used first for searching planes is 2m (default value) [10].

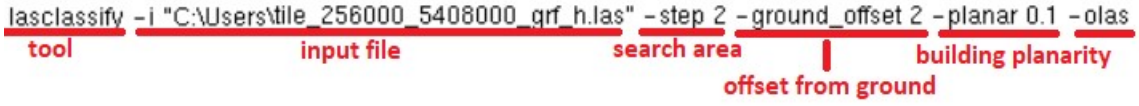


Figure 11.17 Command line for classification

Values I used for initial classification:

search area size: 2 m
building planarity: 0.1 m
ground offset: 2 m

Very noisy data or not properly aligned flight lines can cause problems finding planar regions therefore it´s important to eliminate possible low noise and check for alignment, Figure 11.10. After the classification, it is good to perform a visual check to spot any gross errors in the process. When any misclassified areas are detected there is an option to reclassify them manually, but the focus is on finding the best parameters to correctly identify all present buildings by adjusting the algorithm parameters.

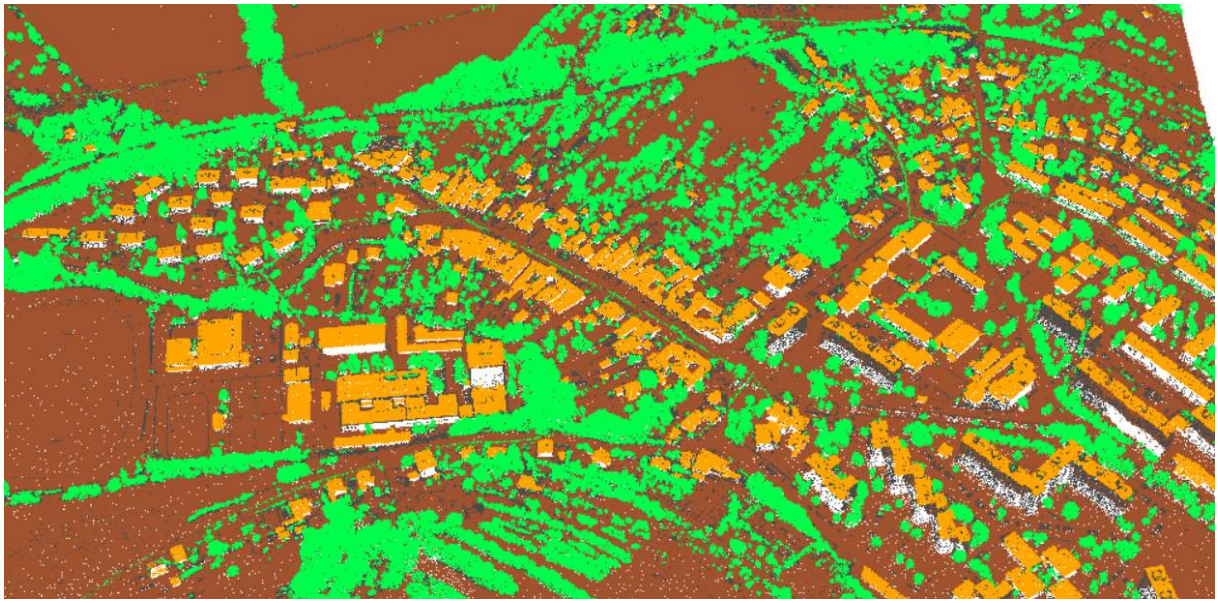


Figure 11.18 Classification of points into ground (brown), vegetation (green) and building (yellow) classes

11.7 Boundaries

Figure 11.19 Command line for creating polygons

Input for this action is a classified point cloud with separate building layer. Boundaries tool creates single polygon where "islands of points" are connected by edges that are traversed in each direction once.

The controlling value is '-concavity' that can be specified in the command line. This parameter specifies that voids with distances of more than specified value in meters are considered the exterior (or part of an interior hole), so the polygons representing two objects have to be at least specified distance apart to be considered separate polygons [10]. Examples of two concavity values 1m and 10m are in Figure 11.20.

Concavity value was set to 1 meter based on the estimated minimal distance between buildings in reality which is defined by the cadastral law. So, buildings that are closer together will be drawn in a united polygon and buildings that are further apart will be drawn separately.

To use only point classified as buildings for creating polygons I will use command '-keep_class buildings' to specify which points will be used for the boundaries [10].

Output is a file exported in shapefile format, containing polygons representing the boundaries of identified builds in LiDAR data. This is the final output of the LiDAR processing which will be used later in the comparison with VCM.

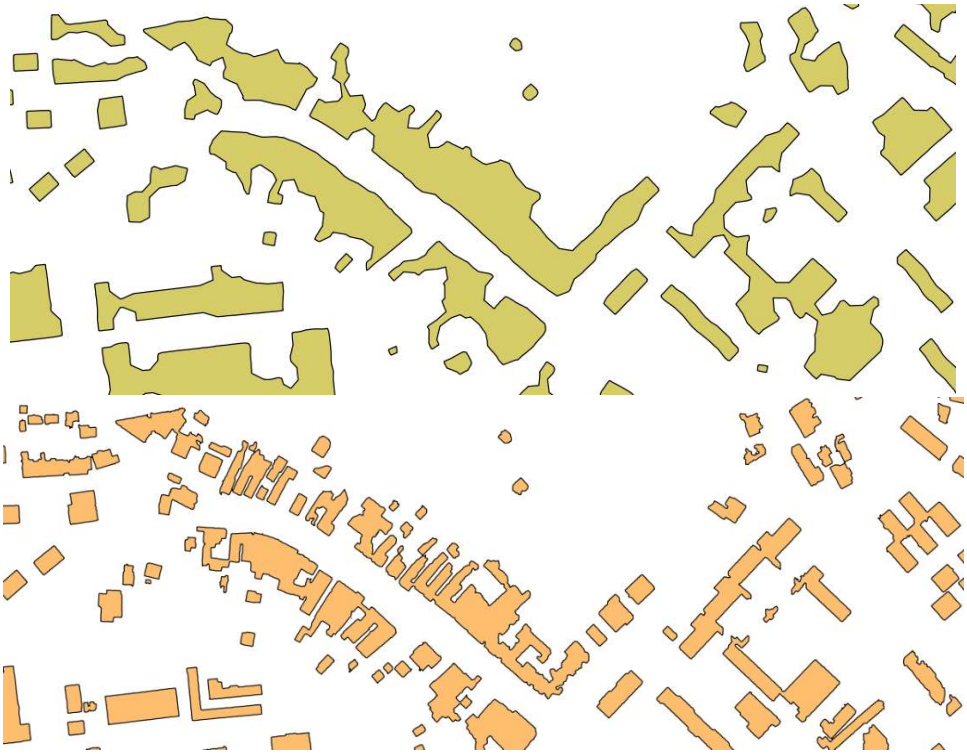


Figure 11.20 Concavity value 10 m – top, 1 m - bottom

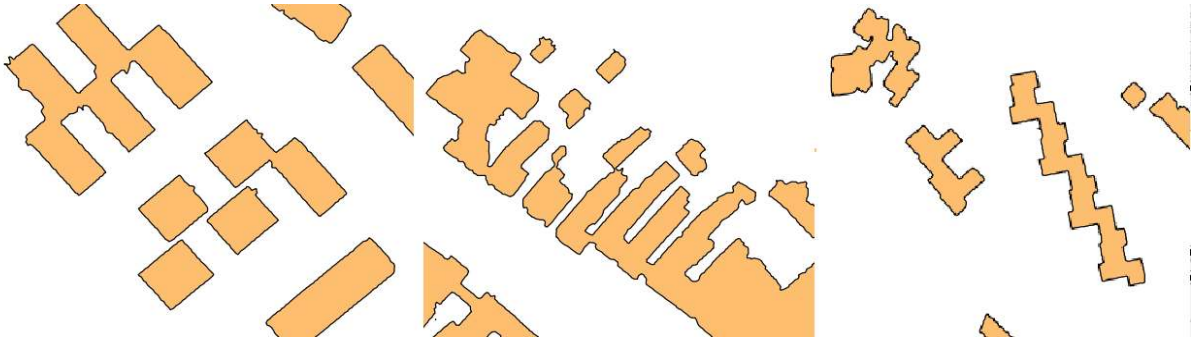


Figure 11.21 Details of polygons with concavity 1 m

Evaluation of the processing

After the output file is produced it has to undergo evaluation to find out the quality of the process. This will be based on estimating the statistic numbers of correctly, incorrectly or partially identified buildings.

As a part of the process I will perform a visual check for incorrectly classified object which should be excluded from the data set before performing the comparison with VCM. This control should be done preferably in automated way. Search algorithm could be based on assumption that building polygons have usually shape of perpendicular rectangles and misclassified objects have irregular shape. Another possible attribute might be comparison of position between polygons since most of the misclassified were located in remote areas, away from the other polygons. Due to limited time for this project decided to focus on finishing the experimental part and performed only simpler visual check.

If the misclassified objects are not eliminated, they will be otherwise identified as outliers from the VCM and as illegal buildings therefore is important to remove as many of these as possible to achieve plausible results when comparing with the VCM data.

Figure 11.22 ZBGIS with polygons (grey) overlap

Visual check will be done by using an image from ZBGIS as a background and overlapping it with the polygons from LiDAR. In ZBGIS I can see the buildings outlined by color where the outliers are clearly visible. This is simplified by the fact that misclassifies are mostly present in less urban or rural areas and their polygons have irregular shape therefore are easier to spot.

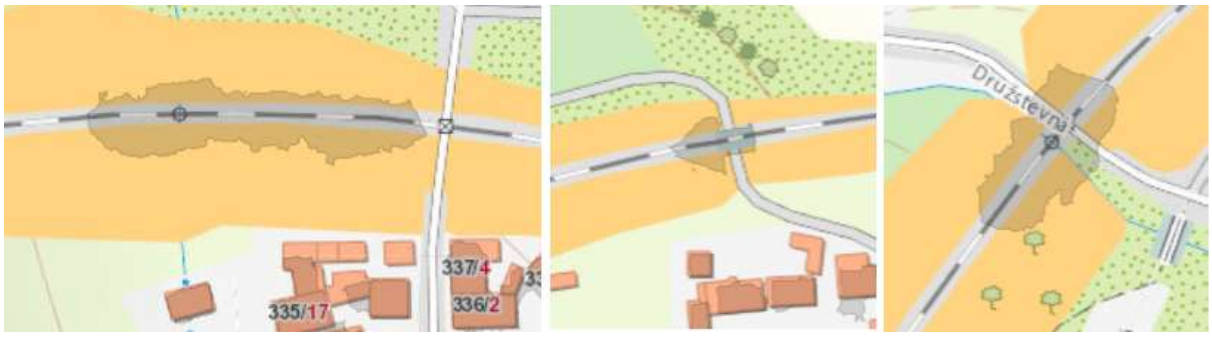


Figure 11.23 Misclassified polygons - grey

Statistics

From total number of 268 polygons were 15 identified as misclassified. That is 6% of the total number of classified polygons. Number of unclassified buildings along with other statistics will be estimated after the VCM is prepared, section 12. and the number of polygons compared to the values from LiDAR data. Remaining statistics will be computed after analyzing the results of data comparison in section 14.

12. Preparing the VCM

Preparing of the VCM consists mainly of addressing the problem with buildings not being stored in a separate layer but in two separate layers along with other structures present. This issue is described in sections 4.2.2. and 8.

Process will be described in consequential steps where bout automated software tool and manual editing were required. Input is a file containing the VCM of Stará Tura downloaded from UGKK portal where I have granted access to these data as a registered surveyor. The file was first converted to dwg format using a national transformation service and opened in software MicroStation which is a CAD type software. Here I can access different layers of the map, edit drawings and create desired exports. This process will consist of following steps:

- Load VCM into MicroStation, keep two layers containing building boundaries and delete the rest
- Set ZBGIS image as background to identify building boundaries and delete the remaining drawing
- Automatically convert all closed line strings to polygons using “Polygonize” tool
- Manually convert all the remaining line strings to closed polygons using “Create complex shape” tool
- Export polygons into shapefile

First, I open the file in MicroStation, locate the two layers, described in section 4.2.2, which I will be using: Plot borders and boundary layer of plot usage and remove all of the rest layers.

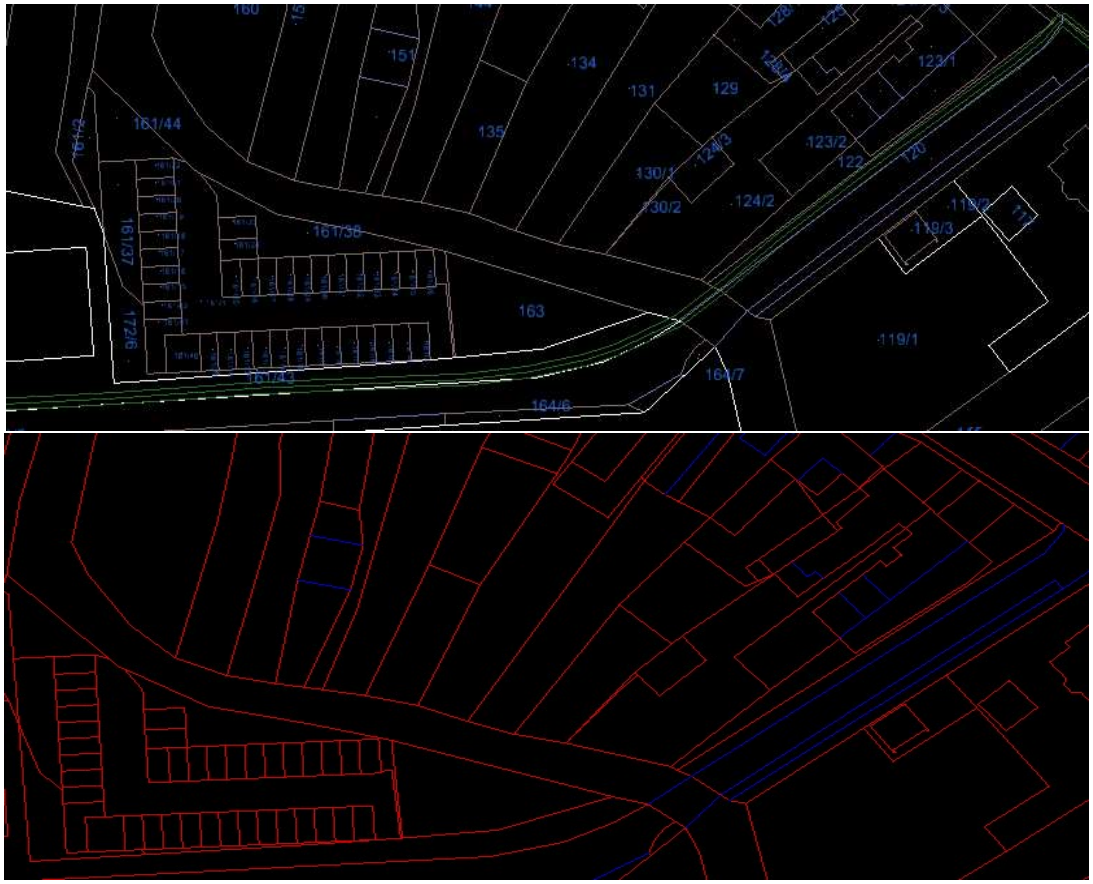


Figure 12.1 VCM top – all layers, bottom – only two layers containing buildings

Afterwards, ZBGIS image is used as background to identify the building structures. This is achieved by connecting the ZBGIS image as a reference file. Since the image could be only saved as a pgn file it has to be manually scaled and fitted to the VCM.

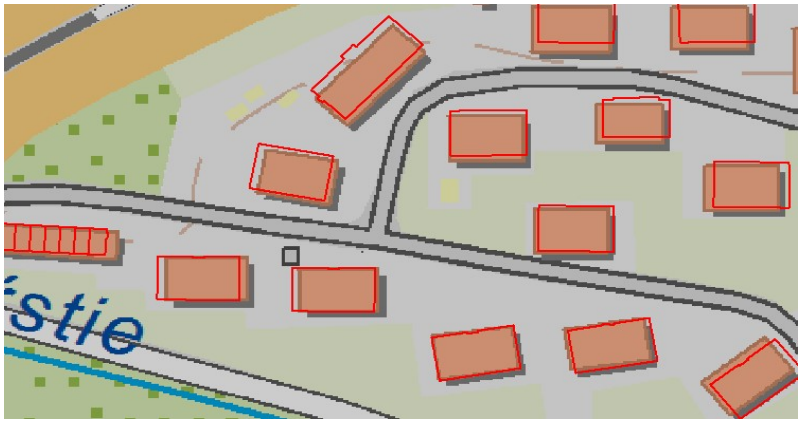


Figure 12.2 Building boundaries with ZBGIS image as background

Now I can clearly see which lines represent building boundaries and which are for other structures as plots, roads etc. Building boundaries are kept and the rest of the drawing is removed simply by selecting the lines and deleting them from the layers. This process was not ideal and also time consuming therefore I chose to edit only a part of the VCM corresponding to the size of one tile so the comparison of the two data sets can be executed. This editing was simplified by the fact that building boundaries have usually shape of perpendicular rectangles and also that many of the building plots are drawn separately (described in section 4.2.2.) what made distinguishing of the building's drawings easier. After the editing was done, I saved the remaining drawings into single layer instead of two, so the layer now contains only lines representing the boundaries of buildings.

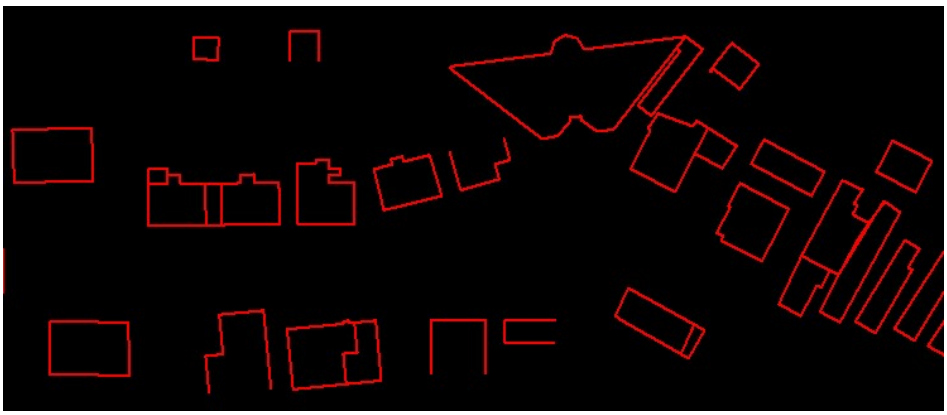


Figure 12.3 Layer with only building boundaries

At the current state all building boundaries are saved in single layer and drawn by line elements. Now I need to convert these lines into closed polygons. As seen in the figure above some of the boundaries are not closed where the lines are missing. This is caused by the structure of VCM data where some of the buildings are drawn as a supplement to existing plot

and not as a separate object. Therefore, after the abundant plot drawing is removed these buildings are left with missing lines. Creating polygons is performed in two steps.
First, closed boundaries are treated. MicroStation provides a tool “Polygonize” which creates polygons from the closed line boundaries. Here all the line elements are converted to polygons one for each closed boundary. Afterwards I addressed all of the remaining boundaries which were not closed and therefore not processed by the tool “Polygonize”. Using tool “Create complex shape” I manually closed all of the remaining boundaries one by one. After the boundary was closed by the tool it was converted to a polygon.

This way all of the line boundaries were closed and converted to polygons. At last I exported the layer containing all of the polygons in shapefile. This edited VCM output file will be used as an input for the comparison with LiDAR data.

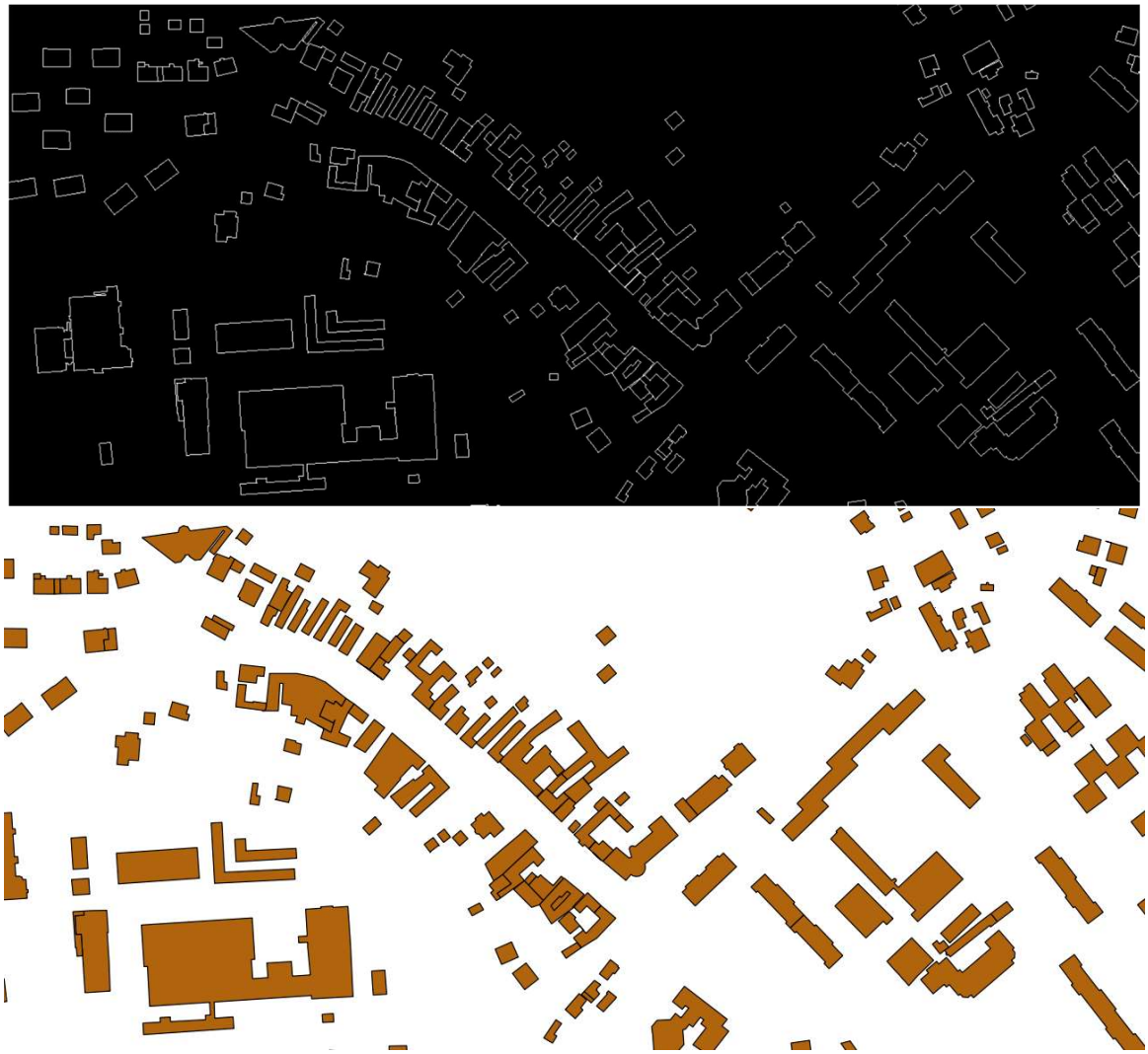


Figure 12.4 Prepared polygons in MicroStation (dwg format) – top, in QGIS (shapefile format) - bottom

Number of separate polygons in the VCM after it was edited is **250**. These polygons represent all the buildings present in the chosen area with valid construction permit and listed in Cadastre of real estate.

13. Comparison of the data sets

After the VCM and LiDAR data were prepared to desired form in processes described in sections 11. and 12. I will compare then with the purpose of finding outliers which will be analyzed to detect illegal buildings. This process is based on assumption that buildings present in the reality (represented by LiDAR data) and not found in the VCM (representing the lawful state) might be considered illegal. Inputs are two files containing polygons representing building boundaries. Goal is to overlap the LiDAR file with VCM file and locate outliers between them.

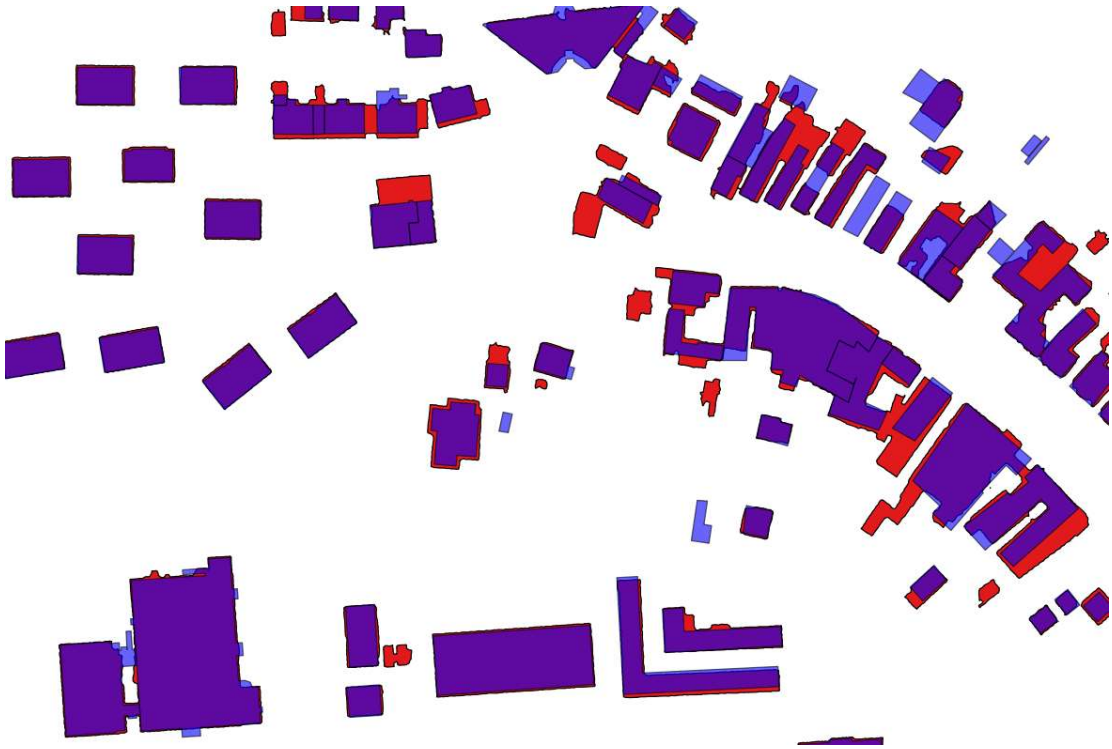


Figure 13.1 Layers overlap LiDAR - red, VCM - light blue, overlapping area – dark blue

The method I chose for this step is to compare polygons representing the buildings from VCM and LiDAR. This decision was based on knowledge obtained when working with the data sets and research. Polygons can describe the shape of a building boundary well along with other benefits. These are the reasons supporting using polygons:

- Polygons are closed shapes that can represent the shape of a building well
- Most of the boundaries in VCM are already drawn as closed objects which can be easily converted to polygons
- Buildings in LiDAR data can be drawn as polygons by creating point clusters and creating polygons from the edges
- The overlap of VCM and LiDAR is expected not to be perfect, but the touching or intersecting polygons still can be compared
- Inside of the polygon can count as part of the polygon what can be beneficial to the comparison

Polygons are good representation of the building geometry for this experiment. With the closed structure the inside can count as a part of the polygon and this can be helpful when comparing two data sets that are not expected to be perfectly equal. The VCM data are created from vector lines which are mostly straight with well-defined edges and LiDAR comes from a processing the roof image from a point cloud where the edges of the polygons can be distorted. Using polygons instead of for example lines should handle these problems.

Work environment chose for this step is an open source software QGIS widely used for GIS applications providing useful tools for manipulation with vectors and spatial analysis, capable of opening shapefiles and other formats.

First, the data are imported to the QGIS and stored in two separate layers. To avoid problems with the coordinate systems inside the software I set the coordinate system of the first loaded file as a default one for the opened project, so all the all the data loaded afterwards will be displayed in this system. It does not matter which data set will be loaded as first since they were both transformed in the same system but when working with the software, I found out that sometimes it tends to load the data in different systems so this is a good precaution to avoid related problems.

In Figure 13.1, I can see the actual overlap of the polygons. No visible shift between them of other visible deformation is detected and I can proceed further. Comparison will be performed in with the use of research tool for vector analysis “select by location” which makes a selection of the polygons in one layer based on their corresponding location to second layer, meaning that polygons in the two layers that are overlapping, touching etc. will be selected. This is controlled by setting criteria based on which the polygons will be selected.

The criteria were based on assumptions about relations between the data sets. Important thing is to compare polygons corresponding to the same buildings in both data and also take in account that the shape of polygons will be different and possible shift might be present as well. The assumptions were:

- Polygons representing the same building might have different shape
- Slight shift in position can be present between the polygons corresponding to the same building
- Polygons from LiDAR data might represent only partially detected building
- Only polygons from the two data set representing the same building in reality should be compared

These assumptions must be taken in account when deciding on the criteria for selection tool. The criteria are limited by the options provided by software tool. These are the options that were selected to fulfill all of the above and to perform a successful comparison.

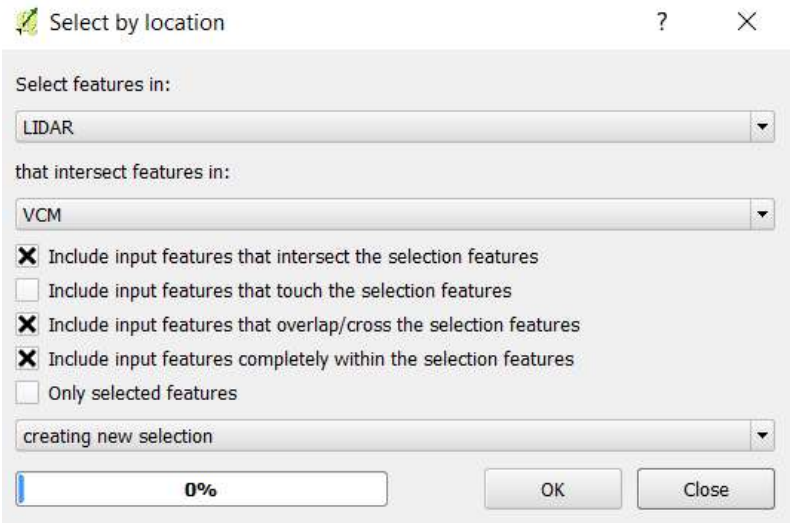


Figure 13.2 Tool Select by location options

In the Figure 13.2 are the selected parameters. The features present in LiDAR layer that intersect with feature in VCM layer will be selected based on chosen criteria:

- Include features that intersect the selection features - to take in account the different shapes of polygons
- Don't include the features that touch - to avoid selecting polygons not corresponding to the same building
- Include features that overlap/cross – to take account for possible shift and different shape of the polygons
- Include features completely within the selection features – to include also polygons representing only partially identified buildings in the LiDAR data

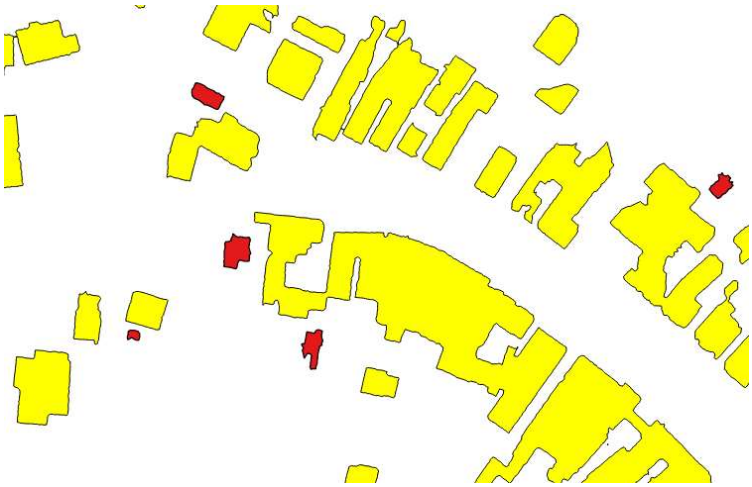


Figure 13.3 Selected polygons - yellow, remaining polygons - red

The result is a selection with all the overlapping features chose on picked criteria. Now I need to invert the selection to find the polygons not present in both data. This is done by simply deleting the current selection of polygons. First, I allow editing of the LiDAR layer, open the attribute table, Figure 13.4. Where I can see list of the polygons that were selected and delete them, Figure 13.5. This leaves me with LiDAR layer containing only the polygons which are not present in the VCM layer.

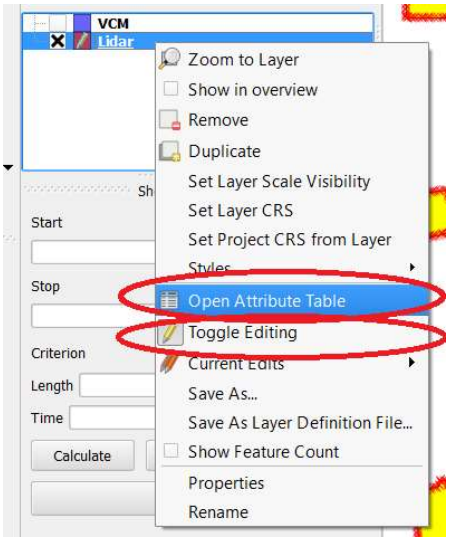


Figure 13.4 Toggle editing, open attribute table commands

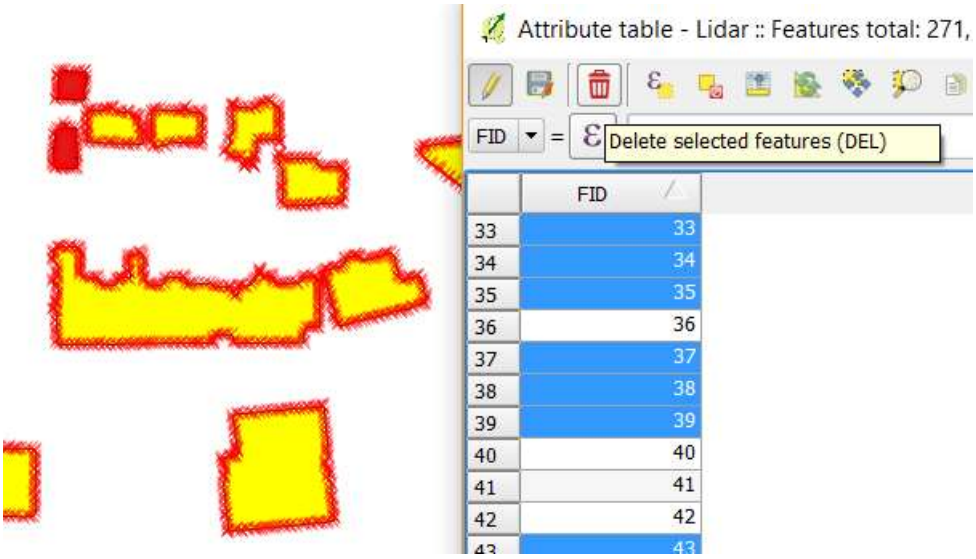


Figure 13.5 Polygons enabled for editing - left, attribute table with selected polygons - right

From the total number of **253** polygons in the LiDAR layer entering this step, now I have **32** remaining.

When looking at the results I can identify a visible complication: many small polygons with area smaller than 25 m^2 were left in the layer. Caused by the fact that buildings with lesser area were not drawn in the VCM, explained in section 4.2.2. and therefore, not overlapped with LiDAR polygons. I will use another QGIS tool and perform a "Selection by area" where I set the threshold to select polygons with area from 0 to 20 m^2. I chose the upper value 20 instead of 25 m^2 to avoid losing buildings with area close to the bottom limit that might be not fully identified to prevent of losing such polygons. Polygons with area smaller than 20 m^2 were selected and removed from the layer.

From **32** polygons after removing the ones with small area I was left with **15** polygons.

These 15 polygons might represent either misclassified objects that were not excluded before or buildings that are present in LiDAR but not in VCM and can be considered illegal. They will have to be analyzed one by one to find out what are they representing. This will be done by loading the layer with remaining polygons into ZBGIS which will be used as a background. In addition, this time I will use satellite images from Google Maps and Street View to substitute for the field examination which would be the best option to perform if possible.

I analyzed each of the polygons using the ZBGIS image as background which contains the buildings of VCM, on Google Maps satellite image to see if any structure can be identified and if available also on Street View which contains actual photos of the area. Problems with identification were that the satellite images have poor resolution when trying to identify small objects with size length up to 20 meters and also that Street View images are present only on the main streets and therefore are covering only approximately half of the area.

Analysis of these polygons can have 3 possible outcomes:

- polygons were not present in VCM because they represent outliers and structures not listed in VCM
- polygons cannot be properly identified
- polygons were not present in VCM because they represent illegal buildings

Results of the analysis of 15 polygons:

- 5 polygons were identified from Street View as garden sheds or wooden structures that don't require construction permit, therefore are not listed in VCM, Figure 13.6.
- 9 polygons couldn't be identified due to small resolution of satellite images, Figure 13.7.
- 1 polygon was identified as illegal building, Figure 13.8.

First group of **5 polygons** contains buildings that are not present in VCM because they don't require construction permit since they are not considered permanent structures with fixed connection to the ground, explained in section 4.3. These are usually garden sheds, small wooden buildings, auxiliary structures etc. as can be seen in the figures below.

Figure 13.6 Examples of identified structures

Second group of **9 polygons** could not be properly identified because of insufficient identification options as small resolution of the satellite images, figure below. I can assume these polygons can represent either misclassified structures, buildings not required permit as in the group before or illegal buildings. For this group would be best to perform a field check where the site would be examined by surveyor to find out the real matter of these structures.

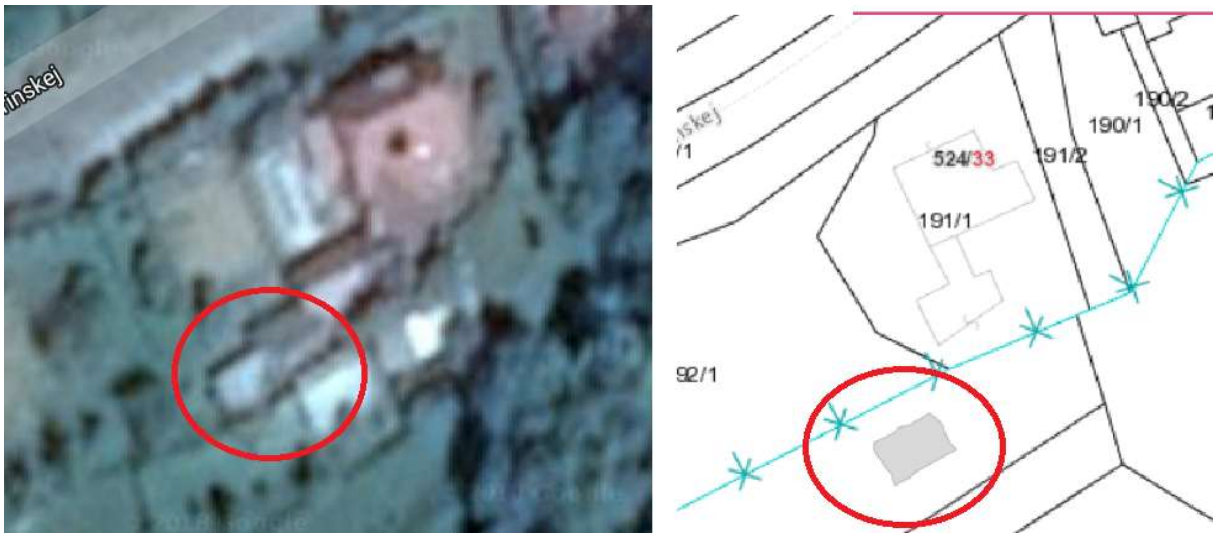


Figure 13.7 Example of unidentified polygon

Last group holds **1 polygon** which was identified as illegal building. This building was identified in the LiDAR data but is not present in Cadastral map as can be seen in Figure 13.8. what means this building is not listed in Cadastre of real estate.

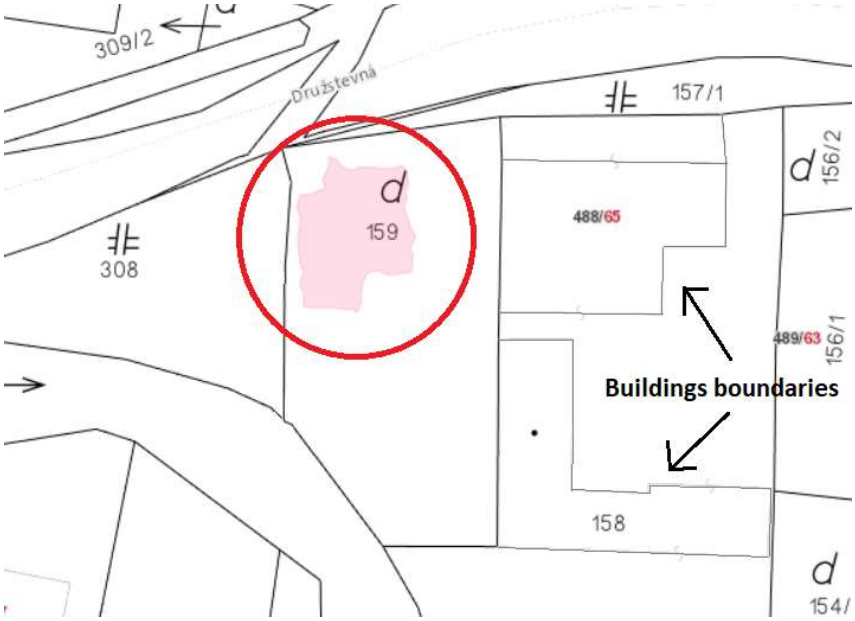


Figure 13.8 LiDAR polygon displayed on Cadastral map

This claim is supported by images from Street View which show the presence of this building in reality, Figure 13.9. These pictures also show the brick foundation of this building what makes it fixed to the ground and therefore requiring construction permit. To make sure I found the ownership document for this plot (plot number 159 in Cadastre region of Stará Tura) which did not contain any information about the present building. Based on these facts I can claim this building is missing a valid construction permit and can be considered illegal.

Figure 13.9 Images of identified illegal building

Summary of data comparison

Goal of this section was to compare two sets of data to identify illegal buildings in the chosen area of interest. Both sets contained polygons representing the building boundaries. LiDAR data set contained 253 polygons that were compared with the 250 polygons in VCM which represents the legal state of the Cadastre. This comparison resulted in 32 outliers in the LiDAR data compared to VCM polygons. From these 17 were excluded for having area smaller than 20 m^2 (referred to section 4.2.2.) leaving a group of 15 polygons. This group was examined one polygon at a time by comparing to Cadastral map, satellite images from Google Maps and images from Street View what resulted in following findings:

- 5 polygons identified as buildings not present in Cadastral map (don't require construction permit according to construction law, section 4.3.)
- 9 polygons were not identified due to insufficient resolution of satellite images and no available images from Street View
- 1 polygon was identified as illegal building, what was supported by missing information regarding the building in the ownership document for the plot where building was found

Polygon type	No. of polygons
Input polygons VCM	250
Input polygons LiDAR	253
Outliers between data sets	32
Excluded small polygons	17
Polygons for identification	**15**
Not present in VCM (due to permit)	5
Unidentified	9
Illegal buildings	**1**

Table 13.4 Summary of compared polygons

Estimating number of unidentified buildings in LiDAR data set

When knowing the classification count of all polygons in LiDAR I can also estimate the approximate percentage of buildings not identified in LiDAR data. LiDAR´s 253 polygons compared to VCM 250 polygons resulted in 32 outliers in LiDAR set. This value consists of polygons present only in LiDAR (small polygons, illegal building, those not present in VCM) plus polygons present only in VCM (buildings unclassified by LiDAR). Used values are from table above.

Input polygons LiDAR – Small polygons – Not present in VCM – Illegal buildings =

= polygons LiDAR = 253 – 17 – 5 – 1 = **230**

Input polygons VCM – polygons LiDAR = Unclassified building polygons = **20**

The polygons were sorted in classes based on their status and the count for each class is listed in Table 13.5, along with a value of percentage. This value was computed by comparing the number of polygons from each class to the total number of building polygons found in LiDAR data set. The percentage shows the proportion of the class compared to sum of building polygons. Building polygons were computed as total number of created polygons minus the number of misclassified polygons leaving only polygons which should represent buildings.

	Number of polygons	Compared to total [%]
Total of building polygons	253	100
Misclassified polygons	15	5.9
Unclassified buildings	20	7.9
Polygons common with VCM	221	87.4
Outliers from VCM	32	12.6
Small buildings	17	6.7
Polygons for identification	15	5.9
Identified illegal building	1	0.4

Table 13.5 Polygons count with proportion percentage

14. Conclusion

The project work started with defining the initial problem statement:

How to identify illegal buildings using geodetic data gathering?

With the goal to find solution to the initial problem statement by answering all the related sub-questions I conducted research on the available geodetic data gathering approaches reflected in Pre-analysis chapter. Geodetic options were divided in the means of produced data and method used. The research led to conclusion that data captured by LiDAR will be most suitable for the experiment because of advantages compared to other possibilities like data coverage beneath vegetation or point cloud with 3D coordinates. After the method was picked, new problem statement was formulated:

How to identify illegal buildings using LiDAR technology?

Following was the experimental part dedicated to finding answer to the problem statement. Experimental part included more detailed description of the data sets along with auxiliary research and workflow overview. Following were three main processing steps:

- Processing of LiDAR data
- Preparation of VCM
- Comparison of two data sets

Processing of LiDAR point cloud holding 16 400 000 points led to a file with 268 polygons where 15 were identified as misclassified leaving 253 polygons representing building boundaries. The edited VCM file contained 250 polygons, describing the lawful state of cadastre.

Comparison was performed on picked area of interest covering town Stará Tura with size 1.1km^2. Polygons from the two data sets were compared in QGIS environment with the result of 32 polygons identified as outliers of LiDAR from VCM. This selection was subtracted by other 17 polygons which were identified as small buildings. The remaining 15 polygons were analyzed with following results:

- 5 polygons identified as buildings not present in Cadastral map.
- 9 polygons were not identified due to insufficient resolution of satellite images and no available images from Street View.
- 1 polygon was identified as illegal building.

The presence of the building in the reality was documented by images in Street View and LiDAR measurement and the fact that it was identified as illegal what was supported by the fact that it was not drawn in the Cadastral Map and also by missing information regarding the building in the ownership document for the plot where building was found.

The values for different polygons that were classified are stored in Table 14.1 and the polygon count along with the percentual proportion from the total number of building polygons are listed in Table 14.2.

Polygon type	No. of polygons
Input polygons VCM	250
Input polygons LiDAR	253
Outliers between data sets	32
Excluded small polygons	17
Polygons for identification	**15**
Not present in VCM (due to permit)	5
Unidentified	9
Illegal buildings	**1**

Table 14.1 Summary of compared polygons

	Number of polygons	Compared to total [%]
Total of building polygons	253	-
Misclassified polygons	15	5.9
Unclassified buildings	20	7.9
Polygons common with VCM	221	87.4
Outliers from VCM	32	12.6
Small buildings	17	6.7
Polygons for identification	15	5.9
Identified illegal building	1	0.4

Table 14.2 Polygons count with proportion percentage

I want to conclude that I managed to find a solution to the problem statement through research and experimental part, where I successfully identified illegal building and described the process in this report.

15. Discussion

Despite many complications that occurred during the experimental part of the work I am satisfied that I managed to achieve the goal and identify illegal structure from the LiDAR data set. This project brought interesting topics in consideration and broaden my spectrum of knowledge.

I am convinced there is space for improvement of the overall process which could lead to better results. Different approach to resolving the problems that occurred during the work on project as could bring better results. One of these could be different parameter and workflow for building classification resulting in less missing structure and unclassified buildings.

Evaluation of the results achieved by the comparison of data set could be treated by automated processing which would require less time than manual control. Such evaluation might be based on processing algorithm comparing the rectangular shape of identified polygons with another data source as orthophoto for example.

Major complication was caused by the structure of VCM where the building boundaries were not store separately but had to be manually edited. For the future experiments it would be very useful to obtain data containing separate building boundaries from ZBGIS portal or other data source.

Last though is related to the possibilities of use of classified LiDAR data. With the outlined structure of buildings and vegetation which can be exported in multiple formats, these can be used to create various kinds of applications for urban or environmental planning, mapping or implementing graphical designs in the surroundings in real time. I see great potential in using LiDAR scanners and hope to work with this technology soon.

Bibliography

1**. Office of Geodesy, Cartography and Cadastre of the Slovak Republic. 1995.**
Law no. 162/1995 Z.z. National Law Of Cadastre Of Real-Estate.
http://www.zakonypreludi.sk/zz/1995-162, 2018

2. **Satellite Imaging Corporation.** Satellite images.
https://www.satimagingcorp.com, 2015

3. **American Congress on Surveying and Mapping. 1994.** Glossary of the Mapping Sciences, American Society of Civil Engineers, ISBN 9780784475706

4. **Geoportal.** https://www.geoportal.sk/sk/geoportal.html, 2018

5. **UGKK SR**, Office of Geodesy, Cartography and Cadastre of the Slovak Republic.
http://www.skgeodesy.sk/sk/ugkk, 2018

6. **Morin. 2002.** Calibration of Airborne Laser.
http://www.ucalgary.ca/engo_webdocs/NES/02.20179.KrisMorin.pdf, 2015

7. **Abdullah. 2012.** Methodology for processing raw lidar data for urban flood modeling. Delft. ISBN 978-0-415-62475-6.

8. **NOAA. 2013.** LIDAR - Light Detection and Ranging - remote sensing method used to examine the surface of the Earth.
http://oceanservice.noaa.gov/facts/lidar.html, 2013.

9. **Meng, Currit. 2010.** Ground filtering algorithms for airborne Lidar data. ISNN 2072-4292

10. **Isenburg. 2015.** Processing point clouds collected by LiDAR scanner.
https://rapidlasso.com/complete-lidar-processing-pipeline-from-raw-flightlines-to-final-products, 2015

11. **Zhang, Whitman. 2005**. Comparison of three algorithms for filtering airborne LIDAR data. Photogramm. Eng. Remote Sens.
http://www.academia.edu/7620432/Comparison_of_Three_Algorithms_for_Filtering_Airborne_Lidar_Data, 2005

12. **Aijazi. 2013**. Segmentation based on classification of 3D urban point clouds.
https://www.researchgate.net/publication/258811391_Segmentation_Based_Classification_of_3D_Urban_Point_Clouds_A_Super-Voxel_Based_Approach_with_Evaluation, 2013

13. **Vosselman, Sithole. 2004.** Recognizing structure in laser scanner point clouds.

14. **Applied knowledge services. 2015.** Urban migration.
http://gsdrc.org/topic-guides/urban-governance/key-policy-challenges/urban-migration, 2018

15. **Construction law. 2015.** Appendix
https://petras.blog.sme.sk/c/121446/Novy-stavebny-zakon-1Dodatocna-legalizacia-ciernych-stavieb-ostava.html, 2015

16. **SME newspaper. 2014.** Illegal cottages in region Oravska priehrada.
https://myorava.sme.sk/c/20223297/cierne-stavby-lemuju-breh-oravskej-priehrady-zvalit-ich-nema-kto.html, 2018

17. **SME Bratislava. 2015** Bratislava: Illegal constructions.
https://bratislava.sme.sk/t/2137/bratislava-cierne-stavby, 2015

18. **SKgeodeti. 2017.** http://skgeodeti.sk

19. **Doneus, Miholjek. 2015.** Airborne laser bathymetry for documentation of submerged archaeological sites in shallow water. ISPRS – International Archives of the Photogrammetry, Remote Sensing and Spatial Information

20. **Short, Nicholas. 2010.** Elements of Aerial Photography. Remote Sensing Tutorial Page 10-1. NASA.

21. **University of Colorado Boulder. 2011.** Aerial Photography and Remote Sensing.
https://www.colorado.edu/geography/gcraft/notes/remote/remote.bak5, 2011

22.**Graham, Roger. 2000.** Manual of Aerial Photography.
London and Boston, Focal Press. ISBN 978-184995-286-6

23. **Fox. 2011.** Advanced Remote Sensing. https://www.leefoxadvremote.blogspot.com, 2011

24. **Gisgeography. 2018.** A complete guide to Lidar: Light Detection And Ranging.
https://gisgeography.com/lidar-light-detection-and-ranging, 2018

25. **UGSS. 2018.** Earth Resources Observation and Science Center
https://www.usgs.gov/land-resources/earth-resources-observation-and-science-center?qt-programs_l2_landing_page=0#qt-programs_l2_landing_page, 2018

26. **Weih, Riggan 2012.** Reason and Rigor. Conceptual framework.
https://www.researchgate.net/publication/278961764_Theoretical_and_Conceptual_Frameworks_in_the_Social_and_Management_Sciences, 2012

27. **Chiu, Cheng-Lung. 2015.** National Airborne Lidar Mapping and Examples for applications in deep seated landslides in Taiwan. Geoscience and Remote Sensing Symposium (IGARSS), 2015

28. **Elmqvist. 2002.** Principles Of Object-Oriented Modeling And Simulation. https://books.google.sk/books?isbn=1118859162

29. **Markham. 2014.** Simple guide to confusion matrix terminology https://www.dataschool.io/simple-guide-to-confusion-matrix-terminology, 2014

30**. Blanschke. 2012.** A review of supervised object-based land-cover image classification https://www.sciencedirect.com/science/article/pii/S092427161630661X, 2012

31. **Trimble. 2016.** Trimble Harrier 68i Corridor Mapping System, Data Sheet. https://www.trimble.com, 2016