Classification of Tartu's street network

1. Task. Setting up

In our GitHub repository, we will upload the code used to carry out the tasks. Due to the nature of this project, based on geographical representations, we used QGIS to visualise and create points and polygons, thus, we cannot show these processes on the scripts.

Link to the repository: https://github.com/MrKri03/IDS_project

2. Business understanding

The Tartu City Government is currently using a structured geographical database containing line shapes to describe the Tartu road network. These lines are constantly updated in the main database, giving a meaning to each segment of the network with some descriptors of traffic roads, such as width, speed limit or cover. Nevertheless, the Tartu City Government is restricted to use only these lines to represent the traffic and road status on maps. In order to use a more meaningful geographical unit, they require polygon shapes containing and describing the roads with the same feature descriptors as the lines. By this way, the Tartu City Government can be more transparent showing the roads on maps in different case scenarios, for instance, during roadworks, roads blocked for special events or sections of roads blocked due to snowstorms.

To obtain a polygon shape, the Tartu City government proposes to carry out road recognition from aerial orthophotos taken by Maamet. There are three possible ways to perform this task: two of them use Machine Learning classification using the pixel values and the third one uses a Deep Learning classification. The first one is based only at the pixel level, where a Machine Learning classifier is trained with pixel values and classifies the pixels. The second one is similar but instead of classifying pixels, it uses the Geographical Object-Based Image Analysis (GEOBIA) where neighbour pixels with similar characteristics defined by a segmentation algorithm are classified in one label. The GEOBIA methodology can overcome the limitations of pixel classification because the gradient of values of pixels in images of urban areas is heterogeneous, leading to a mixture of labels in the classification. If these pixels share a common edge, they will be classified as an homogeneous unit.

The third approach is based on the novel U-Net deep neural network architecture. The main disadvantage of this approach is that it needs several images as training samples, and also data augmentation to be able to recognise the roads.

This project addresses the road extraction problem with the second approach (figure 1) with a robust segmentation algorithm and machine learning classifier developed for open-source libraries to obtain classified polygon segments (hereafter PS). Theoretically, roads can be distinguished from other features by exploiting spectral and spatial information but this can depend on the classification, depending on the variables (colours of the bands and ancillary data), post-classification and visual inspection. Thus, both the segmentation and classification algorithms play a key role to discriminate features on the orthophotos and get the roads.

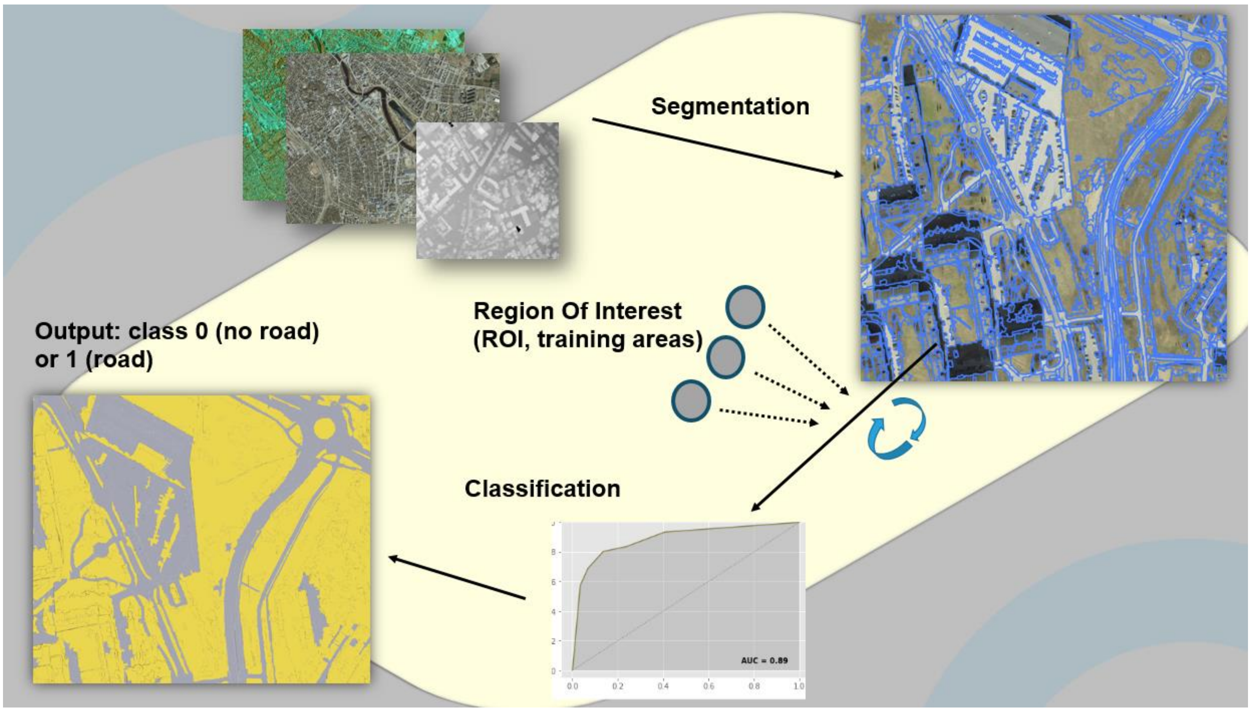


Figure 1. General workflow to follow in this project.

The results will show how feasible it is to perform a road extraction using a rapid methodology with the public geodata available. The assessment of the Machine Learning classifiers will be done on a binary classification and based on the results, the same classifiers will be used to extrapolate the road extraction to other areas. For our goal, we test the methodology on three areas of Tartu with three different degrees of complexity: different levels of roads, different sizes of buildings, more or less vegetation hiding the roads, and different sizes of shades coming from different objects on the surface, such as buildings.

To perform these analyses, the team includes two members who have the knowledge in Remote Sensing, Geographical Information Systems and Data Science. This implies that, from a selection of variables based on remote sensing criteria, the team builds a suitable data frame for classification and assessment. Only open-source software is used, reducing costs of private software. Specifically, Python programming libraries and QGIS (Quantum GIS) are used to process images and visualise the training areas and results, because here the spatial component is essential. Because the source data comes from official data warehouses, the team assumes that it is correctly preprocessed (orthoimages are georectified correctly, lines do not have geometry errors…). The only constraints are the processing speed, as the team is not using a powerful cloud processing but own machines with limited resources. The images are typically heavy to process.

Final results will be presented in a poster format showing the whole workflow, similar to figure 1, and the assessment of classifiers, as well as a final map of the road extraction. Also, the geographical shapes will be produced, preserving, as requested by the responsible person of geodata of Tartu City Government, the unique identifier (Global ID) of each segment in the road network.

3. Data understanding

3.1 Gathering data

Our project's goal is to classify the Tartu City Government roads into 5 hierarchical categories, based on various parameters. The spatial data required for the project is:

a. Tartu city roads
   i. geometries (.shp)
   ii. road attributes (columns) also from the .shp file.
   iii. datamodel (explanations of the attributes)
b. Orthophoto of Tartu
   i. RGB raster (.tif) data.
   ii. CIR raster (.tif) data.
c. Digital surface model (DSM)
   i. raster (.tif) data.

The shapefile and raster data should be close in time to make sure that the data would match. It's also necessary that the orthophotos are done in the spring where the roads aren't covered with vegetation shadows. We want to be able to see as much road on the orthophoto as we can. Orthophotos and DSM are open data that can be downloaded from Maa-amet website (Geoportaal, i.a). Tartu is covered mainly with two orthophotos in scale 1:10 000. We are going to use map sheets with numbers 54761 and 54752. We are also going to use DSM data which should help us to distinguish between roads and rooftops. Using the RGB values from the orthophotos only to train the classifier might not discern between roads and rooftops, thus, the DSM data was added to help with the model training. We are going to use a 1 m resolution DSM with a scale 1:2000 to get the most accurate data. We downloaded 10 images (.tif): 473658 ,474658 ,474659 ,472657 ,471658, 471657, 472658, 473659, 474662, and 473662.

Tartu roads are also open data that can be downloaded as a shapefile from Tartus GeoHUB website. We downloaded "LI_soidutee.shp" from 29.10.2022. Tartus GeoHUB data model was sent to us by Tiina Arras who works at Tartu City Government. The initial downloaded data can be seen on figures 2 and 3.

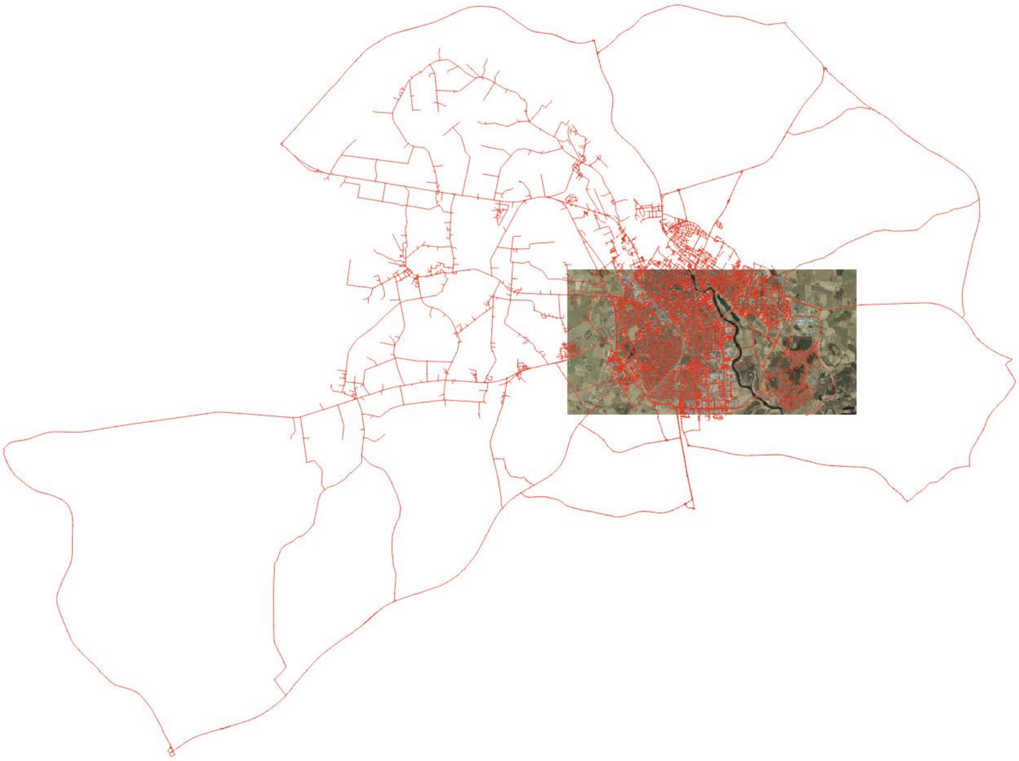


Figure 2. Downloaded data for the project. Red lines are Tarty city roads in from the Tartus GeoHUB database and raster images are downloaded from Geoportaal.

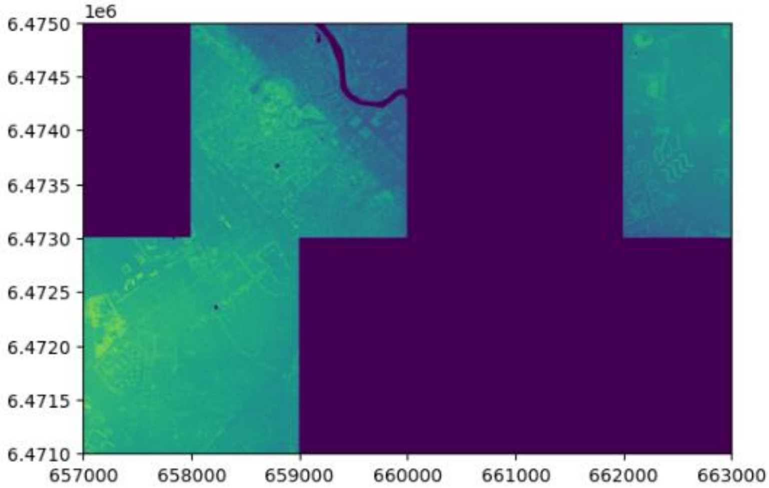


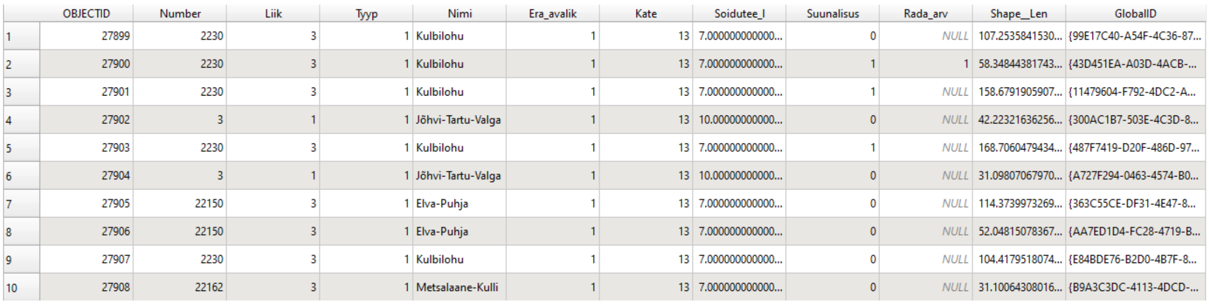
Figure 3. Downloaded DSM images merged together.

In addition to the RGB orthophotos , we download the Coloured Infrared (CIR) orthophotos of the same date to get the values only of the near infrared band. This band was added to the RGB .tif file, because this band could also help to train a better classification model.

3.2 Describing data

The orthophotos are from different dates. Map sheet 54761 orthophoto was taken on 04.04.2019. Sheet number 54752 orthophoto was taken on 03.04.2019. Two orthophotos are different but we also couldn't use the newer orthophotos because those images were taken in June (07.06.2021) and the vegetation was covering the roads a lot more than on the images taken in 2019. Map sheet 54761 and 54752 orthophoto resolution is 20 cm and both are in Lambert Conformal Conic Coordinate Reference System. DMS data is with 1 m resolution and the data has been updated 29.11.2019.

Road shapefile attribute table includes 11 fields. Fields are road number, type (“liik”), special type (“tyyp”), name, property, cover, width, number of lanes and global ID. There’s also object (“OBJEKTID”) and shape length that are usually automatic fields in shapefiles. All the fields are shown on figure 3.



	OBJECTID	Number	Liik	Tyyp	Nimi	Era_avalik	Kate	Soidutee_l	Suunalisus	Rada_arv	Shape__Len	GlobalID
1	27899	2230	3	1	Kulbilohu	1	13	7.000000000000...	0	NULL	107.2535841530...	{99E17C40-A54F-4C36-87...
2	27900	2230	3	1	Kulbilohu	1	13	7.000000000000...	1	1	58.34844381743...	{43D451EA-A03D-4ACB-...
3	27901	2230	3	1	Kulbilohu	1	13	7.000000000000...	1	NULL	158.6791905907...	{11479604-F792-4DC2-A...
4	27902	3	1	1	Jõhvi-Tartu-Valga	1	13	10.00000000000...	0	NULL	42.22321636256...	{300AC1B7-503E-4C3D-8...
5	27903	2230	3	1	Kulbilohu	1	13	7.000000000000...	1	NULL	168.7060479434...	{487F7419-D20F-486D-97...
6	27904	3	1	1	Jõhvi-Tartu-Valga	1	13	10.00000000000...	0	NULL	31.09807067970...	{A727F294-0463-4574-B0...
7	27905	22150	3	1	Elva-Puhja	1	13	7.000000000000...	0	NULL	114.3739973269...	{363C55CE-DF31-4E47-8...
8	27906	22150	3	1	Elva-Puhja	1	13	7.000000000000...	0	NULL	52.04815078367...	{AA7ED1D4-FC28-4719-B...
9	27907	2230	3	1	Kulbilohu	1	13	7.000000000000...	0	NULL	104.4179518074...	{E84BDE76-B2D0-4B7F-8...
10	27908	22162	3	1	Metsalaane-Kulli	1	13	7.000000000000...	0	NULL	31.10064308016...	{B9A3C3DC-4113-4DCD-...

Figure 3. Attribute table of Tartu city roads.

Roads are provided as segments and there are 13484 road segments in the shape file. Each segment has an unique ID ("GlobalID"). In the assignment we are requested to keep the global ID because it is used to join data in Tatu City Governments MSSQL database.

3.3 Exploring data

Raster images width and height is 25000. The images are with 3 bands (RGB). Map sheet 54752 had band 1 values from 33 to 225, band 2 values from 30 to 228 and band 3 values from 22 to 219. Map sheet 54761 band 1 values are between 30-221, band 2 between 26-233 and band 3 from 19 to 226. These digital numbers represent a scaled range of intensities of reflected light to the sensor. The higher the value, the brighter the colour is.

DSM images width and height is 1000 m and they are all with one band. The band values range between 0-106.19.

Road shapefile includes most of the necessary fields that are needed for the analysis. Those fields are road type and cover. Field "liik" is an integer data type that contains hierarchical road type. Values range between 1-18 and code 99 (road that's in the planning process). Field cover ("kate") is also an integer field that contains codes from 0 to 49 and 99. The data model describes all the code values. Both fields have no missing data (0 or Nan values).

3.4 Verifying data quality

Orthophotos are high quality and quite big files. Map sheet 54752 is 364080 KB and map sheet 54761 is 368618 KB. It's more than we can actually process so for the analysis we have decided to create three study areas. Two areas are in map sheet 54752 and one area in 54761 (figure 3). The DSM data is also really detailed and value 0 is used only for the water bodies.

The shape file is quite good quality. The roads are digitised mainly properly and are also snapped to create a network. Although there are a few roads that are of poor quality. For example we can see on figure 4 that the roadline (north-south) direction has been drawn next to the road, not on it. We have cut out the road in the study areas and checked that all the roads in the study areas are digitised on the roads.

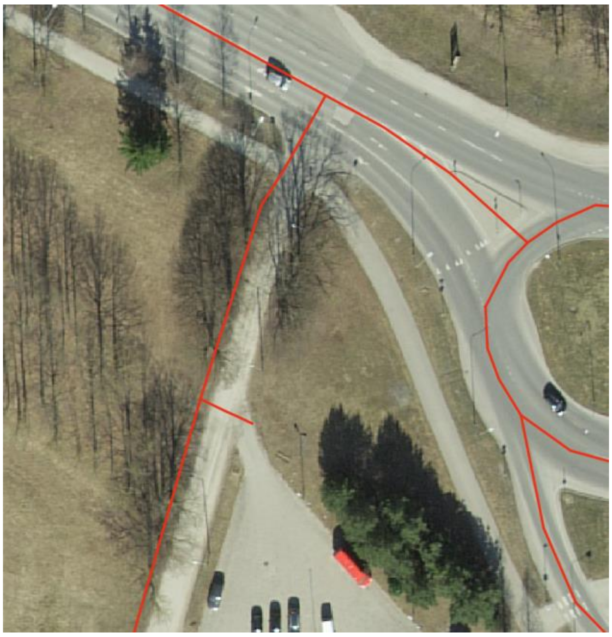


Figure 4. An example of a road that's now digitised correctly on the road.

The shapefile is missing road width and speed limit parameters which was one of the parameters that the Tartu City Government wanted to classify the roads with. The road width is a parameter that we can’t find from anywhere else and it's also a parameter that the Tartu City Government doesn't have. The city is also missing the speed limit data. We should able to get it according to Tiina Arras form the Tark Tee website (https://tarktee.mnt.ee/#/en) but we can't find it. So at the end we are planning to use road type and road cover type to classify the roads into 5 hierarchical categories.

In general we have created 3 study areas and clipped all the given data (roads, orthophotos and DSM) with the study areas. We have also added a fourth band (near-infrared) to the RGB images. The final data is shown on figure 5.

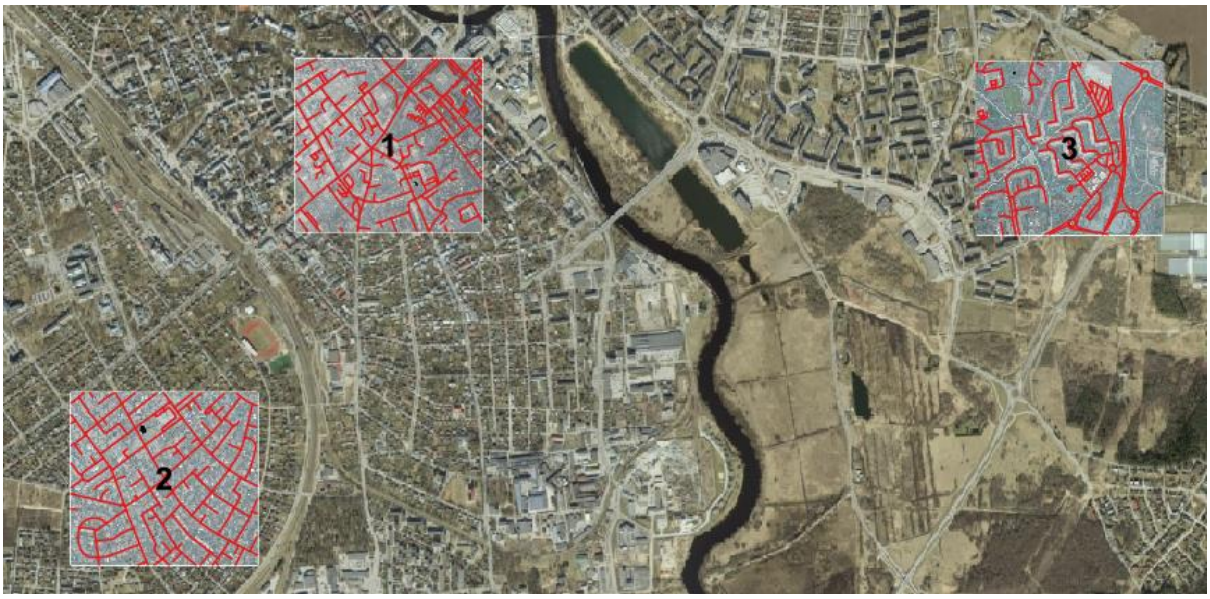


Figure 5. Three chosen study areas that are going to be used for this project.

4. Planning our project

We followed a methodology combining GIS techniques to overlay, intersect and mask spatial data, as well as value extraction from rasters (georeferenced array of numbers) using a zonal statistic tool within polygon shapes. These polygon shapes are defined by : 1. Buffer areas surrounding points to use as training dataset. Each polygon of buffer is located in specific regions of the orthophotos that we considered to have significant values to describe the labels (0- no road ; 1-road). Based on our expertise, we placed training polygons on rooftops, roads, shades and other type of surfaces (parks or vegetation). 2. PS obtained from segmentation, which, having the same attributes as the training buffers but no labels, will be predicted.

The goal of this methodology is to build two different geodataframes `using pandas, numpy, rasterio` and `geopandas` libraries. One to train a classifier (buffers) and the other to predict from extracted features (PS). These PS are the result of image segmentation using the `scikit-image` package, specifically the Felsenszwalb's algorithm (Felzenszwalb & Huttenlocher, 2004) because it is fast and requires only three hyperparameters : scale, size and sigma. The first two are related to the level of detail to split the clusters of pixels and the third one, the tolerance of the Gaussian filter used to group pixels during the preprocessing. The Machine Learning algorithm is implemented using the `scikit-learn` package, specifically Random Forest (Breiman, 2001) for non-parametric distribution of values and its high performance for image data (Maxwell et al., 2018).

All other functionalities are run using the most popular packages built in Python programming language.

List of tasks:

1. Planning the work and collecting data: 3h
2. Data preparation: 40h (20h each)
3. Finding the best model and training the data: 5h
4. Data interpretation: 6h
5. Poster and summarising the results: 5h

The time can be divided by two because every team member is involved in all the steps.

References

Breiman, L. (2001). Random Forests. *Machine Learning* 45, 5–32.

Felzenszwalb, P.F, Huttenlocher, D.P. (2004). Efficient Graph-Based Image Segmentation. *International Journal of Computer Vision* 59, 167–181.

Geoportaal. (i.a). Lastly visited 26.11.2022. https://geoportaal.maaamet.ee/eng/Spatial-Data-p58.html.

Maxwell A.E., Warner T.A., Fang F. (2018). Implementation of machine-learning classification in remote sensing: an applied review. *International Journal of Remote Sensing* 39:9, 2784–2817.

Tartu GeoHUB. (i.a). Lastly visited 26.11.2022. https://geohub.tartulv.ee/pages/avaandmed#avaandmed.

Tartu GeoHUB datamodel. (i.a). Lastly visited 26.11.2022. https://tartulv.sharepoint.com/:x:/s/GISmeeskond/EYW2juvZ8ShJjKxV5Wp56tEBfa-fBawXi0Sr5uQ17FTIfg?rtime=9Ydwc2bM2kg.