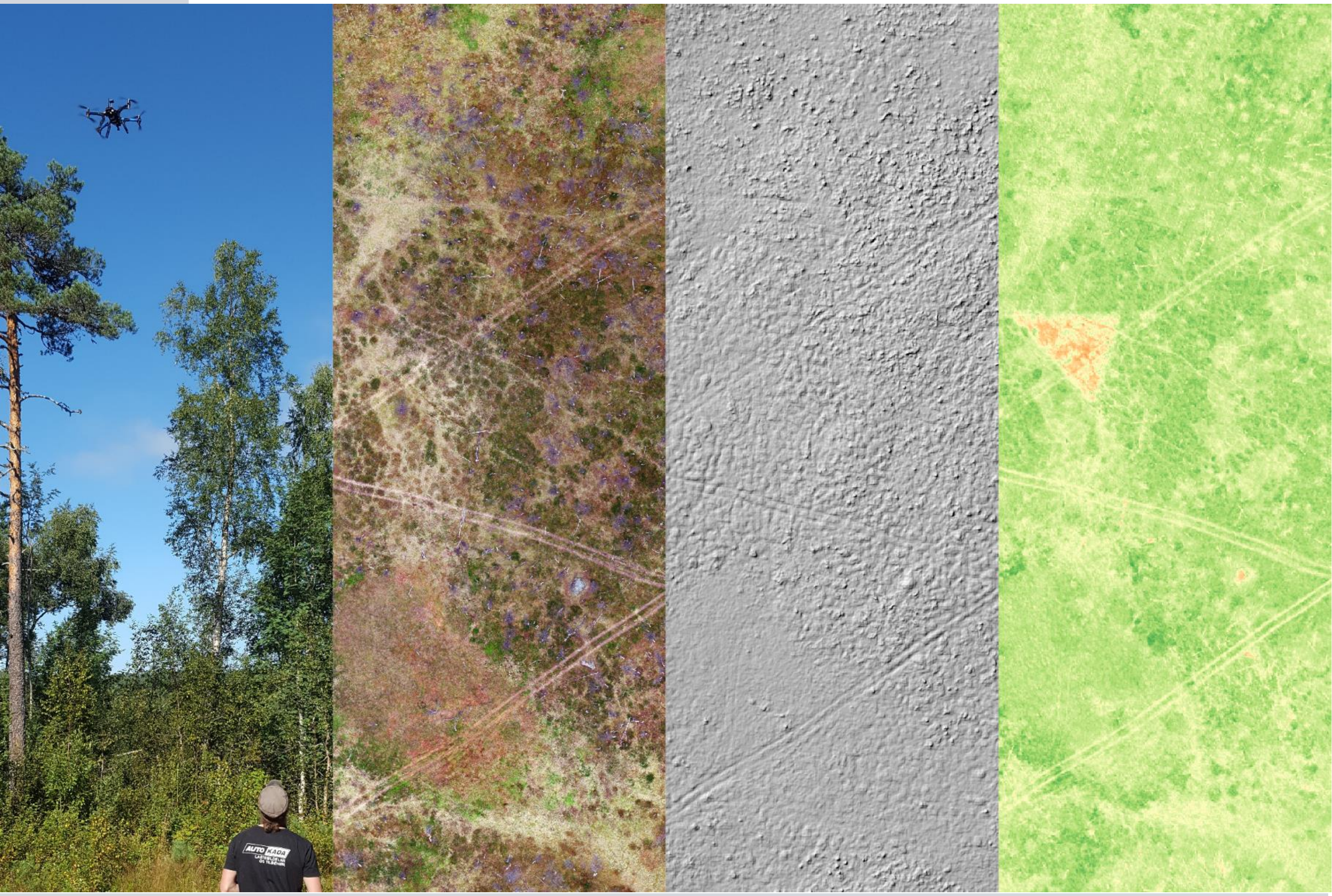
www.nina.no

2137

NINA Report

Wheel rut mapping with high resolution ortho-imagery – a comparison of data and methods

Stefan Blumentrath, Stefano Puliti (NIBIO), Sindre Molværsmyr, Øyvind Hamre



NINA

Norwegian Institute for Nature Research

NINA Publications

NINA Report (NINA Rapport)
This is NINA's ordinary form of reporting completed research, monitoring or review work to clients. In addition, the series will include much of the institute's other reporting, for example from seminars and conferences, results of internal research and review work and literature studies, etc. NINA

NINA Special Report (NINA Temahefte)
Special reports are produced as required and the series ranges widely: from systematic identification keys to information on important problem areas in society. Usually given a popular scientific form with weight on illustrations.

NINA Factsheet (NINA Fakta)
Factsheets have as their goal to make NINA's research results quickly and easily accessible to the general public. Fact sheets give a short presentation of some of our most important research themes.

Other publishing.
In addition to reporting in NINA's own series, the institute's employees publish a large proportion of their research results in international scientific journals and in popular academic books and journals.

Wheel rut mapping with high resolution ortho-imagery – a comparison of data and methods

Stefan Blumentrath
Stefano Puliti
Sindre Molværsmyr
Øyvind Hamre

Norwegian Institute for Nature Research

Blumentrath, S., Puliti, S., Molværsmyr, S. & Hamre, Ø. 2022. Wheel rut mapping with high resolution ortho-imagery – a comparison of data and methods. NINA Report 2137. Norwegian Institute for Nature Research.

Oslo, April 2022

ISSN: 1504-3312
ISBN: 978-82-426- 4925-6

AVAILABILITY
Open

PUBLICATION TYPE
Digital document (pdf)

QUALITY CONTROLLED BY
Benjamin Cretois

SIGNATURE OF RESPONSIBLE PERSON
Research director Kristin Thorsrud Teien (sign.)

CLIENT(S)/SUBSCRIBER(S)
Environmental management agency (Miljødirektoratet)

CLIENT(S) REFERENCE(S)
M-2289 | 2022

CLIENTS/SUBSCRIBER CONTACT PERSON(S)
Ragnvald Larsen

COVER PICTURE
UAVs for monitoring wheel ruts © Øyvind Hamre, Bård Gunnar Stokke

KEY WORDS
UAV, drone, ortho photo, image analysis, deep learning, wheel rut, monitoring, GIS, method

NØKKELORD
UAV, drone, ortofoto, bildeanalyses, dyp læring, kjørespor, overvåkning, GIS, metodeutvikling

CONTACT DETAILS

NINA head office
P.O.Box 5685 Torgarden
NO-7485 Trondheim
Norway
P: +47 73 80 14 00

NINA Oslo
Sognsveien 68
0855 Oslo
Norway
P: +47 73 80 14 00

NINA Tromsø
P.O.Box 6606 Langnes
NO-9296 Tromsø
Norway
P: +47 77 75 04 00

NINA Lillehammer
Vormstuguvegen 40
NO-2624 Lillehammer
Norway
P: +47 73 80 14 00

NINA Bergen:
Thormøhlens gate 55
NO-5006 Bergen.
Norway
P: +47 73 80 14 00

www.nina.no

Abstract

Blumentrath, S., Puliti, S., Molværsmyr, S. & Hamre, Ø. 2022. Wheel rut mapping with high resolution ortho-imagery – a comparison of data and methods. NINA Report 2137. Norwegian Institute for Nature Research.

The number of registered All Terrain vehicles (ATV) in Norway has been constantly increasing over the last decade. Driving these vehicles off-road in nature requires special permission because of the potentially severe damage it can cause in nature. The number of registered vehicles however suggests that illegal driving occurs. In order to be able to better monitor this issue, efficient mapping techniques are required to cover the large areas where this can be relevant. Remote sensing has been explored as an efficient technique for that purpose earlier, however recent technological advances like the availability of Unmanned Aerial Vehicles (UAV) or deep learning for image analysis may further increase the potential of remote sensing.

The aim of this project has therefore been to develop a coherent workflow to detect wheel ruts in drone and / or plane-based aerial imagery that can serve as a starting point for practical monitoring tools. Method-development was conducted for two case study sides, one in northern and one in southern Norway, where drone imagery from autumn 2020 was available. The developed workflow covers all relevant steps of image analysis from preparation of input data to postprocessing of modelling results and was made publicly available as a set of Python scripts.

Results show that deep-learning performed better than more traditional image analysis techniques and the initial deep-learning models developed in this project produce fair to good results for both plane- and drone based imagery in both study sites. Models utilizing drone data perform slightly better than models based on aerial images with regards to correctly capturing wheel ruts. Models based on drone imagery capture more details but currently also show a larger degree of noise and scattered false positive classifications. Models from aerial images perform best in open areas while they struggle more in forested areas. The developed post-processing routine improves the quality of the final products and can produce condensed and more usable representations of the results. However, the classification of the results during post-processing with regards to the severity of damages to esp. soil / terrain should undergo systematic evaluation and re-adjustment if needed.

Together with the modelling results, re-processing of the raw drone images illustrate that the main benefit from using drone imagery is the timely data acquisition, both in terms of time of the year but also with regards to e.g. local urgency to monitor an area in more detail. Drone data can therewith be seen as an on-demand technology that is complementary to aerial images that are taken on a regularly basis for the Norwegian orthophoto program every 5th to 10th year.

Other potential benefits from using drone imagery like the possibility to capture photogrammetric terrain models or multispectral imagery in and of themselves, currently do not seem to justify the extra effort necessary to acquire drone imagery because their contribution modelling accuracy may even be negative due to data quality issues. In particular, monitoring of soil impact of wheel ruts by means of repeated collection of photogrammetric terrain models seems to be hard if not impossible to conduct at the extent and scale used in this project with study area of ~6km^2.

Due to the limited amount of situations that are covered by the current models, further training under different seasons, light conditions, vegetation types and so on would be necessary to make the models more transferable and thus applicable in practical management. Since models for both drone and aerial imagery show comparable performance, images from the Norwegian orthophoto program should thus be a natural starting point in order to increase the amount of training data and therewith the number of conditions the model is trained with. In that context, a more systematic evaluation of the effect of image resolution should be conducted and other available data sources like ultra-high resolution satellite images (with up to 30 cm resolution)

may be considered. It should also be investigated whether it is feasible and adequate from an end-user point of view to consolidate the deep learning models that currently are different for drone and aerial imagery, into one coherent model in order to reduce the maintenance effort and at the same time increase the amount of both training data and imagery the model could be trained with. To that end, also recent methods to limit the required amount of test- and training data, like Few-Shot Learning (see e.g. Wang et al. 2020) should be explored in order to increase the practical applicability in a monitoring context. Finally, even if the developed workflow is usable already, technical improvements can further improve the practical applicability

Stefan Blumentrath, NINA, Sognsveien 68, 0855 Oslo, stefan.blumentrath@nina.no
Stefano Puliti, NIBIO, Høgskoleveien 8, 1433 Ås, Stefano.Puliti@nibio.no
Sindre Molværsmyr, NINA, Thormøhlensgate 55, 5008 Bergen, sindre.molværsmyr@nina.no
Øyvind Hamre, NINA, Sognsveien 68, 0855 Oslo, oyvind.hamre@nina.no

Contents

Foreword

We would like to thank our contact persons at the Norwegian Environmental Management agency (Miljødirektoratet): Ragnvald Larsen, Marit Johanne Birkeland, Line-Kristin Larsen and Agnès Moquet-Stenback for good and constructive discussions and valuable feedback during progress meetings. Another thanks goes out to the development teams of the OpenDroneMap and GRASS GIS projects who swiftly addressed issues we encountered with the software during the project. Thanks also to Benjamin Cretois for a thorough quality check of the final report.

Oslo, April 2022
Stefan Blumentrath

1 Introduction

The number of registered All Terrain Vehicles (ATV) in Norway has constantly increased over the last decade, with growth rates significantly above those of e.g. passenger cars (see **Figure 1**).

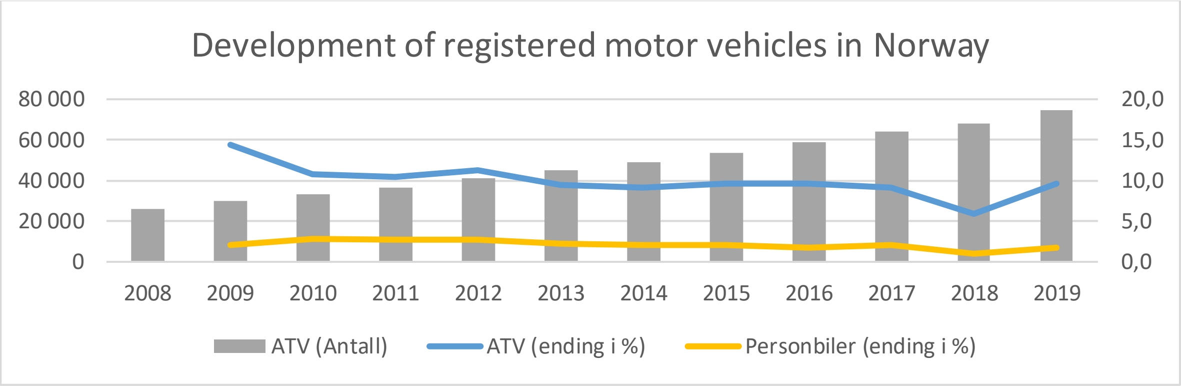


***Figure 1**. Development of registered All Terrain Vehicles (ATV) in Norway in comparison to passenger cars as a reference (Opplysningsrådet for veitrafikken 2022)*

In principle, it is forbidden in Norway to drive ATVs off-road in nature, without special permission. However, the increasing number of registered vehicles suggests that such illegal use occurs and that it may increase as well. The extent to which such illegal activities occur is however unknown.

The risk of damage to vegetation and soil associated with increasing legal and illegal use of ATV in the terrain is an important background for the project. Wear and tear of vegetation as an effect of traffic in open country and monitoring of this, has for a long time been an important topic related to, among other things. management of protected areas and national parks (see e.g. Tømmervik et al. 2005, Evju et al. 2010). Non-motorized traffic is also increasing, and this leads to similar challenges related to monitoring (Evju et al. 2020). Knowledge of the scope and location of the problem is, however, a prerequisite for effective handling in environmental management.

Due to the large and inaccessible areas where the problem can occur, there is a need for effective methods to capture relevant data and shed light on the problem. Remote sensing is therefore a natural data source to address this and try to get a better overview over the problem.

Attempts have previously been made to identify wheel ruts with the help of remote sensing from aircraft and high-resolution satellites (Tømmervik et al. 2005). However, the possibilities for automating the analyses have been limited (ibid). Soil damage from motor vehicles has also been an important issue in forestry for a long time (see Dale 1995). Access to new data sources, here especially from drones, and analysis methodologies such as image segmentation, object-based image analysis, machine learning and deep learning provide opportunities for more accurate and effective monitoring (see eg Pierzchała, Talbot & Astrup 2016, Ćwiąkała et al. 2018, Rodway-Dyer & Ellis 2018, Ancin-Murguzur et al. 2020, Eagleston & Marion 2020). Based on this, NIBIO has worked for several years to develop a sophisticated system for detecting wheel ruts from harvesting machines using drones in connection with logging (see, for example, Talbot, Rahlf & Astrup 2018), which is still under development (Kildahl 2020). The new methodological approaches have also been tested in connection with similar topics such as habitat type mapping, mapping of road edges (Senchuri 2020) or identification of road construction in connection with wind power development (Due Trier & Salberg 2020).

2 Aim and objectives

In order to address the demand of managers to be able to monitor the extent and potential effects of off-road usage of ATVs in nature, the aim of this project is to develop a prototype of coherent set of remote sensing tools that can be combined in a workflow that provides managers at the end with useful information on and insight into the extent of vegetation and soil damages from ATV usage in nature. That workflow is supposed to cover all relevant image analysis steps from preparation of training- and test data, training and retraining of machine learning models, detection of wheel ruts in imagery as well as post-processing of classification results. In addition to locating wheel ruts from ATV usage, desired output should also cover - as far as possible - an estimation of severity of wheel ruts e.g. in terms of depth of wheel ruts or degree of vegetation damage. Finally, the workflow should where relevant and possible include ancillary data to improve classification results.

A focus during development of that workflow is to assess opportunities, advantages and disadvantages of as well as trade-offs between available data sources. Here, emphasis is put on the transferability of the models between sensors, in particular between image data acquired by

- Plane, which is happening in Norway on a regular basis as part of a national program[1] conducted by the National Mapping Authority (Kartverket) with a repetition cycle of ~ 5 to 10 years. Data from that program will become available regularly without additional acquisition effort. This data is, however, usually limited to the RGB spectrum.
- Drone, which will require to conduct targeted, more or less resource demanding acquisition campaigns. Drone imagery can however also provide terrain information, multispectral channels or even high resolution laser scanning data.

Other possibly relevant sensors in that context, like newly available very high resolution satellite imagery are not scope of the project, but findings regarding plane based ortho imagery may give an indication about the applicability of that data source for the detection of wheel ruts. The aim of that assessment is to help identifying an appropriate balance between detection quality and considerations regarding practical applicability with regards to resource requirements for processing (compute) and labor for acquisition of data. A sub-objective in that context is to study to what extent a two-fold approach is feasible and appropriate, that uses plane-based orthophotos and drone imagery complementary, where aerial images are used for “screening” of lager areas and drone campaigns targeted for in depth monitoring of relevant areas identified during screening.

In order to identify a cost-efficient monitoring approach, another sub-goal of the project is to determine a set of minimum requirements with regards to image / data quality and methodology.

The developed software solutions will be based on Free and Open ource technology and software libraries, so that the source code and models that will be made available as a result of the project can serve as a vendor-neutral, usable prototype, that opens for further development and improvement e.g. through application in new areas with either targeted data collection (from drones) or independent data acquisition (national aerial image program).

[1] https://www.kartverket.no/geodataarbeid/program-for-omlopsfotografering

3 Study Area, Data and Methods

3.1 Study areas

Study areas for this project have been two sites, one in Balsfjord, Troms (Northern Norway) and one in Rjukan, Telemark (Southern Norway) (**Figure 2**). The extent of the study area in Balsfjord is 2 km in east-west direction and 2.3 km in north-south direction, while the Rjukan study area is 3.77 x 2 km respectively. The landscape at Balsfjord is mostly birch forest and wetland areas, while Rjukan is located at and above forest line and mostly covered by low alpine vegetation.

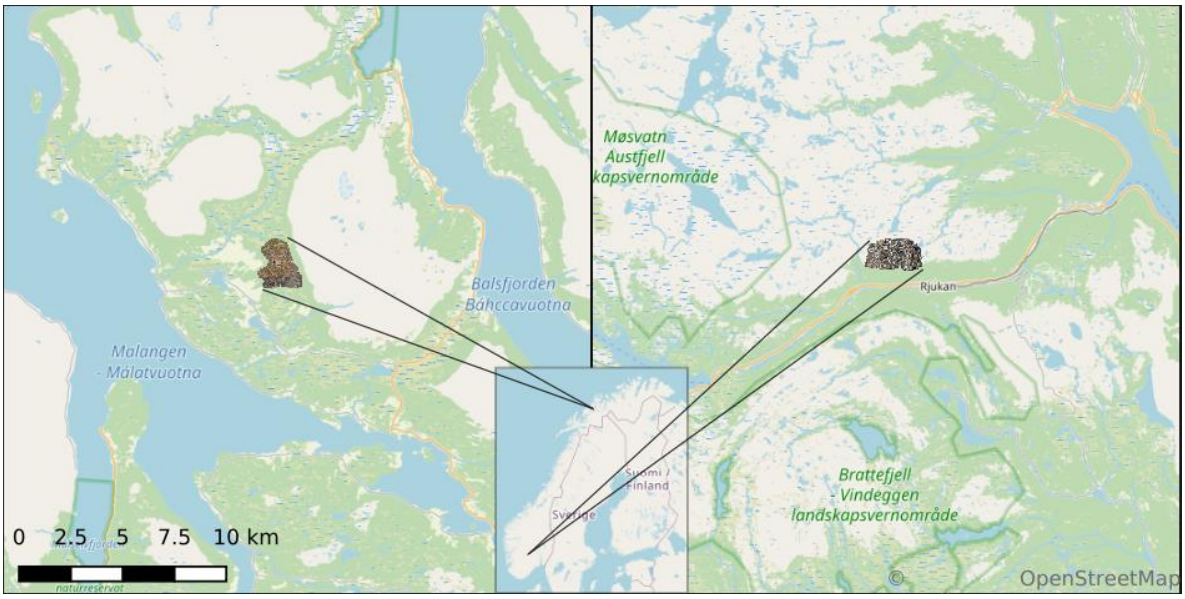


***Figure 2**. Overview over the location of the study areas in Norway*

3.2 Data

Data utilized in this project consists of different types of imagery data collected for the two study areas with UAV- and plane-born sensors (the latter referred to as “aerial images” in the following), training data for image analysis as well as ancillary geospatial data for masking possible mis-classifications.

3.2.1 Imagery data

3.2.1.1 Raw drone image data

Drone photos were captured by Andøya Space Center in 2020 using both DJI Phantom 4 Pro (P4P) and DJI Phantom 4 Multispectral (P4M). The drones have been flown in a grid pattern with a maximum altitude of 120 meters.

***Table 1**. Basic metadata for the raw drone imagery available to the project*

Area	Date	Sensor	Size (in km^2)	Approximate ground sampling distance (GSD in cm / pixel)	# of photos	Size (GB)
Skutviksvatnet	2020.10.08	P4M Multispektral	3,3	2.6	35.988	122
	2020.10.08	P4P RGB	3,3	2.8	3.311	26,8
Rjukan	2020.10.31	P4M Multispektral	9	3.5 x 7.0	73.755	250
	2020.10.31	P4P RGB	9	3.7	4.189	29,3
					117.243	**428,1**

3.2.1.2 Processed drone image data

Andøya Space Center processed the raw drone images into two RGB mosaics. Digital Surface Models (DSM) or Digital Terrain Models (DTM) that can be produced with photogrammetric algorithms from raw drone imagery had not been produced and was not available. Neither had the acquired multispectral drone imagery from the DJI Phantom 4 Multispectral (P4M) been processed, that contains infrared and red edge bands that are regularly used for vegetation monitoring.

3.2.1.3 Orthophotos from Norge i Bilder

For the two study areas, orthoimages were downloaded from Norge i Bilder in order to be able to study the effect of data sources on detection quality. The study site at Balsfjord was covered by a mosaic of two acquisitions with 0.25m resolution for one part of the area and 0.1m resolution for the other part (see **Table 2**). The study site at Rjukan was covered completely by a single project. For both study areas aerial-orthoimages were downloaded as LZW compressed Geotiffs, resampled upon download to the lowest common denominator of 25 cm resolution.

***Table 2**. Most recent orthoimages for the study sites available from Norge i Bilder*

Study site	Project name	Acquisition date	Resolution	Pixel depth
Balsfjord	Troms 2016	2016-08-19	0.25m	24bit
Balsfjord	Balsfjord Målselv 2017	2017-07-23	0.1m	24bit
Rjukan	Tinn 2019	2019-06-17	0.1m	24bit

3.2.2 Data annotation and training data

Data annotation was carried out for a selection of tiles for both Balsfjord and Rjukan using either the drone or the aerial images. For the study site at Balsfjord the entire area was annotated. However, due to time constraints only a portion of the study site at Rjukan was annotated. The coverage of the annotation is visualized in **Figure 3** where the grey areas indicate where annotation was conducted, while white areas have not been annotated. Annotation was carried out using tiles of the study area, where - within each tile – all visible, relevant objects (wheel rut damage, hiking trails, …) were registered. The partial annotation of the Rjukan study site thus leads to the chess-board-like pattern visible in **Figure 3**.

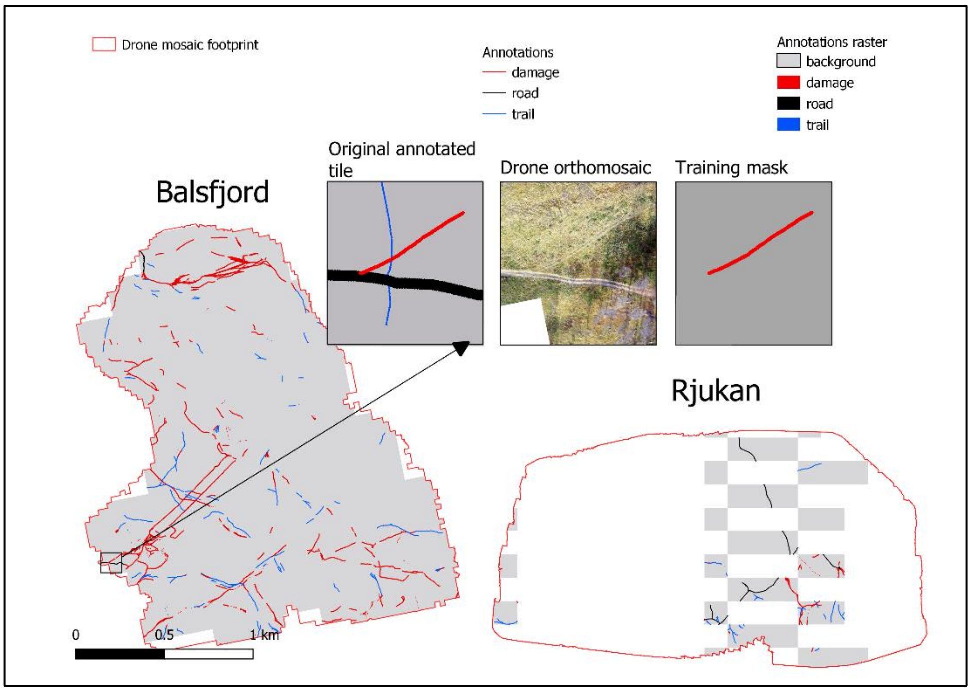


***Figure 3**. Visualization of the annotated data (grey areas: annotated, white areas: not annotated) for the drone datasets in Balsfjord and Rjukan. With a detail on a tile used for training the model.*

The annotations consisted of drawing polylines in the center of the following linear elements: wheel rut damage (damage), hiking trails (trails), and dirt roads (road). The polylines were then buffered based on the category (i.e. damage= 0.6 m; trails= 0.3 m; road=2 m) and the resulting polygons rasterized (see **Figure 3**). Due to the different points in time at which the drone and aerial images were taken, the two datasets had to be annotated separately. **Table 3** summarizes the length of the annotated tracks between the two data sources (i.e. drone and aerial images). The differences in total length were largely due to the lack of substantial wheel rut damage at the time of acquisition of the aerial images and in some cases due to their poor visibility because of the coarser resolution of the imagery or vegetation cover. The annotated data[2] is together with a script for pre-processing[3] available in the projects code repository on gitlab.

***Table 3**. Summary of the annotated datasets split by study area and data source.*

AOI	Dataset	Damage (m)	Trail (m)	Road (m)	Tot (m)
Balsfjord	**Drone**	15888	5821	283	21992
	Aerial images	7545	1342	389	9276
Rjukan	**Drone**	1490	1964	1998	5452
	Aerial images	1393	1615	2078	5086

[2] https://gitlab.com/ninsbl/wheel_rut_detection/-/tree/main/deep_learning/1_training_preprosess_bash/ground_truth_gpkg
[3] https://gitlab.com/ninsbl/wheel_rut_detection/-/blob/main/deep_learning/1_training_preprosess_bash/burn_fasit.sh

3.2.3 Ancillary data for masking possible misclassifications

In order to be able to mask out potential misclassifications of wheel ruts in a post processing workflow, ancillary geospatial data is used, meaning available data that has been mapped in other contexts. Here, especially data from “Felles Kart Database” (FKB[4]) is utilized. In areas where FKB data is lacking national 1:50,000 data (N50[5]) may be used alternatively, with somewhat reduced accuracy. From all potential layers available in the FKB dataset only the layer with road polygons and water lines were deemed relevant in this context.

Existing hiking trails may be exploited and extended by ATV users, so that using hiking trails as a mask dataset did not seem appropriate. Gravel roads were included in the FKB road polygons (at least for the study areas) so that using FKB line features was not considered here. Smaller streams and shores may spectrally appear to be similar to wheel ruts and confused with those during predictions. Therefore, line representation of water in FKB is utilized as an additional mask dataset. The FKB water layer is significantly more detailed than equivalent data in N50.

3.3 Methods

3.3.1 Reprocessing of raw drone images

The pre-processed mosaics provided by Andøya Space Center did neither contain a Digital Surface Model or Digital Terrain Model nor was a multispectral (including the Near Infrared (NIR) and Red Edge band) of the Phantom 4 Multispectral drone available. In order to be able to assess the information content of digital surface or terrain models in addition to multispectral imagery, the available raw drone images were re-processed using OpenDroneMap (ODM[6]) version 2.6.7.

The available raw drone imagery contains data from two consecutive days for both study sites, taken with the same drone model and by the same operator. This gives a unique opportunity to look at the potential to use repeated drone acquisitions to monitor changes in the landscape. Earlier studies in Norway by Ancin-Murguzur et al. (2020) showed high accuracy and reliably comparable terrain data. The authors claim that the application of drones is suitable and beneficial to monitor soil erosion in and around hiking trails. However, their study covered only a very limited area of 200 x 90 m photographed by a low-flying drone (10m) resulting in a ground sampling density of ~0.5 cm. It is an open question whether their conclusion remains valid when the application of drones is scaled up by a factor of ~100. In the context of reprocessing the raw drone images it was therefore also assessed how well digital surface models from drone acquisitions from two consecutive days at this scale are to detect eventual changes in the terrain. This was done by subtracting the two surface models to generate a difference map. Ideally, only minimal differences would be visible.

The specific settings for reprocessing in OpenDroneMap were chosen in order to optimize the quality of resulting digital surface and terrain models. Best practice recommendations to that end in the OpenDroneMap documentation were followed and the taken steps during reprocessing are documented in a Unix shell script that is available in the public GitLab repository[7] for the project. The resulting products were processed with 5 cm resolution, as this seems to strike a good balance with regards to the GSD in the raw imagery on the one hand and the amount of data and thus requirements with regards to resources for processing on the other hand.

[4] https://www.kartverket.no/geodataarbeid/geovekst/fkb-produktspesifikasjoner
[5] https://register.geonorge.no/register/versjoner/produktspesifikasjoner/kartverket/n50-kartdata
[6] https://www.opendronemap.org/
[7] https://gitlab.com/ninsbl/wheel_rut_detection/-/blob/main/drone_image_processing/run_odm_split.sh

3.3.2 Detection of wheel ruts in drone images and aerial photos with deep learning

For the scope of training the deep learning model only the damage class was considered, thus all the other classes in the annotation raster were set to zero (i.e. background). Two separate models were trained based on the input imagery and annotations (drone or aerial images) from both study sites:

- Drone model
- Aerial images model

For training the model we selected a Deeplabv3 model (Chen et al. 2017) with a ResNet-101 backbone available from the PyTorch hub[8]. DeepLab is a semantic segmentation architecture that overcomes the loss of information due to the input's size reduction in traditional convolutional neural networks / pooling layers In contrast, the DeepLab architecture allows the segmentation of objects at multiple scales by employing dilated convolutions and Atrous Spatial Pyramid Pooling (ASPP) modules.

The model training requires batches of images of shape [N, 3, H, W], where N is the number of images, H and W are the height and width of the image in pixels and should be > 224 pixels. Based on the hardware available for training the network, N, or the batch size was set to 20. To ensure batch sizes of at least 20, we converted the original RGB images to 16 bits. H and W were set to 300 pixels as this was the largest possible size given the available hardware. Thus, the wall-to-wall imagery and annotation rasters were tiled[9] into image tiles of 300 pixels x 300 pixels. When multiplied by the ground sampling distance of each image data source the number of pixels corresponds to squares with sides of approximately 21 m and 60 m for the drone and the aerial images, respectively.

Amongst all of the available tiles, tiles were selected that were intersecting with the annotated damage tracks. Amongst these, 70% of the tiles were randomly selected for model training, while the remaining were used for final validation of the model. In total 625 and 137 tiles were selected for training drone and aerial image models, respectively.

Training was performed for a maximum of 8000 epochs, without hyperparameter tuning. No early stopping strategy was adopted and the best model, identified as the one with best Intersection over Union (IoU) for the damage class was stored at the end of the process[10]. IoU is the area of overlap between the predicted segmentation and the ground truth divided by the area of union between the predicted segmentation and the ground truth.

During training, the best performing models were identified using Intersection-over-Union (IoU) metric and consequently selected for prediction of wheel ruts. The selected trained models were then applied and validated throughout the entire area. For validation the training tiles were excluded from the accuracy assessment as these data were seen by the model.

Because the reprocessed drone imagery became available only late in the project, only the mosaic produced by Andøya Spacecenter was used for wheel rut detection with deep learning.

[8] https://pytorch.org/hub/pytorch_vision_deeplabv3_resnet101/

[9] The tiling of the original images is documented in the following scripts: https://gitlab.com/ninsbl/wheel_rut_detection/-/tree/main/deep_learning/2_training_preprocess_python

[10] The model training is documented in the following scripts: https://gitlab.com/ninsbl/wheel_rut_detection/-/tree/main/deep_learning/4_train

3.3.3 Edge detection, object-based image analysis for detection of wheel ruts in drone images and aerial photos

In order to substantiate methodological choices, a sub-goal in the project has been to assess the performance of classical image analysis techniques like edge detection, texture analysis and object based image analysis (OBIA) in comparison to deep learning methods described in chapter 3.3.2.

Therefore, drone and aerial image mosaics were also processed using Object Based Image Analysis (OBIA) workflow loosely oriented on the approach by Lennert et al. (2019). RGB channels were used to compute image texture measures (e.g. the Grey Level Cooccurrence Matrix), edge detection (using zero crossing) and, based on all the aforementioned, image segmentation (using mean shift algorithm). Segmentation was conducted in a way that avoids over-segmentation, meaning that segments or image objects do not get too large so that pixels within the narrow wheel rut structures do not get merged with neighboring segments.

However, the superiority of deep-learning based methods for the given problem became evident already when comparing deep learning classification with segmentation results. In consequence, the OBIA methodology was not further pursued. An example of that comparison is presented and discussed in chapter 4.3.

3.3.4 Post-processing of classification / detection results

Prediction results from deep learning models described in chapter 3.3.2 are expected to contain misclassifications and noise that to some degree can be filtered using a post-processing routine (see **Figure 4**). Model output is a binary image classification that detects occurrence of wheel ruts. Applying post-processing to those binary classification results also allows to distinguish linear damages in single tracks from more sheet-like damages, where there are multiple crossing or parallel tracks, that have not been part of the image analysis in described in chapter 3.3.2.

Furthermore, prediction results are initially returned in raster format and can be transformed into a more lightweight and user-friendly format (i.e. vector lines and polygons). Because image analysis thus far has been limited to localizing wheel rut damages, further description of the generated spatial objects is applied during post-processing along with further data cleaning by means of utilizing filtering and transformation algorithms, application of ancillary data as well as data derived from drone imagery like Digital Surface Models (DSM) or vegetation indices (e.g. Normalized Difference Vegetation Index (NDVI)).

Finally, occurrence of wheel ruts and the density of such is a descriptive parameter in the classification system Nature in Norway (NiN), named “7TK[11] Spor etter ferdsel med tunge kjøretøy” in Norwegian. That parameter has a defined scale for the extent of damages to vegetation by wheel ruts.

The GIS workflow to be developed in this project consists therefore of the following four steps (see also **Figure 4**).

1. Masking of known, non-relevant structures like roads or streams that have some likelihood to be confused with wheel ruts during image analysis. Data used for this step of the post-processing routine are described in chapter 3.2.3.
2. Cleaning and noise removal by filtering detected objects of a size or geometry (length or length/boundary relationship)
3. Subdivision of resulting classification into homogenous objects (lines / areas) that can be further described in terms of:

[11] https://artsdatabanken.no/Pages/181998

   a. Measurable effect of wheel ruts on the soil, as depressions and channels detectable in the Digital surface model (DSM) derived by photogrammetry from overlapping images acquired by drones
   b. Measurable effect on the vegetation as reduction of values in the vegetation index derived from multispectral (here only drone) imagery
   c. Geometrical properties of the detected image objects like length and width

4. Aggregation of the filtered results into the scale defined by the 7TK parameter from NiN

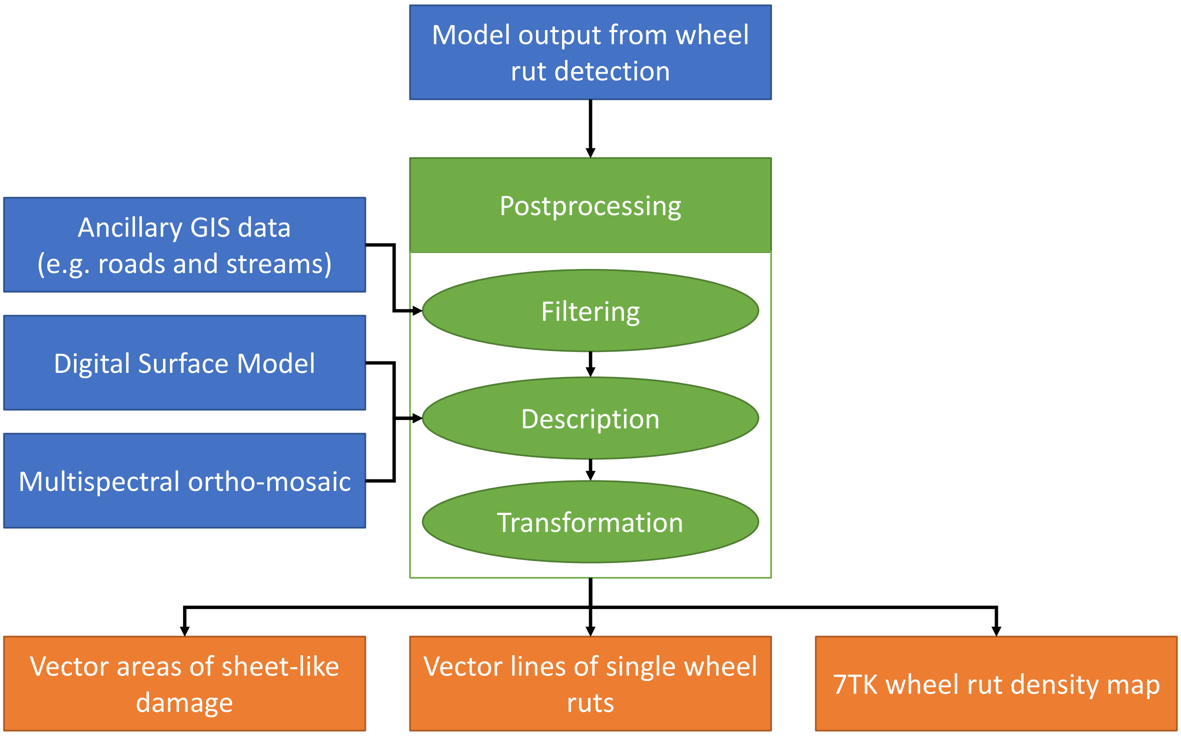

Model output from wheel rut detection
Postprocessing
Ancillary GIS data (e.g. roads and streams)
Digital Surface Model
Multispectral ortho-mosaic
Filtering
Description
Transformation
Vector areas of sheet-like damage
Vector lines of single wheel ruts
7TK wheel rut density map

***Figure 4**. Overview over general concept of the post-processing workflow*

Technical requirements for the workflow development have been to implement parallel, tiled processing in order to be able to a) utilize multiple cores for efficient processing or b) process the relatively large drone data also with limited resources. In addition, parameters used during processing, that may vary between data sources or detection results should be possible to be specified by the user. Default values for those settings, as well as hardcoded script internal settings, are defined based on tests in several subregions and development iterations.

GRASS GIS 8.0 is used as the main library for the post-processing of detection results because it provides the required functionality like conversion of data formats from raster to vector, filter algorithm, map calculator and so on and these tools are technically efficient to handle the significant amounts of data coming from drone-based sensors. Finally, GRASS GIS comes with a Python Application Programming Interface (API) that amongst others includes solutions for tiled, parallel processing.

3.3.5 Numerical and visual assessment of detection and processing results

To evaluate the models' ability to detect wheel rut damage from ATV we adopted a twofold approach consisting of both a numerical and a visual assessment of the quality of the automatic detection and post processing.

The numerical approach consisted of computation of the precision (P), recall (R), and F1-score. The precision score is a measure for the ability of the model to minimize the number of false positive samples, the recall score describes the ability of the model to predict all positive test samples correctly, while the F1 score represents the harmonic balance of the precision and recall score. Finally, as an overall accuracy measure, balanced accuracy is computed as the average of the F1-score of all classes (wheel rut damage (1) and background (0)).

While the numerical assessment describes the general performance of the model quite well, it is important to understand the spatial distribution and pattern of detection results, in order to identify and address possible issues of the produced models. That is especially true when, such as in this case, in the resulting models should be understood as a first prototype that can be improved in further development. The purpose of visual assessment of the classification accuracy is thus to identify cases where model predictions either fail or perform well. This can give indications about possible further corrections as well as uncertainties in the modeling results. Visual assessment is conducted as a qualitative, expert judgement where selected cases are highlighted.

4 Results and discussion

4.1 Re-processing of drone images

An important motivation to reprocess the available raw drone imagery has been to assess the information content of both multispectral drone imagery and photogrammetric digital terrain and surface models (DTM / DSM).

4.1.1 Multispectral imagery and vegetation indices

Results of the re-processing of the multispectral data show patchy patterns (see **Figure 5**). That is likely due to lack of calibration between images during processing in OpenDroneMap. However, relations of reflectance values between bands within the mosaics are mostly preserved during processing so that e.g. values of the Normalized Difference Vegetation index (NDVI) derived from those patchy images is much less affected by the patchiness of the mosaics, though not entirely free from it (see **Figure 5B**).

It is important to note that multispectral mosaics (left hand side in **Figure 5A**) in OpenDroneMap are returned with reflectance values between 0 and 1 instead of digital numbers in the classical spectrum of RGB images (0-255). Thus, images do not render like e.g. RGB-orthophotos in Geographic Information Systems out-of-the-box. That means they are not immediately suitable for visualization.

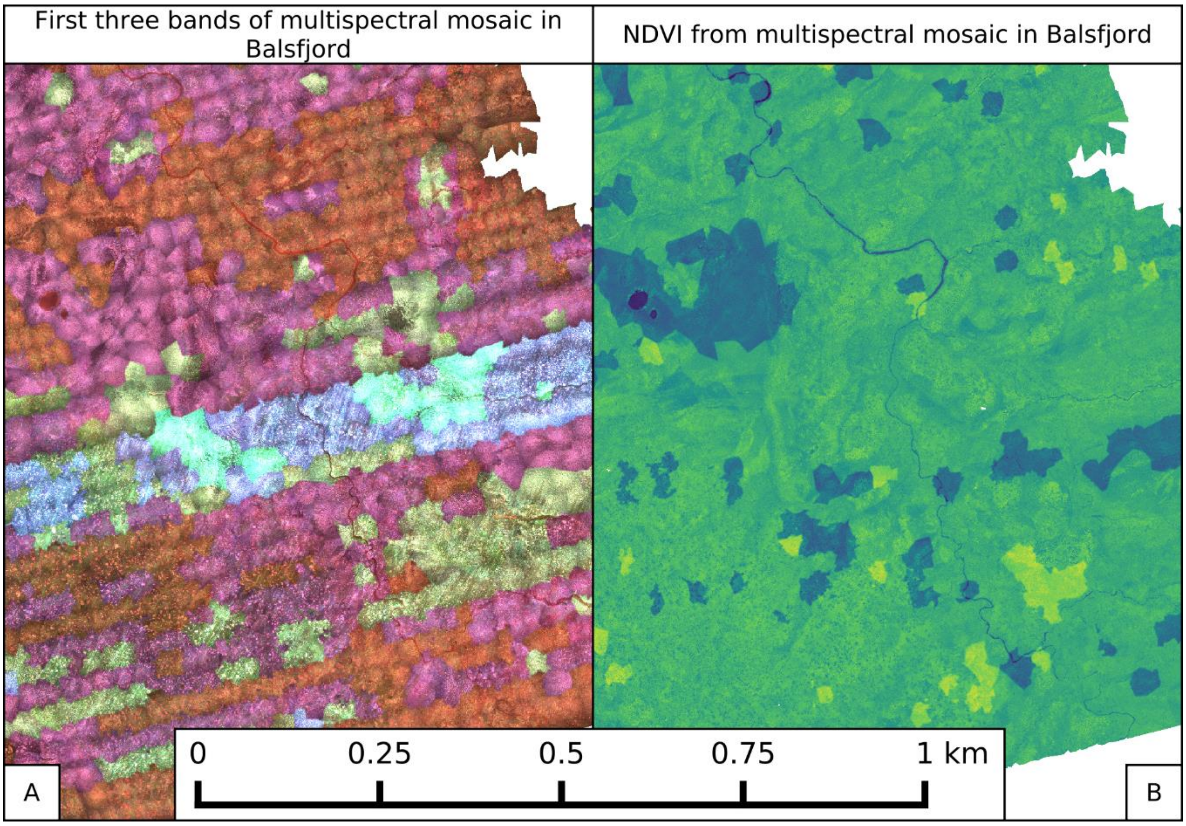

First three bands of multispectral mosaic in Balsfjord
NDVI from multispectral mosaic in Balsfjord
0
0.25
0.5
0.75
1 km
A
B

***Figure 5**. Multispectral mosaics and Normalized Difference Vegetation Index (NDVI) from P4M-drones processed with Open Drone Map 2.67*

Due to the lack of comparison with multispectral imagery from other drone models other or processing with other software, the issues visible in multispectral images cannot be pin-pointed to a single source of error with absolute certainty. However, calibration of camera parameters and intercalibration across single images is ongoing work in ODM for the drone model that has been used to acquire the available raw data in this project. Related software changes are about to become available[12]. Thus, increased quality of multispectral image products from ODM can be expected in future version, how significant the improvements will be remain to be seen. Still, the patchy characteristics of the output from the current algorithms in OpenDroneMap with the provided multispectral imagery clearly hamper the usability of these data in a wall-to-wall application where not only local (within pixel value relations and statistics) but also global (across pixel value relations and statistics) become important.

4.1.2 Photogrammetric terrain and surface models

Thanks to the overlap between images from systematic drone survey campaigns it is possible to generate very high-resolution terrain models from those data using photogrammetry. For both study areas digital surface and terrain models were produced with data from both the multispectral and RGB sensors.

Resulting photogrammetric terrain and surface models from the multispectral raw data (see **Figure 6A**) show an overall lower quality, compared to such models derived from the higher resolution RGB imagery (see **Figure 6B**). Tracks from ATVs become visible in significantly more detail from the latter data source. This indicates a trade-off between using higher-quality terrain information (feature matching from high resolution images) and vegetation information (infrared and red edge bands from multispectral imagery).

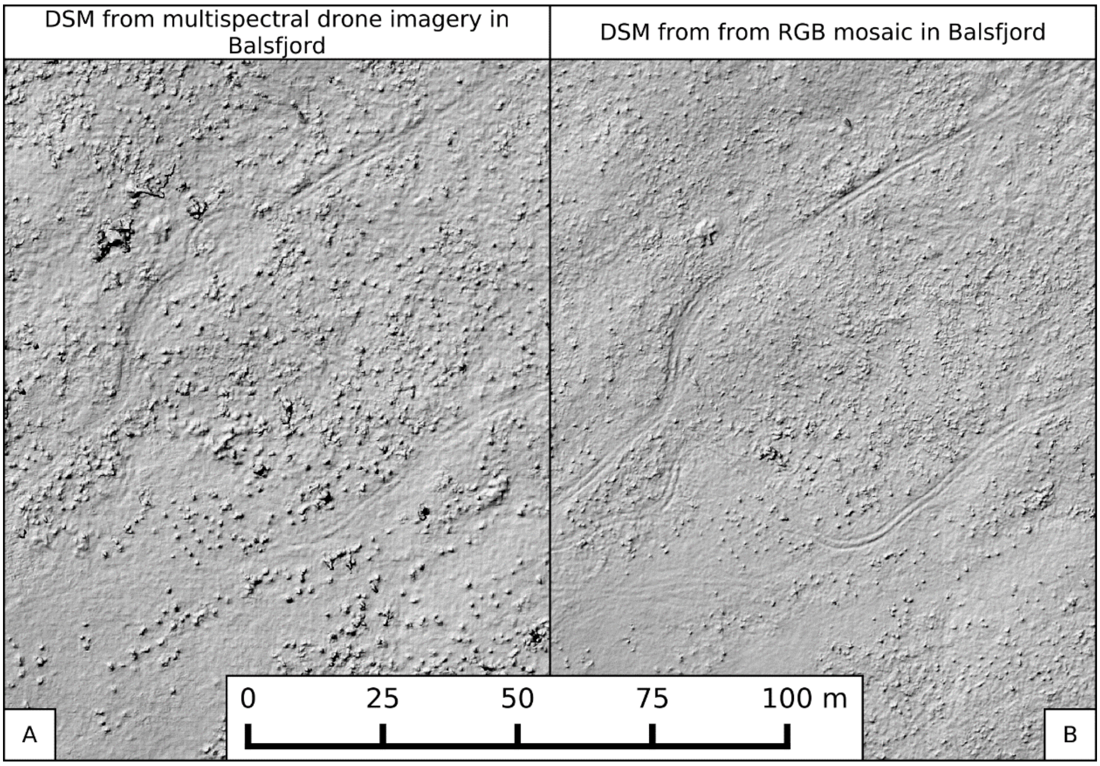

DSM from multispectral drone imagery in Balsfjord
DSM from from RGB mosaic in Balsfjord
0
25
50
75
100 m
A
B

***Figure 6**. Comparison hillshade representation of Digital Surface Models (DSM) from multispectral (P4M drone, left) and RGB imagery (P4P drone, right) processed with Open Drone Map 2.67*

[12] https://github.com/OpenDroneMap/ODM/pull/1392

Both datasets contained a notable number of artifacts and areas with corrupted height information, likely due to a lack of overlap between single images and/or problems for the software to identify common features in image pairs within more homogenous vegetation. This underpins the requirement for adequate flight planning for capturing imagery that can also produce high quality terrain information.

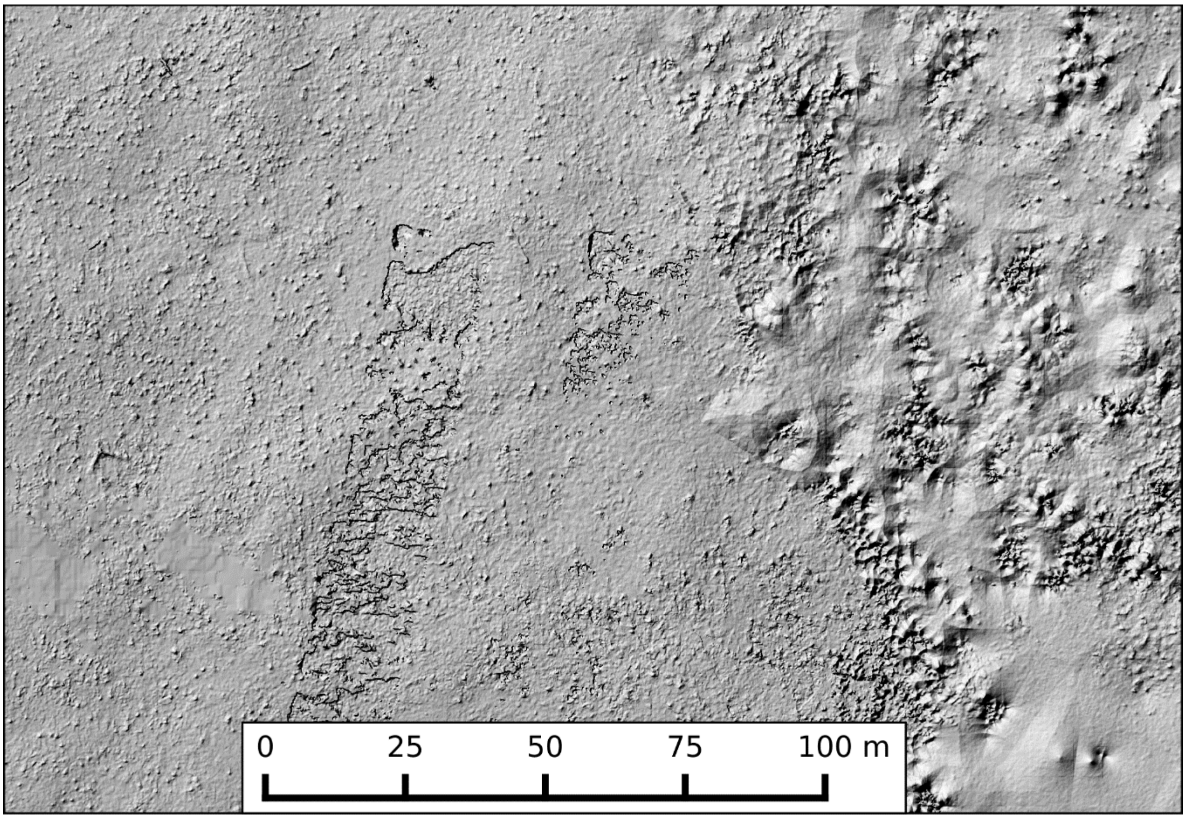


Figure 7*. Example of artefacts and corrupted height information in the DSM created from high resolution, RGB drone imagery in Balsfjord*

4.1.3 Monitoring severity of surface damage over time

For photogrammetric terrain and surface models from drone imagery to be useful for monitoring the depth of the same trails over multiple years, there need to be high repeatability in the creation of DTMs to begin with. Among the available dataset for both study areas there was a smaller patch that was flown on two consecutive days. Using this area, comparability of terrain models from two consecutive days were assessed to identify concerns one should have in mind when designing future monitoring.

Ideally the datasets from both days should be almost identical. However, a difference map of the generated 2.5D models from the multispectral drone images of the Balsfjord study area, which had the lowest GPS error, shows significant mismatch between data generated from the two independent days. The average difference in Z-direction was ~3m and standard deviation of differences was ~0.7m. That is a multiple of the depth of wheel ruts that was measured manually and in the post-processing routine of up to ~ 0.2 - 0.3m (see also chapter 4.4.4). Error in that dimension renders a direct, pixel-wise comparison of height values in detected wheel ruts using photogrammetric terrain models ruts inappropriate.

The main reason for the significant overall differences between the two independent terrain models are that there is both a notable offset in x- and y- direction as well as a general tilt of the two datasets towards each other in the horizontal plane (see **Figure 8**). These issues could be circumvented to some degree by means of using reference points, measured with high-precision GPS. Such measurements however increase the time required for data acquisition.

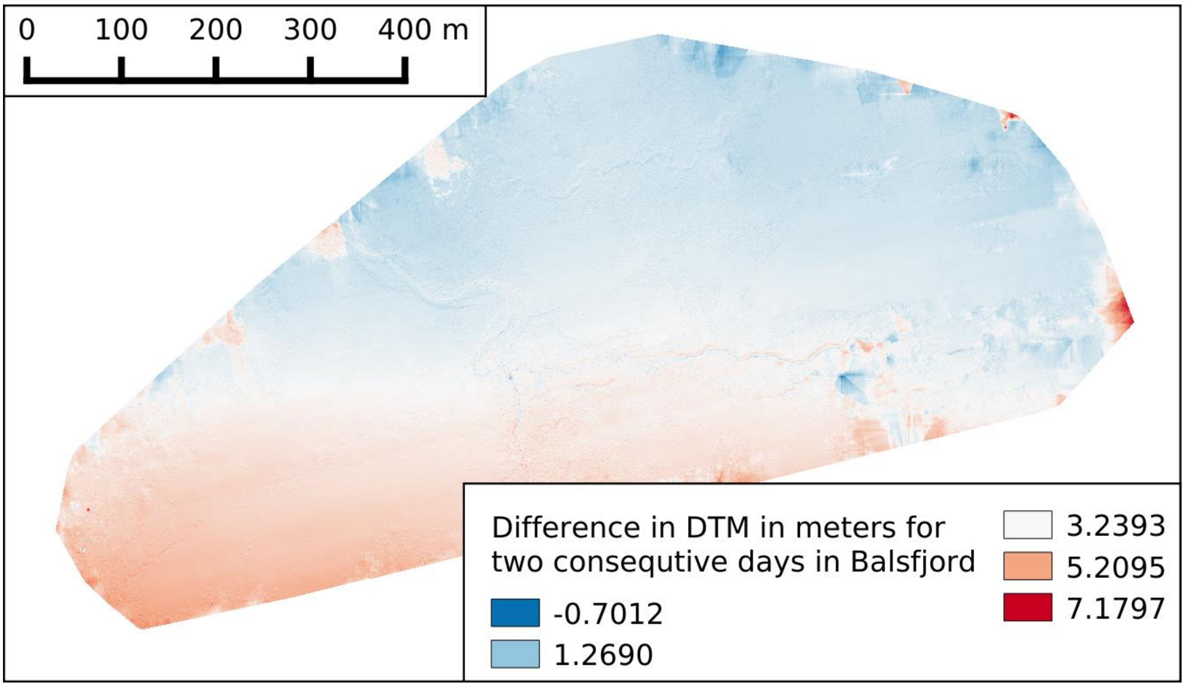

0
100
200
300
400 m
Difference in DTM in meters for two consequtive days in Balsfjord
-0.7012
1.2690
3.2393
5.2095
7.1797

Figure 8. *A pseudo-color representation of the difference between the DTM from Balsfjord day 1 and day 2. Both missions were flown with Phantom 4 Multispectral and have a ODM calculated GPS error of 0.48 and 0.75 meters.*

The dataset collected by the multispectral drone (P4M) is corrected with RTK GPS. In theory, this should achieve an absolute spatial accuracy around 0.1 meters or lower[13]. The most likely reason for the GPS error to be higher in this case is insufficient image overlap in some parts of the image, which usually drives up the overall error slightly.

However, retaining equally high quality of the data for such large areas is in any case challenging as there are many factors to consider, flight time being the main one. Even a GPS error of just 0.1 meters is not much smaller than the width of an ATV wheel. Thus, comparing track depth on a pixel-by-pixel basis with the available data does not seem a realistic endeavour at the given scale in general. So, findings from Ancin-Murguzur et al. (2020) that drones can be a very valuable tool for monitoring e.g. soil track erosion needs to be understood as highly scale dependent and mostly feasible for very detailed drone campaigns at smaller extents.

The monitoring of depth from year to year is in other words not a straightforward process. If reasonably possible at all at the given scale here, more effort would have to be put into the more exact spatial matching of terrain models from two points in time.

Within a single day drone imagery is clearly able to capture impact of off-road driving ATVs on soil and terrain (see **Figure 6**). Given the artefacts visible in the photogrammetric terrain models of the available data, an obvious challenge is to capture images in a way that allows to create terrain models of homogenous and high quality.

[13] It should also be noted that we have only looked at the reported GPS errors from OpenDroneMap, and not used the ideal method with points collected in the field to compare against.

4.1.4 Resources consumption for re-processing of drone images

Resource consumption for handling drone images is an important practical consideration for a wider application of a monitoring approach. Processing the large amounts of drone imagery is quite resource demanding. Processing times for the 8 given datasets (2 days from 2 sensors in 2 study sites) varied between a few hours of processing for smaller areas of multispectral imagery to several days for the larger areas with high resolution RGB images. And even on powerful server hardware datasets need to be split into chunks in order to be processable with the software library applied here (OpenDroneMap). When it comes to both peak and average resource consumption, processing the high-resolution RGB-imagery in OpenDroneMap is significantly more resource demanding than processing of the multispectral imagery. Although multispectral imagery is significantly faster to process, it is much less efficient with regards to storage consumption and spatial coverage. The available multispectral datasets cover less area while occupying 5 to 8 times more storage than the respective RGB imagery.

4.2 Detection of wheel ruts in drone images and aerial photos with deep learning

For the drone dataset the model achieved best performance at epoch 63 (IoU: 91.6% for the background and 30.9% for the wheel ruts), while for the aerial image dataset the IoU was best at epoch 726 (IoU 97.9% for the background and 26.6% for the wheel ruts).

4.2.1 Numerical evaluation of detection quality

The validation of the raw predictions from the trained DeepLABV3 model revealed a good balanced overall accuracy for all models with scores between 66 and 82 (see **Table 4**) The background class was mostly predicted correctly as seen by the large precision, recall, and F1 scores.

Table 4. *Summary results (prior to post-processing of the predictions) of the different combinations of study area and data source in terms of overall accuracy (OA = average of the F1-score of all classes (wheel rut damage (1) and background (0)), precision (P = true positives / (true positives + false positives)) representing the ability of the model avoid false positive detections, recall (R = true positives / (true positives + false negatives)) representing the ability of the model to detect all wheel rut pixels, F1 score (F1 = 2 * (precision * recall) / (precision + recall)) for the two classes object of study. In addition, specific information on the type of commission errors is reported split into roads and hiking trails which were wrongly classified as damaged by wheel ruts.*

			Background			Wheel rut damage			Other commission errors	
Area	**Data source**	**OA**	**P**	**R**	**F1**	**P**	**R**	**F1**	**road**	**Trail**
Balsfjord	**drone**	71.7	97.7	94.1	95.8	38.8	62.0	47.7	68	47.2
	aerial images	66.7	99.1	98.3	98.7	29.4	42.7	34.8	17.6	15.3
Rjukan	**drone**	66.4	97.4	90.0	93.6	28.7	63.0	39.4	67.74	38.75
	aerial images	82.2	99.8	99.1	99.5	52.6	85.3	65.0	49.4	12.7

On the other hand, the prediction of the wheel rut damage class achieved in general relatively low precision scores between 28 and 52, and moderate to good recall scores between 42 and 85. In other words, the models detect wheel ruts at an acceptable rate but also contain a number of false positive detections.

Even though the aerial and drone datasets should be considered as separate datasets as they were collected at different times, with different number of damages (see **Table 4**), the models trained on aerial images showed fewer false positive detections. This was mainly due to the presence of many artifacts in the drone predictions due to the use of relatively small tiles. While this was not possible to solve within the limited time for this specific project, future developments of this methods should aim at developing ways to reduce such edge effects for example by down sampling the drone imagery and using larger tiles as for the aerial images. Further improvement could be achieved by implementing an overlap between neighboring tiles, which could be removed in post-processing and thus reduce the edge artifacts.

Correct detection of wheel ruts shows for both drone and aerial images a moderate to good quality (scores of 30 to 60), with a slightly better performance of drone imagery. The opposite is true in the Rjukan case. However, there were only very few validation data points available for aerial images, making the validation results for this case less reliable. The validation for the wheel rut detection from drone imagery in Rjukan on the other hand mostly confirms the results from the Balsfjord case.

While the road and trails annotations were not used for training the model, we used this information to determine the amount of predicted damage pixels which belonged to roads or trails. Such additional analysis revealed that both models were classifying roads and hiking trails as wheel rut damage to different degrees (see **Table 4**). This was likely caused by the fact that these two classes were not specified in the model and possible ways to overcome such issue are either to train models that include also these classes, or remove these predictions in post-processing based on existing geospatial databases (see chapter 4.4).

4.2.2 Visual assessment of detection results

In the figures below (**Figure 9** to **Figure 16**) it is possible to observe the final output for the predictions from the model for the drone and aerial images and for Rjukan and Balsfjord, compared to the annotated reference data. As visible by the comparison of the predictions for the drone model (**Figure 12** and **Figure 16**) and the aerial images model (**Figure 10** and **Figure 14**), the former was characterized by the presence of artifacts (i.e. commission errors) at the edge of the tiles used for prediction. Note for examples the almost linear occurrence of such artefacts in the eastern part of the Rjukan study site. Predictions from drone imagery are furthermore negatively affected by false positive detections that were relatively evenly distributed across the entire study area and seem to follow the pattern of the tiles used during prediction (see also (**Figure 12** and **Figure 16** for detail).

Some of loss of accuracy is likely caused by the annotation where reasons for misclassifications can be that i) annotated tracks were not clearly visible (i.e. prolongation of existing and visible tracks), ii) annotated lines were not exactly in the center of the linear features, iii) uncertainty of the class the annotated features belong to (e.g. transition zone between different features). These issues are likely to have played a negative role on the models' quality.

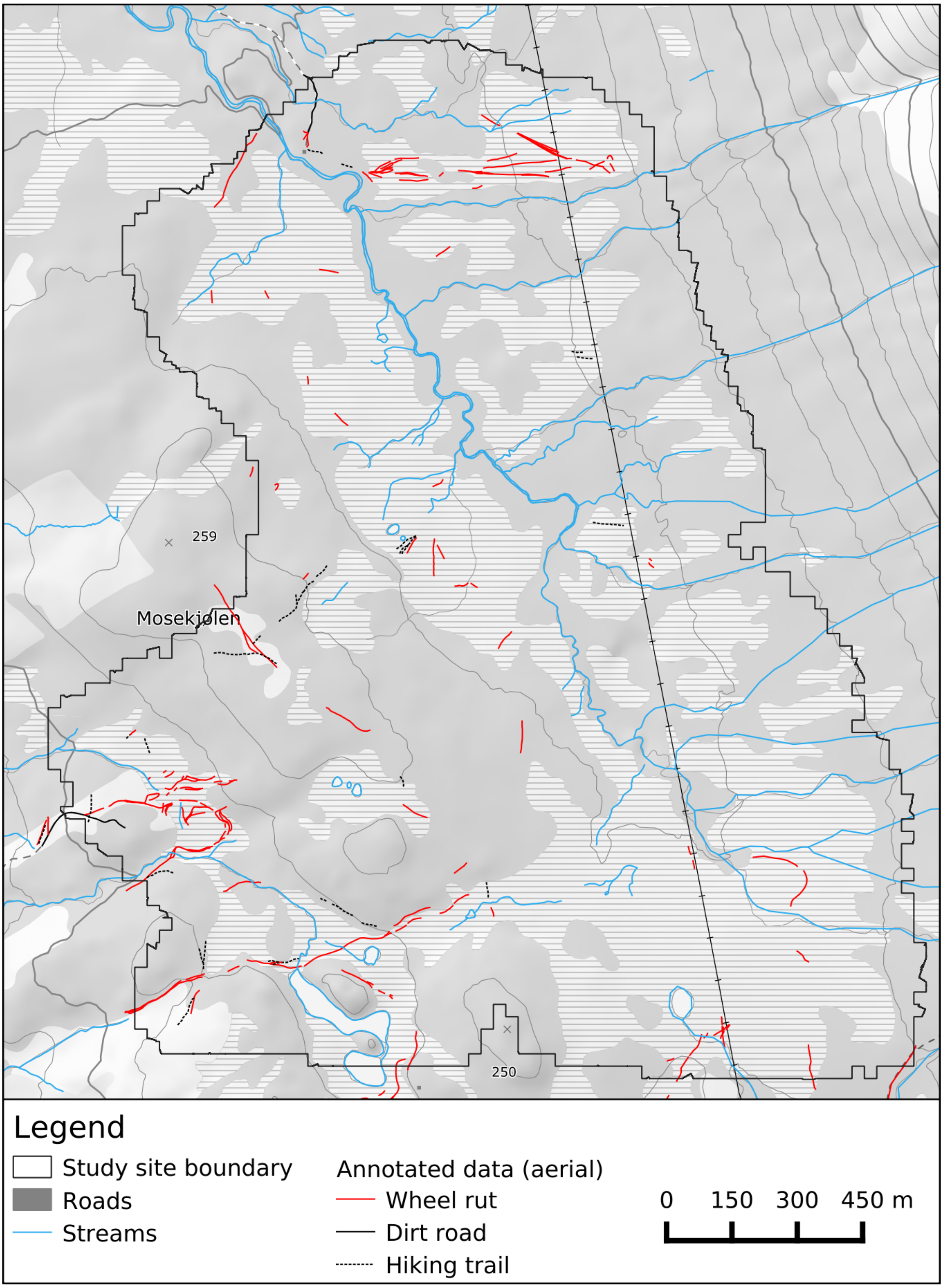


***Figure 9**. Annotated reference data for aerial imagery at the study site in Balsfjord.*

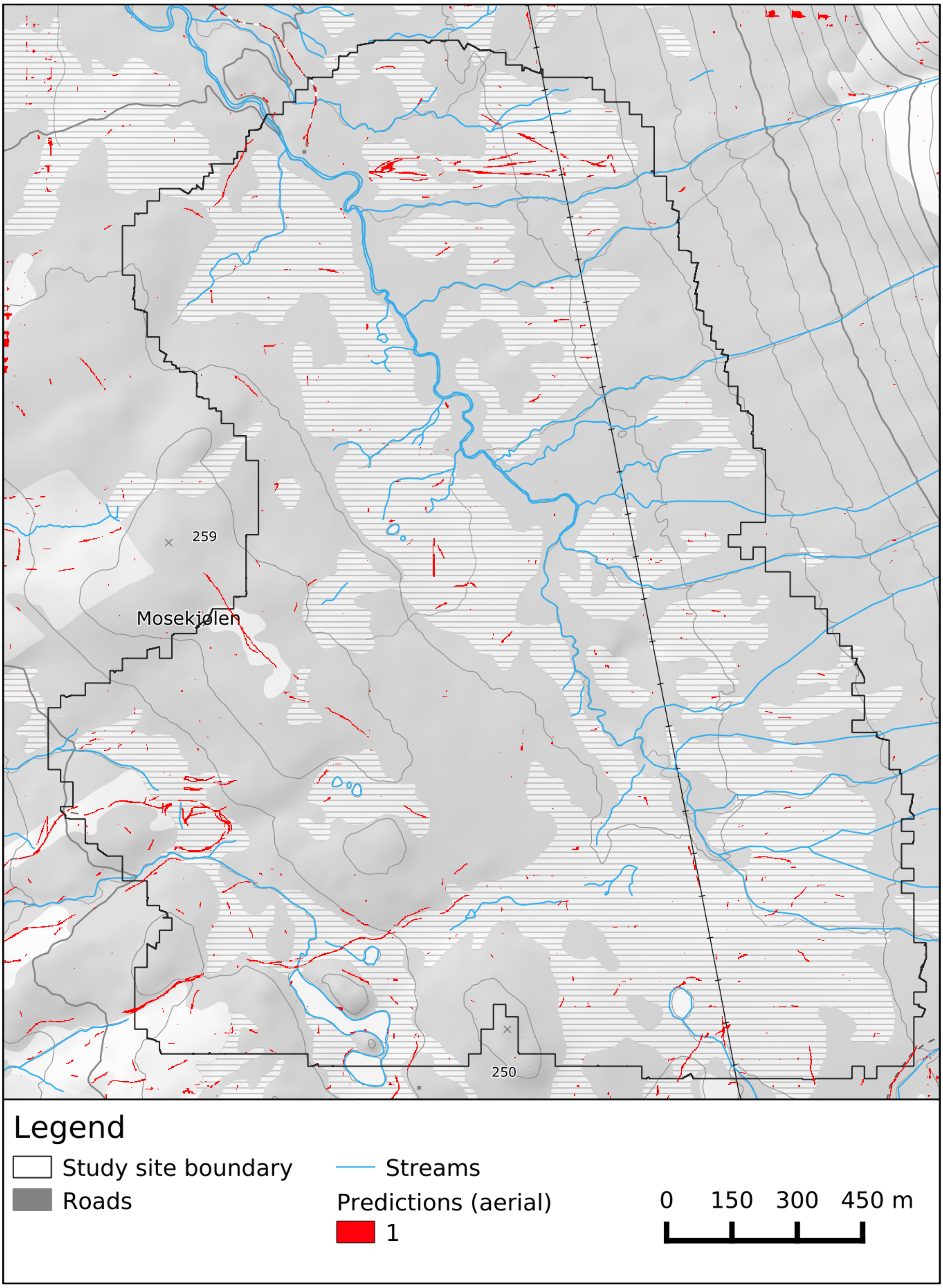


***Figure 10**. Occurrences of wheel ruts predicted from aerial imagery at the study site in Balsfjord.*

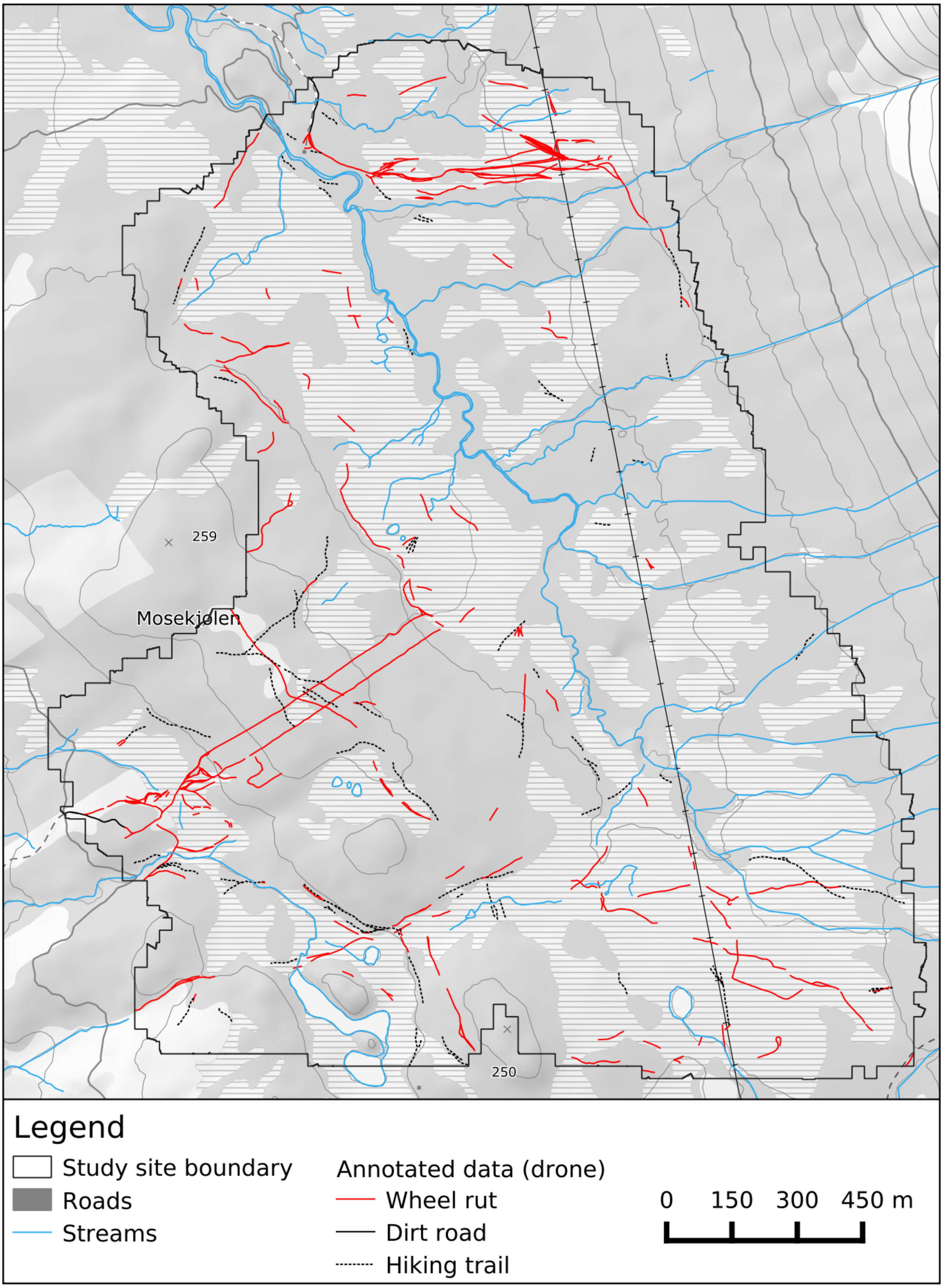


***Figure 11**. Annotated reference data for drone imagery at the study site in Balsfjord.*

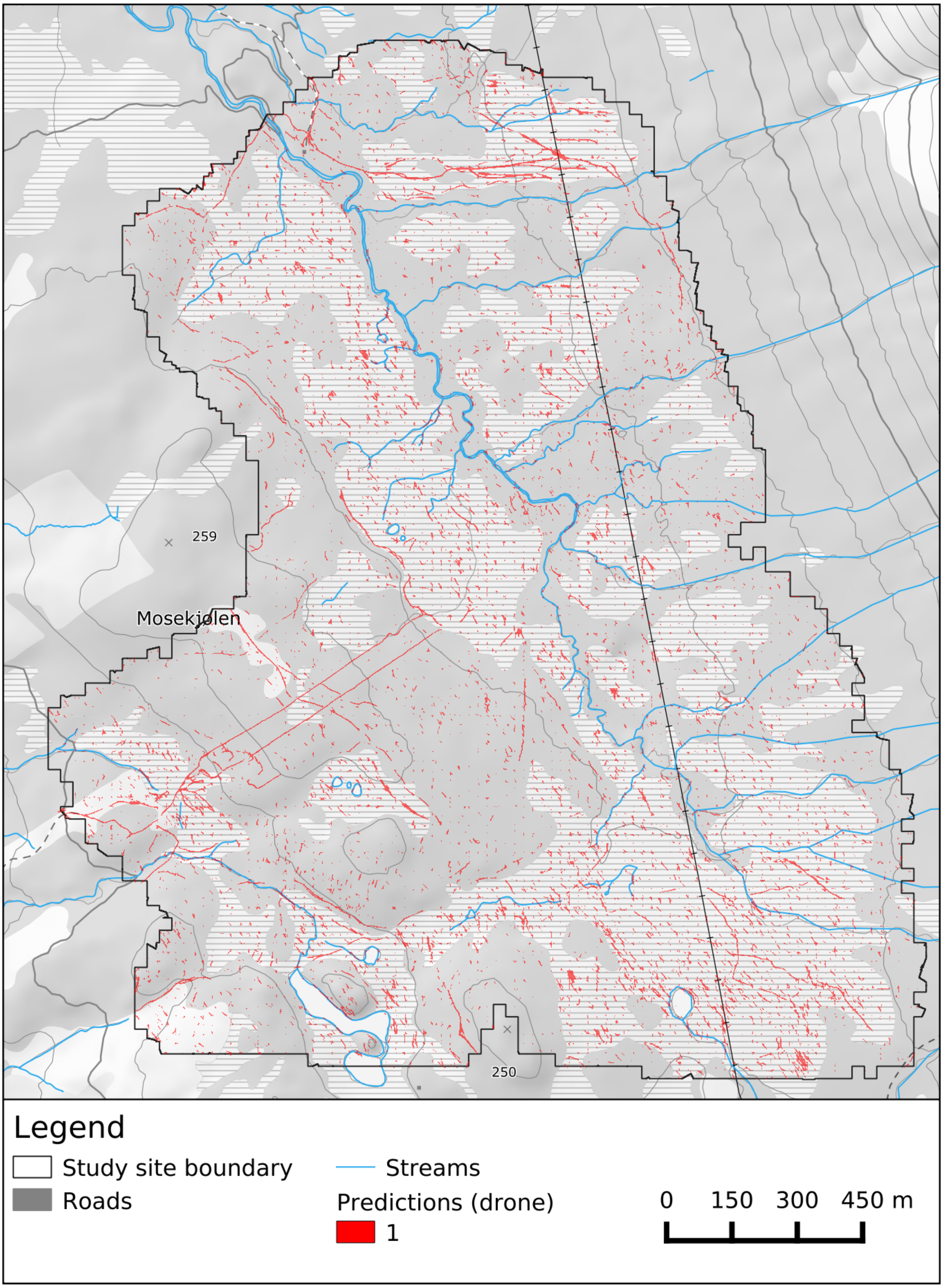


Figure 12. *Occurrences of wheel ruts predicted from drone imagery at the study site in Balsfjord.*

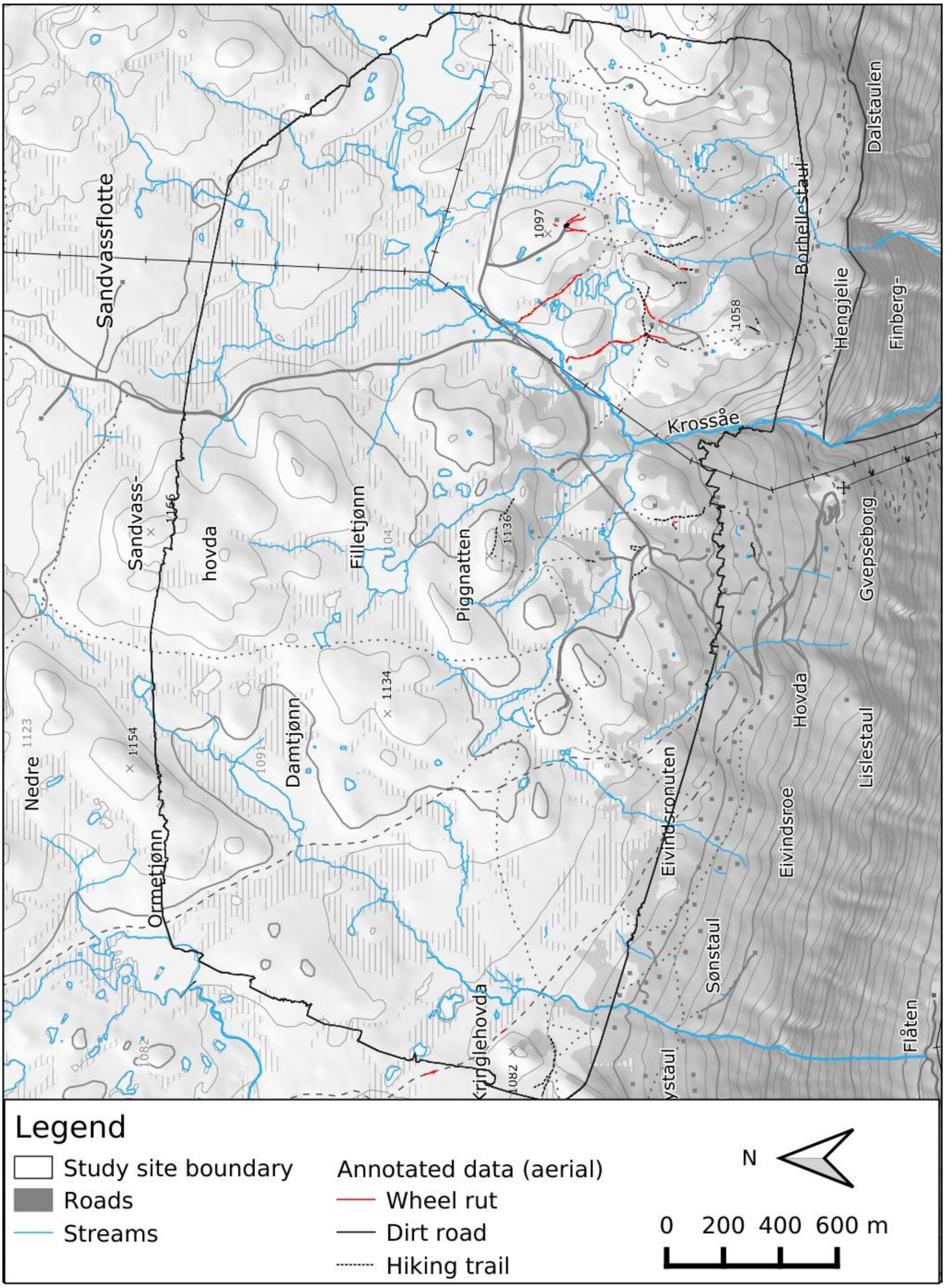


Figure 13. *Annotated reference data for aerial imagery at the study site in Rjukan.*

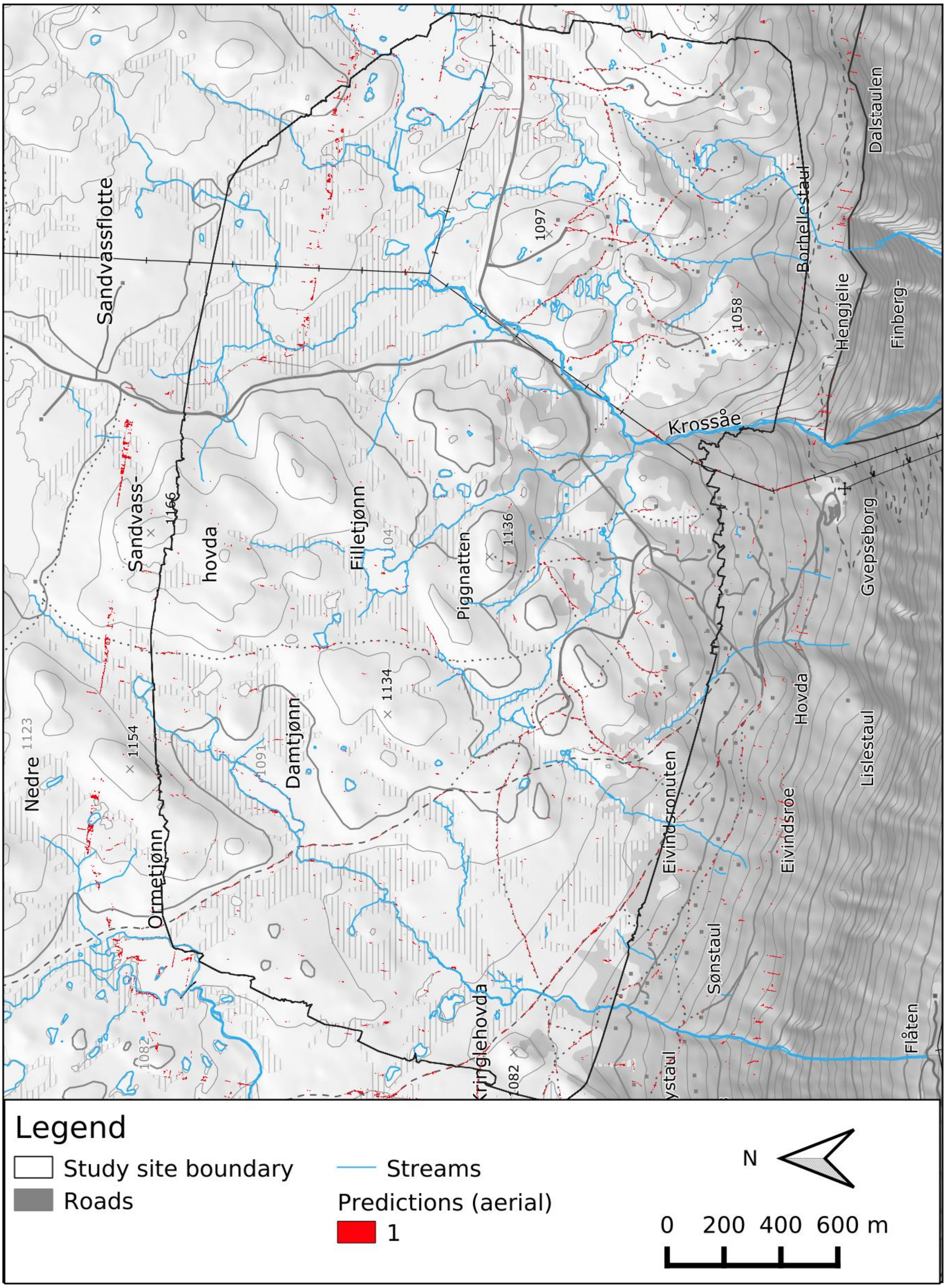


Figure 14. *Occurrences of wheel ruts predicted from aerial imagery at the study site in Rjukan.*

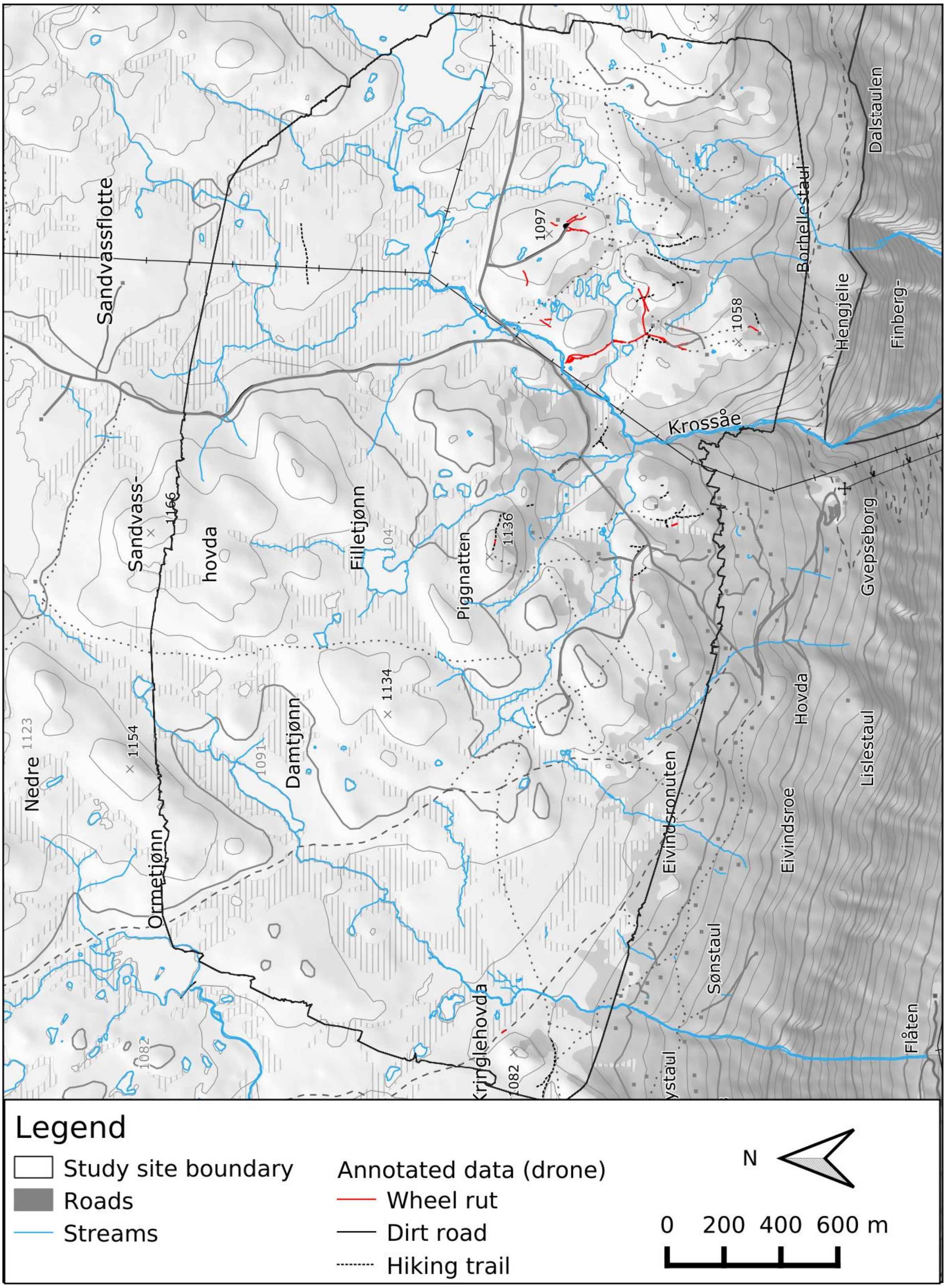


***Figure 15**. Annotated reference data for drone imagery at the study site in Rjukan.*

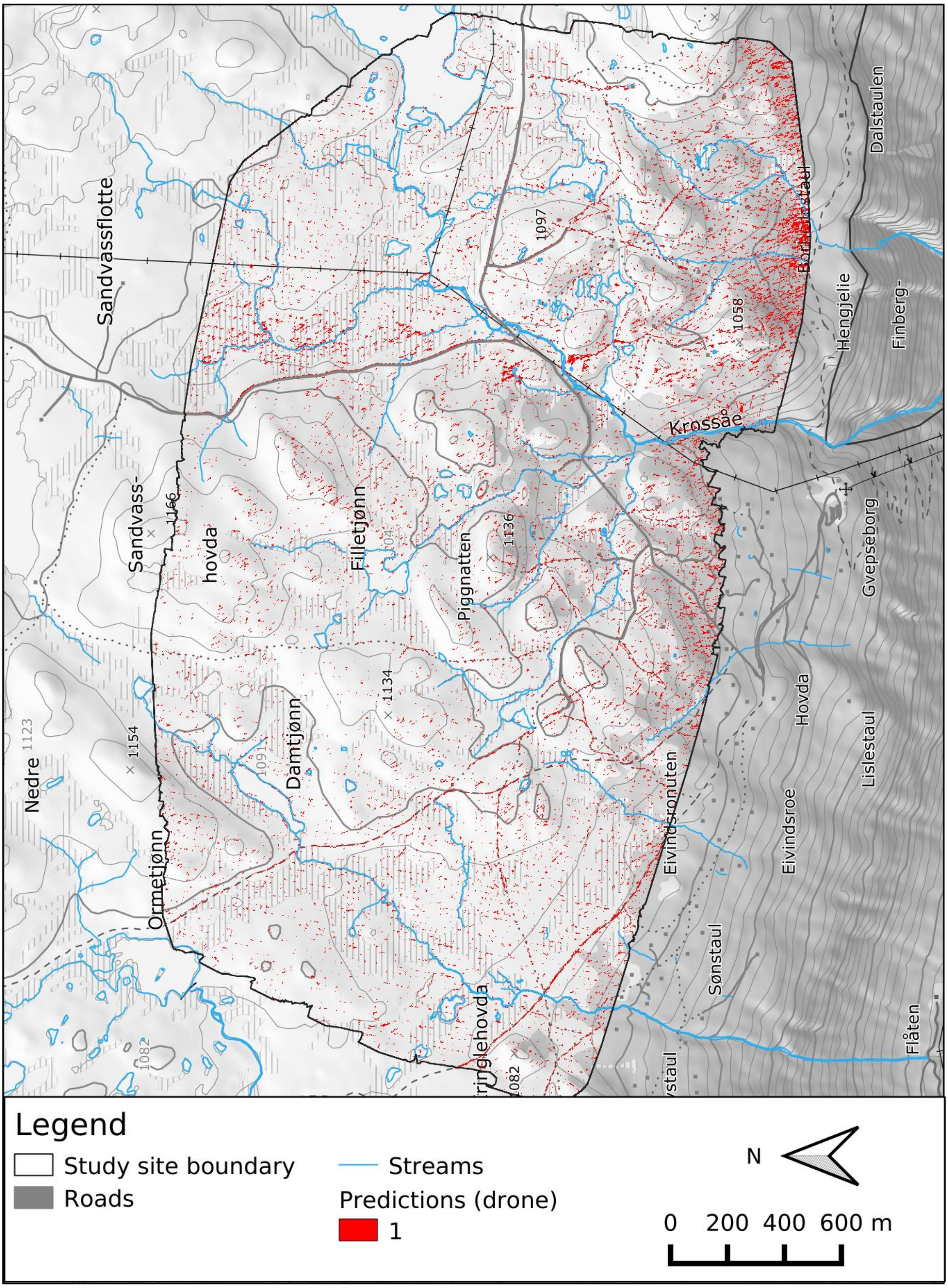


Figure 16. *Occurrences of wheel ruts predicted from drone imagery at the study site in Rjukan.*

In addition to the overall impression, examples can be highlighted from the visual assessment of the modelling results that represent important characteristics of the quality of the output of the developed models.

As an indication for how well the deep learning model can detect wheel ruts in good cases is shown in **Figure 17**. Here, the model is able to correctly detect even a relatively hardly visible wheel rut in a wetland area that may be a remnant of older driving activities.

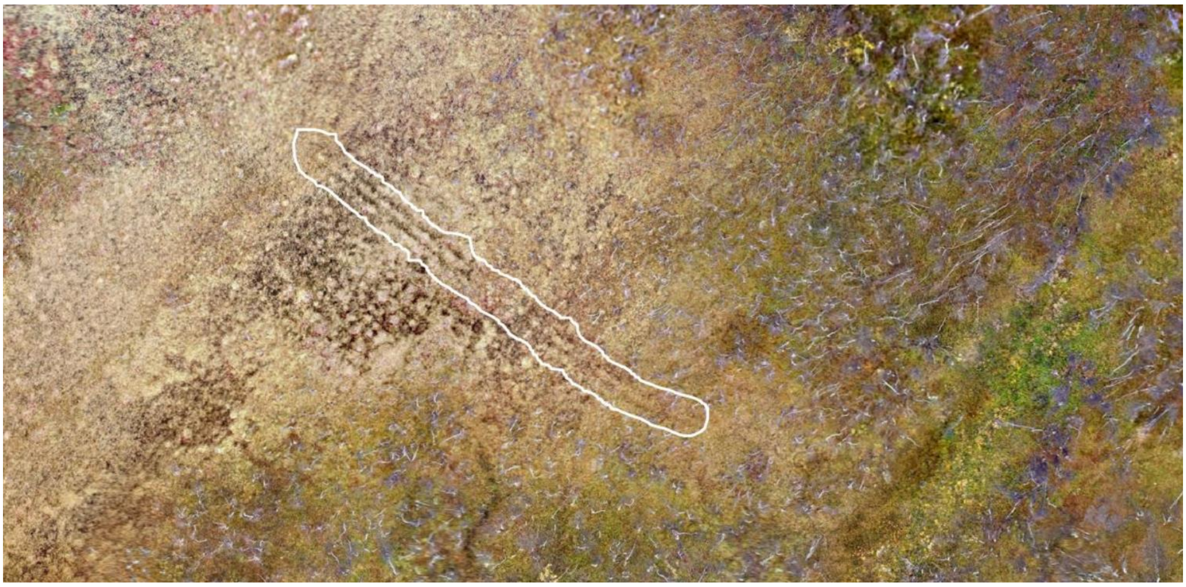


***Figure 17**. Correct detection of a relatively hardly visible wheel rut*

The developed models capture also areas with extensive off-road driving activity quite well (see **Figure 18**). Results from models trained on aerial images seems to be in fact most reliable in exactly these kinds of areas, that can be easily spotted in an overview of the model outputs.

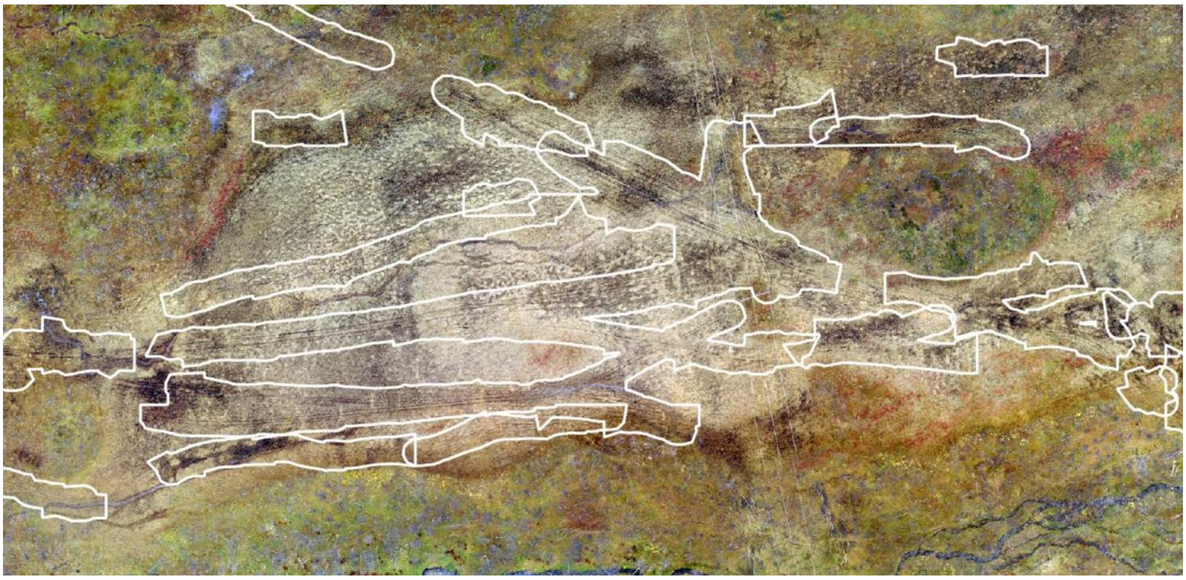


***Figure 18**. Illustration of good detection rates in areas with extensive off-road driving activity*

False positives detection occur mainly in in areas with shadows from vegetation or a more textured vegetation pattern (see **Figure 19**). In the upper image of **Figure 19** also the pattern from image tiles used for training and predictions is notable.

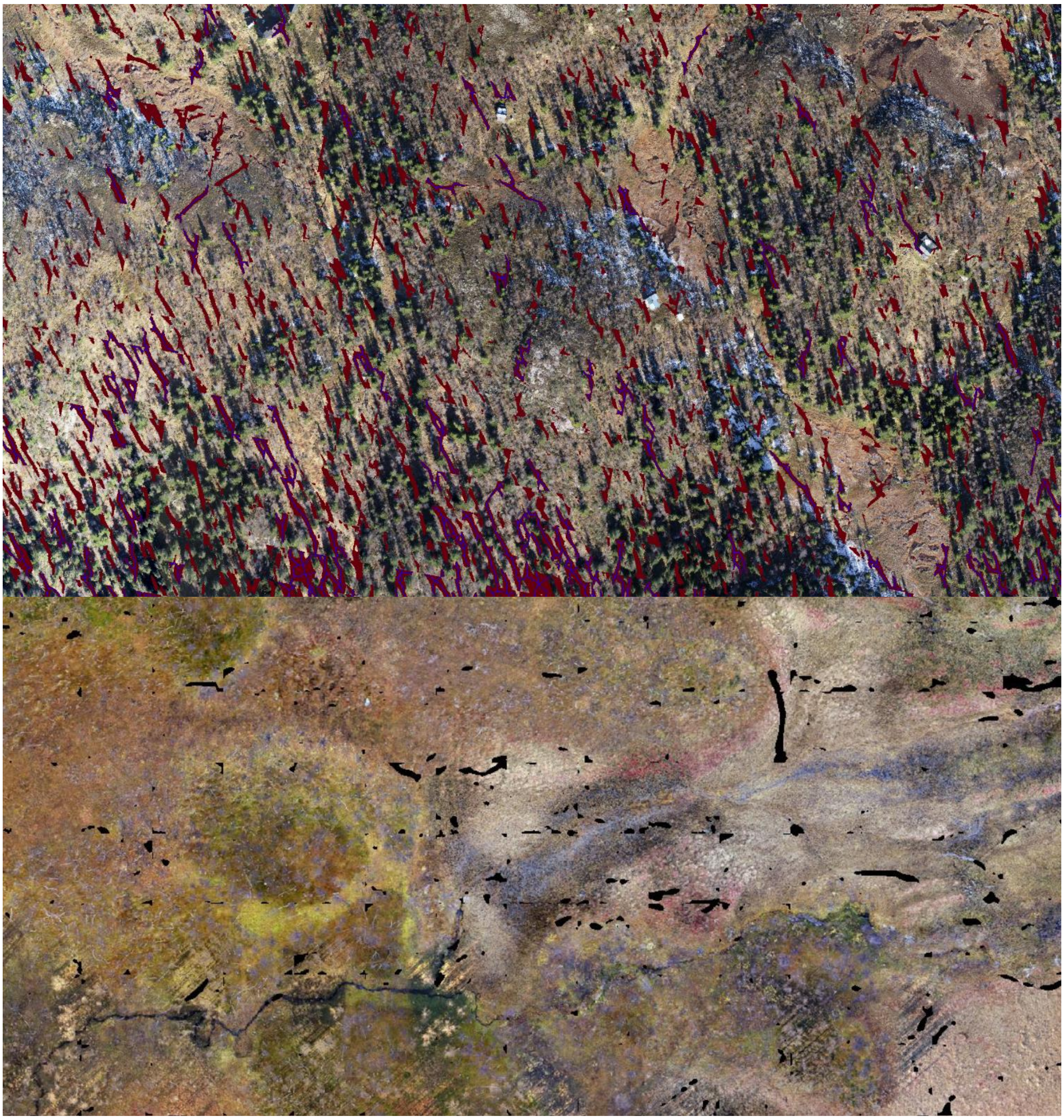


***Figure 19**. Scattered false positive detections in drone imagery in areas with shadows from vegetation in forested areas (top) or a more textured vegetation pattern in wetlands (bottom)*

Other areas where misclassifications notably occur are along water bodies and small streams as well as along trails and or gravel roads (see **Figure 20**). These are kinds of errors that to some extent can and will be addressed during post-processing.

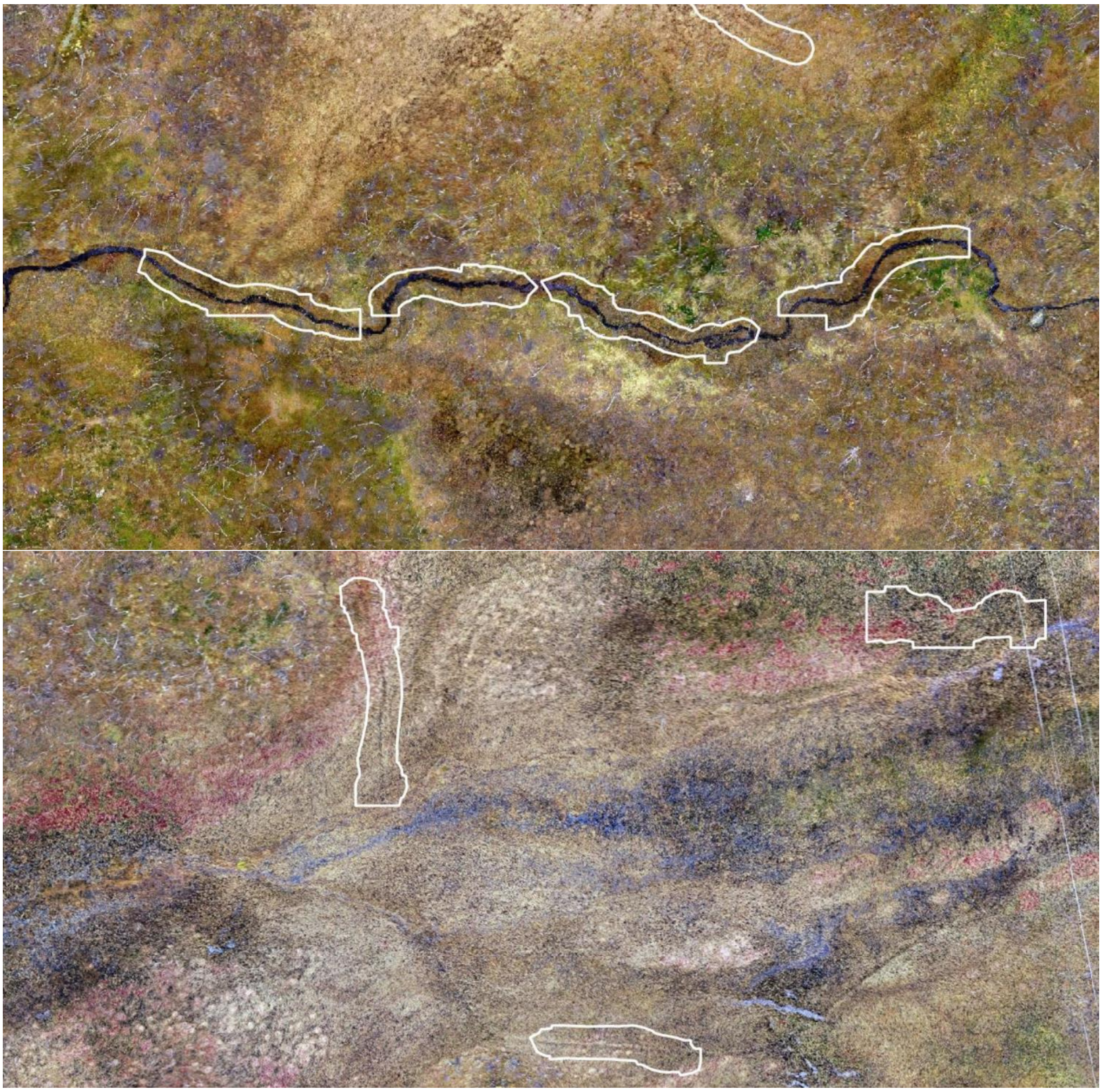


Figure 20*. Examples of false detections along streams and hiking trails in the raw predictions that can – partly – be addressed through post-processing (see chapter 4.4)*

In contrast, areas where wheel rut detection works quite well even when using lower resolution aerial images are especially wetland areas with a rather homogenous, smooth vegetation structure (see **Figure 21**).

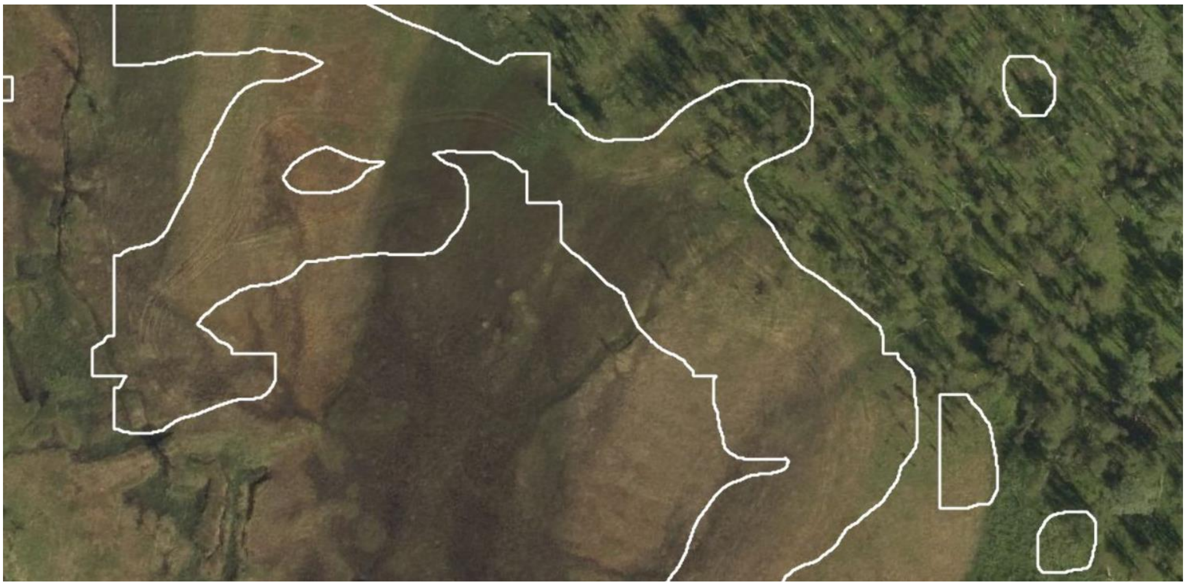


Figure 21. *Example of areas where wheel rut detection works quite well even in aerial images*

4.3 Edge detection, object-based image analysis for detection of wheel ruts in drone images and aerial photos

As mentioned already in chapter 3.3.3, it became evident early in the project from comparison of image segmentation and deep learning modeling results, that the deep learning approach generally performs better for the case of wheel rut detection. Deep learning appears especially advantageous in situations where wheel rut damages are only minor or initial with less clear contrasts to the surrounding vegetation (see **Figure 22**). Image objects and edge detection struggle with picking up boundaries of the wheel ruts, and even with small segment sizes, segment boundaries spill out onto the areas surrounding the actual wheel ruts (see **Figure 22** (C)). This issue will subsequently lead to less clear statistical differentiation and consequently model accuracy.

These findings underpin the work by Guirado et al. 2017, who found that deep learning performed better compared to OBIA approaches when detecting individual shrubs. The fact that other advantages of deep learning are that “it required less human supervision than OBIA, can be trained using a relatively small number of samples, and can be easily transferable to other regions or scenes with different characteristics, e.g., colour, extent, light, background, or size and shape of the target objects” (Guirado et al. 2017) lead to the decision that the OBIA approach was not further pursued in the project and efforts rather focused on post-processing (see chapter 4.4).

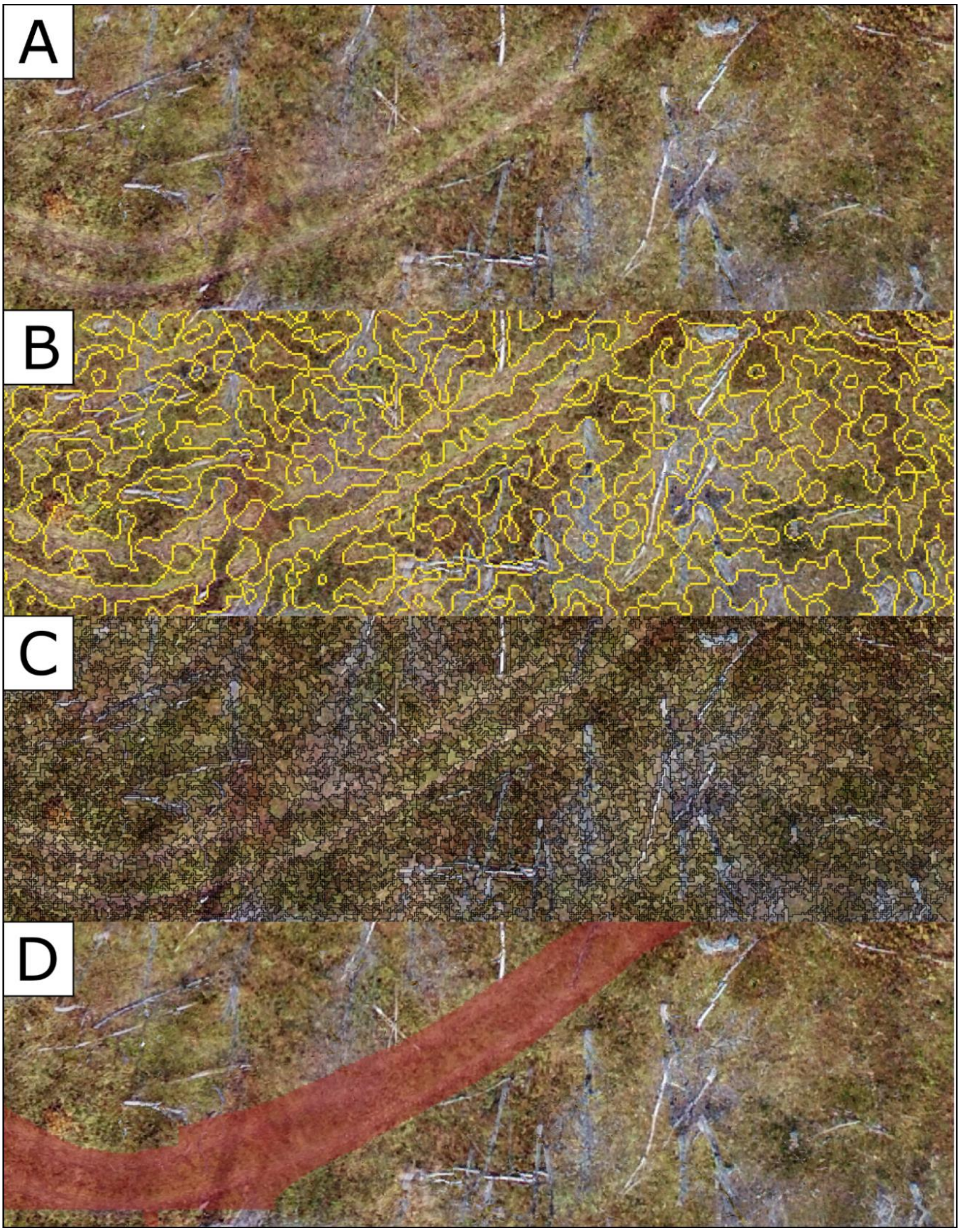


***Figure 22**. Comparison of edge detection with a zero-crossing algorithm (B), image segmentation with mean shift algorithm and settings that produce small but internally more homogenic segments (C), and results from wheel rut detection with deep learning (D)*

4.4 GIS-workflow for post-processing of detection results

The developed GIS-workflow for post-processing of wheel rut detection results aims mainly at extracting the following information:

1) vector line features for single, linear tracs from ATVs
2) vector polygons for areas with sheet-like damages from parallel or crossing tracs from ATVs
3) trac density according to Nature in Norway (NiN) 7TK

Another important aim of the post-processing is to clean artefacts and similar noise in the results is removed from the resulting datasets. A benefit of the vector-based output from post-processing is simplified manual inspection and correction if needed.

4.4.1 Methodological components of the post processing

Most of the elements of the post-processing represent rather standard GIS operations like object extraction, line thinning, map calculator operations, distance measurements, or data type conversions. Filtering of artifacts is then mainly done using size, width and length estimates. The implemented algorithm for assessing the impact of wheel ruts on the soil and vegetation however warrants some more methodological description.

Damage of wheel ruts to vegetation and soil will be digitally visible as linear depressions in vegetation indices (here NDVI) and terrain models respectively with values lower than surrounding pixels. In order to assess the “depth” to which damages occur, first such depressions are identified. For this a geomorphometric algorithm (“r.geomorphon”) is applied that classifies pixels into different types of morphometric structures like channels, pits or ridges. The wheel ruts themselves appear as channels or pits in the resulting maps. Then depth of those channels and pits is measured by subtracting their pixel values from interpolated values of the surrounding cells. Finally, within the identified wheel rut objects the depth is assigned to the vector representation of the mapped wheel ruts as attributes, together with the number of depressed pixels. Assignment of attribute values is done using neighborhood statistics for line features and univariate statistics for areas.

Resulting attribute valued are classified into 4 classes, where the value of

0 refers to zero detectable damage or missing data,
1 refers to minor soil or vegetation damage,
2 refers to medium soil or vegetation damage and
3 refers to most severe soil or vegetation damage.

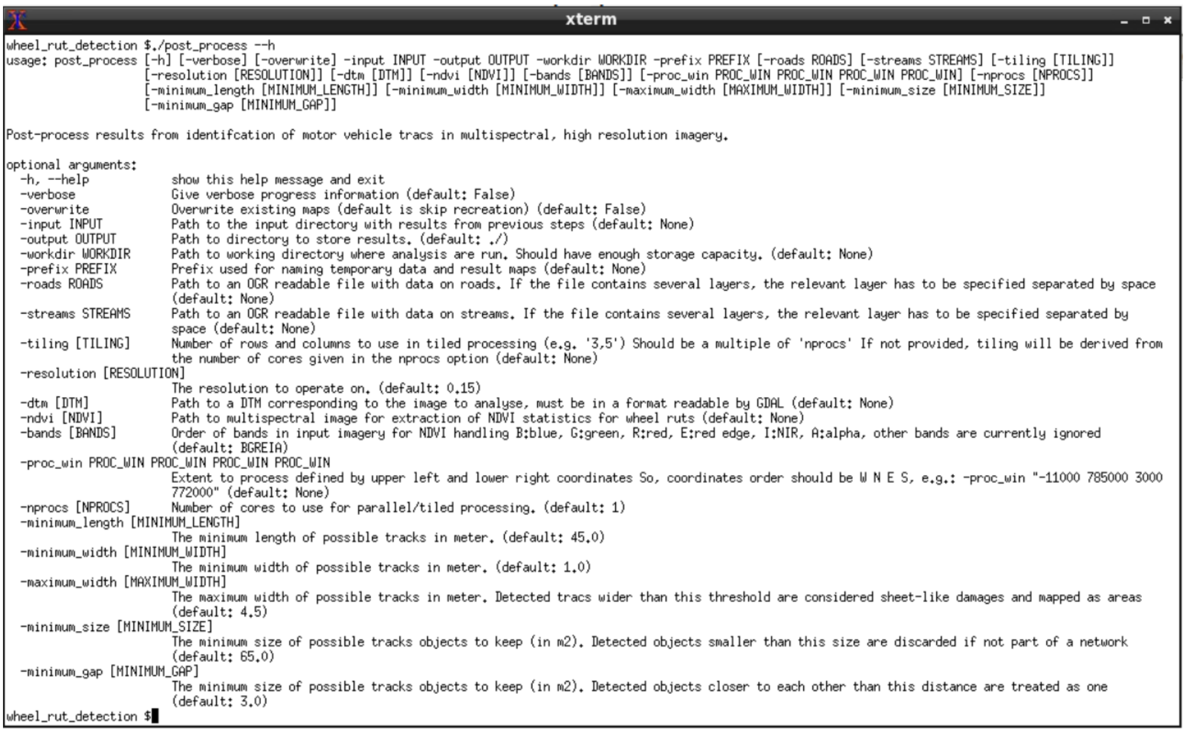
Examples of resulting classifications can be seen in **Figure 27** and **Figure 28**.

Both for filtering and classification, threshold values were defined using visual inspection of the data. Width, length, and size thresholds, these are implemented as parameters to the command line tool and can be adjusted by the user if needed.

4.4.2 Technical documentation

The post-processing algorithms are implemented in a standalone, command line script written in Python 3 (see **Figure 23**). It is tested on Ubuntu Linux 18.04 with OSGeo libraries from UbuntuGIS – unstable repository. On Unix systems the script can be invoked directly with the *./post-process command*, while on MS Windows, the Python interpreter needs to be called first (i.e. *python3 post-process*[14])

[14] With the current libraries in OSGeo4W the workflow does not succeed. However, the workflow utilizes the latest library version, and operating system specific bugs that may occur after major version updates the now hamper the workflow from finishing can be expected to be fixed in upcoming bugfix releases.



```
wheel_rut_detection $./post_process --h
usage: post_process [-h] [-verbose] [-overwrite] -input INPUT -output OUTPUT -workdir WORKDIR -prefix PREFIX [-roads ROADS] [-streams STREAMS] [-tiling [TILING]]
                    [-resolution [RESOLUTION]] [-dtm [DTM]] [-ndvi [NDVI]] [-bands [BANDS]] [-proc_win PROC_WIN PROC_WIN PROC_WIN PROC_WIN] [-nprocs [NPROCS]]
                    [-minimum_length [MINIMUM_LENGTH]] [-minimum_width [MINIMUM_WIDTH]] [-maximum_width [MAXIMUM_WIDTH]] [-minimum_size [MINIMUM_SIZE]]
                    [-minimum_gap [MINIMUM_GAP]]

Post-process results from identifcation of motor vehicle tracs in multispectral, high resolution imagery.

optional arguments:
  -h, --help            show this help message and exit
  -verbose              Give verbose progress information (default: False)
  -overwrite            Overwrite existing maps (default is skip recreation) (default: False)
  -input INPUT          Path to the input directory with results from previous steps (default: None)
  -output OUTPUT        Path to directory to store results. (default: ./)
  -workdir WORKDIR      Path to working directory where analysis are run. Should have enough storage capacity. (default: None)
  -prefix PREFIX        Prefix used for naming temporary data and result maps (default: None)
  -roads ROADS          Path to an OGR readable file with data on roads. If the file contains several layers, the relevant layer has to be specified separated by space
                        (default: None)
  -streams STREAMS      Path to an OGR readable file with data on streams. If the file contains several layers, the relevant layer has to be specified separated by
                        space (default: None)
  -tiling [TILING]      Number of rows and columns to use in tiled processing (e.g. '3,5') Should be a multiple of 'nprocs' If not provided, tiling will be derived from
                        the number of cores given in the nprocs option (default: None)
  -resolution [RESOLUTION]
                        The resolution to operate on. (default: 0.15)
  -dtm [DTM]            Path to a DTM corresponding to the image to analyse, must be in a format readable by GDAL (default: None)
  -ndvi [NDVI]          Path to multispectral image for extraction of NDVI statistics for wheel ruts (default: None)
  -bands [BANDS]        Order of bands in input imagery for NDVI handling B:blue, G:green, R:red, E:red edge, I:NIR, A:alpha, other bands are currently ignored
                        (default: BGREIA)
  -proc_win PROC_WIN PROC_WIN PROC_WIN PROC_WIN
                        Extent to process defined by upper left and lower right coordinates So, coordinates order should be W N E S, e.g.: -proc_win "-11000 785000 3000
                        772000" (default: None)
  -nprocs [NPROCS]      Number of cores to use for parallel/tiled processing. (default: 1)
  -minimum_length [MINIMUM_LENGTH]
                        The minimum length of possible tracks in meter. (default: 45.0)
  -minimum_width [MINIMUM_WIDTH]
                        The minimum width of possible tracks in meter. (default: 1.0)
  -maximum_width [MAXIMUM_WIDTH]
                        The maximum width of possible tracks in meter. Detected tracs wider than this threshold are considered sheet-like damages and mapped as areas
                        (default: 4.5)
  -minimum_size [MINIMUM_SIZE]
                        The minimum size of possible tracks objects to keep (in m2). Detected objects smaller than this size are discarded if not part of a network
                        (default: 65.0)
  -minimum_gap [MINIMUM_GAP]
                        The minimum size of possible tracks objects to keep (in m2). Detected objects closer to each other than this distance are treated as one
                        (default: 3.0)
wheel_rut_detection $
```

***Figure 23**. Screen shot of the command line interface of the post-processing script*

Basic required input are the results of the model predictions provided in the *input* option. Output of the post-processing routine are a GeoPackage file containing vector representations of the extracted and filtered lines and areas as well as GeoTiff with the NiN 7TK map and those will be written into the directory provided in the *output* option. All temporary data is written into the *work-dir* directory, that ideally would be located on a fast storage medium. If ancillary vector data (e.g. from FKB) should be applied for masking, such data can be provided in the *streams* and *roads* option.

Other relevant parameters are;

- *minimum_length*: The minimum length of possible tracks in meter (default: 45.0)
- *minimum_width*: The minimum width of possible tracks in meter (default: 1.0)
- *maximum_width*: The maximum width of possible tracks in meter. Detected tracs wider than this threshold are considered sheet-like damages and mapped as areas (default: 4.5)
- minimum_size: The minimum size of possible tracks objects to keep (in m2). Detected objects smaller than this size are discarded if not part of a network (default: 65.0)
- minimum_gap: The minimum distance between possible track objects to treat as one. Detected objects closer to each other than this distance are treated as one (default: 3.0)

For the case of drone imagery, also a *dsm* can be provided for the assessment of impacts of wheel ruts on soil and multispectral imagery (in the *ndvi* option) for calculation of NDVI and subsequent assessment of the impact of wheel ruts on vegetation. If multispectral imagery is provided, the *band* order has to be specified as a string consisting of band abbreviations, here the default order of bands of the multispectral imagery are set to those of the P4M drone.

Since the process is parallelized to a large degree, it would benefit from distributing the processing tasks over several cores. The number of Central Processing Units (cores) to utilize for parallel processing can be provided in the *nprocs* option. Parallelization is applied using tiles and the *tiling* option defines the number of rows and columns the process should be split into. In order to secure efficient resource utilization with *tiling*, the number of tiles should be set to a multiple of the number of cores in the *nprocs* option. If no tiling is defined but more than one core is allocated, tiling is computed automatically. Tiling can also help to limit the amount of required memory for processing if the number of tiles is larger than the number of cores. Other optional, technical parameters are message verbosity (*verbose*), meaning the level of process information given during computation and whether earlier runs should be overwritten (*overwrite*). If only parts of a mosaic should be processed, a processing window can be defined in the *proc_win* option.

4.4.3 Numerical assessment of post-processing results

Post-processing improves overall accuracy scores for all cases, except predictions for aerial images in Rjukan where only very limited validation data was available (see **Table 5**). As expected improves post-processing the precision score of the wheel rut class and the recall score of the background class respectively, while recall for wheel rut class and precision for background class is slightly reduced. This is because post-processing only removes positive classifications and here especially the number of false positives. The accuracy improvement of classification results for drone imagery is consistently larger than for classifications from aerial images. Classifications from aerial images contain less artifacts from image tiles so that the cleaning part of the post-processing routine has less effect on predictions from that data source.

***Table 5**. Changes in classification accuracy after post-processing (accuracy scores before post-processing are given in parentheses) of the different combinations of study area and data source in terms of overall accuracy (OA = average of the F1-score of all classes (wheel rut damage (1) and background (0)), precision (P = true positives / (true positives + false positives)) representing the ability of the model avoid false positive detections, recall (R = true positives / (true positives + false negatives)) representing the ability of the model to detect all wheel rut pixels, F1 score (F1 = 2 * (precision * recall) / (precision + recall)) for the two classes object of study.*

			Background			Wheel rut damage		
Area	**Data source**	**Overall Accur.**	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
Balsfjord	**Drone**	73.7 (71.7)	97.3 (97.6)	96.1 (94.1)	96.7 (95.8)	46.5 (38.8)	55.5 (62.0)	50.6 (47.7)
	Aerial images	65.1 (66.7)	95.0 (95.2)	97.7 (97.0)	96.3 (96.1)	29.4 (27.9)	16.0 (19.2)	20.7 (22.8)
Rjukan	**Drone**	68.2 (66.4)	96.8 (97.4)	93.6 (90.0)	95.2 (93.6)	34.3 (28.7)	52.1 (63.0)	41.4 (39.4)
	Aerial images	76.8 (82.2)	99.5 (99.8)	99.4 (99.1)	99.5 (99.5)	52.6 (52.6)	56.0 (85.3)	54.2 (65.0)

The effects of cleaning with ancillary data are not reflected in the accuracy scores because these areas were not covered by validation data. Likewise are neither artefacts along the boundary of image mosaics covered with validation data so that these are not reflected in the accuracy scores either.

4.4.4 Visual assessment of post-processing results

The post-processing, conducted with the default settings for the parameters mentioned above (like e.g. *minimum_size*), removes a significant amount of noise and misclassification, keeping correct detections mostly in place (see **Figure 24** and **Figure 25**). Especially for the drone imagery, there still remains a significant number of larger objects with misclassifications (see also **Figure 26**). Cleaning those in a post-processing routine would lead to a disproportional loss of correct detections, and should thus rather be addressed in further model improvements with an increased amount of training data.

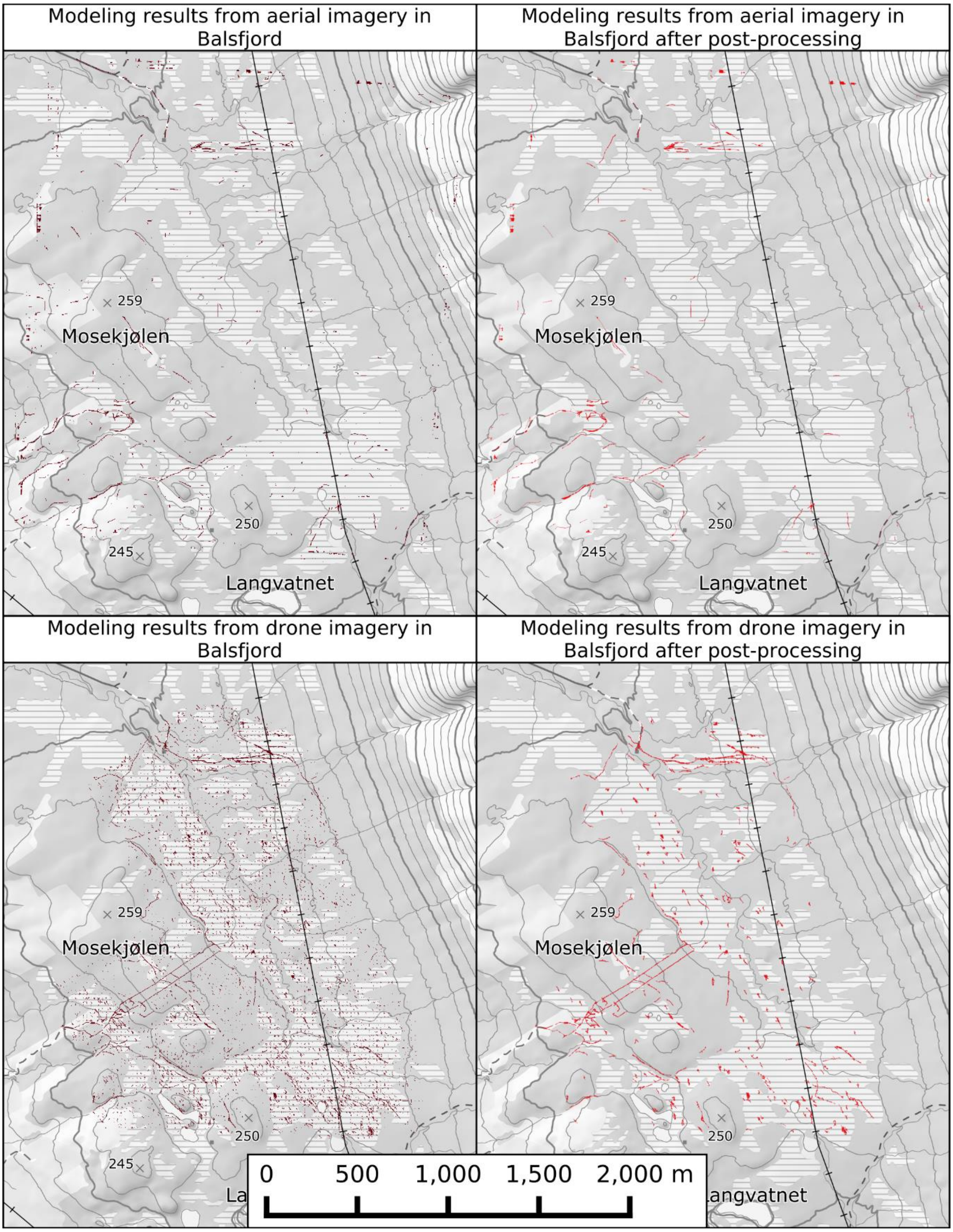

Modeling results from aerial imagery in Balsfjord
Modeling results from aerial imagery in Balsfjord after post-processing
Modeling results from drone imagery in Balsfjord
Modeling results from drone imagery in Balsfjord after post-processing
259
Mosekjølen
250
245
Langvatnet
0
500
1,000
1,500
2,000 m

Figure 24. *Overview over results of post-processing of predictions from drone and aerial imagery at Balsfjord study site*

In Rjukan usefulness of the application of ancillary data becomes more visible where roads are filtered out. However, due to spatial offset between drone imagery and ancillary FKB data not all areas are filtered properly. That offset would ideally be addressed in the processing of the drone imagery and use of high precision GPS.

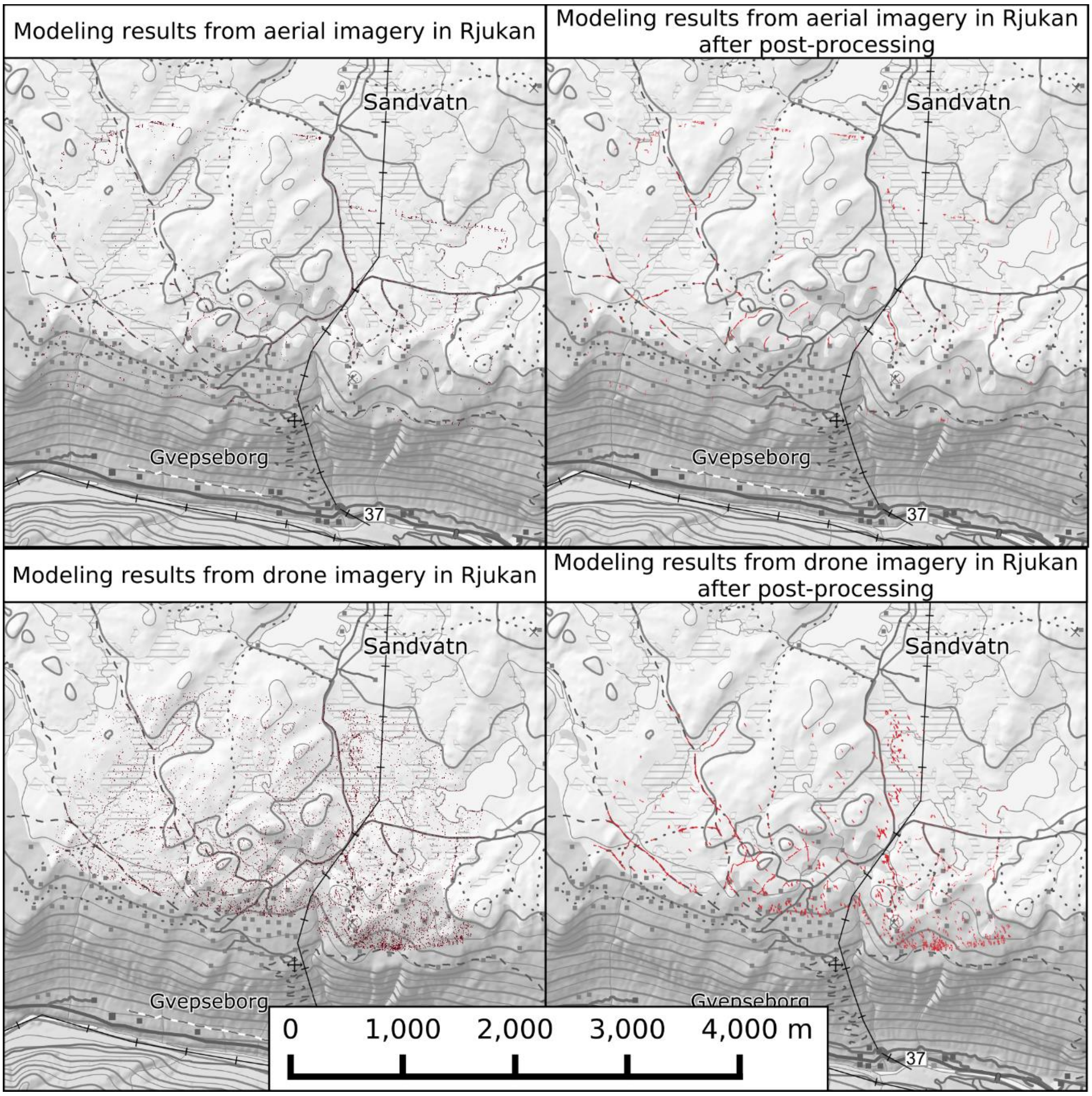

Modeling results from aerial imagery in Rjukan
Modeling results from aerial imagery in Rjukan after post-processing
Modeling results from drone imagery in Rjukan
Modeling results from drone imagery in Rjukan after post-processing
Sandvatn
Gvepseborg
37
0
1,000
2,000
3,000
4,000 m

***Figure 25**. Overview over results of post-processing of predictions from drone and aerial imagery at Rjukan study site*

Post-processing does currently not clean dense artefacts at image boundaries. This is an issue could be included relatively simply in a future version of the post-processing routine e.g. using a buffer around the alpha channel or NoData areas in the input images. For square image data this could already now be addressed by defining an adequate processing window in the respective command line option.

Figure 26 shows cases where post-processing often fails to clean misclassifications are areas in wetlands or forest with vegetation shows pattern in reflections that can be visibly similar to wheel ruts in the training data. Such, more complex cases should be addressed in a next iteration of model improvement and re-training, rather than in rule- or threshold based post-processing routine.

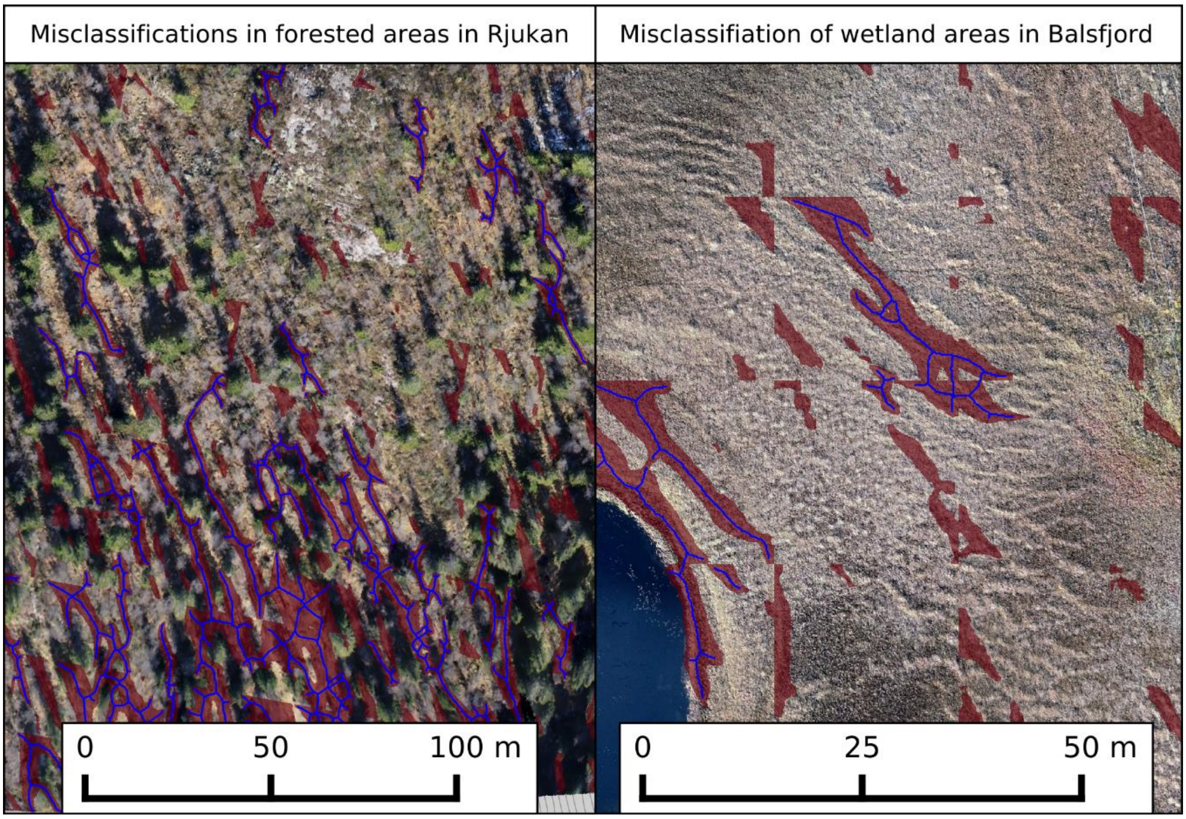

Misclassifications in forested areas in Rjukan
Misclassifiation of wetland areas in Balsfjord
0
50
100 m
0
25
50 m

***Figure 26**: Examples of misclassifications of drone imagery in forested and wetland areas where post-processing fails to clean. Original predictions in transparent red and vector lines from post-processing in blue. Background is the RGB mosaic from drone imagery used for prediction.*

Vector data that is returned from the post-processing can during post-processing be classified with regards to impact of wheel ruts on soil or vegetation, if a DSM or multispectral imagery is provided as input. A disadvantage with regards to evaluating to which extent the classification captures significant and relevant differences on the ground is hampered by a slight spatial mismatch between the model predictions and terrain or multispectral data. That spatial error is due to the fact that the predictions were generated from the original ortho-mosaic from Andøya Spacecenter (see chapter 3.2.1.2) – due to time constraints – and that this does not fully correspond to the terrain models and multispectral mosaics produced later in the project (see chapter 3.3.1).

Both the assessment of soil and vegetation damage rely on the availability of drone imagery from which the required input can be computed. Because of that as wheel as because of mentioned spatial and also temporal mismatch between possible input data sources, these post-processing options are only meaningful for drone imagery.

Figure 27 and **Figure 28** illustrate results of the classification of the severity of wheel rut damages on soil / terrain and vegetation. The upper row shows the raw predictions (transparent red areas) and post-processing results (vector lines) over the drone ortho-mosaic, the row in the middle covers the classification of the impact of driving activity on the soil and terrain in that same areas and the row at the bottom the classification effect on vegetation represented by NDVI values.

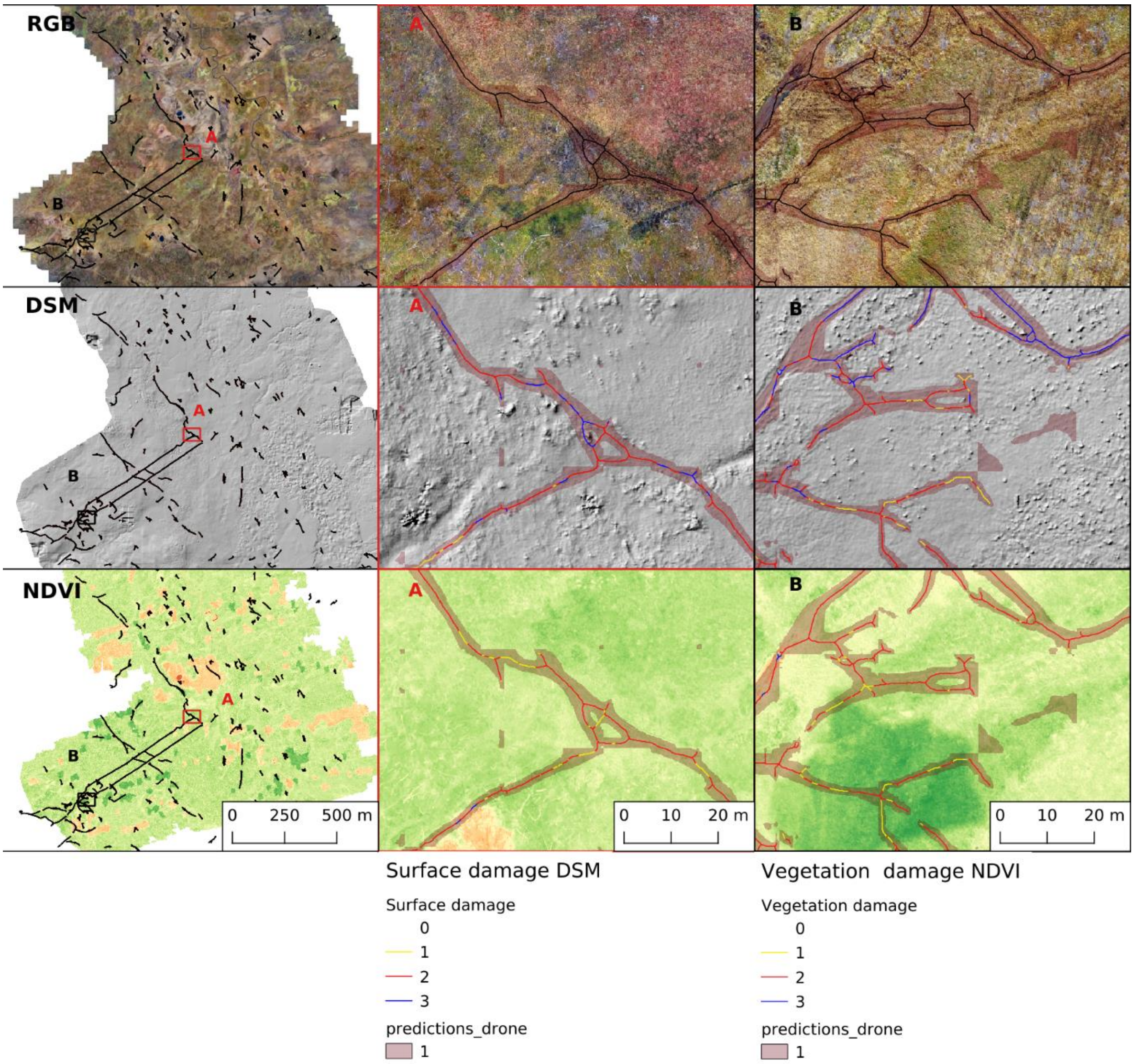


***Figure 27**: Examples of the assessment of soil / terrain impact and vegetation damage of wheel ruts at the Balsfjord study site (damage classes: 0 = zero detectable damage or missing data, 1 = minor soil or vegetation damage, 2 = medium soil or vegetation damage, 3 = most severe soil or vegetation damage.).*

Thresholds for the binning of the underlying continuous depth- and NDVI-depression values into the four classes mentioned in chapter 4.4.1 were chosen based upon visible differences of impact on the terrain and NDVI at wheel rut locations and the different coloring in **Figure 27** and **Figure 28** reflects variation in the underlying values. From the visual assessment, classification of surface damage appears to be more slightly reliable than classification of vegetation damage. Unfortunately, the spatial mismatch between mosaics used for detection of wheel ruts with deep learning, and mosaics used to compute NDVI as well as the produced terrain models leads to errors in the classification and thus makes the assessment difficult. In any case, to what degree visible changes in the terrain model or NDVI values correspond to relevant changes on the ground could not be studied in this project.

In consequence, thresholds that mark the upper and lower bounds of the four classes should be studied and evaluated in further follow up work to ensure they represent relevant information value.

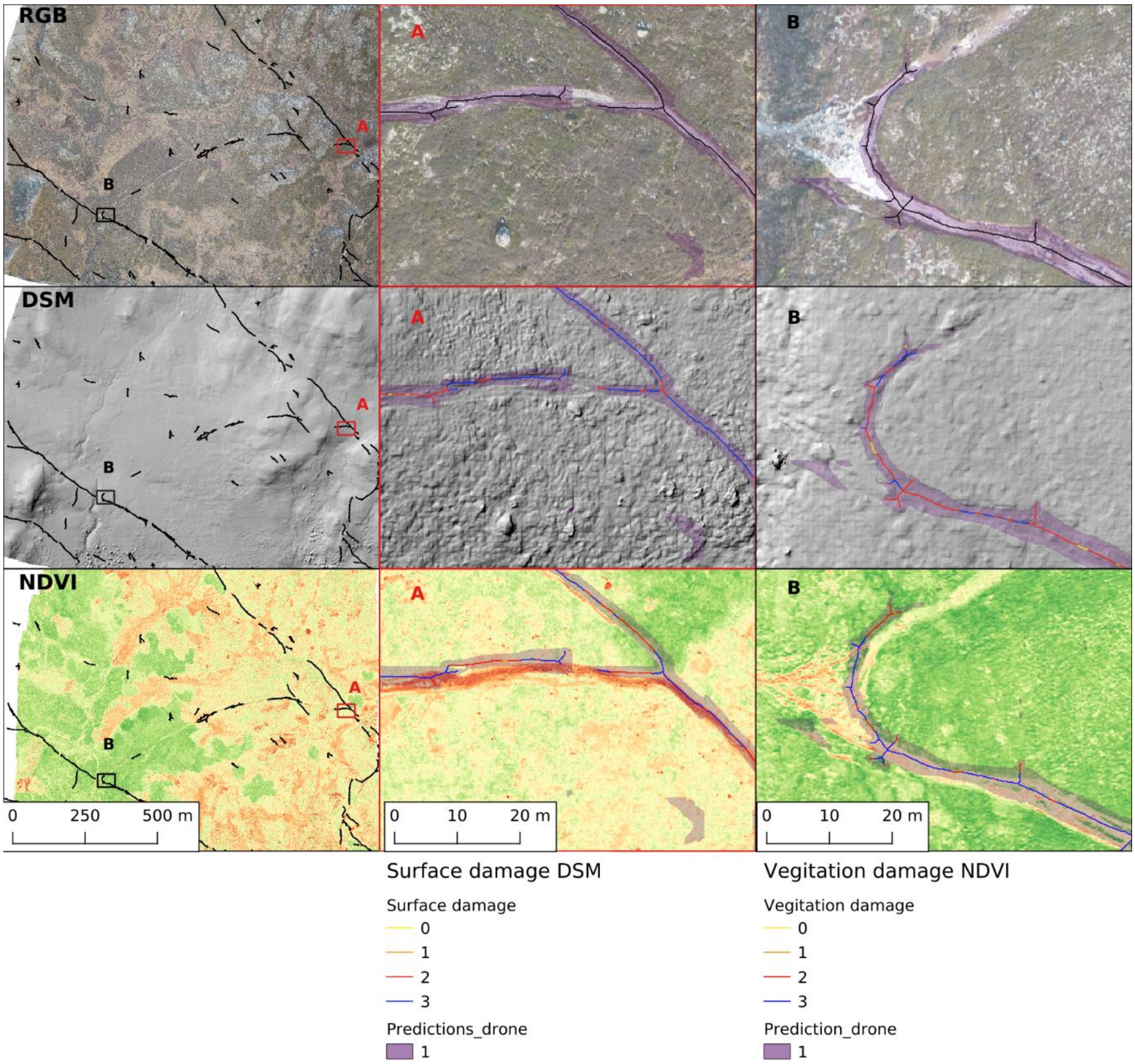


Figure 28*: Examples of the assessment of soil / terrain impact and vegetation damage of wheel ruts at the Rjukan study site (damage classes: 0 = zero detectable damage or missing data, 1 = minor soil or vegetation damage, 2 = medium soil or vegetation damage, 3 = most severe soil or vegetation damage.).*

An additional output from post-processing is a map of wheel rut density that reflects the NiN 7TK variable. **Figure 29** shows those post-processing outputs for the Balsfjord study area for wheel ruts detected from aerial images from 2016 and 2017 on the left-hand side and based on drone imagery from 2020 at the right hand side. The 100m pixel size of the 7TK maps reduces the amount of noise in the map products. Because it is based on counts of 10m pixels with wheel ruts it is however sensitive for false positive detections, so that, esp. for this kind of results more aggressive filtering of false positive predictions would be advised.

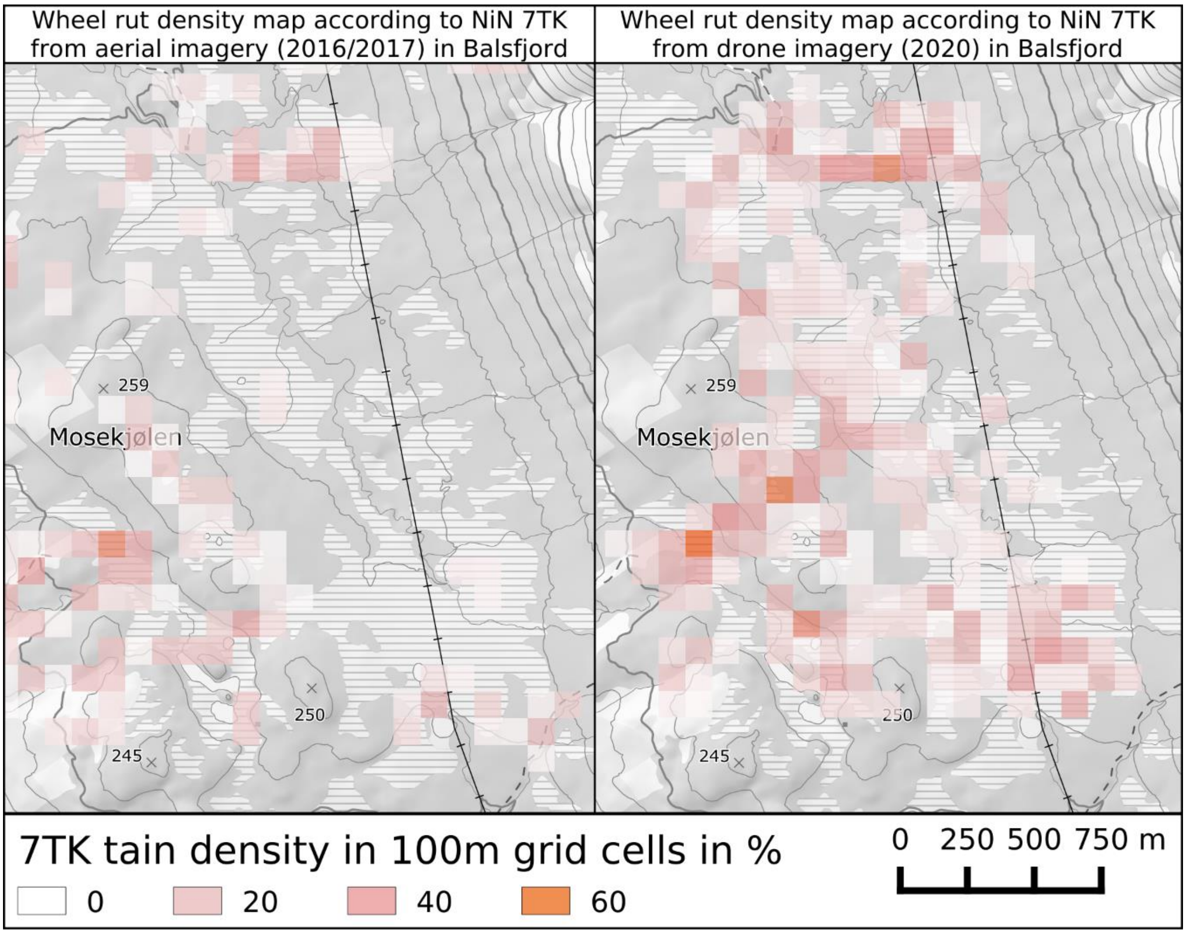

Wheel rut density map according to NiN 7TK from aerial imagery (2016/2017) in Balsfjord
Wheel rut density map according to NiN 7TK from drone imagery (2020) in Balsfjord
259
Mosekjølen
250
245
7TK tain density in 100m grid cells in %
0
20
40
60
0 250 500 750 m

Figure 29. *Track density according to NiN 7TK computed with the developed post-processing routine for the Balsfjord study site*

4.4.5 Recommendations and needs for further development

Main needs for improvements of the post-processing routines are to make it less dependent on the chosen resolution (currently the process is optimized for 0.15m resolution). Furthermore, default values for different settings that have been based on visual data inspection for now, should undergo a more systematical evaluation and re-adjustment. Another relevant area for improvement is the handling of spatial off-set between ancillary data (e.g. through buffering) and a solution for cleaning of misclassifications at the image boundaries. Finally, given the sensitivity of the 7TK maps for false positive predictions, the production of those maps should probably be separated out and applied to the resulting vector data after that received a manual check and a relatively efficiently final tweak.

Further improvements in terms of performance can be expected with the release of the 8.2 version of GRASS GIS where a couple of tools have been parallelized that have not been possible to parallelize in this project, like e.g. the computation of univariate statistics for trail or area units.

4.5 Effect of flight planning and data processing on the detection accuracy and suitability of drone data compared to aerial data

The raw drone photos were re-analyzed and rectified using ODM (see chapter 4.1), resulting in multiple layers that in the future could be captured for new areas and used in the model if they show promising results. The original drone dataset consisted of photos captured using two separate drones, Phantom 4 Pro and Phantom 4 Multispectral. The layers that were analyzed to compare data are the RGB orthophoto, NDVI (created from the multispectral data), digital terrain model (DTM) and aerial photos from the Norwegian orthophoto program captured 23.7.2017.

The multispectral data covered a slightly smaller area, meaning that there are some areas without NDVI in the following comparisons. The drone photos were captured 7.10.2020, more than three years after the aerial photos. This makes a direct comparison between drone photos and aerial photos more difficult, as the ground-truth might have changed significantly between the photos.

In this project the machine learning model has only been trained, tested and validated using either orthophoto captures by drone or aerial photography captured in the Norwegian orthophoto program. In addition, we have created NDVI and DTM to see if they have potential to give useful information to the models in the future.

Figure 30 shows a comparison of drone photos and aerial captured at separate times. There are new tracks produced after the aerial image was taken. The tracks are clearly visible also on both orthophoto and NDVI, and neither one is likely to give information to the model that the other one doesn't also give. However, here the DTM is of decent quality, making is possible to extract a depth measure.

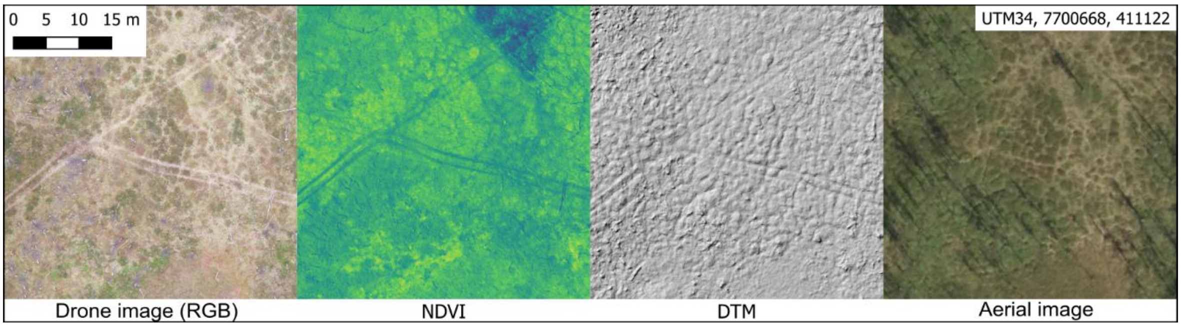


***Figure 30**. Comparison between different layers. Orthophoto from drone, NDVI and DTM produced from drone imagery. Aerial image from the Norwegian orthophoto program.*

The following **Figure 31** illustrates differences between drone imagery and data from the Norwegian orthophoto program with regards to their ability to capture wheel ruts in detail. Between the photos parts of the tracks have also been filled with gravel. In the grass-covered area in the middle of the photos there are tracks visible in both datasets. Surprisingly, the different layers visualize different parts of the tracks with differing clarity. In this case the DTM even shows a clear track to the top right which is difficult to spot in any other layer. In the middle part the NDVI and orthophoto visualize different parts better, but in general the drone based orthophoto shows the better accuracy overall. The tracks in the middle section are partly covered by gras, which is a likely reason for why it is not so clearly picked up by the NDVI signal.

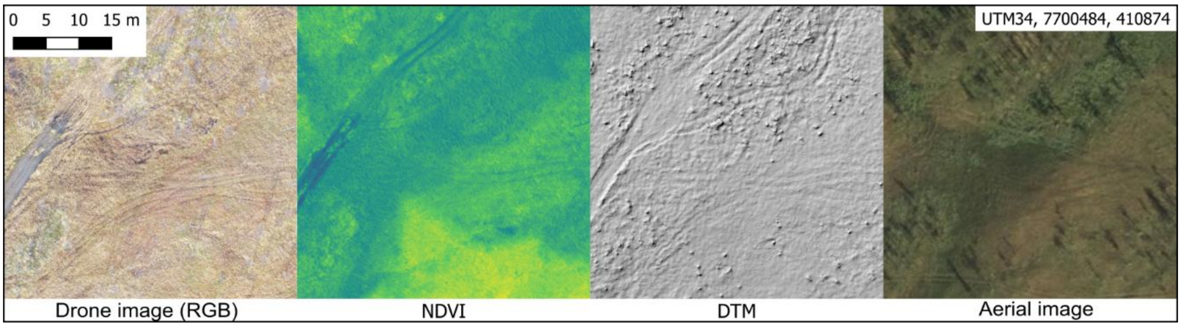


Figure 31*. Example of differences between the available data sources*

A good example for where aerial imagery can be adequate for detecting wheel ruts is shown in **Figure 32**. A large open area with well-established tracks. This view was outside of the multi-spectral coverage, so we don't have NDVI. The DTM however is very poor and full of artifacts. Most likely this is due to too low overlap between drone photos in addition to few recognizable features for the orthorectification algorithm. This will result in too few usable photos covering each pixel for the algorithm.

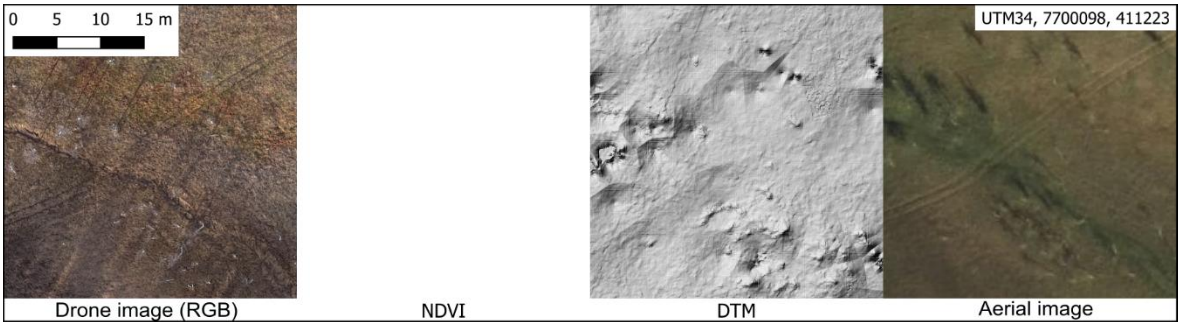


Figure 32*. Example both drone and aerial imagery capture wheel ruts quite well*

Another example of an area with poor quality in the DTM is shown in **Figure 33**. The issues are again likely due to too little overlap between neighbouring drone images and few features in these open areas that can be used in the processing algorithm to identify common points in the images.

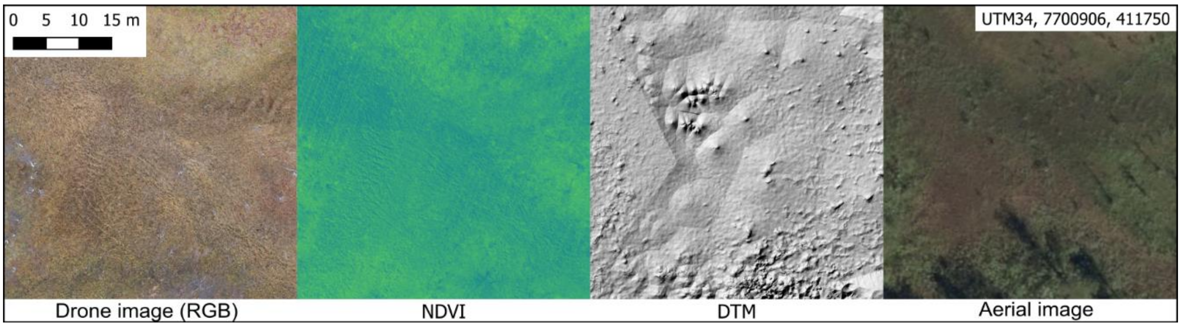


Figure 33*. Example of artefacts in the created DTM*

A location where all data sources visualise tracks quite well is shown in **Figure 34**. The DTM is of medium quality with quite a few artifacts but still captures at least arts of the tracks. One take-away from this example how trees affect the ability to detect wheel ruts depending on the timing of image acquisition. The aerial photos were captured during the summer, while the drone flights were conducted during the autumn. This means that there are more leaves on the trees on the aerial photographs, which both disrupts views in some areas and also casts bigger shadows in sunny conditions especially with lower sun angles.

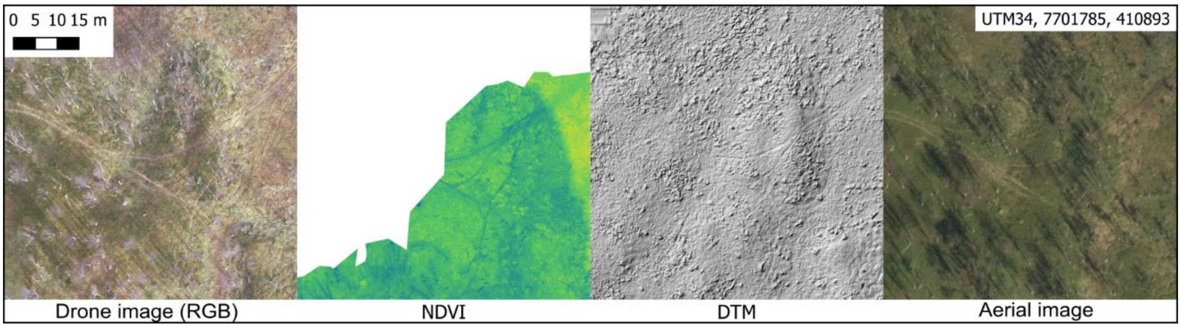


Figure 34*. Example of another location where all data sources capture wheel ruts quite well*

Images in **Figure 35** show a trail instead of a track. The DTM is again of poor quality, but here also the orthophoto shows a blurred area. This is a kind of artifact that is quite common in areas with to low overlap between drone photos or where the orthorectification model cannot find enough recognizable features to stich enough photos together. In this area, there are also a lot of trees, obstructing views of the trail from the aerial photos. Even though the orthophotos are blurred, it is possible to see the trail accurately both in or in the orthophoto and in NDVI. The NDVI is produced from another flight with a different drone, so the orthorectification might be better for it.

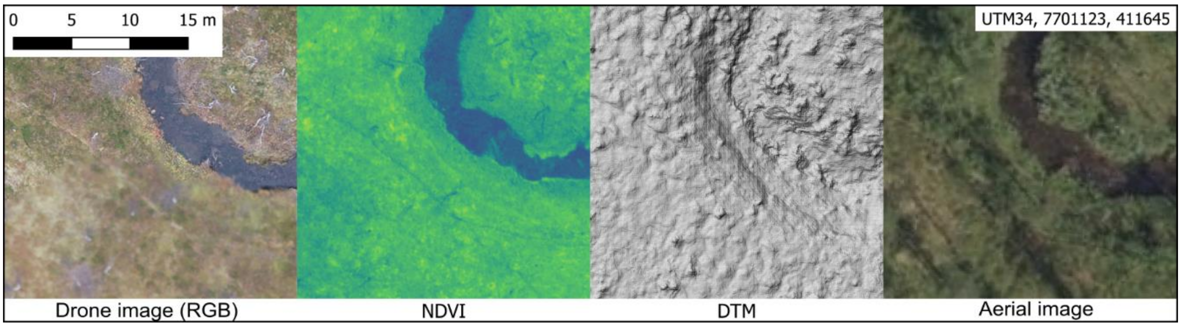


Figure 35*. Example how lack of overlap between neighbouring drone images affects the quality of generated orthophotos less than quality of respective terrain models.*

An example of an area where the tracks are only visible on the drone base orthophoto is shown in **Figure 36**. Tracks are also present in the aerial photos, but significantly harder to spot. The NDVI signal is again obstructed by gras covering the tracks, and DTM does not provide any useful signal for the present track.

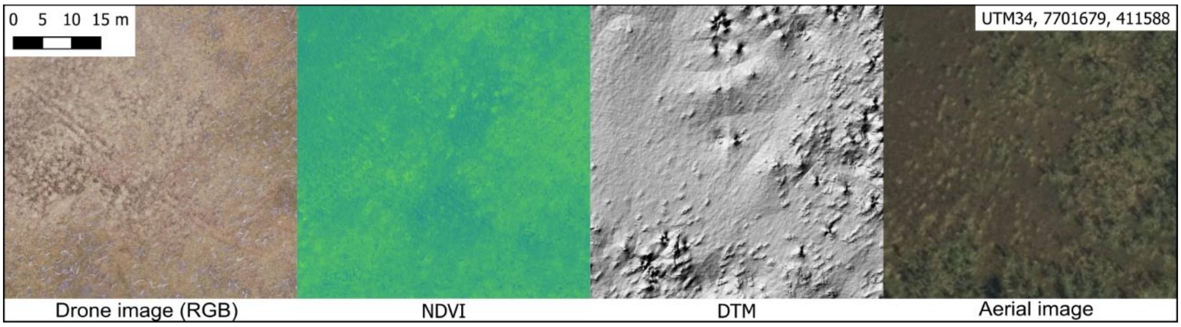


Figure 36*. Example of an area where wheel ruts are more visible on the aerial image than on the drone orthophoto*

The examples above (**Figure 30** - **Figure 36**) illustrate that, both NDVI and DTM can provide some added value depending on the area and quality of the acquired data.

A limiting factor for the usefulness of terrain models however is clearly the lack of homogenous quality in the available data. As discussed in chapter 4.1.2, in order for terrain models to be useful more effort would have to be put into safeguarding quality of the input data during image acquisition and flight planning. With the quality of the available data in this project, adding terrain information into modelling has to be considered rather an additional source of error than a means to improve modelling quality. In that context it is also important to keep in mind that terrain models will not be available for aerial images so that adding those as an additional layer of information in a model will prevent possible transferability of a model between plane-based and drone imagery. At locations with good quality of the terrain models, there is however a possibility to use it to measure the depth of the tracks. Height of pixels in the tracks can be compared to pixels adjacent to the tracks and thereby give a crude depth measurement.

In principle, NDVI from multispectral imagery may provide more frequently additional information compared to orthophoto that can be useful for identifying the location of tracks. However, as discussed in chapter 4.1.1, heterogeneity of quality within the produced mosaics limits also the usefulness of multispectral imagery in modelling. In addition, a different camera/drone is required to capture multispectral imagery. Those sensors regularly have a lower resolution so that the acquisition of multispectral images comes with a trade-off with regards to resolution of the RGB imagery (and in turn also possible DTM) and / or required time to cover comparable areas.

Overall, even though terrain models and multispectral data from drone images can be beneficial in some cases, the additional effort in takes in collecting them with adequate quality makes it questionable to base the motivation for using drones on the availability of any of those additional layers of information. There are however two major reasons to conduct image acquisition with drones for monitoring wheel ruts: 1) drone imagery can provide more detailed and more timely data and 2) data acquisition can be conducted on-demand and at more suitable times of year compared to the Norwegian orthophoto program where images are usually taken during summer.

4.6 Recommendations for drone type and mission planning for wheel rut detection

Detection of wheel ruts are sensitive to both weather conditions and time of year. Both drone photos and aerial photos are harder to use when acquired at low sun-angles and with a lot of shadows in the photos. In summer, both high grass and leaves might obstruct the view on wheel ruts, making them more difficult to recognize. It is also likely that wheel ruts to a large extent are created in connection with moose hunting in the autumn, and therefore are fresher and easier to discover in photos from that time of year. Timing of the flight might therefore be the most important than details regarding mission planning.

That said, for future drone missions, it is advisable to fly as high as possible while ensuring a sufficient ground sampling distance. We have showed that a GSD of 5-7 cm is more than sufficient as long as the models are well trained. There is no need to go for cheap multispectral cameras which generally do not provide much added information in our use-case and one has to compromise heavily either on the resolution or the overall flight time.

If one wants depth measurements from the DTMs however, the quality of the DTM is very important. The quality in the data analysed here varied wildly within the dataset, usually due to lack of convergence for some areas in the orthorectification. The underlying reasons for this is likely slightly to low overlap between the original photos, in combination with large areas with little vegetation and therefore few(er) recognizable features. The quality of the DTM is also likely to decrease with decreased GSD, so to ensure good enough DTM one should ensure that the GSD is at least 5-7 cm and that the overlap between photos are at a minimum 70% in both directions.

This is also important to retain the same degree of high overlap to ensure that the georeferencing using RTK is as precise as possible.

Practically there are many ways to achieve this. Below we describe one way to achieve these results for a larger area.

Suggested drone: While the use of copter type of drones is somewhat more straightforward as these type of UAVs are more commonly available, for larger area coverage we would suggest the use of fixed wing UAVs. Amongst these, of particular interest are the vertical take-off and landing the Quantum System Trinity F90+ UAV is an example. This UAV has an autonomy of up to 90 minutes and allows for vertical take-off and landing which is useful in situations where there is limited space for take-off and landing. Furthermore this UAV allows for deploying a variety of payloads ranging from high resolution cameras to lidar sensors.

Suggested sensor: The suggested sensor depends of the possibility to perform UAV operations at altitudes > 120 m. When this is possible it is advisable to utilize very high resolution sensors such as the Sony RX1R II, which with its 42.4 MP allows for obtaining high resolution imagery even at high altitude (e.g. resolution of 5 cm at 400 m of altitude). In situations where the drone operators are limited to flying below 120 m above ground then the use of such high resolution sensors is not advisable as it would produce excessively detailed imagery with negative consequences on storage space and processing speed. In such situations one could rely on cheaper sensors like the Sony UMC-R10C RGB camera (20.1 MP), which would produce satisfactory results at 120 m while limiting the data volume.
An interesting alternative to RGB cameras would be the use of a laser scanning sensor such as the Qube 240, which would provide active 3D measurements of the surface of greater quality compared to the photogrammetric 3D models from the RGB camera. This would on the other hand require the develop a new type of AI model to be trained on these new data.

Suggested flight parameters: as for the previous it all depends on the operators' licence and his/her possibility to fly above 120 m above ground. Where that is possible the efficiency of the operations would be greatly enhanced as with the same amount of time one can cover a substantially larger area. Given that the interest is to survey wheel rut damages and not vegetation, to boost the flight plan efficiency it is suggested to fly with a 70% forward and 70% lateral overlap. This would produce enough imagery to ensure the reconstruction of a suitable ortho-mosaic. No need for double grid pattern. A orthodiagonal grid pattern could be used to further increase the quality of the DTM, but as it halves the area covered it is not recommended. A more time efficient method to ensure a better performing DTM is to increase the overlap to around 80%. Higher altitude will also reduce the details in the DTM, regardless of the sensor used.

5 Conclusions, recommendations and further development

Results show that the initial models developed in this project produce fair to good results for both plane- and drone based imagery in both study sites. Models utilizing drone data perform slightly better than models based on aerial images with regards to correctly capturing wheel ruts, where drone imagery based models capture more details but currently also show a larger degree of noise and scattered false positive classifications. Models from aerial images perform best in open areas, they struggle though in forested areas.

While the current models provide an initial understanding of the quality of the detection from drones and orthophotos, it is important to keep in mind that the amount of training data used in this study case was limited to two study areas and to two images taken under seasonal, uniform atmospheric and lighting conditions. A full deployment of the models developed in this project does not ensure the transferability to new image data acquired in different seasons, and under varying atmospheric and light conditions. Thus, further development of the current work would need to focus on the expansion of the images and annotations used to train the model. This would have a twofold effect: 1) improve the model detection accuracy since it would be trained on a larger set of data; and 2) improve the transferability of the model to new data.

Based on the results from reprocessing of the raw drone images, it can be concluded that features that are specific to targeted image acquisition with drones like the production of photogrammetric terrain models or multispectral images pose challenges with the regards to creation of data with a homogenous quality and sufficient collocated geometry. Reliable monitoring of trail depth over time by means of comparing photogrammetric terrain models appears to be an undertaking that for larger extents like in this study does not seem to yield in reliable results, at least not without significant extra effort during data acquisition and processing. The main motivation for using drones for monitoring wheel ruts from ATVs should thus be the timely generation of monitoring data. Annotated training data for the Balsfjord study area showed an increase of wheel ruts in that area by ~100 % (or 8 km in total) in the four years between 2016 to 2020. This underpins the value of drone campaigns for timely response in case of known or reported off-road driving activities, where the 5 to 10 years repetition cycle of the Norwegian orthophoto program might not yield timely enough data. The benefit of drone imagery lies thus in a more detailed understanding of the current situation as well as a finer resolution analysis. Drone data can therewith be seen as an on-demand technology that is complementary to aerial images that are taken on a regularly basis for the Norwegian orthophoto program. Under an operational scenario, once the critical areas have been identified, drones could be deployed to capture updated imagery and the drone model used to predict the current situation of ATV damages. Results of the project give reason to believe that images from the Norwegian orthophoto program in the other han can valuable to perform an initial screening of areas where ATV damages are present. Such screening process can help identifying the geographic location of areas that are under threat of potential new damages.

The fact that prediction results for higher resolution drone images perform only slightly better in detecting wheel ruts compared to aerial images suggest that in further improvement a more systematic evaluation of the effect of image resolution should be conducted. In that context also ultra-high resolution satellite images that now a days can deliver up to 30 cm resolution (and up to 15 cm resolution after software based image enhancement) may be considered as a third, additional data source.

Another, though related branch of further research in the subject should investigate whether it would be feasible and adequate from an end-user point of view to consolidate the deep learning models that currently are different for drone and aerial imagery, into one coherent model in order to reduce the maintenance effort and at the same time increase the amount of both training data and imagery the model could be trained with. In addition, recent methods to limit the required amount of test- and training data, like Few-Shot Learning (see e.g. Wang et al. 2020) should be explored in order to increase the practical applicability in a monitoring context.

Images from the Norwegian orthophoto program should thus be a natural starting point in order to increase the amount of training data and therewith the number of conditions the model is trained with. When collecting new training data a critical aspect to be considered is the potential variation in aerial imagery in respect to seasonality, atmospheric, and illumination conditions. For this, one would have to collect a large enough sample of ortho-mosaics that can cover this range. A potential approach could be to start by selecting 10-20 ortho-mosaics collected during spring, summer, and autumn with both leaf on and leaf off conditions. Given that it is impossible to define a specific number of annotations required to obtain a satisfactory model, one would have to adopt an iterative approach based on the following steps:

- Annotated a batch of 5 ortho-mosaics covering a range of variation in terms of seasonality and covering different geographical areas. Concerning the proportion of the area to be covered (i.e. amount of annotated data) one would ideally want to annotated the entire area. In situations where the funding for annotation is limited then one would adopt some sampling techniques to select areas for annotation. In such case, however it is important to ensure that the areas to be annotated are large enough to ensure the presence of ATV wheel rut damages.
- Split the annotated data into a training and validation batch
- Retrain the model using the existing training data plus the new training data from the previous step.
- Evaluate the model's performance using the old plus the new validation data.
- In case more precise results are needed then repeat the steps above.

If it was possible to consolidate the modeling approach to one cross-sensor model independent of the sensor, also flight planning for drone campaigns could be slightly adjusted to produce data that is more comparable to data from the Norwegian orthophoto program. This also holds the potential to increase the efficiency with that drone campaigns can be conducted, both with regards to data capturing and processing, as larger areas can be covered.

Results of the project show that post-processing of the modelling results is both needed, able to improve the quality of the final products and that it can produce condensed and more usable representations of the results. Further technical improvements of the post-processing should cover remaining filtering issues like those at image boundaries along with technical details mentioned in chapter 4.4.5. In terms of methodology, the classification of the results with regards to the severity of damages to esp. soil / terrain should undergo systematic evaluation and re-adjustment if needed.

Technically, in order to make the results more directly applicable by managers, the workflows for the deep-learing steps should be further wrapped or consolidated in order to further reduce the need for manual interaction. To that ends, utilization of publicly available frameworks or even multi-framework wrappers (like Pesek 2022) can be considered.

6 References

Ancin-Murguzur, F. J., Munoz, L., Monz, C., & Hausner, V. H. 2020. Drones as a tool to monitor human impacts and vegetation changes in parks and protected areas. Remote Sensing in Ecology and Conservation 6(1): 105–113. https://doi.org/10.1002/rse2.127

Ćwiąkała, P., Kocierz, R., Puniach, E., Nędzka, M., Mamczarz, K., Niewiem, W., & Wiącek, P. 2018. Assessment of the Possibility of Using Unmanned Aerial Vehicles (UAVs) for the Documentation of Hiking Trails in Alpine Areas. Sensors 18(1): 81. https://doi.org/10.3390/s18010081

Chen, L., Papandreou, G., Schroff, F. & Adam, H. 2017. Rethinking Atrous Convolution for Semantic Image Segmentation https://doi.org/10.48550/arXiv.1706.05587

Dale, Ø. 1995. Omfang og årsaker til hjulsporskader etter skogsdrifter. En feltregistrering fra fem regioner i Norge. Rapport fra Skogforsk, 7/95 1995. pp 27

Due Trier, Ø. & Salberg, A.-B. 2020. Kartlegging av naturinngrep. Sluttrapport – revidert utgave. NR rapport. SAMBA/10/2020. https://www.miljodirektoratet.no/sharepoint/downloaditem?id=01FM3LD2QIB7AZX37HJJFJFRY6Q2KAWMIT

Eagleston, H., & Marion, J. L. 2020. Application of airborne LiDAR and GIS in modeling trail erosion along the Appalachian Trail in New Hampshire, USA. Landscape and Urban Planning 198: 103765. https://doi.org/10.1016/j.landurbplan.2020.103765

Evju, M., Hagen, D., Blumentrath, S. & Eide, N.E. 2010. Verdi- og sårbarhetsvurdering i Børgefjell nasjonalpark. Med spesiell fokus på utvalgte lokaliteter og utfordringer knyttet til ferdsel. - NINA Rapport 543. 111 pp. Norsk institutt for naturforskning. http://hdl.handle.net/11250/2564908

Evju, M., Hedger, R., Nowell, M., Vistad, O.I., Hagen, D., Jokerud, M., Olsen, S.L., Selvaag, S.K. & Wold, L.C. 2020. Slitasje og egnethet for stier brukt til sykling. En feltstudie og en GIS-modell. NINA Rapport 1880. Norsk institutt for naturforskning. https://hdl.handle.net/11250/2683833

Guirado, E., Tabik, S., Alcaraz-Segura, D., Cabello, J., & Herrera, F. 2017. Deep-learning versus OBIA for scattered shrub detection with Google earth imagery: Ziziphus Lotus as case study. Remote Sensing, 9(12): 1220

Kildahl, K. 2020. Forskningsmiljøer og skognæringa satser stort sammen. Nyhetssak NIBIO. https://www.nibio.no/nyheter/forskningsmiljoer-og-skognaeringa-satser-stort-sammen

Johansen, B., Karlsen, S. B. & Riise, T. 2019. Bruk Av Sentinel-2 Satellittdata Innen Forvaltning Av Øvre Dividalen Nasjonalpark Og Dividalen Landskapsvernområde. Tromsø: NORUT

Lennert, M., Grippa, T., Radoux, J., Bassine, C., Beaumont, B., Defourny, P., & Wolff, E. 2019. CREATING WALLONIA'S NEW VERY HIGH RESOLUTION LAND COVER MAPS: COMBINING GRASS GIS OBIA AND OTB PIXEL-BASED RESULTS. ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences XLII-4/W14: 151–157. https://doi.org/10.5194/isprs-archives-XLII-4-W14-151-2019

Neteler, M., & Mitasova, H. 2007. Open Source GIS: A GRASS GIS Approach : A GRASS GIS Approach. Springer, New York, NY, United States.

Opplysningsrådet for veitrafikken 2022. Kjøretøybestanden 2008 til 2020. https://ofv.no/kjoretoybestanden

Pesek, O. 2022. Possibilities of convolutional neural networks use for remote sensing image classification. https://github.com/ctu-geoforall-lab-projects/phd-pesek-2022

Pierzchała, M., Talbot, B., & Astrup, R. 2016. Measuring wheel ruts with close-range photogrammetry. Forestry: An International Journal of Forest Research 89(4): 383–391. https://doi.org/10.1093/forestry/cpw009

Rodway-Dyer, S., & Ellis, N. 2018. Combining remote sensing and on-site monitoring methods to investigate footpath erosion within a popular recreational heathland environment. Journal of Environmental Management 215: 68–78. https://doi.org/10.1016/j.jenvman.2018.03.030

Senchuri, R. 2020. Road edge detection using hyperspectral and LiDAR data based on machine and deep learning. Norwegian University of Life Sciences, Ås. Master thesis. https://nmbu.brage.unit.no/nmbu-xmlui/handle/11250/2725447

Talbot, B., Rahlf, J., & Astrup, R. 2018. An operational UAV-based approach for stand-level assessment of soil disturbance after forest harvesting. Scandinavian Journal of Forest Research 33(4): 387–396. https://doi.org/10.1080/02827581.2017.1418421

Tømmervik, H., Bakkestuen, V., Erikstad, L. & Strann, K.B. 2005. Motorisert ferdsel i utmark. I: NINAs strategiske instituttprogrammer 2001-2005: Landskapsøkologi: arealbruk og landskap. Sluttrapport. – NINA Temahefte 32. pp. 59-67. https://hdl.handle.net/11250/2726145

Wang, 2020. Generalizing from a Few Examples: A Survey on Few-Shot Learning. ACM Comput. Surv. 1(1): 1-34. https://arxiv.org/pdf/1904.05046.pdf

www.nina.no

2137

NINA Report

The Norwegian Institute for Nature Research, NINA,
is as an independent foundation focusing on environmental research, emphasizing the interaction between human society, natural resources and biodiversity.

NINA was established in 1988. The headquarters are located in Trondheim, with branches in Tromsø, Lillehammer, Bergen and Oslo. In addition, NINA owns and runs the aquatic research station for wild fish at Ims in Rogaland and the arctic fox breeding center at Oppdal.

NINA's activities include research, environmental impact assessments, environmental monitoring, counselling and evaluation. NINA's scientists come from a wide range of disciplinary backgrounds that include biologists, geographers, geneticists, social scientists, sociologists and more. We have a broad-based expertise on the genetic, population, species, ecosystem and landscape level, in terrestrial, freshwater and coastal marine ecosystems.

ISSN: 1504-3312
ISBN: 978-82-426-4925-6

Norwegian Institute for Nature Research
NINA head office
Postal address: P.O. Box 5685 Torgarden,
NO-7485 Trondheim, NORWAY
Visiting address: Høgskoleringen 9, 7034 Trondheim
Phone: +47 73 80 14 00
E-mail: firmapost@nina.no
Organization Number: 9500 37 687
http://www.nina.no

NINA

Cooperation and expertise for a sustainable future