Evaluating the Potential of Machine Learning to Automate Deforestation Mapping in Guyana

MATTHEW GREGORY WIECEK

Evaluating the Potential of Machine Learning to Automate Deforestation Mapping in Guyana

By Matthew Wiecek

Supervised by

Danny Donoghue

Isabella Bovolo

Camila Caiado

Dissertation submitted

For the PhD in Geography

Department of Geography

Durham University

ACKNOWLEDGEMENTS

Many thanks to the Guyana Forestry Commission for their support. They have provided invaluable data, they have provided technical support, and they have provided logistical assistance in the field. Thanks to the Norway International Climate and Forests Initiative for their financial support to Guyana.

The European Space Agency's Third Party Mission Programme provided funding for two Radarsat-2 images, and DLR provided funding for the acquisition of TerraSAR-X images.

My thanks go to my supervisors for their years of guidance and patience, for the funding opportunities that they have provided, and for the industry contacts and work experience that they have provided.

I also thank the examiners, Nick Cox and Pete Watt, for taking the time to read my work in depth and provide valuable feedback, both to improve this dissertation and suggestions for future work.

Many thanks go out to my parents for their support both during and before my PhD.

Many thanks go out to the volunteers who made NumPy, Pandas, Scikit-Learn, Scikit-Optimize, RasterIO, and their dependencies. For years to decades, they have poured a vast amount of work into creating the foundation of modern data science, and a world of research has been made possible by the numerical and machine learning software packages that they have created.

Finally, I thank all of the friends I met for their part in making this such an overall positive time.

ABBREVIATIONS

ART: Architecture for REDD+ Transactions

GFC: Guyana Forestry Commission

REDD+ : Reducing Emissions from Deforestation and Forest Degradation

LiDAR: Light Detection And Ranging

SAR: Synthetic Aperture Radar

TREES: The REDD+ Environmental Excellence Standard

TABLE OF CONTENTS

ABSTRACT

Global forest cover has decreased by 12% from 2001 to 2022, and the rate of loss is increasing (Global Forest Watch), which has consequence for carbon flows and ecosystem health. A UN framework, Reducing Emissions from Deforestation and Forest Degradation (REDD+), seeks to provide financial support to developing countries to measure and mitigate deforestation at the country-scale using satellite images. Deforestation measurement in the context of a REDD+ program would benefit from automation via machine learning – that would be able to measure deforestation in countrywide data quickly and accurately.

Existing attempts at using image classification to measure deforestation in satellite images have yielded mixed results with regard to REDD+ reporting requirements. This was done using a limited range of satellites, and a very limited range of classification algorithms. However, the remote sensing literature has much research on the effectiveness of and combinations of different satellite sensors and algorithms that can be applied here. This research compares a broader range of classification algorithms and satellite sensors (both multispectral and SAR) and combinations of satellites, tested in a range of deforestation drivers in Guyana, a country with a large proportion of mixed tropical rainforests and an advanced, experienced REDD+ program.

When comparing three algorithms (Random Forest, Gradient Boosted Trees, Naïve Bayes), with default probability thresholds for prediction decisions, on Sentinel-2 satellite data, the overall accuracies were 96%, 95% and 93%, respectively. When comparing satellites, the overall accuracies were 77% (ALOS-2), 69% (Sentinel-1), 90% (Landsat), 87% (Sentinel-2), 70% (RapidEye), 89% (Planetscope). The overall accuracy when classifying sedimented rivers was 96%, 99% for clear rivers, and 68% when the model was trained on sedimented rivers and tested on clear rivers. When classifying mining sites with no vegetation inside the mine, the Consumer's Accuracy is 91% for forest and 94% for mining, but when classifying mining sites with vegetation inside the mine, the Consumer's Accuracy is 98% for forest and 48% for mining. In all cases, manually tuning the class probability threshold for prediction decisions away from 50% created a map of deforestation that followed the true labels very closely. These results indicate that, in addition to choosing the algorithm and sensor, image classification of deforestation in a REDD+ context must account for the optimal probability threshold and the within-class variation.

CHAPTER 1: INTRODUCTION TO THE MEASUREMENT OF DEFORESTATION

SECTION 1.1 INTRODUCTION TO THE TOPIC AND OBJECTIVES

From 1990 to 2009, deforestation is estimated to have released ~1.4 PgC/yr into the atmosphere, and the bulk of this deforestation has been in the carbon-dense tropical rainforests of South America and Africa (Houghton, 2012; with a broader review in IPCC, 2015). Tropical forest ecosystems have carbon stocks of 306-324 petagrams of carbon (Mackey et al., 2020), and potentially have the ability to sequester 120 petagrams of carbon by 2100 (Houghton and Nassikas, 2018). Tropical land use changes, such as increases in agriculture, logging and mining, are causing extensive deforestation and forest degradation (Pan et al., 2011). As trees are cut down, solid carbon is released into the atmosphere as carbon dioxide, flooding and water-related diseases increase, and agriculture and ecosystem services are harmed (Bovolo et al., 2018, Gentine et al., 2019). Deforestation and forest degradation are currently the second largest anthropogenic source of carbon after fossil fuels, at 20% of global emissions (Van der Werf et al., 2009), and tropical deforestation and forest degradation are key parts of this (Houghton, 2012). This is why deforestation is a vital policy matter: forests provide essential services, and deforestation deprives people of those benefits, and deforestation has the potential to destabilize the climate system by reducing carbon sequestration, releasing large amounts of carbon very quickly, and altering bio-physical processes such as albedo and evapotranspiration. Reforestation, on the other hand, has the potential to help mitigate climate change (Psistaki et al., 2024).

A UN framework, Reducing Emissions from Deforestation and Forest Degradation (REDD+), seeks to provide financial support to developing countries to measure and mitigate deforestation at the country-scale using satellite images. Deforestation measurement in the context of a REDD+ program would benefit from automation via machine learning – that would be able to measure deforestation in countrywide data quickly and accurately.

Existing attempts at using image classification to measure deforestation in satellite images have yielded mixed results with regard to REDD+ reporting requirements. This was done using a limited range of satellites, and a very limited range of classification algorithms. However, the remote sensing literature has much research on the effectiveness of and combinations of different satellite sensors and algorithms that can be applied here. This research compares a broader range of classification algorithms and satellite sensors (both multispectral and SAR) and

combinations of satellites, tested in a range of deforestation drivers in Guyana, a country with a large proportion of mixed tropical rainforests and an advanced, experienced REDD+ program.

The overall aim of the study is to identify the classifiers, datasets and label sets that can produce the accuracy needed to automate REDD+ measurement and monitoring. This can be broken down into specific research questions:

1. How does land cover type mapping accuracy change between classifiers? This can start with single classifiers, and then test ensembles of classifiers.
2. How does land cover type mapping accuracy change between satellite sensors and combinations of sensors? This can start with individual sensors, and then move on to data fusion.
3. How does the complexity of the land cover types affect accuracy? This can be tested by classifying two land cover types, then three and four, and then all land cover types.

In the best case, this research will find a combination of classifiers, data and labels that can identify changes in difficult-to-separate land cover types in near real time. The most immediate impact will be on the Guyana Forestry Commission, as they will be the first to see the results. This is also applicable to other REDD+ countries, some of which have tried to implement machine learning without success. This also has applications to land cover mapping in general.

This case study also has implications for machine learning on satellite images. It evaluates classifier performance using real world data from a variety of satellite sensors, and the class labels are based on higher quality reference data than most research, which will lead to an improved ability to identify and interpret errors.

The rest of Chapter 1 will introduce the problem of tropical deforestation measurement and reporting in the REDD+ context, and introduce the value of automating this process using machine learning. Chapter 2 will explain how this fits into the Guyana Forestry Commission's REDD+ reporting requirements. Chapter 3 will introduce the data that will be used. Chapters 4 to 8 will report results. Chapter 4 is a comparison of classification algorithms. Chapters 5 and 6 will test multispectral and radar images at two location in Guyana. Chapter 7 will test combinations of multispectral and radar images. Chapter 8 will test the ability of machine learning to identify a range of deforestation drivers. Chapter 9 will discuss the issues raised in the previous five chapters, and their implications for machine learning and deforestation measurement.

SECTION 1.2 MOTIVATING RESEARCH ON DEFORESTATION MEASUREMENT

Accurate and efficient forest monitoring is vital, because forests are important regulators of the global climate, and form critical parts of the global carbon cycle, acting both to sequester and to store carbon. Carbon dioxide is removed from the atmosphere by tree leaf photosynthesis, and sequestered in solid plant matter as the plant grows, which eventually goes into the soil as forest litter, and the carbon returns to the atmosphere by natural decay or by burning (see Malhi and Grace, 2000 for details). Deforestation accelerates the latter part of the cycle when felled trees burn or decay, increasing the amount of atmospheric carbon. Timber harvesting can delay this, if the wood harvested does not decay or burn immediately. Reforestation accelerates the former part of the cycle, removing atmospheric carbon, but over longer periods of time. Forests are therefore an important carbon sink, especially in the tropics, which are estimated to account for 55% of global forest carbon stocks (Pan et al., 2011). Forests have a powerful effect on atmospheric carbon dioxide concentrations and hence global temperature (Reichstein and Carvalhais, 2019, Nunes et al, 2019, Zhao et al, 2022). This is especially true of tropical forests, which in addition to sequestering carbon, also cool the air through evaporation. See Bonan (2008) for an overview, and see Bovolo et al. (2018) for an example of how forests have long-distance effects on climate via evapotranspiration. This makes forests vital to solving social, environmental and climate change problems, and means that forests need to be protected and restored.

In addition to climate regulation, forests provide a wide range of other ecosystem services. Examples include: timber and fuel production, nutrient retention, biodiversity protection, ecotourism, and spiritual and traditional values (Aznar-Sánchez et al., 2018), along with clean water, water availability, landslide prevention, flood mitigation, mitigation of coastal waves, fire resistance, food security, control and avoidance of disease, medicine, and pollution mitigation (Brandon, 2014, Rajasugunasekar et al., 2023). Reducing deforestation therefore has additional benefits beyond climate change alone.

An important international effort to help reduce the rate of deforestation is REDD+ (**R**educing **E**missions from **D**eforestation and Forest **D**egradation). REDD+ is a United Nations Framework Convention on Climate Change (UNFCCC) framework for financing, agreements between a developed country and a developing country, adopted in 2013 at COP19, where developing countries receive financial support through carbon-credits for maintaining standing forests and

avoiding deforestation and degradation. This can provide funding from the international community to maintain forest ecosystems and their services for the global benefits of climate change mitigation, instead of promoting the clearing of forests for economic development and the income it brings at the national scale. In order to receive carbon-credit funds, countries are responsible for monitoring and quantifying the extent of deforestation prevented through the Measurement, Reporting and Verification (MRV) process. As part of this process, countries find it beneficial to measure and report changes in forest cover using remote sensing data which covers wide areas, much larger areas than it is possible to monitor on foot, at a much cheaper price, in a standardised and timely manner.

Investments in REDD+ were initially mostly between two countries as bilateral agreements, or as multi-lateral agreements through third parties (e.g. World Bank Carbon Fund, or the Green Climate Fund), which helped set up the financing framework. More recently, in order to meet more general carbon-credit forest financing markets, a new international carbon crediting program, has been set up for financing national REDD+ programs, **A**rchitecture for **R**EDD+ **T**ransactions (ART, 2021), which was set up in 2018. ART's stated mission is to finance global-scale action on climate change, incentivize governments to achieve results, and serve as a global quality benchmark to provide confidence in the integrity of REDD+ reporting and financing (ART, 2021a). This has recently come to include provisions for financing both restoration of destroyed forests in countries that have seen extensive deforestation, and preservation of existing forests in **H**igh **F**orest **L**ow **D**eforestation (HFLD) countries, where deforestation has been minimal.

The **R**EDD+ **E**nvironmental **E**xcellence **S**tandard (TREES) is ART's standard for Monitoring, Reporting and Verification. TREES was first published in 2020, and TREES version 2.0 was published in 2021 (ART, 2021a), and maintains compatibility with earlier guidelines, agreements and frameworks. Under ART-TREES, a REDD+ country creates a standard operating procedure that meets the requirements of ART and accepted standards for emissions reporting based on remote sensing. ART-TREES accounting requirements include activity data, emission factors and removal factors, a monitoring plan, uncertainty quantification, and environmental, social and governance safeguards.

Examples of those earlier good practice guidelines are the **G**lobal **O**bservation of **F**orest **C**over and **G**lobal **O**bservation of **L**and **D**ynamics (GOFC-GOLD) of the Global Forest Observations Initiative (GFOI). REDD+ countries have a lot of flexibility in how they design their standard

operating procedures, but these guidelines will help a country ensure that it passes audits. Important aspects of remote sensing practice standards include (GOFC-GOLD, 2016):

1. Selecting the forest definition and designating the forest area. There is variation in the definition of forest, which determines what changes are counted. Section 1.2 of this research goes in depth into this.
2. Selecting satellite data. Concerns include sensor type (multispectral, SAR, etc.), spatial resolution (the smallest changes are not visible in lower resolution images), cost, availability, etc. Section 1.3 explains what factors go into selecting data, and what options are available.
3. Decision on sampling areas or complete coverage of a country/jurisdiction.
4. Processing and analyzing the satellite data.
5. Accuracy assessment.

Some countries have used machine learning to automate deforestation mapping (such as Mexico and Indonesia, in Sections 1.5 and 1.6), and some countries use humans to interpret satellite images (such as Guyana) – but which is better? Humans have one major advantage over machine learning: humans can use context clues to identify the location and cause of land cover change, and can understand the objective and make decisions about how to achieve that. This is especially important because errors have financial and political consequences. False positives – finding deforestation where there is none, so overcounting – will reduce a country's financial support and political standing. False negatives – failing to find deforestation events – will reduce a country's ability to respond to deforestation, and will reduce the perceived integrity of REDD+ work in the eyes of the world.

But machine learning has its own advantages:

- Machine learning can classify land cover types across an entire country very quickly, far more quickly than a human can. This may enable a country to detect deforestation events in near real time and respond quickly to stop the activity causing the land use change. In the case of illegal activity, the government may be able to respond while the activity is ongoing, instead of detecting it the year after it happened.
- Humans are limited to viewing what a computer screen can display. In remote sensing, bands are the slices of the electromagnetic spectrum that a sensor can record. Colour photographs have three bands: red, green, and blue. A camera's sensor records those

three slices of the visible light spectrum. Satellite sensors have other bands (Red Edge, Near Infrared, Shortwave Infrared, Yellow, SAR polarizations, etc) that record other slices of the electromagnetic spectrum, but only three can be used at a time. Therefore, a human can see only part of the total amount of information recorded in remote sensing data, or has to cycle through a series of band combinations. Machine learning can work with all of the available information. Every band of an image can be processed at once, whether the satellite has four bands or eight or more. Not only that, but images from different satellites can be combined to allow the machine learning algorithm to process all information and combinations of information that would not be feasible for a human operator.

- Machine learning can output both class predictions and class probabilities. Low probabilities indicate high uncertainty, and high probabilities indicate high certainty. Machine learning can therefore quantify uncertainty in a basic way as part of the process.

The three key components of machine learning are:

- Model/Classifier: The model tells the computer how to make predictions from data. The classification algorithm determines the characteristics of the model and how the model is applied to the data.
- Data: The data are the images recorded by a satellite sensor. The image can be a multispectral image (visible light and infrared radiation), or it can be a SAR image (an active sensor that uses microwaves to image a surface).
- Labels: The land cover type that each pixel belongs to. In the "true" labels, each pixel is assigned a true class (based on interpretation of reference data, aided by ground truthing). The "predicted" labels are those generated by the model. In classification in general, each data point has a class associated with it, and in this case, the classes are the land cover types.

Reference data: This is data for a specific purpose, and this supports the creation of class labels and accuracy assessment. Reference data are data of the highest available quality that can be used to create labels that accurately reflect the ground truth.

Some notes about terminology:

- In the context of neural networks, image classification assigns one label to a whole image, and image segmentation assigns a label to each pixel in an image. In other machine learning contexts, image classification assigns a label to each pixel in an image. In this dissertation, image classification refers to the latter; i.e. image classification will assign a label to each pixel in an image.
- In tutorials and guides on machine learning, the term classifier can be used to describe both the model and the algorithm. They are not the same thing, though: the model is the rule that is created using data and an algorithm. Here therefore, the term classifier will be used to describe the algorithm.

The basic steps in the machine learning process are:

1. Class label creation: Using the reference data as a guide, each pixel in the data is assigned a label that indicates its true class.
2. Training/fitting: This step uses the data allocated for training, and the labels associated with that data, to build a classifier. The step-by-step process is determined by choosing a classification algorithm.
3. Validation: This step uses data that were not seen in the training step, and the labels associated with that data, to measure the accuracy of the model/classifier. The classifier is run on the validation data, and outputs predictions. Those predictions are then compared with the labels, and the accuracy is measured.

Accurate machine learning requires high quality training and validation data, labels based on high quality reference data, and the model must use an algorithm that is appropriate to the data and the research question. All three must be present; good data and labels will not work with a poorly chosen model, and a model is only as good as the data and labels given to it.

The literature on classification algorithms and the literature on data quality in the context of remote sensing of forests are extensive already, with many classification algorithm comparisons and satellite sensor comparisons already published. Reviews of data fusion can be found in Pohl and van Genderen (1998); Wang et al. (2005); Pandit and Bhiwani (2015); Ghassemian (2016); Kulkarni and Rege (2020). Examples of classifier studies include Maxwell et al. (2018) on the contradictory results of such studies, Rogan et al. (2008) and Lippitt et al. (2008) comparing neural networks with decision trees, and several studies comparing Random Forest with Support Vector Machines (Pal, 2007, Adam et al., 2014, Zhang and Xie, 2013; Maxwell et al.,

2014a; Maxwell et al., 2014b; Maxwell et al., 2015). There is also research on the importance of the quality of reference data that the results are measured against (Stehman and Foody, 2019; Foody 2009, 2010, 2011, 2013; Foody et al., 2016). This means that the core components of machine learning (algorithm selection and high quality class labels) are well studied.

What the literature currently lacks is research that tackles some specific challenges of REDD+ monitoring and reporting head-on. These challenges include:

- The classification algorithm tests in the previous paragraph tend to study one spectrally homogeneous study area. However, country-wide deforestation measurement will require a classifier that can recognize the same land cover types in many variations. Rivers can be clear or turbid. Mining sites can be active or abandoned. These have varying amount of new vegetation, varying amounts of bare surfaces, and mining ponds of many colours. This study will test classifers and remote sensing satellites in not just one study area, but multiple, and report on the differences in results.
- The creation of class labels for each land cover type and activity. REDD+ land cover classes are spectrally complex and heterogeneous, which results in a large amount of overlap that results in confusion between land cover classes. The mix of confusion matrices and maps of class predictions and probabilities will provide a detailed understanding of the nature of this problem.

When quantifying forest change, it is critical to define what counts as forest, as this will affect the measurement, and there are variations in definitions. The Food and Agriculture Organization of the United Nations defines forest as half hectare area with at least 10% canopy cover and tree height greater than 5 meters. The United Nations Framework Convention on Climate Change allows for some flexibility: individual countries can decide that a hectare must have 10% to 30% canopy cover to be forest, with a tree height of 2-5 meters, as per the Marrakesh Agreement.

This variation in definitions means that the interpretation of forest change figures depends on what each country is counting as forest. Romijn et al. (2013), using Indonesia as a case study, found that forest stock levels are 18% higher when using a "natural forest definition" rather than the local definition and 27% higher when using the FAO's definition of forest cover.

Designating a forest area means deciding what areas will and will not be mapped for deforestation. For example, areas designated for indigenous peoples may be excluded because the practice of shifting cultivation (clearing small areas of forest land for cultivation and then allowing it to revert to natural forest) is considered as sustainable and not a change in land use (Heinimann et al., 2017).

At first glance, the process of deforestation mapping is straightforward: Using remote sensing data, map areas that are forest and areas that are not forest. However, the previous two points indicate that forest loss figures, such as the Global Forest Watch figure cited above, require a close look at the data and assumptions behind the figures is needed. Other issues that create uncertainty are:

- The spatial resolution (ground area covered by each pixel) of the image determines the amount of deforestation that will be detected – small deforestation events will not be detected at Landsat's spatial resolution/pixel size (which is what Global Forest Watch primarily uses), but many of these small events add up to a large discrepancy.
- The date of acquisition of the image matters due to seasonality of forest ecological processes, such as fruiting and leaf shedding.
- Cloud cover in very cloudy areas can mask terrain cover, and therefore create gaps in the data by hiding the area under the clouds.

The extensive literature on data fusion may provide a way to increase the amount of information available to the classifier to help it separate complex classes, and the extensive literature on classifier comparisons will provide options to find the appropriate algorithm for this problem. This study will start by examining the problem of separating classes in simple and complex situations. It will then examine the effect of data fusion, and end with a classifier comparison, using real-world REDD+ data.

In summary, forests are essential to the future prosperity of humanity and the environment. Forests are being destroyed rapidly, but mitigation is possible by measuring and responding to land use change, if accurate and precise data can be collected on the location and extent of deforestation. The path toward this includes tests of classifiers and sensors singly and in combination (which have been done extensively) in the wide range of study areas that the classifier will encounter (which has not been done extensively). Chapters 2 and 3 will explain why Guyana, with its REDD+ program, is a suitable place for such research. The chapters after

that will provide results on classifier tests using several study sites across Guyana, eight satellite sensors, and several classification algorithms across two platforms (Google Earth Engine and Python).

SECTION 1.3 DATA USED IN DEFORESTATION MAPPING

For compatibility with GOFC-GOLD and ART-TREES standards, the data used to map deforestation consists of satellite images, along with some aerial photography for use as reference data. This section will examine factors to consider when selecting satellite images for classification purposes: the type of sensor, the spatial resolution, availability, and cost.

DIFFERENT SENSOR TYPES

There are a wide variety of different satellite remote sensing sensor types, which record different information, and therefore their usefulness in deforestation mapping varies.

First, some elaboration on different sensor types is needed. Satellite remote sensing uses electromagnetic radiation to interact with a surface on Earth and record the reflectance or backscatter with an appropriate sensor. Different materials have different interactions with electromagnetic radiation (EM), which makes different information about the surface visible to different sensors. The different sensors are:

- Multispectral sensors record visible and infrared light that comes from the Sun and is reflected by vegetation and other surfaces on the Earth's surface. Different wavelengths provide information on different aspects of the surface.
- SAR is an active sensor type. This means that SAR emits its own microwave energy, and then records the intensity of the microwave backscatter from the ground surface and return to the sensor. Microwaves emitted and received by SAR sensors are sensitive to water content, molecular structure and dielectric properties.

Here are some important electromagnetic interactions in remote sensing of forests:

- Plant pigments such as chlorophyll a & b and carotene absorb EM radiation most strongly at blue and red wavelengths.
- Near Infrared EM radiation is strongly reflected by vegetation due to refraction at cell wall interfaces. Leaf cell structure has low reflectance in visible light and very high reflectance in at near infrared wavelengths. Water, on the other hand, absorbs strongly

at near and shortwave infrared wavelengths. In Marais Sicre (2020), the near infrared band had the best performance of the optical bands in separating crops in agriculture remote sensing.

- The Red Edge bands are in the transition between the visible and near infrared wavelength regions.
- Vegetation moisture content has two absorption regions, corresponding with the two shortwave infrared bands. In Ferreira et al. (2019), the separation of tree species in Brazil was improved when shortwave infrared bands were added to visible light and near infrared bands. Including shortwave infrared bands improved detection of non-photosynthetic forest litter that varies from one species to another. They also note that shortwave infrared has unique absorption features for biochemical markers such as nitrogen, water, cellulose and lignin (Kokaly et al., 2009).

Both multispectral sensors and SAR sensors are therefore useful, and they are useful in different ways. This opens up the possibility of using both in combination.

DIFFERENT SPATIAL RESOLUTIONS

Satellite sensor images are made up of various pixels which have a particular size, the ground sample distance (GSD) that it covers, which is consistent across the image after the raw data has been processed. This determines the size of the details that can be resolved, and the size of the data. Mixed pixels happen when a land cover boundary goes through the middle of a pixel, which is more common at lower resolutions.

Low resolution sensors (such as AVHRR, MODIS, MERIS) have pixel GSD that cover a very large area, meaning that an area of up to 1 km^2 only has one data point and therefore average values are derived for that pixel. These satellites are therefore unable to resolve small areas, and miss small objects but can span the entire world, staying within the computational resources available to most researchers.

Medium spatial resolution sensors (such as Landsat, Sentinel-1, Sentinel-2) have pixels that cover a much smaller GSD than the low resolution sensors and can resolve much smaller details. The Global Forest Watch figure cited above is based on Landsat data with 30 m GSD.

High resolution satellites have small enough pixels that they can resolve smaller and narrower features like roads. Examples include RapidEye (5 meter pixels), Planetscope (3 meter pixels), and SPOT (1.5 meter pixels after pan-sharpening). Pixel sizes are for the surface reflectance product.

Very high resolution sensors (Worldview, SkySat, Pleiades, IKONOS, etc.) can see the smallest details, and can be used as reference data, but the file sizes are very large, and the camera may be angled away from looking straight down, creating distortion in the image. These factors mean that higher resolution is not always better. Increasing the resolution increases the computational requirements exponentially, and if all relevant details are visible at lower resolution, then lower resolution data may be more suitable for analysis.

COST

Some remote sensing data products are available for free, and some data products are commercial and require funding to access them. The availability of sufficient funding is a major factor in determining which satellites can be used, as some are prohibitively expensive. Some satellite data is available free of charge, such as Landsat and the Sentinel missions. Satellite data products from commercial providers, such as Planet, Airbus and Maxar, require resources in order to task acquisition and to retrieve already acquired imagery from their commercial archive. Developing countries need to mitigate deforestation but have limited financial resources, which will affect the recommendations at the end of this research. When free image products are equal in accuracy to commercial products, the free products will be recommended. Commercial products will need to be higher in accuracy to justify the added expense.

DIFFERENT TIME PERIODS

Civilian Earth observation began in 1972 with the launch of ERTS-1, the satellite that would come to be known as Landsat 1. In the 53 years since, many satellites have come and gone. For contemporary mapping, satellites that are currently operational are needed, and for historical mapping, the satellites that were operational at the time are needed. Table 1 shows which satellites are available now and have been available in the past.

Table 1: Time periods covered by some commonly used satellites.

Medium/High Resolution Multispectral Satellites		**SAR Satellites**	
Landsat Multispectral Scanner	1972-2012	Radarsat-1	1995-2013
Landsat Thematic Mapper	1982-2012	ALOS PALSAR	2006-2011
SPOT-1	1986-1990	Radarsat-2	2007-present
SPOT-2	1990-2009	TerraSAR-X	2007-present
SPOT-3	1993-1997	ALOS-2	2014-2024
SPOT-4	1998-2013	Sentinel-1	2014-present
CBERS-1	1999-2003	Radarsat Constellation	2019-present
MODIS-Terra and MODIS-Aqua	1999-present	COSMO-SkyMed	2007-present
Landsat Enhanced Thematic Mapper	1999-2021		
CBERS-2	2003-2009	**Very High Resolution Multispectral**	
SPOT-5	2003-2015	IKONOS	2000-2015
Resourcesat-1	2003-present	QuickBird	2001-2015
ALOS AVNIR	2006-2011	Worldview-1	2007-present
RapidEye	2008-2020	GeoEye-1	2008-present
Resourcesat-2	2011-present	Worldview-2	2009-present
SPOT-6	2012-present	Worldview-3	2014-present
Landsat Operational Land Imager	2013-present	Worldview-4	2016-2019
Planetscope Dove Classic	2014-2022	SkySat	2013-present
SPOT-7	2014-2023		
CBERS-4	2014-present		
Sentinel-2 Multispectral Imager	2015-present		
Planetscope Dove R	2019-2022		
CBERS-4A	2019-present		
Planetscope SuperDove	2020-present		

SECTION 1.4 REDD+ AS A SOLUTION TO DEFORESTATION

Implementing remote sensing-based tools to the problem of deforestation requires policy infrastructure that will enable developing countries to acquire the financial, material and technological resources required for monitoring deforestation. The United Nations Framework Convention on Climate Change has responded to this problem by creating the REDD program (Reducing Emissions from Deforestation and Forest Degradation), which was later expanded to REDD+ (the + is about sustainable forest management and restoration of forest carbon stocks) (UN-REDD, 2020). This is a global program where developed countries provide results-based financial support to developing countries for the purpose of limiting deforestation. Here are the key points in the development of REDD+:

- In 1994, the UNFCC pledged to 'stabilize green-house gas concentration', 'within a time frame sufficient to allow ecosystems to adapt naturally', formalised under the 1997 Kyoto Protocol, and 2013 Doha Amendment, which established carbon-trading protocols.
- In 2005, Costa Rica and Papua New Guinea led discussion of REDD (Reducing Emissions from Deforestation and Degradation) at COP11.
- In 2007, the Bali Action Plan was adopted at COP13 (UNFCCC, n.d.). It includes a shared long-term goal for emissions reduction, as well as components for mitigation, adaptation, technology and financing.
- In 2010, the Cancun Safeguards were adopted (UN-REDD, 2024). They include:
    - That actions are complementary or consistent with the objectives of national and international programs.
    - Transparent and effective governance structures.
    - Respect for knowledge and rights of Indigenous and local communities.
    - The full and effective participation of stakeholders.
    - Actions are consistent with conservation.
    - Actions to address the risk of reversals.
    - Actions to reduce displacement of emissions.
- In 2013, the Warsaw Framework was adopted at COP19 (UNFCCC, 2024). This is a series of decisions (some adopted at COP21 in 2015) concerning national Monitoring, Reporting and Verification (MRV) systems.

- In 2015, Article 5 of the Paris Agreement called for parties to conserve and enhance carbon sinks and reservoirs, and to implement results-based payments for reducing deforestation and forest degradation (United Nations, 2015).
- In 2022, COP26 initiated **A**rchitecture for **R**EDD+ **T**ransactions (ART), "a standalone, independent program that develops and administers standardized procedures for crediting emission reductions and removals from national and large sub-national REDD+ programs" (ART, 2021). Its standard for monitoring and verification of emissions removals is The REDD+ Environmental Excellence Standard (TREES).
- The LEAF Coalition was established to bring together forest governments, donor governments and corporate donors to enable REDD+ financing through the sale of carbon credits (LEAF Coalition, n.d.). These credits must meet the standard set by ART-TREES.

By 2019, 39 countries were participating in REDD+ agreements. As of 2020, the status is (Forest Carbon Partnership, 2023):

- One country where ART Trees has been approved at time of writing – The Cooperative Republic of Guyana
- 6 countries that are eligible to receive results-based financial support.
- 12 countries that have submitted a national strategy or action plan.
- 15 countries that have submitted a summary of information on how safeguards are being addressed and respected.
- 50 countries that have submitted a REDD+ forest reference level or forest reference emission level for technical assessment to the UNFCCC, covering more than 70% of the total forest area in developing countries.

When analyzing satellite images to map forest change, the default analytical method is visual interpretation. This is the easiest to implement, and experienced operators can classify images with high accuracy. However, visual interpretation is very time consuming and would benefit from automation using machine learning. Classification and segmentation of images is a well-established task with a large scientific literature documenting research and industry experience which can be applied to the problem of deforestation mapping. The rest of this chapter will introduce three examples of land use change mapping from Mexico, Indonesia, and Guyana.

They will illustrate the challenges of automating land use / deforestation mapping, as well as the potential for solutions.

SECTION 1.5 MEXICO'S AUTOMATED DEFORESTATION MAPPING SYSTEM

One example of an automated REDD+ MRVS is Mexico. In the early 2010s, Mexico adopted a fully automated deforestation detection system, called Monitoring Activity Data in Mexico, or MAD-Mex. The methodology looks like this: (Gebhardt, n.d.)

- The historical baseline was made using Landsat, which had 40 years of data in 2012. In 2012, their plan for the future was to incorporate high resolution RapidEye images into the workflow.
- The image processing pipeline ends with cloud masked Landsat images, four vegetation indices, a digital elevation model with aspects and slope, and descriptive metrics for each band in the time series (Minimum, Maximum, Average, Range, Standard Deviation).
- The prepared images are then segmented using object labels and vectorization, and object classes are mapped with the help of reference vector cartography. Objects are classified using 10-fold boosted decision trees.
- Accuracy assessment is done using reference field inventory points. Validation was performed at the national level because there were too few samples per Landsat frame. Only those forest inventory points were used where each sub-plot has the same land cover type. Random area-weighted stratified sampling was done, using the land cover classes as strata.
- The reference data comes from the National Institute of Statistics and Geography of Mexico (INEGI), which has published national vegetation maps at a scale of 1:250.000 with a 25 ha minimum mapping area.
- The mapping process includes manual preliminary interpretation of satellite images, field verification in selected sites, reinterpretation based on field verification and comparative analysis, analysis of and integration of information to object polygons, a map which has layers on vegetation types, vegetation ecosystems, agricultural usage, and other information.

Before moving on, a word about measuring accuracy using confusion matrices:

- Interpretation of Confusion Matrix: The rows and columns represent the true labels and predicted labels. In data science in general, which one represented which is chosen by the researcher and labelled, and in remote sensing, the convention is to have the rows represent true labels and the columns represent predicted labels. In this research, the remote sensing convention is followed. The diagonal of the matrix is the correct predictions, and the upper and lower triangular matrices around it are the incorrect predictions, with the true class and falsely predicted class indicated.
- Usefulness of Confusion Matrices: The goal of this research is to understand the ability of different classifiers and sensors to classify deforestation drivers, and a confusion matrix provides the most detailed class-by-class breakdown of the errors.
- Interpretation of Accuracy: In terms of a confusion matrix, the accuracy is the number of pixels in the diagonal divided by the number of pixels in the confusion matrix. Accuracy is strongly influenced by class imbalance. A classifier where the majority class is predicted very accurately, and the minority class is predicted very inaccurately, will have high accuracy because the majority of pixels were classified accurately.
- Usefulness of Accuracy: Accuracy is a simple way of summarizing the overall performance of a classifier, and accuracy scores are easy to compare. Accuracy scores do not provide details of which classes the classifier handled well and which it didn't, and accuracy is sensitive to unbalanced data.
- Interpretation of Consumer's Accuracy: This indicates how many predictions in each class were correct.
- Interpretation of Producer's Accuracy: This indicates how many examples of each class were found.
- Usefulness of Producer's and Consumer's Accuracy: These are two ways of summarizing the per-class performance of a classifier, and they can be easily added to a confusion matrix as an extra row and column. In this way, they show the number of predictions by class that were correct, and the number of samples of each class that were found, with the confusion matrix entries that they're summarizing.

Table 2: The components of a confusion matrix.

		True Labels		Consumer's Accuracy
		Forest	Mining	
Predicted labels	Forest	Correctly identified forest	Mining identified as forest	Percentage of forest predictions that are correct
	Mining	Forest identified as mining	Correctly identified mining	Percentage of mining predictions that are correct
	Producer's Accuracy	Percentage of true forest pixels that the model found	Percentage of true mining pixel found	Overall Accuracy

Table 3: A demonstration set of predictions. This is the data source for the example confusion matrix in Table 4 below.

True Label	Forest	Mining	Forest	Mining	Forest	Mining	Forest	Mining
Prediction	Forest	Mining	Mining	Mining	Forest	Forest	Forest	Mining

Table 4: A confusion matrix derived from the data in Table 3.

		True Labels		Consumer's Accuracy
Predicted labels		Forest	Mining	
	Forest	3	1	75%
	Mining	1	3	75%
Producer's Accuracy		75%	75%	75%

Gebhardt et al. (2014), representing four Mexican government agencies, assessed the accuracy of its mapping using the data and workflow described above. Three types of accuracy were measured:

- The overall accuracy is the percentage of predictions that were correct. This ranged from 62% to 65.7%.
- The Producer's Accuracy is the percentage of pixels in each class that were predicted. This ranged from 53.4% (Temperate Mixed Forest) to 77.5% (Tropical Evergreen Forest).
- The Consumer's Accuracy is the percentage of predictions in each class that were correct. This ranged from 30.5% (Wetland) to 91.1% (Scrubland).

One thing to note about that accuracy is that this is not sufficient to fully automate deforestation mapping. The number of errors is significant enough that the output would need a lot of manual inspection and correction. This could negate the labour saving that automation would ideally provide.

Another issue is that the accuracy assessment is biased, because it is based, not on a random sample distributed across the country, but on government field work (specifically, the National Institute of Statistics and Geography in Mexico). Government fieldwork is not distributed randomly across the country, and so the sample is likely to be biased and not reflect the true population. Government fieldwork is instead driven by policy priorities and accessibility of sites, which introduces bias. For example, the reference data is biased toward large-scale permanent agriculture, and a lack of representations of small-scale agriculture. Mas et al. (2014) redid the analysis in Gebhardt et al. (2014) and found that, when a truly random sample was used, the accuracy was much lower.

Mas et al. (2014) also point out that the change figures presented are implausible. The change over the 18 years from 1993 to 2008 is 40%, and the changes from 2002 to 2005 and 2005 to 2008 are 43%. This is the consequence of building a classification model using low quality reference data and biased samples.

Finally, Mas et al. (2014) argue that a fully automated approach will not produce sufficiently accurate results. It relies exclusively on spectral data, leading to much confusion. They argue that automation must assisted rather than replace human decision making.

That last point is a valid argument against relying too much on the system presented in Mexico, and other systems that rely exclusively on multispectral data. However, machine learning is fully capable of incorporating data from other sources, such as SAR, Lidar, elevation models, hyperspectral sensors, and ancillary GIS data. The last problem that Mas et al. identify is there one that can easily be solved with further research. The next two sections will look at two other case studies in automation of deforestation mapping: Indonesia and Guyana.

SECTION 1.6 INDONESIA'S AUTOMATED DEFORESTATION MAPPING SYSTEM

Another country that has automated REDD+ MRV work is Indonesia. Based on the summary in Kustiyo et al. (2015), Indonesia has built its National Carbon Accounting System, which has two components: the remote sensing component and the emissions estimation component. The satellite data they use is a mix of Landsat 8 and SPOT 5/6/7. This was supplemented by information from experts, along with GIS data, maps and ground information. The ground truth data was satellite images with 1-2 meter resolution where the boundary between forest and non-forest is obvious.

Their workflow starts with image selection and preparation. One cloud-free image from each year is selected. One clear image is preferred, but rarely possible in a place as cloudy as Indonesia. Orthorectification, radiometric calibration, cloud masking and mosaicking are then applied.

The change mapping process starts with the creation of a forest base probability image using supervised classification, done with the help of expert input on land cover, and statistical procedures including canonical variate analysis. Stratification was used because of the variation in land cover across Indonesia. It was noted that a single classifier does not work on all areas around Indonesia. After the base mapping is done, forest cover probabilities for the other years are derived using a semi-automated matching process. In the multi-temporal classification stage, individual year classifications are combined and refined using Bayesian networks, using the entire time series, with weights related to accuracy, temporal change and neighbouring values.

So far, two efforts to automate deforestation mapping have been presented. They have shortcomings, related to the quality of reference data, methodological issues, and the problem of separating spectrally similar land cover types. There is a clear need for research to solve the

latter two problems, in a region that has higher quality reference data. This chapter will end with an introduction to such a region.

SECTION 1.7 GUYANA'S REDD+ PROGRAM

The first REDD+ program to implement results-based payments was the Guyana-Norway REDD+ Agreement (Guyana Forestry Commission, 2020). In the original agreement, Norway agreed to provide up to USD 250 million in results-based payments to help Guyana transition to a low-carbon green development-based economy.

The goal of the MRVS project is to contribute to Guyana's green development pathway by implementing the MRVS, reporting on the REDD+ Interim Indicators, and streamlining the REDD+ indicators. The Guyana Forestry Commission has approached this as a continuous learning project, and Guyana's MRVS has been expanded and refined considerably over the last ten years.

Guyana is a High Forest Low Deforestation country, where about 85% of the country is forest, and deforestation from 1993 to 2013 was 0.02% to 0.079% per year. The most common driver of deforestation is mining, which accounts for 60% of deforestation from 1990 to 2009 (pre-REDD+) and 90% from 2009 to 2012 (REDD+ period) (Bholanath and Cort, 2015). In a world where many parts of the tropical rainforest around the Amazon River are being rapidly destroyed, releasing large amounts of carbon into the atmosphere, Guyana has been the opposite. Their mining and other activities have destroyed a small fraction of their forests, and the Guyana Forestry Commission (GFC) has been dedicated to preserving the extensive forest that was never lost.

An automated workflow based on machine learning needs to be validated against a high quality benchmark before it can be deployed. Guyana is able to serve as this benchmark because it has not attempted to automate deforestation measurement and reporting process. Instead, Guyana has built a robust REDD+ program based on human interpreters who have 15 years of accumulated experience and institutional knowledge.

The benefits go both ways. Machine learning research would benefit from the resources and experience that the Guyana Forestry Commission has. Research on machine learning in remote sensing of forests benefits from a case study that offers high quality reference data, sufficient examples of different deforestation drivers, and opportunities for industry applications. Guyana

is a prime example of this. Guyana and Norway signed a REDD+ agreement in 2009, one of the first in the world, to be managed by the Guyana Forestry Commission (GFC). Since then, GFC has acquired thirteen years of experience in REDD+ measurement, reporting and verification. GFC currently holds six years (2017, 2018, 2019, 2021, 2022, 2023) of aerial photography and very high resolution satellite images of deforestation sites, which provides a first-rate set of reference data. GFC also stands to benefit greatly from the results of this research – detailed results on the optimal machine learning workflow will save a lot of labour and money in their efforts to monitor and mitigate deforestation.

In summary, Guyana's REDD+ program has the resources needed to do high quality machine learning resources and the ability to implement solutions that come out of this research. This means that Guyana is the ideal case study in an effort to solve the problems that have made automation difficult so far. This research is being done in the context of a successful REDD+ program. In October 2021, Guyana because the first country to have TREES registration documents accepted (ART, 2021b). In December 2022, Guyana received the world first ART-TREES credits (ART, 2022). The next chapter will go into more detail about deforestation and REDD+ methodology in the Guyana context.

SECTION 1.8 SUMMARY AND CONCLUSION

In summary, deforestation is a pressing global problem, with major implications for climate change and the humans who live in tropical forests. Dozens of developing countries are using remote sensing to measure deforestation and receive financial assistance under the REDD+ program and its associated policy frameworks (ART and TREES). Currently, this is done manually by humans who interpret countrywide remote sensing data. If it were possible to automate this work with machine learning, then a lot of labour would be saved in measuring deforestation, labour that could be redirected toward responding to deforestation.

There have been previous efforts to automate deforestation measurement with machine learning. The main problem is that the accuracy has not been high enough for deforestation mapping. The consequences of errors are great. False positives will overestimate deforestation, which will have severe financial and political consequences. False negatives will result in the country not responding to deforestation events that are happening.

There is room for improvement in methodology. The existing accuracy assessments lack predicted land cover maps that show the location and spatial distribution of errors, and they lack class probability images, that would provide the spatial distribution of uncertainty.

Guyana is a suitable country to provide a benchmark to validate machine learning. This includes aerial photography with sub-meter resolution. This would allow very close examination of classifier performance at a variety of locations.

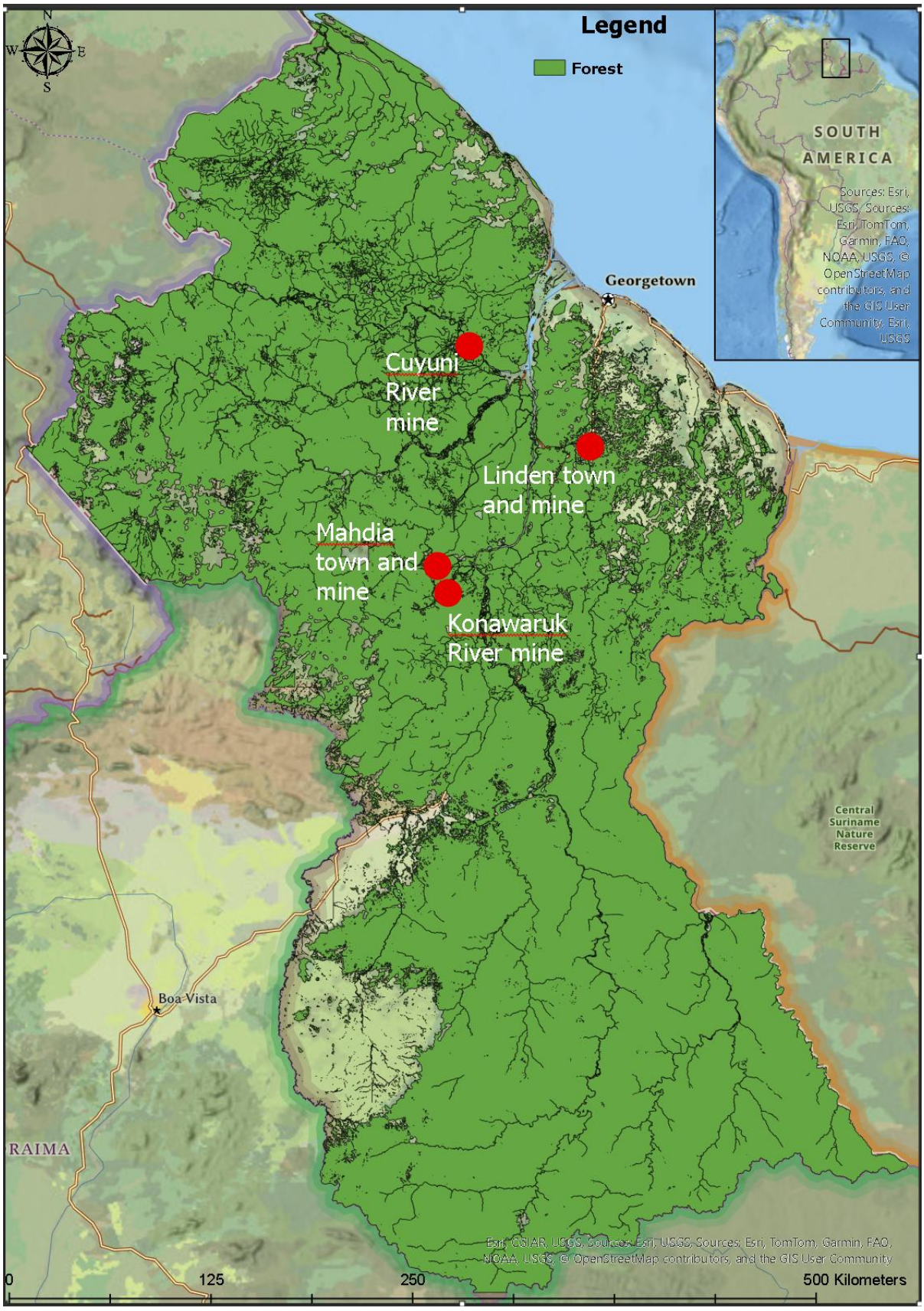


Figure 1: A map of Guyana showing the extent of forest (green) and the locations of the main study sites. Data provided by Guyana Forestry Commission

SECTION 2.1 OVERVIEW OF THE GUYANA REDD+ PROGRAM

Guyana's REDD+ program began in 2009, when The Cooperative Republic of Guyana and the Kingdom of Norway signed a REDD+ agreement. This program is administered by the Guyana Forestry Commission (GFC). GFC is responsible for advising the Minister of the Environment on forest policy, laws and regulations, and for the management of State Forests. GFC's work is guided by the National Forest Plan, which was created to address the National Forest Policy. GFC develops and monitors standards for forest sector operations, develops and implements forest protection and conservation strategies, oversees forest research and provides support and guidance on forest education and training.

As part of the 2009 REDD+ agreement, Guyana became responsible for setting up a Measurement, Reporting and Verification System (MRVS). Payments were tied to the amount of forest that is lost, and no money is provided if deforestation crosses a specified threshold. Recently, Guyana has been transitioning to financing Guyana's deforestation work by selling carbon credits on the private market, under ART-TREES (**A**rchitecture for **R**EDD+ **T**ransactions and **T**he **R**EDD+ **E**nvironmental **E**xcellence **S**tandard).

Guyana's REDD+ program has grown in many ways over the last fourteen years. In 2009, Guyana developed a three-year roadmap as a framework for a national REDD+ MRVS (Guyana Forestry Commission, 2019). This roadmap outlines progressive steps to build a fully implemented MRVS. The purpose of the MRVS is to monitor, report and verify forest carbon emissions from deforestation and forest degradation. The objective of the initial MRVS Road Map was to undertake comprehensive, consistent, transparent and verifiable assessment of forest area from 1990 to 2009 using Landsat-type satellite data that meets the criteria of the IPCC Good Practice Guidelines for LULUCF (Guyana Forestry Commission, 2019).

In conjunction with its REDD+ program, Guyana launched its Low Carbon Development strategy, which seeks to promote economic development while combatting climate change. Its two goals are to transform Guyana's economy to deliver greater economic and social development by following a low carbon development path, and to provide a model to the world on how climate change can be addressed through low carbon development in developing countries if the international community takes the necessary collective action, especially relating to REDD+ (Guyana Forestry Commission, 2019).

From its inception, Guyana's MRVS was designed to include data from multiple sources that can be combined in a consistent manner. There was a stepwise approach allowing time to bridge gaps in capacity allowing the integration of image processing and time series analysis routines. In 2010, forest change mapping was done using a combination of Landsat and DMC images. In 2011, Guyana began using high resolution RapidEye images. In 2012, 2013 and 2014, very high resolution aerial photography of high risk areas was acquired, something not seen in the other two REDD+ case studies (Mexico and Indonesia).

In 2014, Phase 2 of the roadmap was developed to consolidate and expand capacities for national REDD+ reporting and MRV, and support Guyana in meeting evolving international reporting requirements from the UNFCC and continue to fulfill additional reporting requirements. In 2017, the MRVS moved into its second phase in line with tasks set out in the MRVS Road Map (Guyana Forestry Commission, 2019). The initial steps of the REDD+ program allowed for a historical assessment of forest cover to be completed, key database integration to be fulfilled, and for interim/intermediate indicators of emissions from deforestation and forest degradation to be reported (Guyana Forestry Commission, 2019). In the time since, reporting on timber harvesting and illegal logging has been mainstreamed under full emissions accounting using existing methods.

Mining and infrastructure degradation estimates are calculated using new methods as part of a move toward more comprehensive yet sustainable methods for MRVS implementation after 2020. The sustainability means avoiding costly high resolution satellite images or aerial photography, as part of Guyana's desire to move toward a lower cost REDD+ implementation (Guyana Forestry Commission, 2019). In Phase 2, Guyana looked at lower cost satellite options, and found that Sentinel-2 and Landsat provided a good alternative to RapidEye. Sentinel-2's five-day revisit time will be used to develop near real time, continuous monitoring within the MRVS. Using Sentinel-2 and Landsat together allows for 6-7 observations per month. From 2015, a combination of Sentinel-2 and Landsat data was primarily used. In 2017, 2018 and 2019, aerial photography was commissioned. It was recorded by mounting a GeoVantage multispectral (RGB + Near Infrared) sensor to a Cessna and flying over target areas. Starting in 2019, GFC started using Sentinel-1 quarterly composites, along with a mix of Planetscope, RapidEye, Sentinel-2 and Landsat data. Typically, the highest resolution cloud free image is used to find deforestation events. The MRVS has been built to be data agnostic.

One way in which Guyana's REDD+ program has increased in sophistication is that the Guyana Forestry Commission measured forest degradation, and was one of very few countries to do so. After the ART Workbook was introduced, degradation was modelled via a buffer around deforestation events.

In 2018, the reporting was based on several REDD+ Interim Indicators, which allow for reporting while the full MRVS is under development. The framework for that year moved away from Interim Indicators toward the reporting of total forest carbon emissions and removals, with a focus on reporting emissions. The Guyana Forestry Commission's Monitoring, Reporting and Verification System (MRVS) has matured to the point where it reports annual forest carbon emissions and removals by activities caused by deforestation and forest degradation. The forest degradation work includes establishing national carbon conversion values, expansion factors, wood density and root/shoot ratios, and a detailed assessment of key processes affecting forest carbon, including a long-term monitoring plan for forest carbon (Guyana Forestry Commission, 2019).

In Phase 2, local communities became involved in the MRVS. This allows REDD+ activities to benefit from the skills and experience of local communities, in addition to the expert knowledge in GFC. In addition, the list of objectives of the program has grown, and now includes scientific journal publications.

In 2020, Phase 3 of the roadmap was developed to create a path toward a fully operational forest carbon reporting platform, suitable for a potential market-based mechanism while meeting all UNFCCC recommendations.

Recent developments in remote sensing technology have opened up new options for higher resolution optical satellites (such as SkySat and Worldview), higher revisit times (Planetscope), and online platforms that support faster and more efficient use of data (such as Google Earth Engine). GFC has also negotiated with ESRI for access to their full mapping suite, including Web dashboards. From 2021, SkySat and Worldview-3 data has been used as reference data, along with a mix of Planetscope, Sentinel-2, Landsat and Sentinel-1 data.

Figure 2 below shows the components of Guyana's REDD+ program. In Figure 1, my research fits into "Satellite Data – Deforestation", at the top left. Key data in my research was provided by the Guyana Forestry Commission, and my outputs will feed into MRVS work. If this step

received near-real-time inputs with greater accuracy and precision, then this will support new measurement and monitoring products in the future. With infrastructure like this already in place, the outputs of this research will be ready for application immediately.

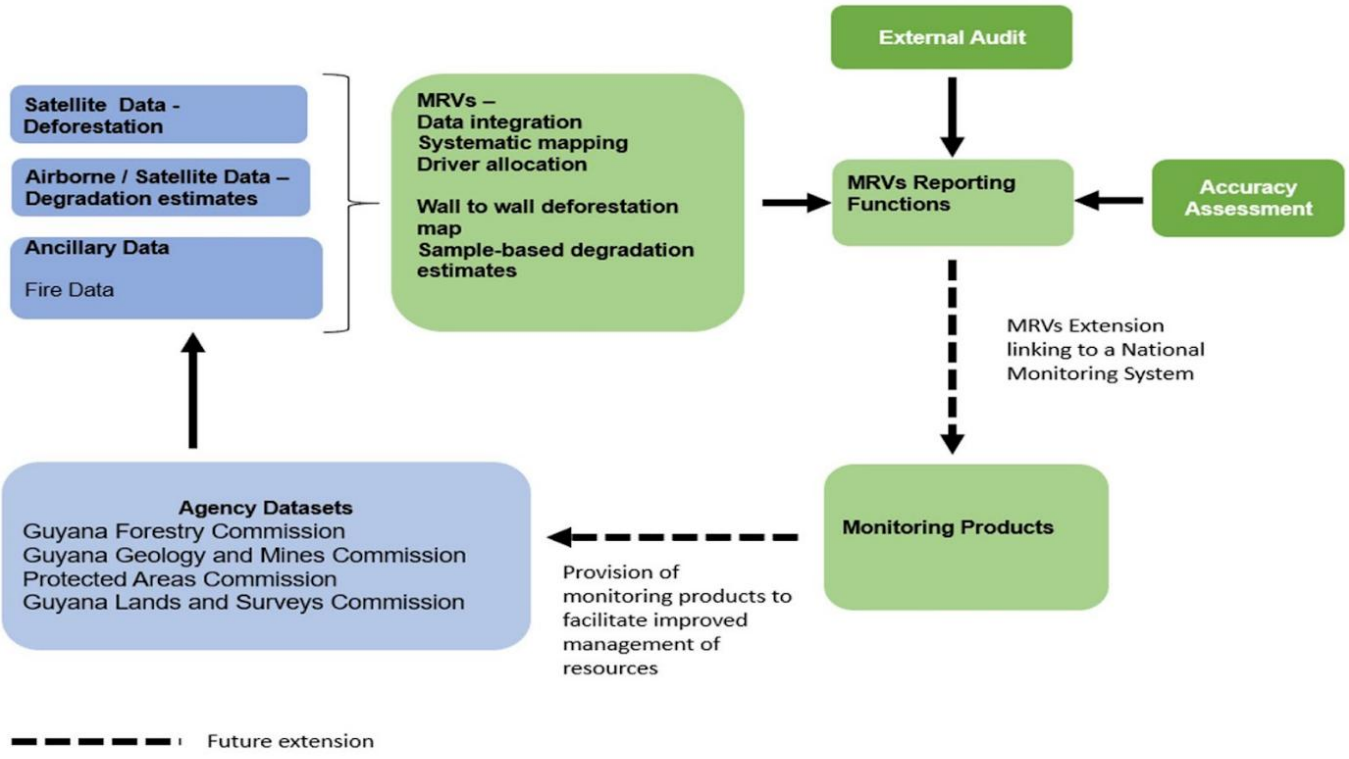


Figure 2: The components of Guyana's REDD+ program and their connections. This research is part of the "Satellite Data – Deforestation" component in the top left. The outputs of this research will feed into the MRVS component beside it. Source: https://forestry.gov.gy/wp-content/uploads/2022/07/Guyana-MRVS-Report-Year-2021-Final-1.pdf

In summary, one reason that Guyana is an excellent case study in remote sensing of forests is that the existing deforestation monitoring program has undergone fourteen years of development and experience. A wealth of data have already been collected, and the issues of mapping forest change in this area are already well understood. The results of the MRVS have applications to a range of policy and decision-making functions in the natural resources sector, especially forest management. Over the past decade, Guyana's MRVS has generated a wealth of data that can be used to understand the multiple uses of forests (Guyana Forestry Commission, 2019). Today, the potential of the data generated through annual mapping of

forest change extends well beyond the intended MRV function into policy, decision-making, integration of compliance functions, and more effective management within the natural resources sector. This research will therefore undertake an in-depth analysis of classifier and sensor performance for remote sensing of forests and relevant land cover changes in the context of Guyana's REDD+ program.

SECTION 2.2 THE GUYANA FORESTRY COMMISSION'S CHANGE MAPPING PROCESS

The process of mapping change is about identifying and measuring permanent changes in land use that results from anthropogenic activities, and then digitizing them into a map. In Guyana, change is mapped by digitizing change areas in satellite images, using a Geographic Information System (GIS). The entire country is divided into 24 tiles, and each tile is assigned to a technician. The operator searches for change within their tile and digitizes it. The current monitoring system is designed to map change events in the year in which they occur, and then monitor any change that occurs over that area each year. When a polygon's status is constant, its land use type is updated to stay consistent with the previous analysis. When there is a change in the land use type, this is recorded using the appropriate driver for that change.

Currently, the satellite sensors being used are Planetscope (Dove Classic sensor), Sentinel-2 (Multispectral Instrument sensor) and Landsat (Operational Land Imager instrument), with Sentinel-2 being the default. Funding for Planetscope data has been provided by the Norway International Climate and Forests Initiative (NICFI). Landsat is being used as a supplement to fill in gaps in Sentinel-2 data due to clouds. This is supplemented by tasking SkySat to acquire very high-resolution data in targeted areas. The Guyana Forestry Commission's standard operating procedure includes these steps:

1. If the Sentinel-2 images are offset from the official UTM grid, apply spatial corrections using control points.
2. The Minimum Mapping Unit in their process is one hectare, except for roads of all types. For roads, the minimum mapping unit is one stretch of road that is a minimum of 10 meters wide and 100 meters long. Smaller roads than this are not counted. This applies to roads made of any material. In this study, the one hectare Minimum Mapping Unit is not used; the mapping area is the pixel size of each satellite sensor. The pixels for each

satellite are 3 meters wide (Planetscope), five meters wide (RapidEye), 10 meters wide (Sentinel-2), and 30 meters wide (Landsat).

3. Mapping Deforestation Events starts with a systematic review of each 1 km grid in a 24 km^2 tile. When change events are found, they are digitized using satellite imagery, previous mapping, and Enhanced Vegetation Index (EVI) data. Land use change is attributed by documenting satellite image evidence, the period that the change occurred in, the change driver and the final land use. Attributes are applied to the Change Master using the Change Toolbar.
4. Quality Assurance and Quality Control (QA/QC) is done by an analyst in the Guyana Forestry Commission's mapping team who systematically checks through 445 24 km tiles, validates the no overlaps rule, reviews the attribution and ensures that the driver and land use type is recorded. Agency reference data and very high spatial resolution images are available to assist with validation.
5. Independent Accuracy Assessment is done yearly by Durham University.
6. After the QA/QC work is done, the data passes through two additional models to produce the final reported figures. The first model ingests the change data created in the previous steps, along with reference data on forest types, the country boundary, land classes, and potential for change and accessibility. It verifies that no change events have been assigned to areas that were confirmed not to have change, and it assigns the information in the reference data list above to each change event. The second model uses the output of the first model to produce a series of tables and charts from templates, for the final report.
7. The final all change dataset is produced using an ArcGIS model, which uses as inputs the change master, the non-forest layer, the country boundary, the land classes layer, and the potential for change and accessibility map. All changes are clipped against the non-forest data to ensure that no change areas have been mapped in confirmed non-forest areas.

The manual change mapping process is labour intensive. The next step for GFC is automated detection of deforestation events, which will reduce the cost of monitoring deforestation, and allow human experts to spend less time finding deforestation and more time responding to deforestation.

SECTION 2.3 CREATING CLASS LABELS BASED ON THE LAND COVER TYPES

The goal of this research is to train a model that takes pixel values as inputs and outputs the correct land cover category or deforestation driver, making this a classification problem. Therefore, a critical step is creating a set of classes that meets REDD+ reporting requirements.

The Guyana Forestry Commission's Standard Operating Procedure describes both land cover classes (Table 5) and deforestation drivers, based on IPCC classes (see Table 5 column IPCC Description). A land use class is the resulting land use based on the driver. For example, bare land is the land use class resulting from mining activity. Anthropogenic activities that drive such land use change include mining, roads, settlements, logging and agriculture.

The way for a computer to automatically identify these land cover types is to recognize their colours, textures, shapes and context clues. Forests have one colour and texture, rivers have another. Mining sites are identified by recognizing a cluster of objects that always occur together – ponds, bare surfaces, roads and vegetation growing inside the mine. One of the identifying features of agriculture is that the fields tend to be rectangular. Settlements are recognized as buildings in a grid of roads, with grass, isolated trees and paved surfaces. Machine learning normally uses colour information (pixel values) only, and Convolutional Neural Networks can also use textures and structures. None of these algorithms use the context clues that a human would use, so one of the challenges of this research is finding a way to separate these classes without using context information.

Table 5: Schema for Guyana and IPCC land cover types. Taken from the Guyana Forestry Commission's standard operating procedure for 2023.

Class	Guyana Description	IPCC Description
Forest	Mixed forest	Forest Land
	Wallaba/Dakama/Muri forest	
	Swamp/Marsh forest	
	Montane forest	
	Mangrove forest	
	Savannah Forest	
	Shifting Agriculture systems	
Non-Forest	Cropland	Cropland
	Tropical unmanaged grassland	Grassland
	Tropical unmanaged shrub-land	
	Open Water	Wetlands
	Herbaceous unmanaged wetland	
	Settlement	Settlements
	Pre-1990 Mining Areas	
	Rock	Other lands
	Bare-land	

The land cover classes being used are based on the set of classes used in accuracy assessment. The classes being used are: Forest, River, Mining, Road, Bare Land, Settlement, Agriculture, Grassland.

From 2017 to 2019, Guyana commissioned high resolution aerial photography of high risk areas. The sensor was GeoVantage, which records Blue, Green, Red and Near Infrared at 25 to 60 cm resolution, depending on altitude. It was mounted on a Cessna aircraft. Unlike satellite images, these are cloud-free multispectral images (though they do have cloud shadow).

The study areas created in this research are areas that have GeoVantage aerial photography. Different air photos have different combinations of classes represented, and in each test in the results chapters (Chapters 4-8), a pair of images with the desired combination of classes was selected. One member of the pair is for training the model, and the other member is for testing the model.

These air photos record details as fine as individual vehicles and the canopy of individual trees, and as such allow for maximum precision in identifying boundaries between land cover types. As part of this research, the labels were created by tracing the land cover boundaries in ArcGIS. Then, a shapefile was created that has the land cover boundaries, and then it was converted to a raster. The very high resolution of the reference data means that very small details are visible, including details that are not visible in the data being used for training and validation. They were therefore used to create class labels. Here's how they fit into machine learning:

- Start with the data and the class labels. The class labels are the correct answers, and the data occupies the place of a question.
- Train a model that uses the data and the labels to figure out how to go from the question to the answer.
- Prepare another dataset and class label set. Run the new data through the model to make predictions. This asks the computer a question, and the computer uses the model to give answers.
- Compare the predictions with the class labels to assess the model's accuracy. If the accuracy is low, then the model is bad, and cannot be used to make reliable predictions. If the accuracy is high enough, then the model is good, and will make reliable predictions.
- The study areas are areas that have GeoVantage aerial photography, and they represent a wide range of land cover type combinations.

SECTION 2.4 STUDY AREAS AND EXAMPLES OF CLASSES

The satellite remote sensing data used in this study comes from the locations where very high resolution aerial photography was acquired. From 2017 to 2019, the Guyana Forestry Commission ordered aerial photography of a sample of high-risk areas. These photographs were taken using an airborne GeoVantage multispectral sensor (with blue, green, red, and near infrared bands) mounted on a Cessna aircraft. Depending on the altitude, the spatial resolution is 25 to 60 cm, which means that very small details are visible. These air photos serve as the

reference data. The class labels for training and testing were derived by tracing over the land cover boundaries in these very high resolution images.

Because this is where very high resolution reference data is available, this is where the data used in this study was selected. For each satellite sensor, an image was clipped out that matches the extent of each air photo. Approximately 70 air photos were acquired in each year over three years, so most sensors have 210 images total. Within these, the air photos that have forest only were not used.

Figures 3-22 have examples of each land cover type, including Forest, River, Mining, Road, Settlement, Agriculture and Grassland. There can be substantial spectral variation within each land cover class; hence each class will have two or more examples. This variation will be important in interpreting the results of classification workflow comparisons – this spectral variation within land cover classes is something that a human technician can deal with effortlessly, but it can have a major impact on classifier behaviour.

In looking at the examples of each land cover type, several sources of confusion are visible:

- Forests have high spectral variability, due to the texture of the canopy, especially in higher resolution data. This creates opportunities for confusion when part of the canopy overlaps with another class.
- Rivers vary in their appearance. Some are clear, and some are laden with sediment. This presents a problem when classifying data nationwide: a model trained on rivers in one area wlll have problems identifying other rivers, and a model trained on all rivers will have such a range of spectral variation that confusion with other classes becomes likely.
- Mining sites have a variety of objects within them, with very different spectral properties. This includes mining ponds, bare surfaces, and new vegetation.
- Roads are very narrow, which means that they have a lot of edge pixels relative to interior pixels.
- Settlements include buildings, roads, bare surfaces, grass, and isolated trees. This creates many opportunities for confusion with the other classes.

Table 6, taken from the Guyana Forestry Commission's Standard Operating Procedure, describes criteria for each class. One important thing to note is that these criteria are based on policy needs, which are not always compatible with the needs of machine learning. Here are some examples:

- Roads are only counted as deforestation when they are more than 10 meters wide. However, the high resolution satellite images and the aerial photography are able to see even the narrowest of roads. This may mean that a classifier will be able to pick up roads that shouldn't be counted.
- Another important thing to note is that, for most activities, the area covered must be greater than one hectare. However, once again, the high resolution satellite images and aerial photography show very small details, much smaller than a hectare. For settlements, this means that there are small areas that should be counted as part of the forest, but will appear as settlements to the classifier.

Another issue is the details that aren't mentioned: where exactly should the line be drawn between land cover classes? The aerial photography has high enough resolution that individual bushes and buildings can be clearly seen. In mining sites and settlements, the edge of the forest was drawn at the edge of the trees, which means that isolated trees and low vegetation are included in the Mining and Settlement classes. Isolated trees and low vegetation are spectrally similar to the forest, which will make classification easier, by categorizing like with like. But they need to be measured as part of the mining site, for policy reasons. Finding a machine learning workflow that meets REDD+ reporting requirements will be part of the challenge of this research.

Table 6: Summary of Activities and Drivers. Taken from the Guyana Forestry Commission's standard operating procedure for 2023.

Drivers	Activity	Criteria	End Land Use Class
Forestry	Sustainable Forest Management	Fall inside state forest area and is a registered concession	Degraded forest by type
			Bare land
	Infrastructure	Roads > 10 m	Settlements
Mining	Deforestation	Roads > 10 m	Settlements
	Infrastructure	Deforestation sites ≥ 1 ha	Bare land
Agriculture	Deforestation	Deforestation sites ≥ 1 ha	Cropland
Fire	Deforestation	Deforestation sites ≥ 1 ha	Bare-land or Cropland
	Degradation	Degraded forest sites	Degraded forest by type
Infrastructure	Deforestation	Roads > 10 m	Settlements
Settlements	Deforestation	Areas deforested for human settlements ≥ 1 ha	Settlements
Shifting Cultivation	Deforestation	Deforestation sites ≥ 1 ha	Cropland
Reforestation	Reforestation	Monitor abandoned deforestation sites for an increase in vegetative cover	Degraded forest by type
Afforestation	Afforestation	Monitor historical non-forest areas for afforestation	Degraded forest by type

Figure 3: Examples of forest in images taken by various satellite sensors. The within-class variation in the forested are can be a challenge to a classifier.

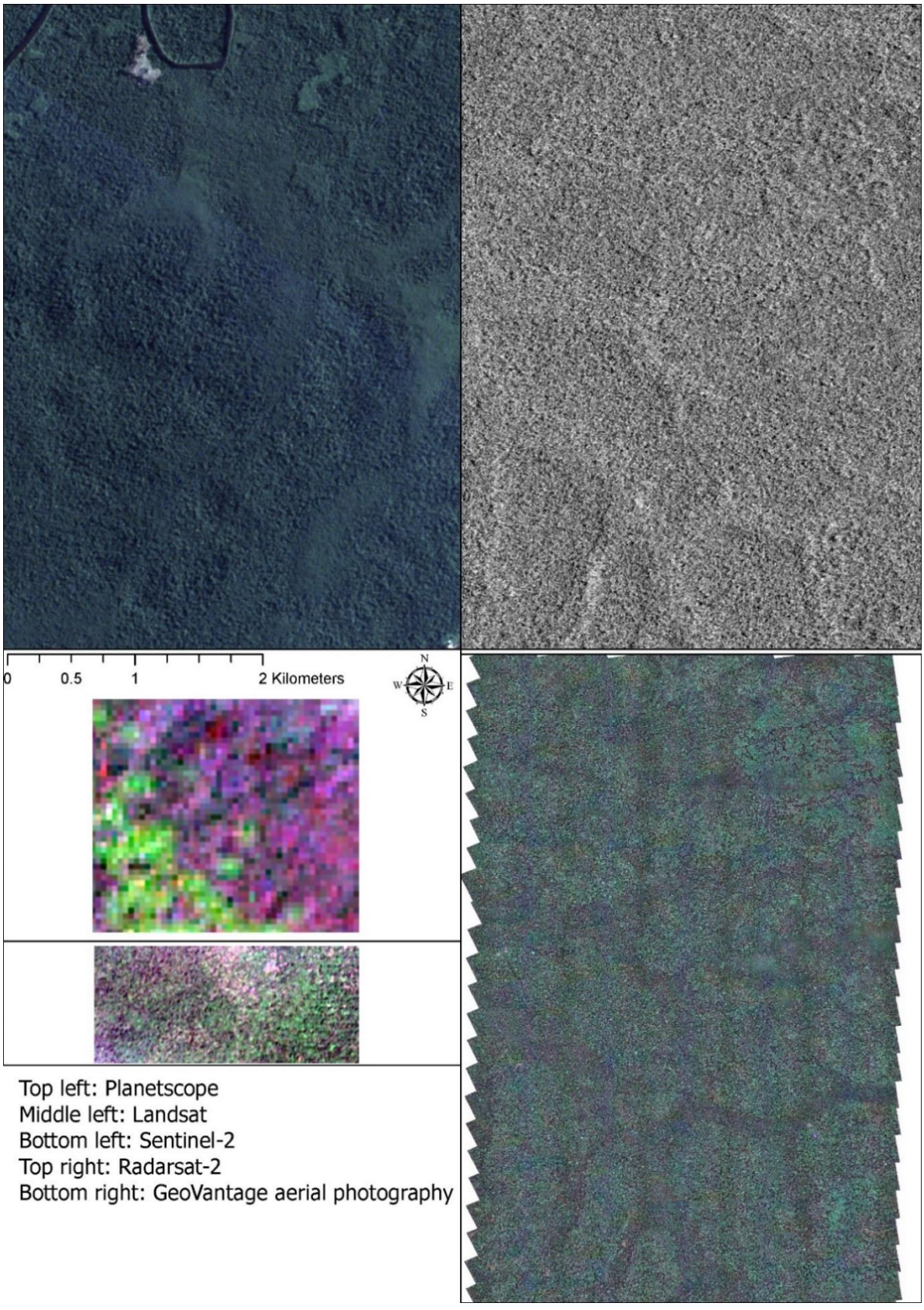


Figure 4: Aerial photograph of forest in northwestern Guyana. Figure 2 shows this land cover type from satellites.

Figure 5: Aerial photograph of forest in northwestern Guyana. Figure 2 shows this land cover type from satellites.

Figure 6: Examples of rivers in images recorded by different satellites and locations. Notice that not all rivers are spectrally similar.

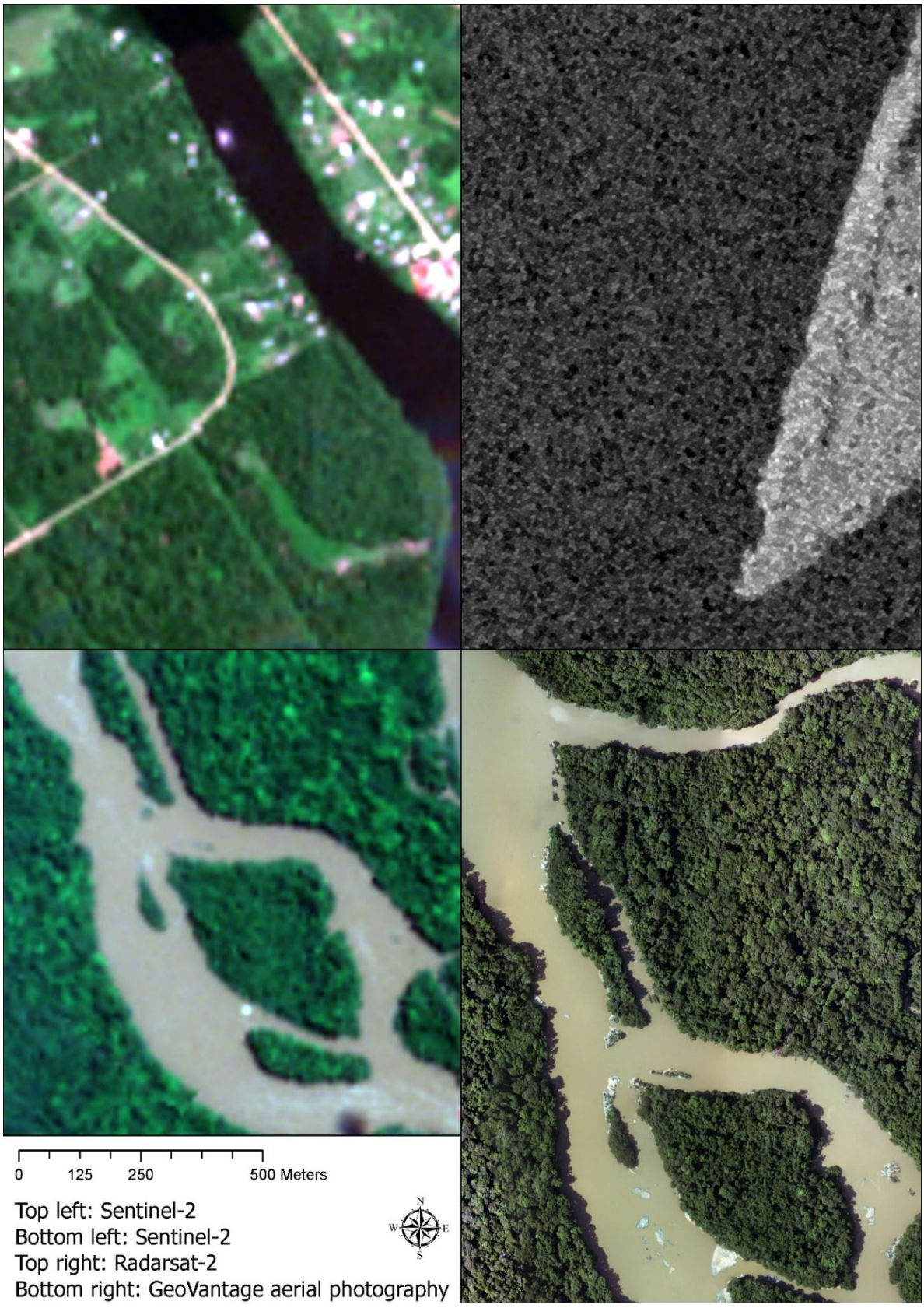


Figure 7: River in a forest in northwestern Guyana, taken from an aircraft. Figure 5 shows this land cover type from satellites.

Figure 8: The Demarara River, near Linden, taken from a bridge. Figure 5 shows this land cover type from satellites.

Figure 9: Examples of a mining site in images recorded by different satellite sensors. Notice the complexity – there are bare surfaces, roads, ponds, and young vegetation, which are all very different from each other and very similar to other classes.

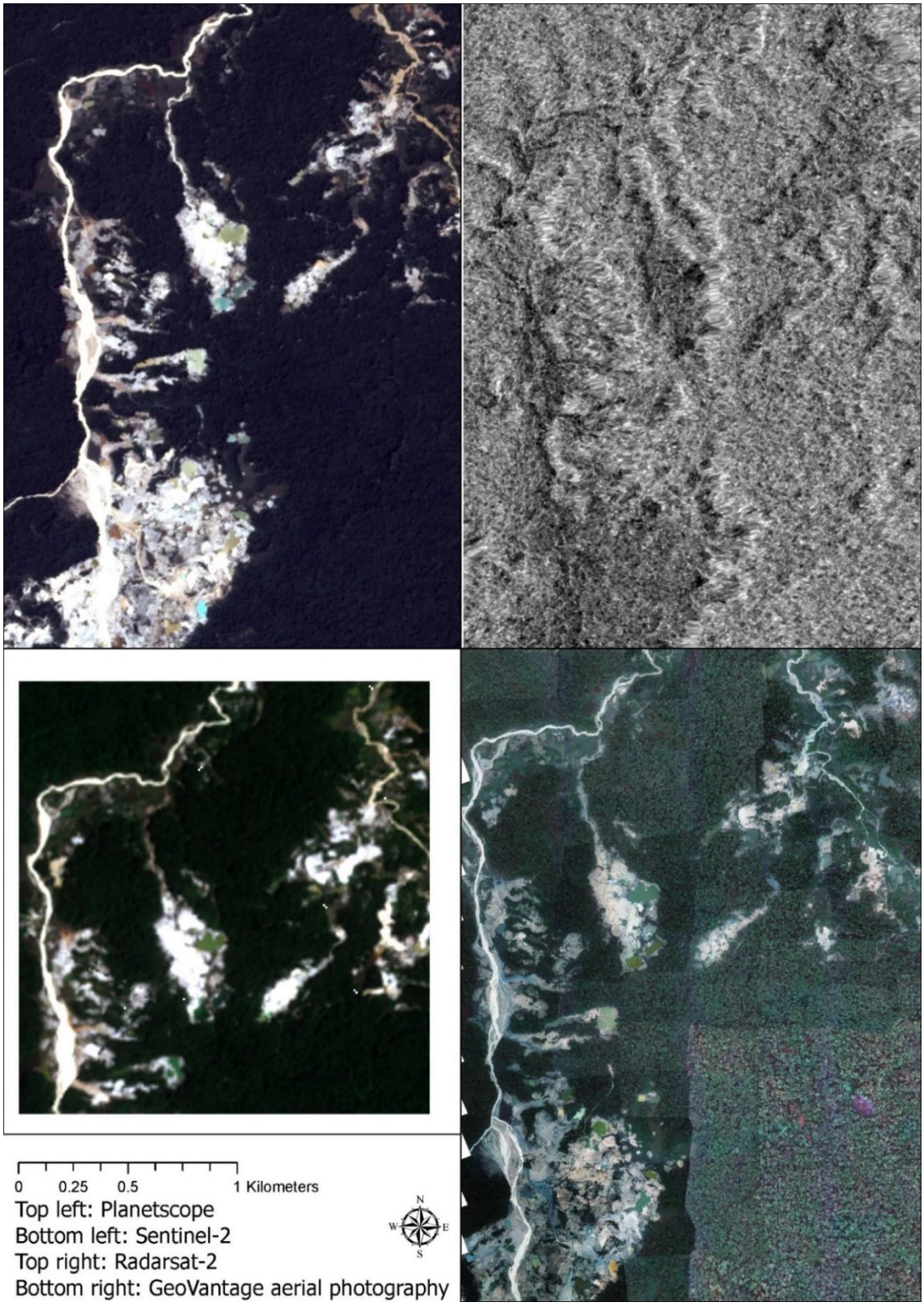
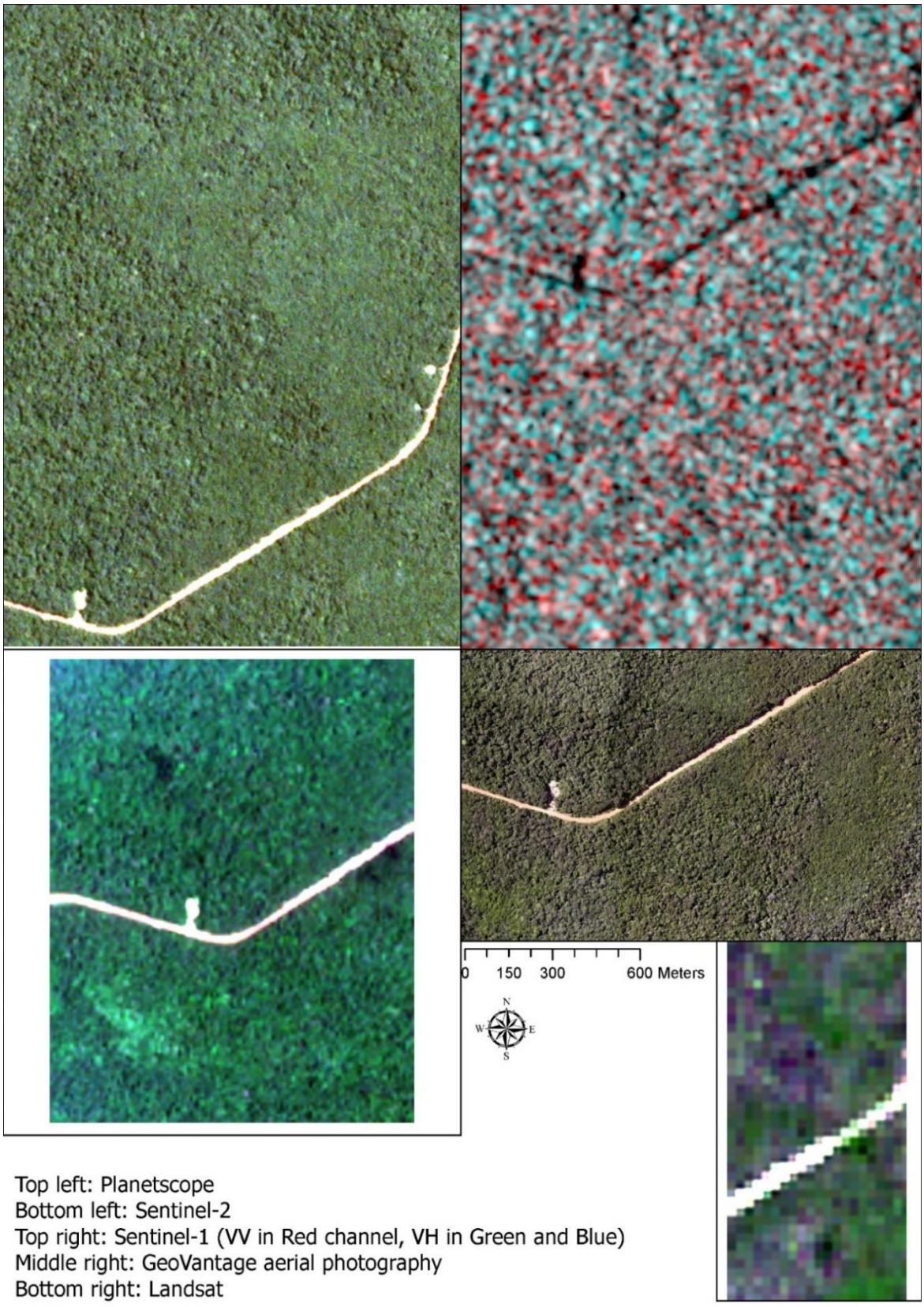


Figure 10: The active bauxite mine near Linden.

Figure 11: An abandoned section of the artisanal gold mine near Mahdia. Figure 8 shows this land cover type from satellites.

Figure 12: Examples of a road in images recorded by different satellite sensors. This class is very thin, with lots of edge pixels.



Top left: Planetscope
Bottom left: Sentinel-2
Top right: Sentinel-1 (VV in Red channel, VH in Green and Blue)
Middle right: GeoVantage aerial photography
Bottom right: Landsat

Figure 13: A road in central Guyana. Figure 11 shows this land cover type from satellites.

Figure 14: A road near Mahdia. Figure 11 shows this land cover type from satellites.

Figure 15: Examples of a settlement near Linden. Like mining sites, Settlements are complex classes, with many spectrally distinct objects within them.

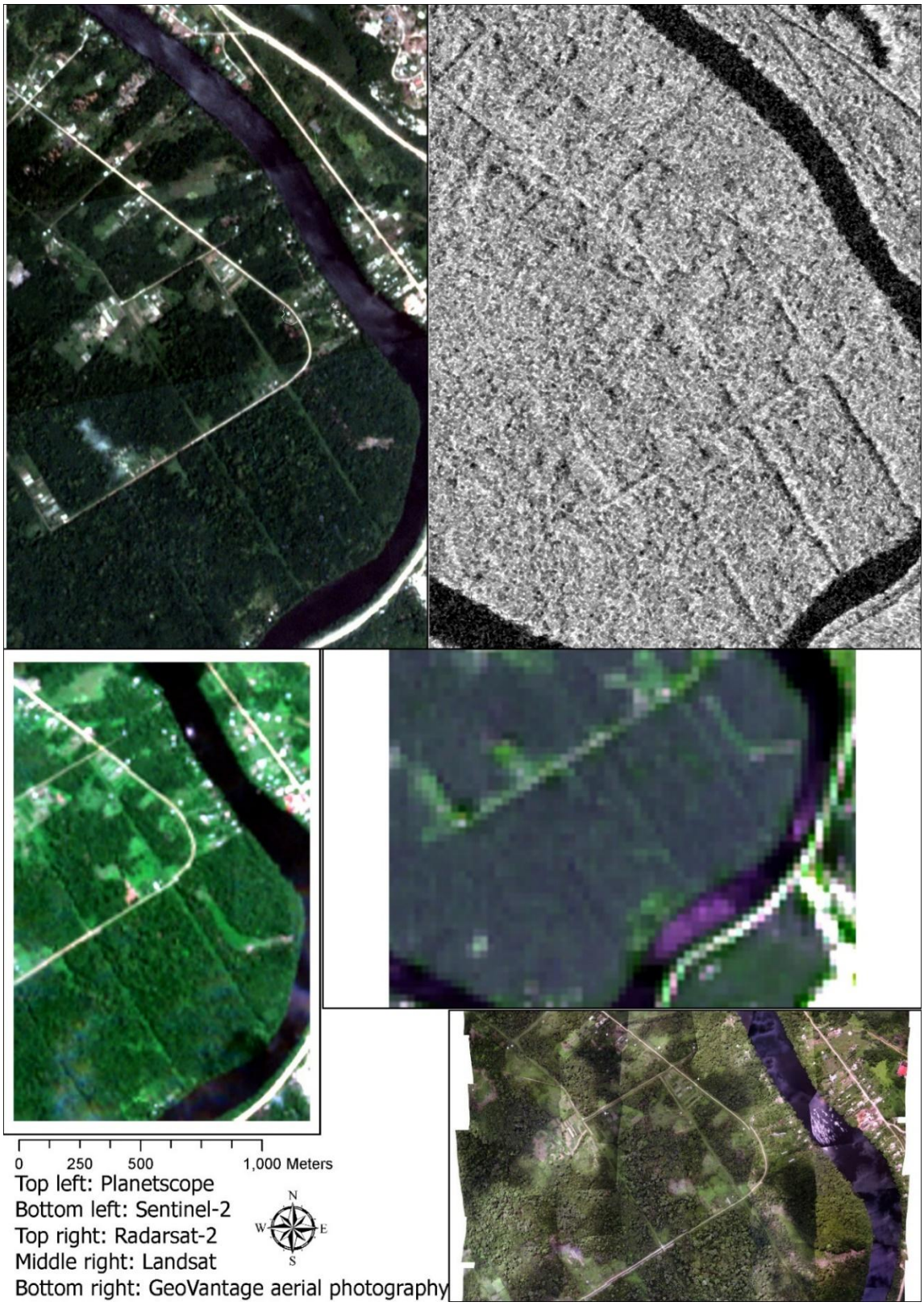


Figure 16: Settlement south of Linden. Figure 14 shows this land cover type from satellites.

Figure 17: The town of Linden. Figure 14 shows this land cover type from satellites.

Figure 18: Examples of agriculture on the Essequibo River, near its mouth.

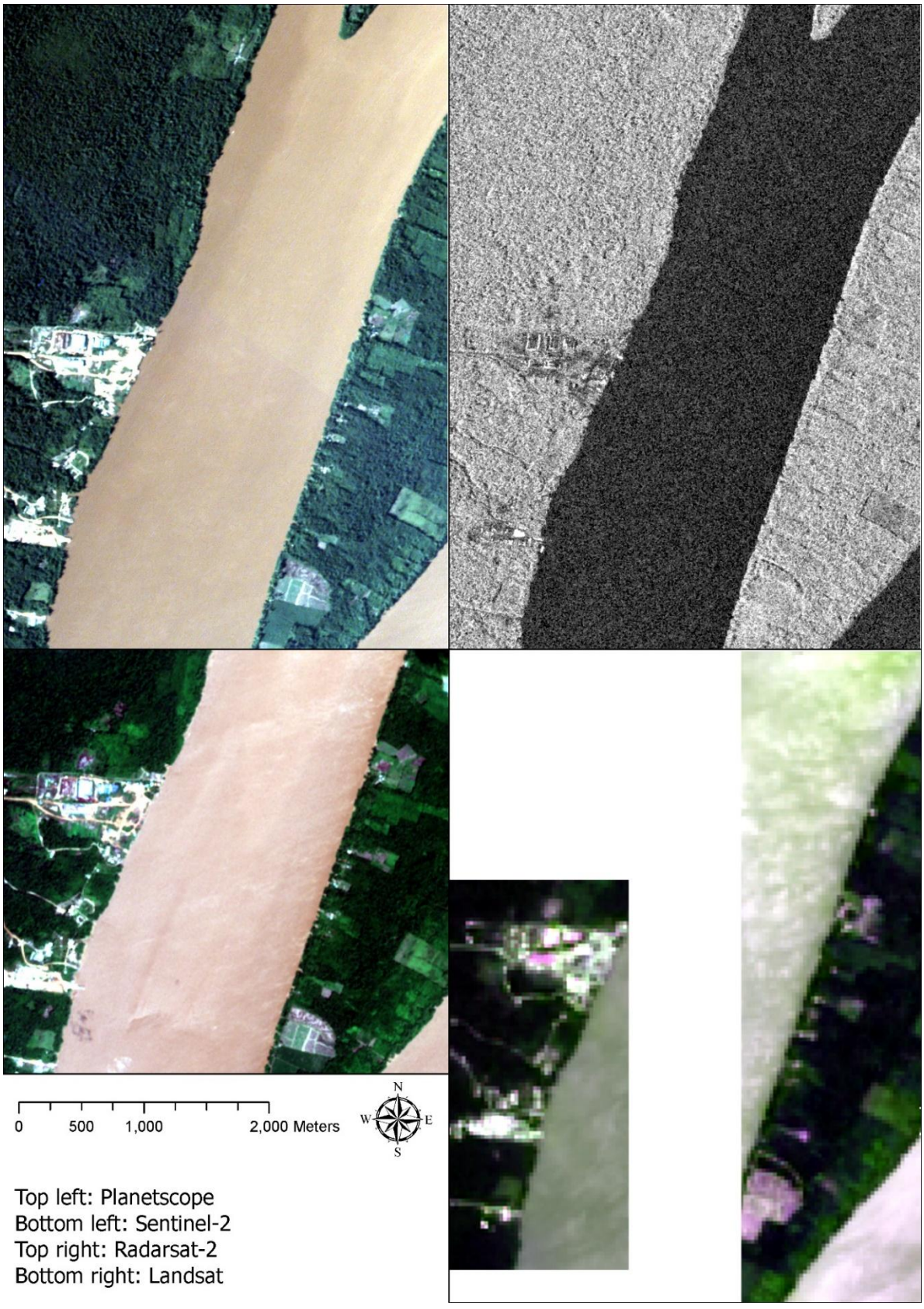


Figure 19: Agriculture near the mouth of the Essequibo River, taken from an aircraft.

Figure 20: Agriculture near the mouth of the Essequibo River, taken from an aircraft.

Figure 21: Examples of grassland. The same surface can look very different in images recorded by different sensors – see the top left corners of each image.

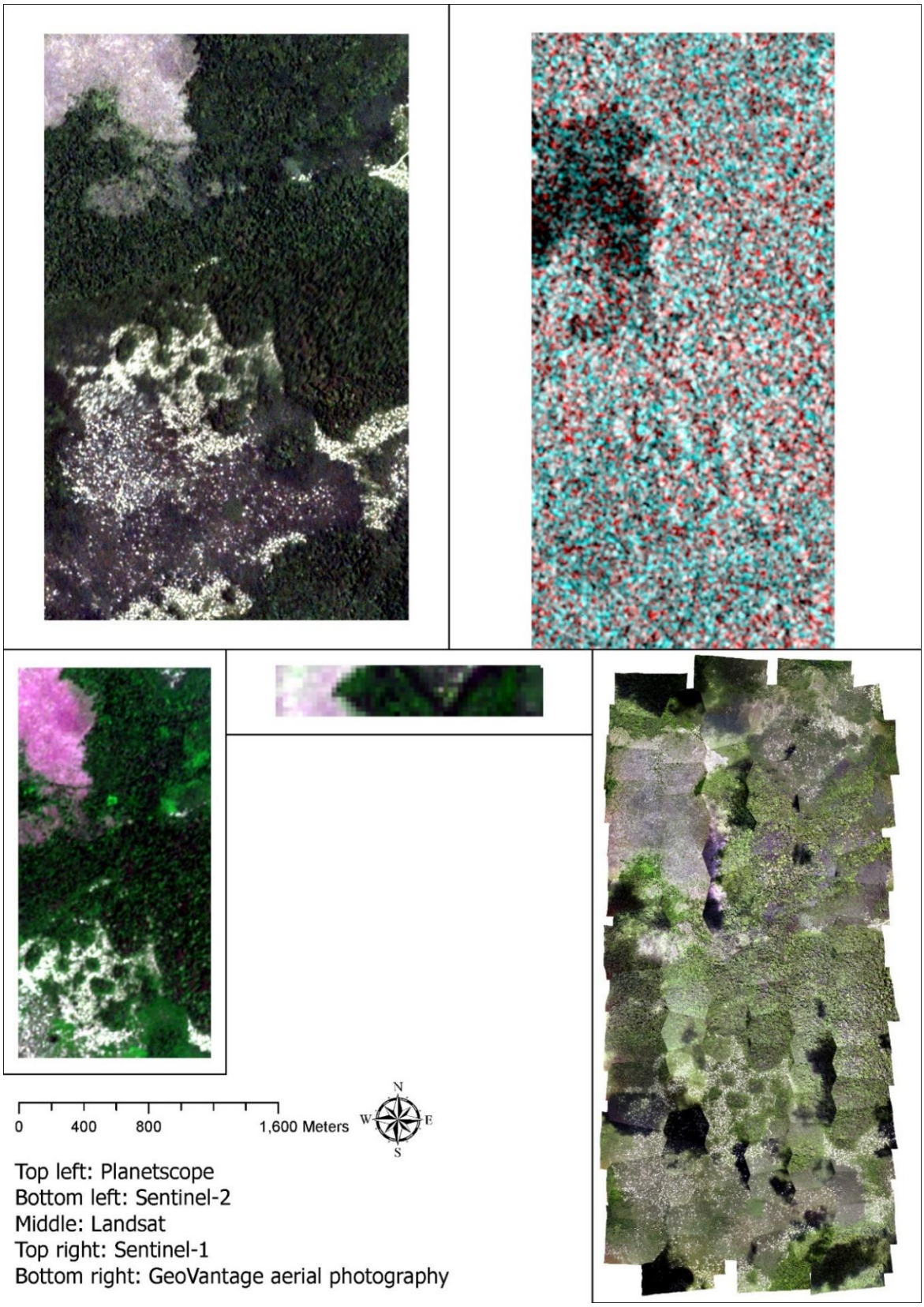


Figure 22: Grassland area north of Linden.

SECTION 2.5 POTENTIAL SOLUTIONS

The fundamental problem in classification of land cover types is the difficulty of separating the land cover classes seen in Figures 3 to 22. The classifier works by separating classes in the data according to the labels provided, and anything that appears identical in the data will not be separable. The separability of classes depends on whether the data values are different or the same in each class in each dataset. When two types of surfaces interact very differently with EM radiation, they're easily separated in the image recorded. When two types of surfaces interact the same way, they can't be separated. Separation is easiest when there is no overlap between classes, and when the gap in distributions of data values between classes matches the shape of the estimator. Separation is difficult when there is a lot of overlap between classes, or when the chosen algorithm is not suited to the data. The task is therefore to find out how to solve the problem of overlap and how to match the estimator to the data.

WHICH DATA ARE BEST?

In Leckie (1990), multispectral and SAR images were compared in their ability to separate tree species near Chalk River in Canada. SAR was better at separating softwood tree species (red pine, white pine, jack pine, spruce) than visible light and infrared bands. Near infrared and mid infrared bands were important for separating hardwood species and softwood species. Visible and mid-infrared bands were useful for separating open and forested areas.

In the context of satellite remote sensing, sensor fusion combines multispectral and SAR images into one image. There are two inputs: A three band multispectral image and a one band SAR image. There is one output: a three-band fused image, that has information from both sources.

This is an extension of pansharpening, which is a method of creating an image with both high spatial and spectral resolution (illustrated in Figure 23). In multispectral images, there is a trade-off between spatial resolution and spectral resolution. Lowering the spectral resolution means widening the range of wavelengths, which increases the number of photons acquired. Increasing the spatial resolution reduces the number of photons that are captured per pixel. This creates a situation where a sensor needs either high spatial resolution and low spectral resolution to get a signal, or high spectral resolution and low spatial resolution.

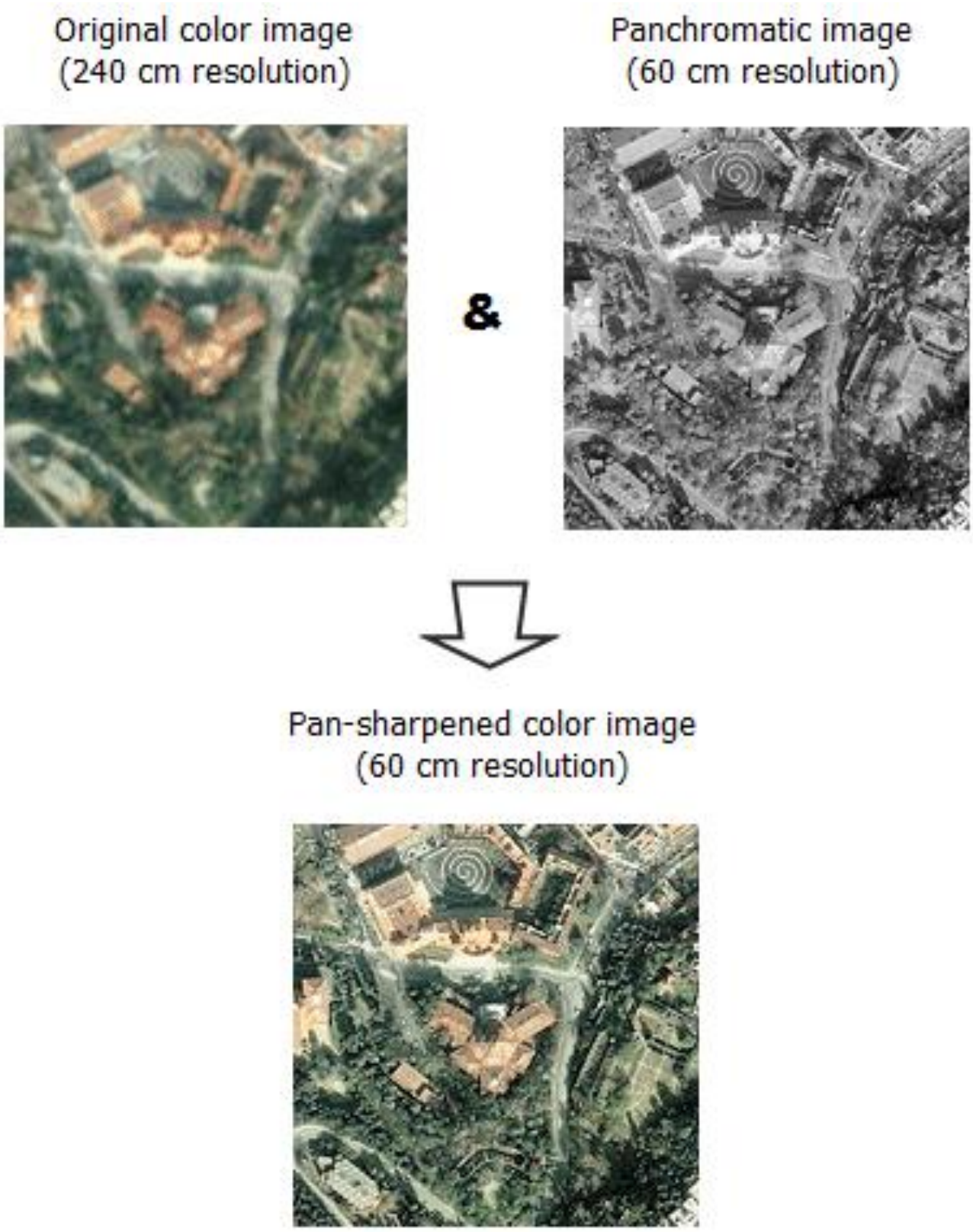


Figure 23: An example of pansharpening. Source: https://pro.arcgis.com/en/pro-app/latest/help/analysis/raster-functions/fundamentals-of-pan-sharpening-pro.htm

The answer to this was pansharpening. A satellite has both multispectral bands, which have high spectral resolution and low spatial resolution, and a panchromatic band, which has high

spatial resolution and low spectral resolution. A variety of mathematical techniques exist to fuse the panchromatic band with the multispectral bands to produce an image that has both high spectral resolution and high spatial resolution.

This has two inputs, a 3-band RGB image and a 1-band panchromatic image, and the output is another 3-band RGB image with higher spatial resolution. The reason is that, until recently, remote sensing mostly meant visual interpretation of digital images. Computer displays are designed to work with three channels exclusively (Red, Green, Blue), and that's also what the human eye sees. Pansharpening was designed to work within this by fusing 4 input bands (low resolution RGB + high resolution Panchromatic) into 3 output bands.

The performance of image classification depends heavily on the amount of information in the data. A higher information dataset can be created by combining the information of multispectral and SAR images. There is a decades-long body of literature (reviews can be found in Pohl and van Genderen (1998); Wang et al., 2005; Pandit and Bhiwani, 2015; Ghassemian, 2016; Kulkarni and Rege, 2020) that shows that accuracy can be improved by combining data from multiple sources. In these reviews and the literature they review, the goal of sensor fusion is to combine images from different sources in a way that increases the quantity and quality of information available. A fused multispectral-SAR image has spatial, spectral, textural, and dielectric information in one image. Features not visible in either source image alone can be detected here. Pohl and van Genderen (1998) summarizes this as "1 + 1 = 3".

In satellite remote sensing, different sensors and bands record different information about the same surface. SAR, visible light and infrared data are complementary in this way. Machine learning can see all of these all at once, and this may allow it to separate classes that aren't separable in one sensor alone.

Sensor fusion has been tested in a REDD+ context in Malawi (Hirschmugl et al., 2018), and it found that accuracy increased from 76% (optical only) and 69% (SAR only) to 85% for optical and SAR data based on a Bayesian combination method (from Reiche et al., 2015).

Sensor fusion uses a mathematical process to combine or replace information. In this context, this means spatial and spectral information. A multispectral image's bands (Red, Green, Blue and Near Infrared, and more) contain both spatial and spectral information. SAR images contain spatial information. The algorithm used here is the HSV transformation implemented in Google Earth Engine.

In addition to fusing multispectral and SAR images, there is also the option of stacking the bands. Each image is an array of pixels, with two spatial dimensions (x and y coordinates) and one band dimension (the values of each band at each pixel). There can be two images, recorded by two satellite sensors, that have their pixels at the same xy coordinates, but have different bands. In stacking, the second image's band values are added to the band axis of the first image, keeping the same pixel locations in space, but extending the band axis. If the axis specified is the feature axis, which can also be thought of as the channel or band axis, then a new array will be made that has the combined features of the source arrays. For example, RapidEye has five bands, for an array of dimension (height, width, 5), and Sentinel-1 has two bands, for a dimension of (height, width, 2). Stacking them along axis 2 will produce an image that has dimension (height, width, 7). This differs from sensor fusion in that the exact pixel values of both source images are kept in the new image - in sensor fusion, the fused image has pixel values that are derived from but not identical to the source images.

There is some literature on this approach, though it isn't nearly as extensive as the literature on sensor fusion. Holmberg (2021) stacked Sentinel-1 and Sentinel-2 bands and found a small increase in classification accuracy, though it was judged not to be worth the added computational requirements. Kuplich (2006) used both SAR and multispectral bands in a neural network, and class discrimination increased when both SAR and multispectral bands were used over SAR bands only.

WHICH CLASSIFICATION ALGORITHM IS THE BEST?

Another solution to the problem of insufficient accuracy is to try different classifiers (classification algorithms). There is a lot of literature that compares a small selection of common classifiers. But there's also a gap in the literature here: there are many less common classifiers that haven't been tested. A solution may be found in an overlooked classifier or ensemble of classifiers.

The choice of classifier is critically important to the accuracy of the classification output, and so a classifier comparison will be part of this research. There is a vast amount of literature comparing classifiers. Typically, three or four common classifiers are compared based on overall accuracy or confusion matrices. For example, Zhao et al. (2019) compares Classification and Regression Trees, Random Forest, Support Vector Machines and neural networks. This research

is different from other classifier comparisons in that a much broader range of classifiers will be tested, on multiple platforms (Python with Scikit-Learn and Keras, Google Earth Engine).

The literature is full of contradictory results on which classifier is the best to use (more discussion in Maxwell et al 2018). Neural networks outperform decision trees and their boosted implementations (Rogan et al 2008), and decision trees and their ensembles are more accurate than neural networks (Lippitt et al 2008). Random Forest outperforms Support Vector Machines (Pal 2005, Adam et al 2014), and Support Vector Machines outperform Random Forest (Zhang and Xie 2013, Maxwell et al 2014, Maxwell et al 2014, Maxwell et al 2015). This makes it clear already that there is no straightforward answer in the literature.

Looking at each study, it becomes clear that their data and methodologies are not comparable. Rogan et al 2008 uses two study areas in northern and southern California, the data is Landsat 5 with ancillary elevation data, and reference data was acquired by interpreting forest resource photographs. Lippitt et al 2008 used Landsat 7 data, and the study area was in Massachusetts. Rogan et al 2008 compared the fuzzy ARTMAP neural network algorithm with two decision tree implementations: S-Plus, C4.5. Lippitt et al 2008 compared ARTMAP, Multilayer Perceptrons, Self-Organizing Maps, and decision trees using the Gini and entropy splitting rules. Any one of these differences could result in differences in classifier performance. Lawrence and Moran 2015 used 30 datasets from Montana, Wyoming, Utah, Colorado, New Mexico, Arizona, Nevada and California to compare Random Forest, C5 decision trees, logistic model trees, support vector machines, and Multivariate Adaptive Regression Splines. Random Forest was the best in 18/30 datasets, C5 was the best in 11/30, and Logistic Model Trees was the best in 1/30.

In summary, the best classifier is very context-dependent, and the results of a classifier comparison in one context cannot be applied to a different context. It will therefore be important to test different classifiers in the context of Guyana REDD+ monitoring.

HOW TO MAKE GOOD LABELS?

In the discussion of the land cover classes in Section 2.6, it became clear that the official land cover classes in the Standard Operating Procedure pose several challenges for machine learning:

- The river examples were not consistent – clear rivers and turbid rivers are spectrally very different.

- Mining sites have vegetation that can be confused with forests, ponds that can be confused with rivers, and bare surfaces that can be confused with roads.
- Roads less than 10 meters wide aren't counted toward deforestation, but they will be picked up by a classifier, especially in higher resolution data.
- Settlements have grass, trees, roads, buildings and paved surfaces.
- Deforestation is only counted if the area is greater than a hectare, but a classifier will pick up many smaller details in high resolution data.
- Human operators use context information to separate these classes. However, machine learning relies primarily on spectral information alone, though Convolutional Neural Networks can also use texture and structure information.

These specific problems, in a REDD+ monitoring context, have never been tackled head-on in the literature. In the existing literature, the class labels are simply the land cover types, and the study areas are small areas that have clear and carefully chosen examples. A model that seeks to classify land cover types across an entire country will not be working with a small set of simple examples, but the heterogeneous data described above. This study will test models and satellite sensors using more diverse data than most studies on this topic.

SECTION 2.6 THE KEY RESEARCH QUESTIONS

The previous section shows that there is a lot of research on the data and model aspects of machine learning, including classifier comparisons and tests of data fusion. There is also some research on the importance of good reference data in creating good class labels. What is missing from the research on data fusion is tests in the context of REDD+ and mapping in mixed tropical forest regions. These areas are typically very cloudy and forest to non-forest transition involves identifying and characterising a number of complex of land cover types. That last point is the core of the problem: country-wide classification of deforestation will need to classify a heterogeneous set of sites, a level of spectral diversity that a classifier will struggle with. This research is an opportunity to collect the results needed to develop a strategy that tackles that challenge head on.

In the research on class labels, there is a lack of research on the problem with using official land cover classes as class labels in machine learning. The problem with this is that class labels based on human activities and heterogeneous areas may not be compatible with machine learning. There are multiple activities that can have vegetation inside them (mining,

settlements), and mining ponds can be spectrally similar to a nearby river. Spectral variation within classes leading to overlap between classes will create a situation that a classifier may not be able to handle. Also, the classes are normally separated using context information, and so it may or may not be possible to separate them using spectral information alone.

This leads to the research questions last seen in Section 1.1:

1. How does land cover type mapping accuracy change between classifiers? How is this affected by the spectral variation from one site to the next, and how is it affected by the nature of the classes?
2. How does land cover type mapping accuracy change between satellite sensors and combinations of sensors? Does this address the problems that have been summarized in this section?
3. How does the complexity of the land cover types affect accuracy? This is closely related to the issues discussed in the previous paragraph.

In the best case, this research will find a combination of classifiers, data and labels that can identify changes in difficult-to-separate land cover types in near real time. It evaluates classifier performance using real world data from a variety of satellite sensors, and the class labels are based on higher quality reference data than most research, which will lead to an improved ability to identify and interpret errors. The most immediate impact will be on the Guyana Forestry Commission, as they will be the first to see the results. This is also applicable to other REDD+ countries, some of which have tried to implement machine learning without success. This also has applications to land cover mapping more generally.

The Guyana Forestry Commission has the potential to benefit greatly from human-equivalent automated deforestation detection. The manual change mapping process is labour intensive. Automation will allow technicians to spend less time finding deforestation and more time minimizing deforestation. In addition to deforestation mapping, it's a small step from deforestation to forest degradation, fire monitoring and species mapping, and this technology can be adapted to identify tree species and to map change, health and species in the mangroves along tropical coastlines. Mangroves are currently being assessed for inclusion in the monitoring of Guyana’s forests. In addition to Guyana, there are dozens of countries worldwide that would benefit from this as well.

CHAPTER 3: DATA SELECTION AND PREPARATION

SECTION 3.1 DATA NEEDED TO ACHIEVE THE RESEARCH AIM

There are two types of Earth observation sensor used in this research; these are multispectral sensors and SAR sensors. Multispectral sensors record spectral information: reflected visible and infrared light that comes from the Sun and is reflected by vegetation and other surfaces on the Earth's surface. One drawback of multispectral images is that they cannot see any feature past a closed vegetation canopy. Another drawback is that multispectral optical sensors cannot penetrate clouds, and thus large gaps can appear in the data when there are clouds and cloud shadows present in the imagery. As with many tropical regions, Guyana has frequent cloud cover. These issues hamper work with optical satellites and means that SAR satellites may bring considerable benefits.

SAR is an active sensor type. This means that SAR emits its own microwave energy, and then records the intensity of the microwave backscatter from the ground surface and return to the sensor. Backscatter intensity and polarisation are determined by surface texture and dielectric properties. One advantage of SAR images is that they are not affected by cloud cover, and another advantage is that some SAR configurations can provide structural information in three dimensions. In addition, SAR images are useful in forest mapping, because they are sensitive to forest structure changes even after the canopy has closed (Milodowski et al., 2017).

Different sensor types can provide complementary information. Features that are indistinguishable or invisible to one sensor may be visible to another sensor. Two objects that have the same spectral properties cannot be distinguished in multispectral images. Likewise, two objects that have the same surface texture and dielectric properties cannot be distinguished in SAR images. Sicre et al. (2020) found that classification of crops improves in accuracy when SAR and multispectral data are both used.

This research provides a detailed look at the performance of different sensors and algorithms in image classification, in isolation and in ensembles. In this context, performance means a classifier's ability to separate two classes in a given dataset. Whether classes can be separated depends partly on whether they look similar or different in a given dataset, and this depends on what the sensor records.

There are also logistical and technical concerns, such as the cost of acquiring images and their technical capabilities and computational requirements. This chapter will describe in detail the

options available for data, and the implications for our understanding of image classification and the implications for industry practice.

Machine learning requires training data, validation data, training labels, validation labels and reference data.

- The training data are the data used to fit the model. In classification problems, the training data needs labels to indicate the correct class for each data point. The training labels are the labels used with the training data to fit the model.
- The validation data are the data used to test the model. The validation data are not seen during fitting. The validation labels are used to measure the accuracy of the model's predictions on the validation data.
- The reference data are the data used to create the labels. This problem requires high resolution reference data, because different datasets are being compared, and to see details on the ground not visible in lower resolution products. The reference data must have the highest quality possible – errors in the reference data will impact model accuracy.

SECTION 3.2 REQUIREMENTS

The key problem is that some land cover classes are difficult to separate, and the proposed solutions are to create ensembles of different sensor types and break down the classification process into steps. Answering this question will require data that meets these requirements:

- The data must come from more than one sensor type. Sensor types include multispectral, SAR, hyperspectral, thermal and Lidar.
- The sensor must be available in 2017, 2018 and 2019 – the years covered by the reference data.
- The sensor must have data available in the areas covered by the reference data.
- The data must either be free or funding must be available to acquire the data.

The eight sensors chosen are:

- SAR: ALOS-2 (Figure 12), Sentinel-1 (Figure 13), Radarsat-2 (Figure 14), TerraSAR-X (Figure 15).
- Multispectral: Planetscope (Figure 30), RapidEye (Figure 31), Sentinel-2 (Figure 32), Landsat (Figure 33).

The coming sections will explain in detail why these were chosen.

SECTION 3.3 WHAT DO SENSOR TYPES RESPOND TO?

The two best sensor types for this research are multispectral and SAR. There is already a large amount of research on the topic of combining them. Multispectral sensors are the most well established in remote sensing of forests, with SAR sensors being the second most common.

There is also some research on combining multispectral and hyperspectral data, where the combined dataset has the spatial resolution of the multispectral image and the spectral resolution of the hyperspectral image. This has very high computational requirements because of the large number of features combined with the large number of samples. Therefore, this research will create ensembles of multispectral and SAR images.

Multispectral sensors record slices of the visible light and infrared spectra called bands (Figure 29). Typical bands have names like Blue, Green, Red, Red Edge, Near Infrared and Shortwave Infrared, and their wavelength ranges are chosen based on the reflectance properties of surfaces. In remote sensing of forests, these are the most common uses of multispectral bands:

- **Visible Light**: Plant pigments such as chlorophyll absorb energy most strongly in the Blue and Red bands. All multispectral sensors have Blue, Green and Red bands, with three exceptions: the Landsat Multispectral Scanner, Terra-ASTER and CBERS, which have Green and Red only.
- **Near Infrared**: Near Infrared EM radiation is strongly reflected by vegetation due to refraction at cell interfaces. Leaf cell structure has low reflectance in visible light and very high reflectance in Near Infrared. Water reflects almost no Near Infrared. All multispectral sensors record Near Infrared.
- **Red Edge**: The Red Edge bands are in the transition between the two. RapidEye, SuperDove and Worldview-3 have one Red Edge band, and Sentinel-2 has three Red Edge bands.
- **Shortwave Infrared**: Vegetation moisture content has two maxima, corresponding with the two Shortwave Infrared bands. Sentinel-2 and Landsat have two Shortwave Infrared bands, and Worldview-3 and Terra-ASTER have eight.

In Marais Sicre (2020), the Near Infrared band had the best performance of the optical bands in separating crops in agriculture remote sensing. In Leckie (1990), multispectral and SAR images were compared in their ability to separate tree species near Chalk River in Canada. SAR was better at separating softwood tree species (red pine, white pine, jack pine, spruce) than visible light and infrared bands. Near infrared and mid infrared bands were important for separating

hardwood species and softwood species. Visible and mid-infrared bands were useful for separating open and forested areas.

In Ferreira et al. (2019), the separation of tree species in Brazil was improved when Shortwave Infrared bands were added to visible light and near infrared bands. Including Shortwave Infrared bands improved detection of non-photosynthetic forest litter that varies from one species to another. In addition, shortwave infrared has unique absorption features for biochemical markers such as nitrogen, water, cellulose and lignin (Kokaly et al., 2009).

SECTION 3.4 TIME COVERED

In Section 3.2, one of the criteria was that the data must be from 2017, 2018, or 2019. Table 7 shows the years in which different satellites provide data.

The medium and high resolution satellites that meet that criterion are: Landsat 7 (Enhanced Thematic Mapper), Landsat 8 (Operational Land Imager), Sentinel-2, SPOT-6, SPOT-7, RapidEye, Planetscope with the Dove Classic sensor, CBERS-4, Resourcesat. Of these, Landsat 7 and 8 and Sentinel-2 provide Shortwave Infrared capability, Sentinel-2 and RapidEye provide Red Edge capability, and all others provide visible light and Near Infrared capability.

The SAR satellites that meet that criterion are: Radarsat-2 (Figure 27), TerraSAR-X (Figure 28), ALOS-2 (Figure 25), Sentinel-1 (Figure 26).

Almost all Planetscope images used here were recorded using the Dove Classic sensor. The first Dove R images in Guyana are from the end of 2019, and the first SuperDove data is from 2020.

Table 7: Availability of multispectral and SAR satellites. A suitable satellite for this study has data in the years 2017, 2018 and 2019. More recent satellites are suitable for future work.

Medium to High Resolution Multispectral Sensors		SAR Satellites	
Planetscope SuperDove	2020-present	ALOS-2	2014-present
Planetscope Dove R	2019-2022	Sentinel-1	2014-present
Sentinel-2 Multispectral Imager	2015-present	TerraSAR-X	2007-present
Planetscope Dove Classic	2014-2022	Radarsat-2	2007-present
Landsat Operational Land Imager	2013-present	ALOS PALSAR	2006-2011
RapidEye	2008-2020		
Landsat Enhanced Thematic Mapper	1999-2021		
Landsat Thematic Mapper	1982-2012		
Landsat Multispectral Scanner	1972-2012		

SECTION 3.5 REVISIT TIME

The revisit time of a satellite is the time it takes for the satellite to image the same place on the Earth's surface again (see Table 8 below). Higher revisit range (fewer days between images) means higher temporal resolution in a time series. More importantly, in a place as cloudy as Guyana, a higher revisit rate increases the chance that the satellite will image the area on a rare cloud-free day. The best satellite for revisit rate is Planetscope (daily), and the worst is Landsat (16 days). RapidEye and Sentinel-2 are in the middle (6 and 5 days, respectively).

At the time this research was planned (2020), there were two Sentinel-1 satellites, doubling the revisit time. Partway through this research, one of them failed, and thus Sentinel-1 has less data available from 2021 onward.

Table 8: Revisit time of selected satellites.

Satellite	Revisit Time (days)
Landsat 7	16
Landsat 8	16
Sentinel-2	5
RapidEye	6
Planetscope	1

SECTION 3.6 SPATIAL RESOLUTION

An important factor in making a decision about which satellites to get data from is their spatial resolution. In multispectral images, each sensor has one resolution for all bands, or some bands at one resolution and two bands at a second resolution (see Table 9 below). The best spatial resolution is a pansharpened SPOT image, with 1.5 meter pixels, and the second best is Planetscope with 3 meter pixels. They also have the best revisit rate. However, it should be noted that the lower spatial resolution and revisit time options are also those that offer Shortwave Infrared capability, and in the case of Sentinel-2 and RapidEye, Red Edge capability. This creates a situation where all satellites offer some sort of benefit.

Table 9: Spatial resolution (pixel dimensions) of selected satellites.

Satellite	Bands	Resolution (m)
Landsat 5-9	Multispectral bands Panchromatic	30 15
Sentinel-2	Red Edge and Shortwave Infrared Visible light and Near Infrared	20 10
RapidEye	All bands	5
SPOT-6 / 7	Visible light and Near Infrared Panchromatic	6 1.5
Planetscope	All bands	3

SECTION 3.7 COST

Some satellite image products are provided free of charge, and some are available for a price (see Table 10). This has implications for both research and industry applications. For research, this means that some image products are available immediately, and some will only be available if funding is granted, and the funding may limit the quantity of data that can be acquired. For a government's REDD+ program to use commercial image products, they would need to get funding from the donor country. In the case of Guyana, the Norway Climate and Forests Initiative (NICFI) is already paying for access to Planet image products. To access all other paid

image products, a case for funding above and beyond what is already being provided would need to be made. This means that the best candidates will be Planet image products, as well as Landsat and Sentinel data.

Radarsat-2 data was acquired with funding provided by the European Space Agency's Third Party Mission Programme. TerraSAR-X data was acquired with funding provided by Airbus' Academic Support Programme.

Table 10: Cost of selected satellites, reflecting the minimum order size.

Sensor	Cost
Sentinel (all)	Free
Landsat	Free as of 2009 (United States Geological Survey)
Planetscope RapidEye	Free: 5000 km^2/month with Planet's Education and Research Program €900 per terabyte with a departmental license €540 per terabyte with a campus license
ALOS-2 PALSAR-2	Free: 25 meter resolution images, which are on Google Earth Engine Higher resolution images for a price
TerraSAR-X	€875 to €2450 for archival images, depending on resolution €1750 to €4900 for new acquisitions, depending on resolution Airbus offers a 50% academic discount Funding is available through Airbus' Academic Support Program
Radarsat-2	$3600 (single polarization) to $7800 (quad polarization) Funding is available through the ESA's Third Party Mission Programme

SECTION 3.8 CLOUD MASKING

Multispectral sensors cannot see through clouds, and the information under them is lost. Guyana is one of the cloudiest places in the world, which makes this problem especially acute in this context.

A common way to remove clouds is to create a cloud-free composite. Images from different dates have clouds in different places. If the clouds in each image are masked out, and then the images are composited, then the gaps in one image are filled with data from another image. When this is done, the temporal range needs to be kept to a minimum. If the area changes between images, then there will be discontinuities. As a result, creating cloud free composites in Guyana is often unfeasible. The approach used here is to clip out clear areas between clouds. This still results in more than enough data, especially in higher resolution images. See Figure 24 for an example.

Figure 24: The same area in Planetscope, after (left) and before (right) cloud removal. The image displays the Near Infrared, Red and Green bands.

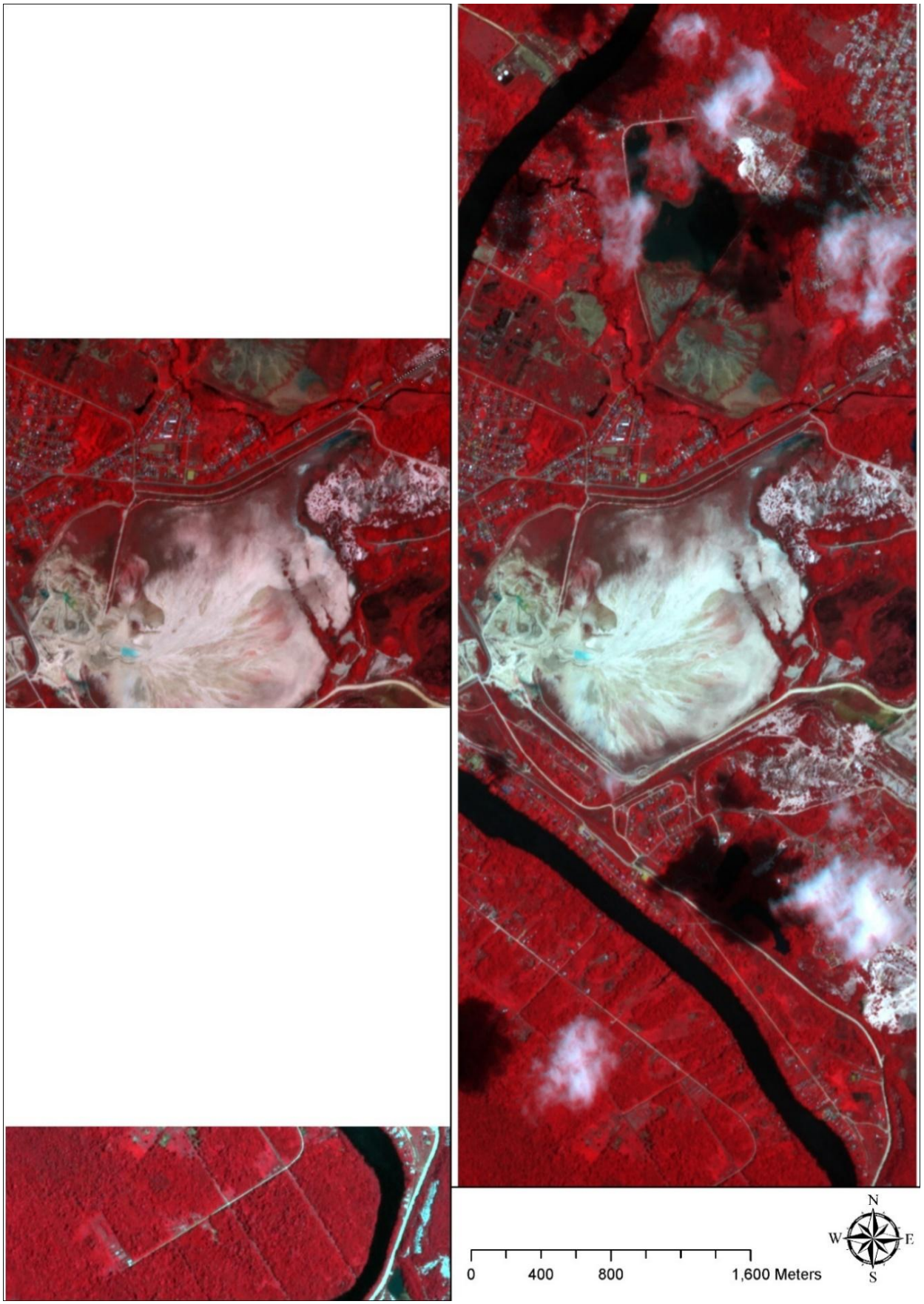


Figure 25: ALOS-2 Level 2.2 images with dual HH/HV polarization, which are ortho-rectified and have terrain correction using the ALOS World 3D 30 meter resolution digital elevation model, downloaded from Google Earth Engine (GEE). Prior to downloading, images were converted from linear units to decibels using the provided equation: γ0 = 10*log10(DN2) - 83.0 dB.

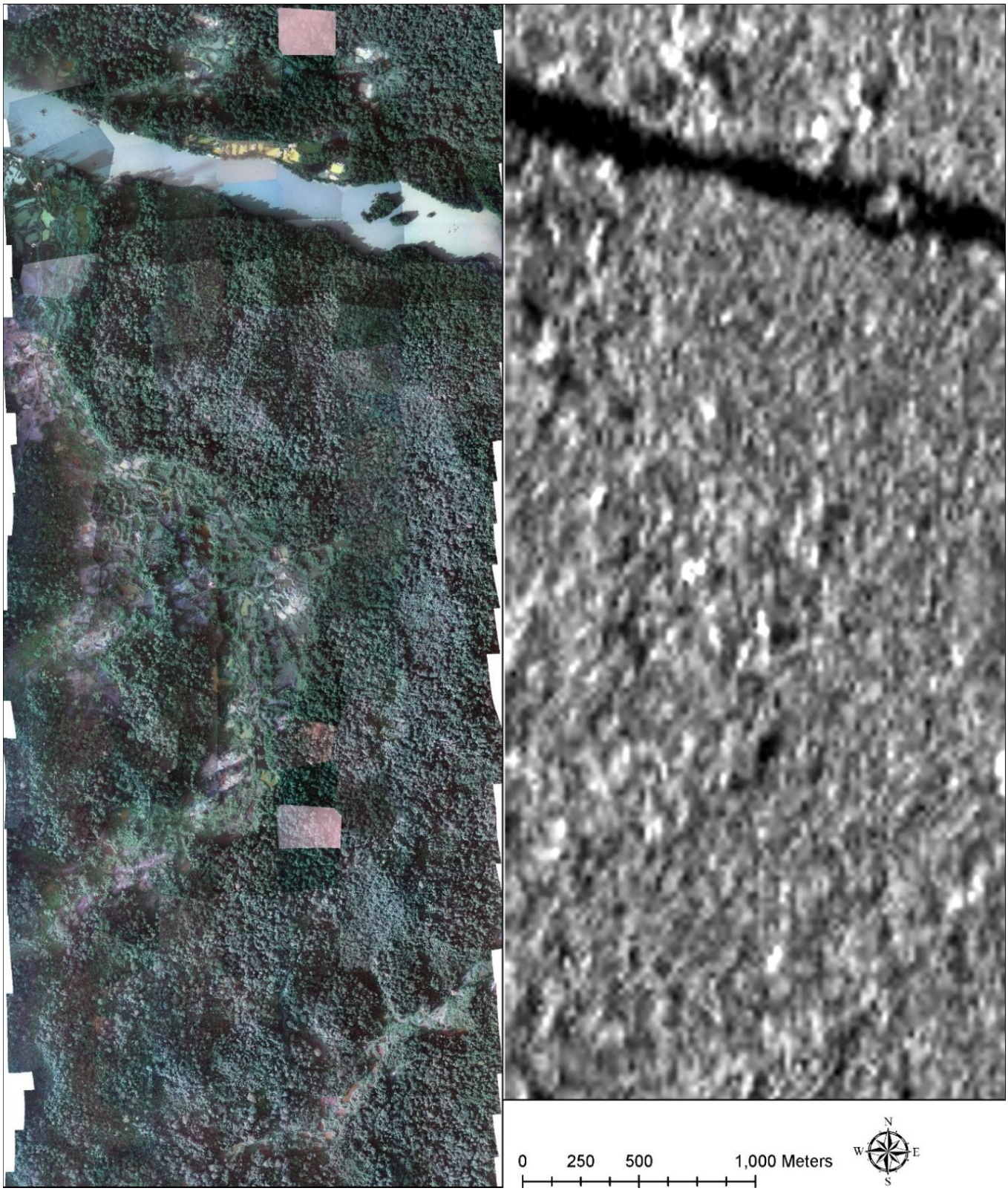


Figure 26: Sentinel-1 data in VV/VH dual polarization was downloaded from Google Earth Engine. GEE uses Sentinel-1 Toolbox apply these processing steps: 1) Thermal noise removal. 2) Radiometric calibration. 3) Terrain correction using SRTM 30 meter resolution. 4) Conversion from linear units to decibels.

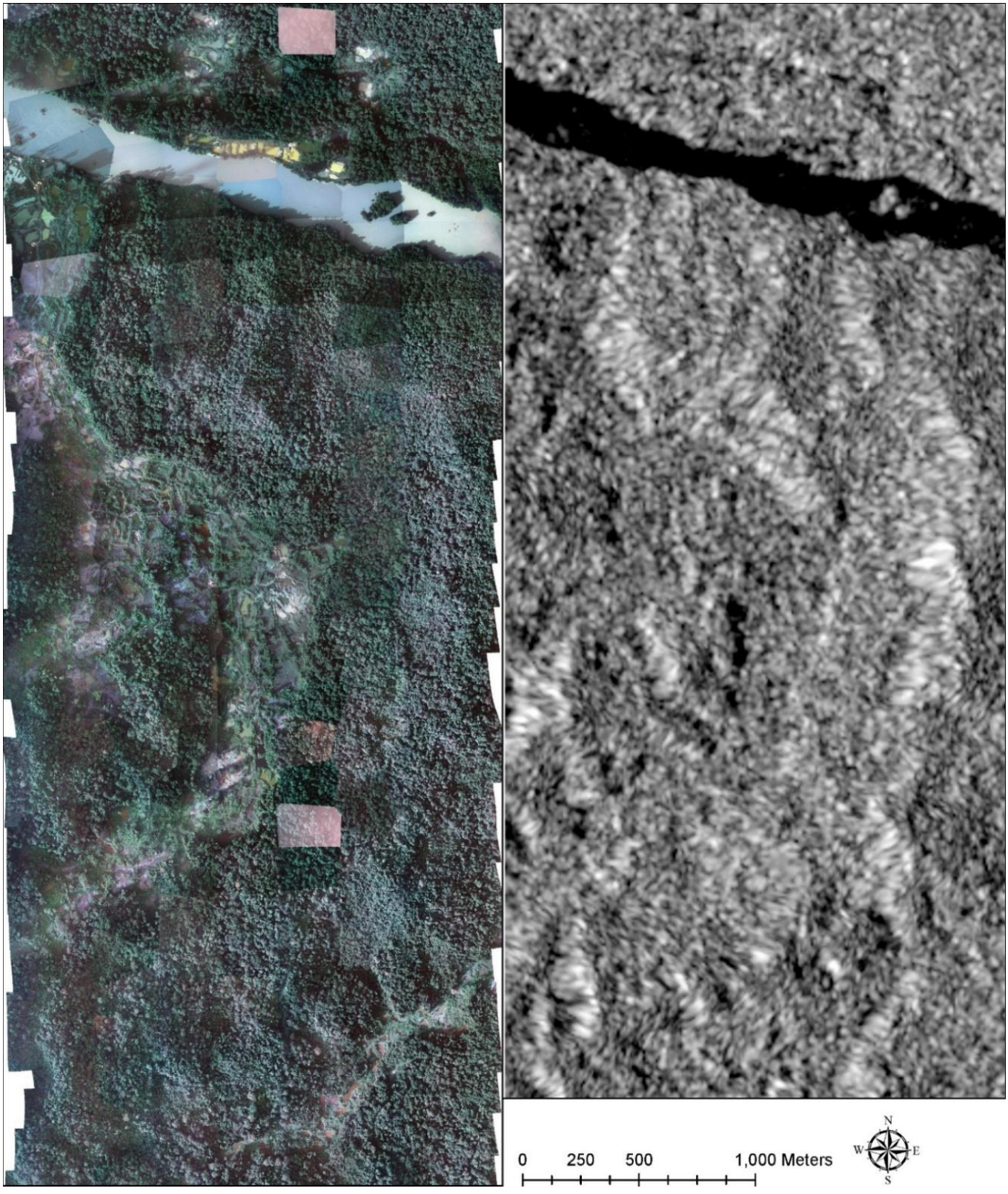


Figure 27: An example of Radarsat-2. This research uses the Extra Fine beam mode in HH single polarization and the SGX image product. SGX images have very fine pixel spacing that preserves all image information and makes images suitable for post-processing (source). Using Sentinel Applications Platform (SNAP), the images were processed for calibration to β0, speckle reduction using Refined Lee, Terrain Correction using the 30 meter Copernicus DEM, and conversion from linear units to decibels.

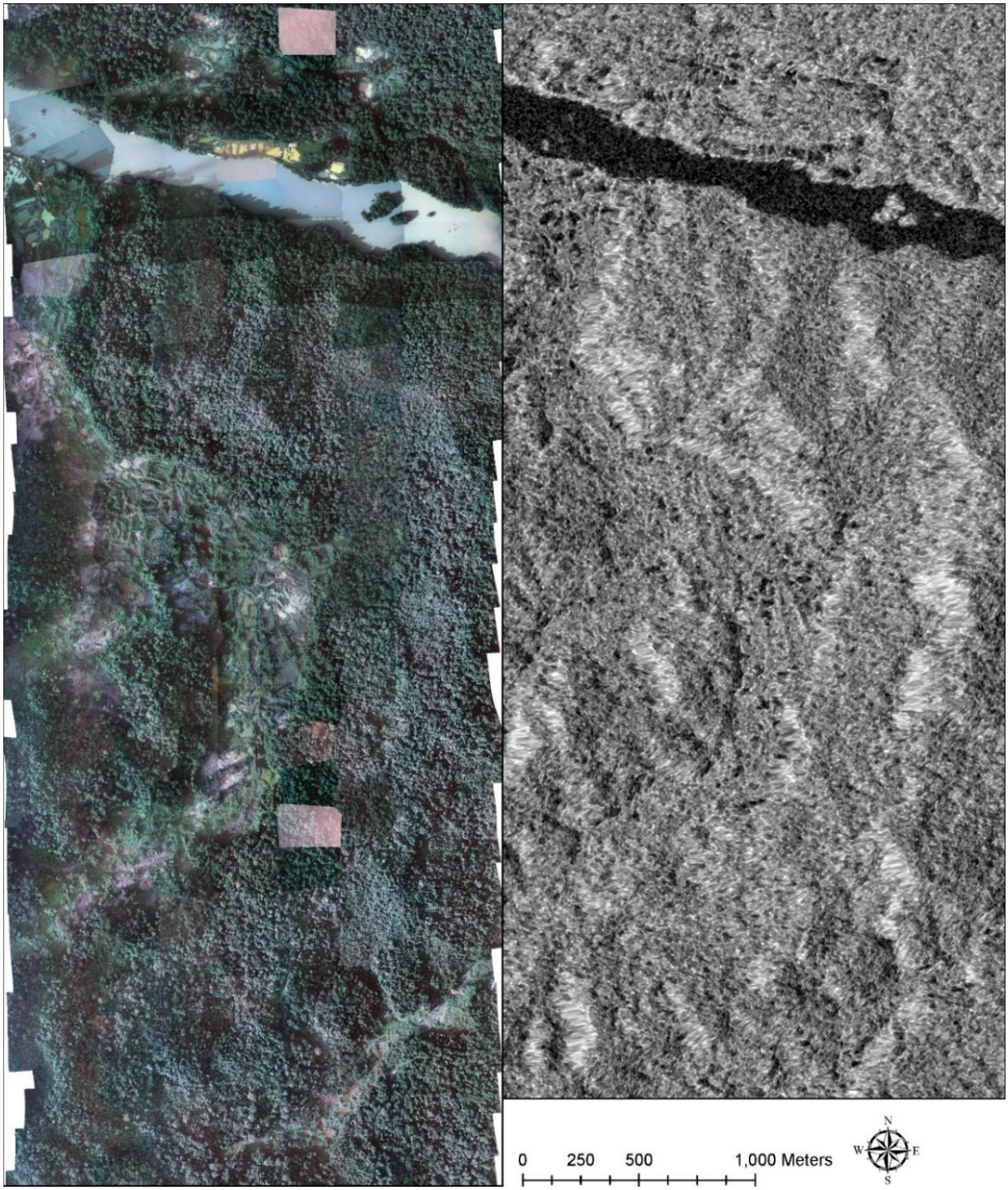
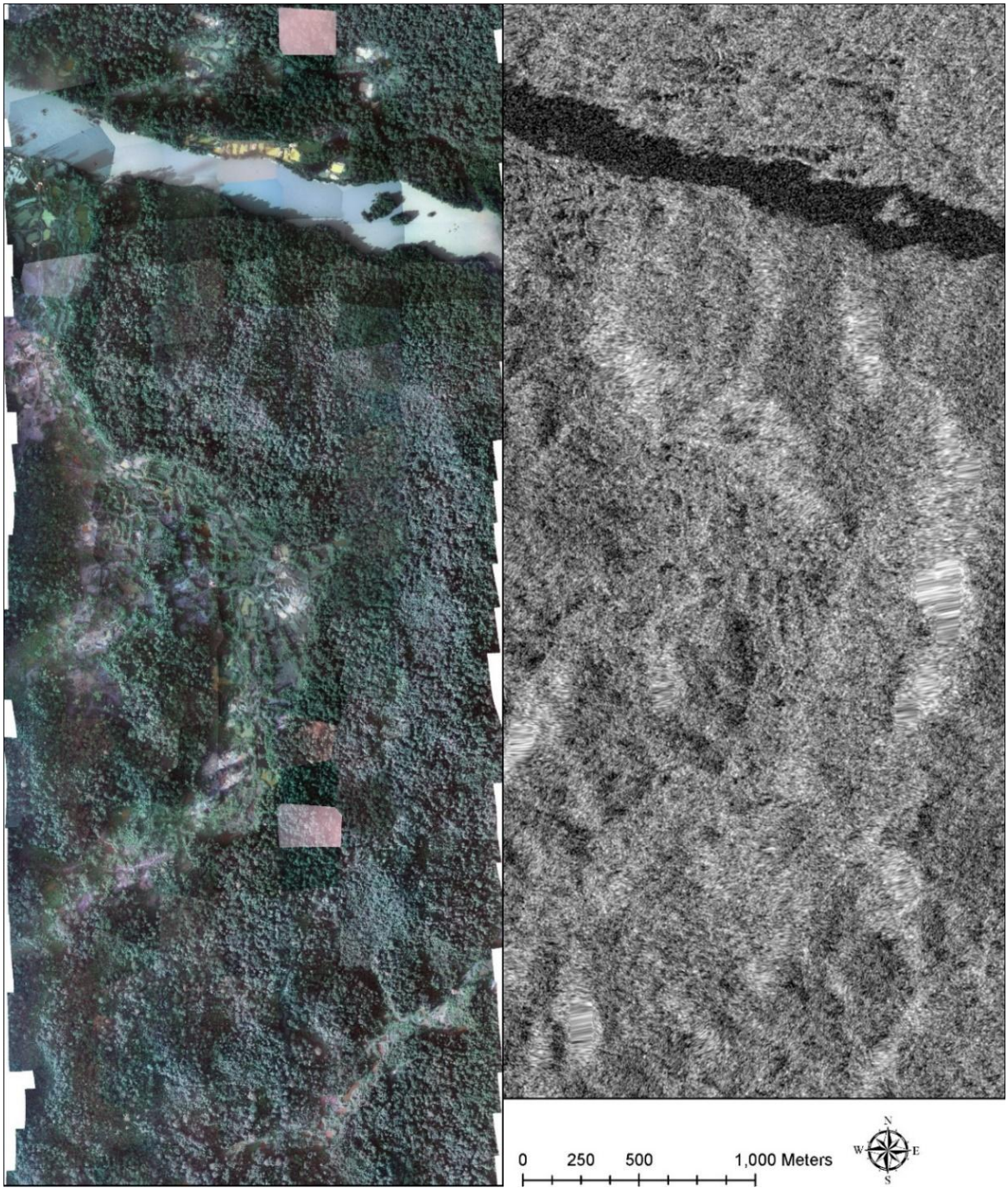


Figure 28: An example of TerraSAR-X. TerraSAR-X images were downloaded as Enhanced Ellipsoid Corrected (EEC) images. This is a multi-look product, projected and resampled to the WGS 1984 ellipsoid using an external digital elevation model. This product also features the highest level of geometric correction available for TerraSAR-X (source). After being downloaded, the only processing applied was conversion from linear units to decibels using the Sentinel Applications Platform (SNAP). The image products are StripMap (2.75 m resolution) and ScanSAR (8.25 meter resolution).



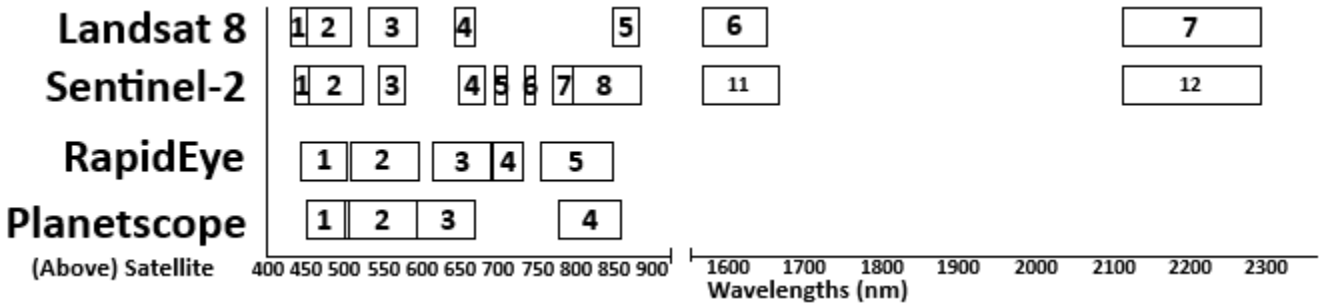


Figure 29: To summarize the spectral information of the multispectral satellites on the next four pages, this figure provides the range of wavelengths captured by each band for each sensor. The band numbers in the figure and the table correspond.

Table 11: Wavelength ranges of the selected multispectral sensors, by band.

Sentinel-2		Wavelength Nanometers	Landsat 8		Wavelength Nanometers
1	Coastal	432 - 453	1	Coastal	430 - 450
2	Blue	459 - 525	2	Blue	450 - 510
3	Green	542 - 578	3	Green	530 - 590
4	Red	649 - 680	4	Red	640 - 670
5	Red Edge 1	697 - 712			
6	Red Edge 2	733 - 748			
7	Red Edge 3	773 - 793			
8	Near Infrared	780 - 886	5	Near Infrared	850 - 880
11	Shortwave Infrared 1	1568 - 1660	6	Shortwave Infrared 1	1570 - 1650
12	Shortwave Infrared 2	2115 - 2290	7	Shortwave Infrared 2	2110-2290
			8	Panchromatic	500 - 680
Planetscope (Dove Classic)		Wavelength Nanometers	RapidEye		Wavelength Nanometers
1	Blue	455 - 515	1	Blue	440 - 510
2	Green	500 - 590	2	Green	520 - 590
3	Red	590 - 670	3	Red	630 - 685
			4	Red Edge	690 - 730
4	Near Infrared	780 - 860	5	Near Infrared	760 - 850

Figure 30: An example of the Planetscope data. Funding has been provided by the Norway International Climate and Forests Initiative via the Guyana Forestry Commission.

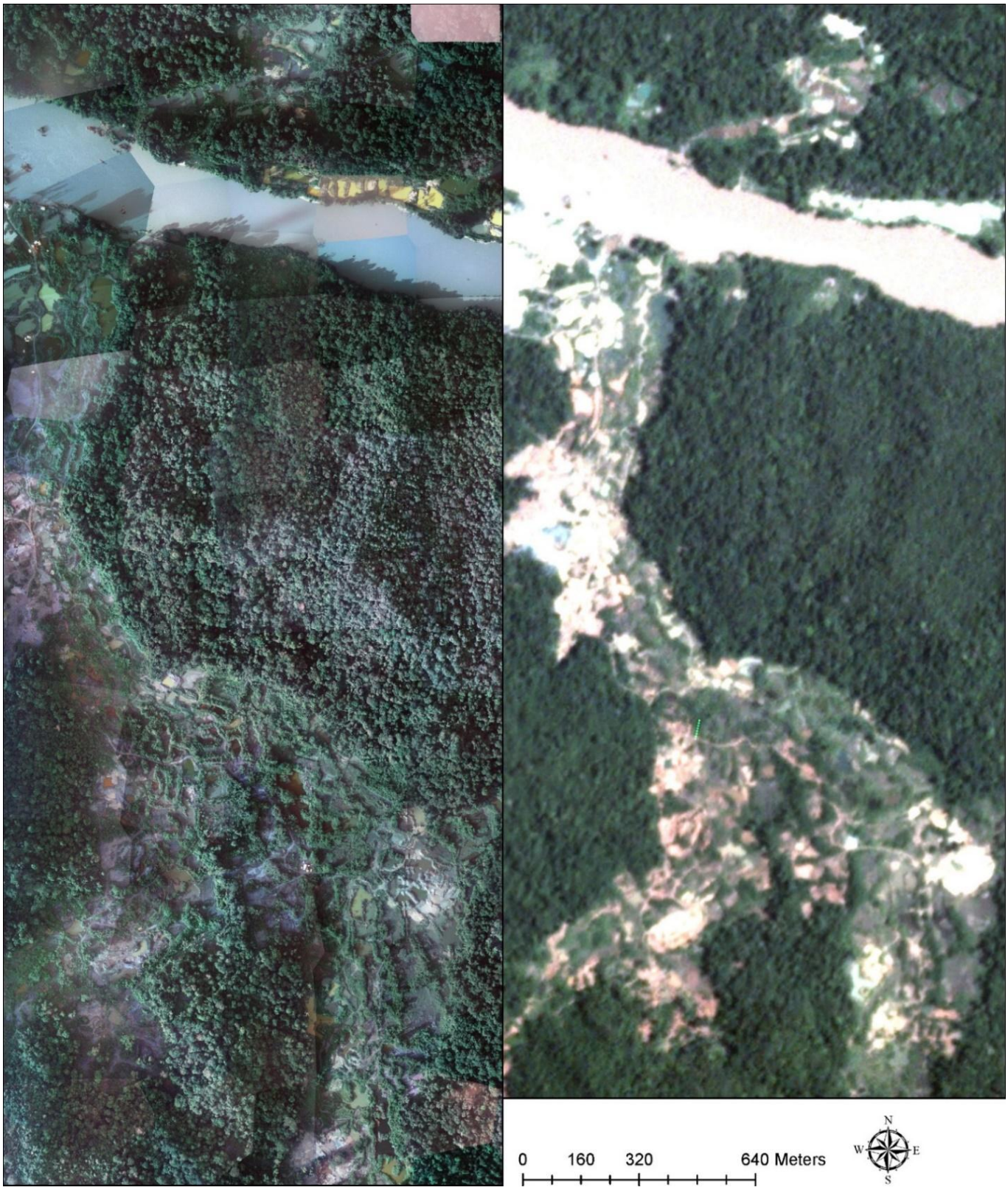


Figure 31: An example of the RapidEye data, showing a settlement. This site was chosen because, due to limited coverage, there is no RapidEye data at the same location as the other seven examples. Funding has been provided by the Norway International Climate and Forests Initiative via the Guyana Forestry Commission.

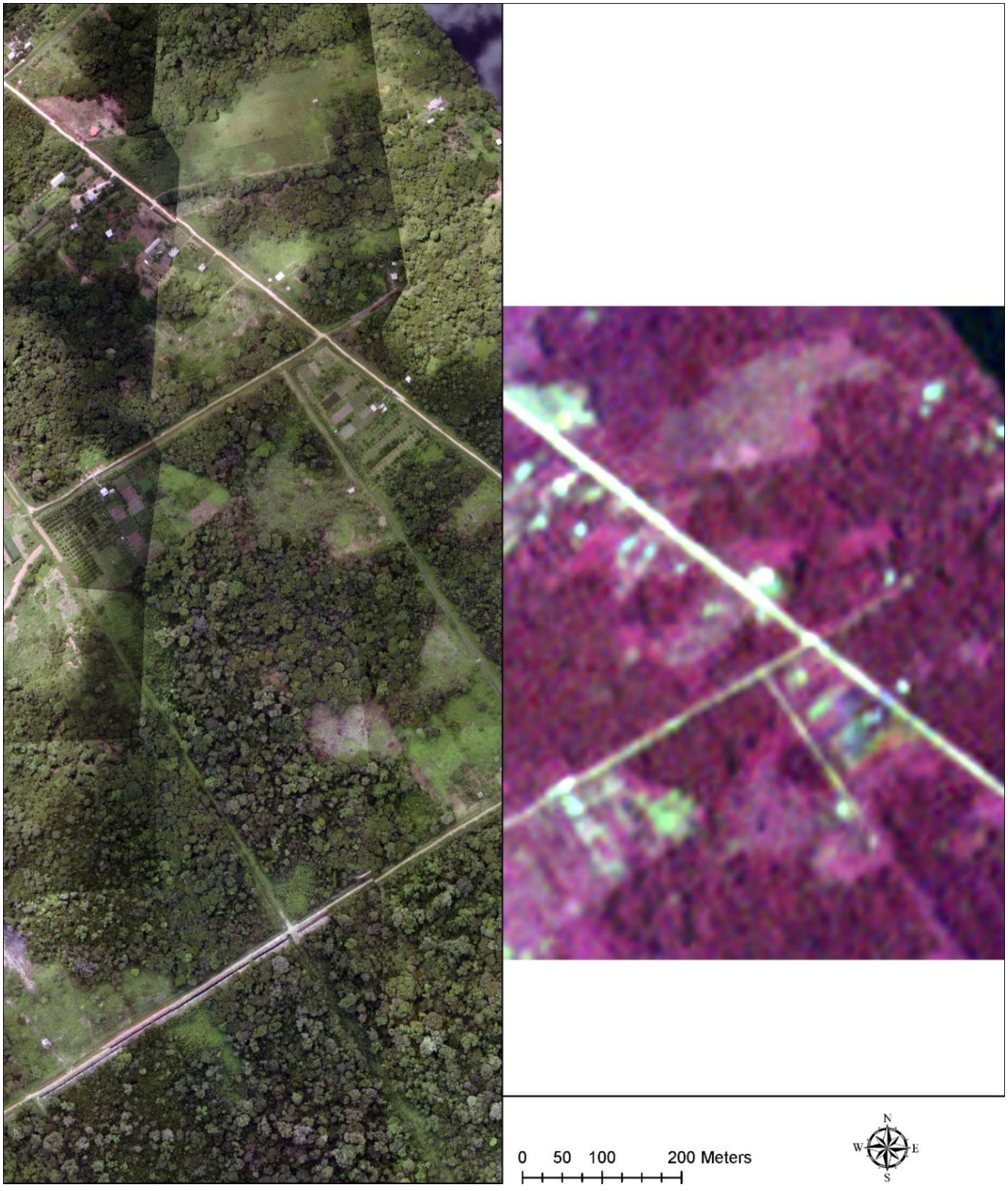
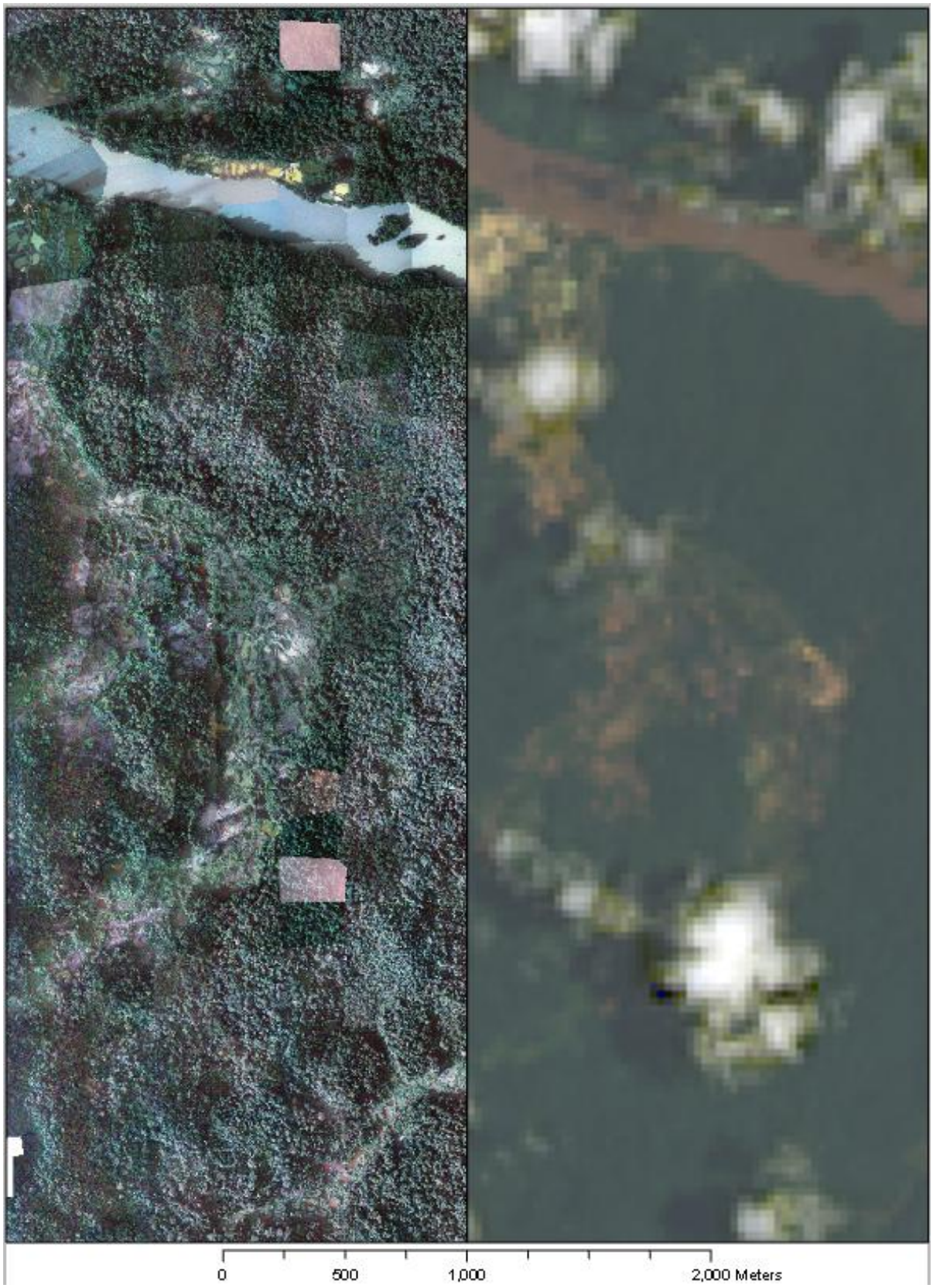


Figure 32: An example of the Sentinel-2 data.

Figure 33: An example of the Landsat data. The data used consists of a clip of the cloud-free part of the mine, in the middle of the image.



SECTION 3.9 REFERENCE DATA

In machine learning, the labels state the correct answer for each data point. The algorithm works under the assumption that every label is correct. This means that training requires high quality reference data that can be used to create accurate labels. For 50 years in remote sensing, high quality reference data has been a vital part of high quality data analysis (Stehman and Foody, 2019). When using machine learning on satellite images, using high quality reference data is vital. Foody (2009, 2010, 2011, 2013) and Foody et al. (2016) found that even a small number of imperfections in the reference data reduces accuracy considerably.

The Guyana Forestry Commission has been acquiring very high quality reference data. In 2017, 2018 and 2019, they ordered aerial photography of a sample of high-risk areas. One example is Figure 34 on the next page; it shows mining on the left side and a river in the top right. The reference data is the GeoVantage aerial photography sensor. These photographs were taken using a GeoVantage multispectral sensor, with Blue, Green, Red, and Near Infrared bands, mounted on a Cessna. Depending on the altitude, their spatial resolution (25 to 60 cm) is higher than any satellite, which means that very small details are visible in them. These air photos serve as the reference data. The class labels for training and testing were derived by tracing over the land cover boundaries in these very high-resolution images.

Because this is where very high-resolution reference data is available, this is where the data was selected. For each satellite sensor, an image was clipped out that matches the extent of each air photo. Approximately 70 air photos were acquired in each year over three years, so most sensors have 210 images total (some sensors have fewer due to limited coverage).

Figure 34: An example of the reference data (left) and the class labels (right). A river is visible in the upper right of the image, and mining is on the left side. Recorded using a GeoVantage sensor on a Cessna. Pixel size depends on altitude and is below one meter.

CHAPTER 4: SELECTING A CLASSIFIER

SECTION 4.1 INTRODUCTION TO THE METHODOLOGY

This chapter will test classifier performance using the most important satellites in an area that provides the best available data, before moving on to more difficult cases. The three most common sensors in remote sensing of forests are Planetscope, Sentinel-2 and Landsat, and they will be the first three to be tested, before moving on to other satellites in later chapters. The first three classifiers to be tested will be three widely used classifiers available in Google Earth Engine: Random Forest, Gradient Boosted Trees and Naïve Bayes.

CLASS LABELS

The mining site has two types of surfaces inside it: bare land, and a mining pond (the example is in Figures 35-38). The process is:

- An essential step in classification is creating good labels, and Figures 22 and 23 provide some clues that will help to decide where the class boundaries should be drawn. Figure 22 shows the Enhanced Vegetation Index (EVI) values for each pixel. EVI uses the Near Infrared, Red and Blue bands to provide a measure of vegetation vitality. EVI is calculated as: EVI = G * ((NIR - R) / (NIR + C1 * R – C2 * B + L)) (USGS, n.d.). In Figure 22, the boundary between dense and healthy vegetation (dark green) and a lack of vegetation (light green to white) is very clear. It shows where the river is, and where the mining site is, including a path to the right of the bare land area.
- This was then used as a guide to creating the boundaries between forest and the non-forest classes.

Figure 38 shows the results of clustering using X-Means in Google Earth Engine. Clustering shows which pixels are similar to and different from which other pixels, which determines how a classifier will behave when given this data. It provides some additional information: the river is spectrally homogeneous, and the mining pond is clustered with the river. The bare land is clustered separately from the river, and from much of the forest. With this in mind, there are a few considerations to keep in mind when drawing class boundaries:

- For a human doing policy work, a mining pond is not a river. It's part of a mining site.
- For a classifier, the mining pond is in this example is spectrally identical to the river, and if a large minority of the mining site is part of the river, then there will be confusion.

- Unbalanced classes can lead to high accuracy in the majority class and low accuracy in minority classes; Appendix D provides a detailed discussion of this matter.

The decision was made to lean toward the needs of the classifier, and label the mining pond as river. This results in three classes that are as spectrally homogeneous and distinct as possible, in the cleanest data available.

CHOOSING A PLATFORM

The platform used for image processing in this chapter is Google Earth Engine. All pixels in each sensor's image were selected, and assigned a random number. 90% of pixels were assigned to the training data, and 10% of pixels were assigned to the validation data. Figure 36 shows the spatial distribution of training and validation pixels, superimposed over the class labels. After that, the results of the comparison begin.

Google Earth Engine is not the only platform capable of machine learning on satellite images – the Python programming language allows the NumPy, Scikit-Learn and Tensorflow module to be used to analyze satellite images (see Table 9 for a summary of the differences). The advantages of Python are the wide range of classification algorithms and metrics, and options for sample size, feature and hyperparameter tuning. Scikit-Learn offers decision trees, three nearest neighbours algorithms, five linear models, five implementations of Naïve Bayes, multilayer perceptrons, bagging and boosting ensembles, and clustering. TensorFlow offers the components needed to built a wide range of neural network architectures. Google Earth Engine offers seven classification algorithms: Classification and Regression Trees, Random Forest, Gradient Boosted Trees, Support Vector Machines, Naïve Bayes, K-Nearest Neighbours and Minimum Distance.

At first glance, it seems that Python is orders of magnitude better than Google Earth Engine in the range of classifiers available. There are a few things to consider:

- Google Earth Engine is a ready-made platform for analysis of satellite data, where much of the data is built into the platform. A Python platform must be built from scratch.
- GEE analysis is done on their supercomputers. Python analysis must be done on the researcher's own computers.
- Measuring the performance of the full range of classifiers is simple: record and report the confusion matrices and other accuracy metrics for each one in a table. In choosing a classifier for a specific task, this offers the widest possible range of options. In gaining

insights into classifier behaviour, however, this wide range of options makes a study much more complicated.

The conclusion is that the most important metrics are Confusion Matrix, Consumer's Accuracy, Producer's Accuracy and Accuracy. In practical terms, Scikit-Learn offers the wider range of metrics, but Google Earth Engine has all of the necessary metrics. The extra metrics in Scikit-Learn do not offer anything indispensable that the Google Earth Engine metrics do not have. For results on the effectiveness of Scikit-Learn classifiers, see Appendix A. Appendix B provides a detailed overview of the classification metrics available.

SAMPLE SIZE

The sample size and strategy needs to be chosen carefully – they affect accuracy greatly. Huang et al. (2002) found that changing sample size and feature set had as great an impact on accuracy as changing the classifier. Rogan et al. 2008 found that reducing the number of training samples by 25% decreased the accuracy of neural networks by 10% and single decision trees by 35%. Pal and Mather (2003) found that, as the sample size increases from 700 to 2700, decision tree accuracy goes from 78% to 84%.

For sample size tuning, Scikit-Learn has a learning curve function that runs the classifier on a list of train sizes or fractions. It can run multiple sample sizes in parallel, and when available, it will use Partial Fit to speed up the process. Google Earth Engine does not have an equivalent function; sample sizes must be tested manually.

In every test, the prediction and probability images were created using the entire image. The confusion matrices were calculated using 10% of the pixels in each image, the 10% not used for training.

FEATURE ENGINEERING

It has been known as early as Serpico et al. (1994) that classification accuracy and computational efficiency depend heavily on the feature set used (in a remote sensing context, the bands). For feature engineering, Scikit-Learn offers several algorithms, which start with one feature and add features until improvement stops, or start with all features and remove features until improvement stops, or select features based on the Feature Importance measure (decision trees) or coefficients (linear models).

Google Earth Engine does not have an equivalent function; different features must be tested manually. Appendix C provides results on feature engineering using this data.

HYPERPARAMETER TUNING

The performance of a classifier depends heavily on the choice of hyperparameters, and different applications call for different hyperparameters. A review of various tuning methods is available in Yang and Shami, 2020. For hyperparameter tuning, Scikit-Learn has three options:

- Grid Search Cross Validation. This is the most exhaustive and the most computationally expensive. Ma et al. (2021) uses a multi-stage grid search process, where the first step uses a small number of samples over a broad search space, and subsequent steps sample a smaller range of values based on the results of previous steps.
- There is also the approach known as "grad student descent", where the researcher manually tries different hyperparameter values and adjusts the range of values using experience, guessing, and the results of previous tests (Yang and Shami, 2020).
- Random Search Cross Validation. This method creates a distribution of hyperparameter combinations in a specific search space, randomly samples this distribution, and tests the sampled hyperparameter combinations. This is a more efficient approach to searching a large search space than Grid Search CV, which tests every possible combination. This is introduced in Bergstra and Bengio (2012).
- Bayes Search Cross Validation via the Scikit-Optimize module. Like Random Search CV, Bayes Search CV creates a distribution of hyperparameters in a specified search space. Bayes Search CV uses the result of each iteration to update its belief about the location of the optimal hyperparameter combination, moving closer and closer with each iteration. Early stopping can be used to halt the search when improvement slows.

Google Earth Engine does not have built-in hyperparameter tuning; hyperparameters must be tested manually. One approach is to do tuning in Python and carry it over to GEE.

CHOICE OF CLASSIFICATION ALGORITHM

This research is a test of the capabilities of multispectral and SAR satellite data, alone and in combination. A large classifier comparison, in addition to that test, would multiply the scope of the project, but to produce useful results, it is important to find a well-suited classification algorithm. This section will test three algorithms: Random Forest, Gradient Boosting, and Naïve Bayes. These are implemented in both Google Earth Engine and Python.

Naïve Bayes uses Bayes' Theorem to estimate the probability that a data point x belongs to a class y. A core assumption of the Naïve Bayes algorithm is that all features are independent of one another (hence the Naïve in Naïve Bayes), so it considers each feature individually, without covariance between features. Features are not independent in many applications, but Naïve Bayes has still been known to be an accurate classifier regardless (James et al., 2023, page 154).

The other two classification algorithms are ensembles of decision trees. As per James et al., 2023, pg. 327-340, decision trees divide the data into segments, where each segment is assigned a class based on the mean or mode. Splitting decisions are made by optimizing a measure of the quality of a split. The set of rules that segments the data can be visualized as a tree of splitting decisions, which makes decision trees easy for a human to interpret. This makes decision trees well suited to data with complex, nonlinear relationships.

One disadvantage of decision trees is high variance: the predictions of the decision tree algorithm can vary greatly, due to a slight change in the input data (James et al., 2023, pg. 340-343). The solution provided is the Bagging algorithm. Bagging is an algorithm that constructs not a single model, but an ensemble of models. Typically, a Bagging ensemble fit dozens or hundreds of decision trees. Each decision tree is fitted not on the whole dataset, but on different random subsamples of the data. The predictions of the individual trees are then aggregated. This reduces the variance of the output.

Like Bagging, the Random Forest algorithm (James et al., 2017, pg. 343-345) is an ensemble algorithm that fits many decision trees and aggregates the outputs. Random Forest keeps the core feature of Bagging: each decision tree in the ensemble is fitted on different subsamples of the data. Random Forest further improves classification accuracy by making a second change. When constructing each split in the decision trees, a random subsample of the features is considered. The reason is that different features have different levels of influence over the construction of the decision tree. It is possible for one highly influential feature to dominate the construction process, such that the decision trees are highly correlated. Averaging the results of many highly correlated decision trees would not reduce the variance of the output by very much. Random Forest is designed to produce decorrelated decision trees that have low variance and high reliability.

Another way of building an ensemble of decision trees is boosting (James et al., 2023, pg. 345-348). Where Random Forest builds many independent decision trees in parallel, boosting builds many decision trees in sequence, and each tree depends on the trees that came before it. At each iteration, boosting fits a decision tree to the data, and the errors/residuals are measured. At the next iteration, another decision tree it trained, and the errors/residuals in the last iteration are given emphasis, so that later decision trees increasingly focus on difficult-to-classify data points. There are multiple implementations where the details of the boosting process vary, and the one used here is Gradient Boosting.

SUMMARY OF THE METHODOLOGY

Table 12 provides a detailed summary of the differences between the two platforms. Here are the key points:

- Google Earth Engine has the majority of the data already built in and pre-processed. Python has no built-in and pre-processed remote sensing data.
- Google Earth Engine has a sufficient number of classifiers for the purpose of this research. Scikit-Learn's wider range of classifiers would complicate the comparison greatly without adding more support to the argument.
- Google Earth Engine has all necessary metrics. Scikit-Learn's extra metrics do not add any necessary information.
- Sample size, feature selection and hyperparameters can be tuned in Python, and the results can then be carried over to Google Earth Engine.

In summary, this chapter will use Google Earth Engine to test three common classifiers on three common satellites. Figure 35 shows the location of these data for this comparison within Guyana. This area is one of the best available for this, for several reasons:

- The area is spectrally simple, relative to many other areas. The mining site is mostly bare earth with a single pond; other mining sites have a very complex mix of objects.
- The satellite images were recorded over a time period of one month. Because Guyana is one of the cloudiest countries in the world, there are areas where multispectral data from different satellites is up to six months apart, due to the difficulty of finding a gap in the cloud cover. Having data from the same season maximizes consistency between images.

- The boundary of the study area was chosen to avoid extreme class imbalance. The proportion of river, mining and forest pixels is as close to balanced as possible.

SECTION 4.2 RESULTS AND DISCUSSION

The overall trend in the results in this chapter is that the classification accuracy for Forest and River is very high, and Mining accuracy is high but lower. In Consumer's Accuracy, Forest ranges from 93.81% (Naïve Bayes on Sentinel-2, see Table 18) to 99.11% (Random Forest and Gradient Boosted Trees with Landsat, see Tables 19 and 20). River accuracy shows more variation: Consumer's Accuracy is 93-97% for Random Forest and Gradient Boosted Trees in Sentinel-2 and Planetscope and Naïve Bayes on Sentinel-2, 85% for Random Forest and Gradient Boosted Trees on Landsat, 80% for Naïve Bayes on Planetscope, and 0% for Naïve Bayes on Landsat.

Looking at the predicted land cover map and class probabilities for Random Forest on Planetscope (Figure 42), in the area labelled Mining, the predicted class boundary follows the true boundary very closely. In the area labelled Path, two thirds is not predicted. In the class probabilities, the path is visible, in an area that has a lower probability of being forest, though the probability of forest is still above 50% there. The predictions are made by setting a threshold of 50% to the class probabilities, and this would suggest that the Path area would be predicted correctly if the threshold were adjusted based on the mixed pixels present.

The second classifier is Gradient Boosted Trees, and in this test the input data are from the same Planetscope imagery as the previous test. The class probabilities (Figure 43) show that the predicted class boundaries are close to the true boundaries, with similar distribution to the Random Forest results in the previous test. However, where Random Forest has high certainty everywhere except for class boundaries, Gradient Boosted Trees has low certainty everywhere. Despite this, in the predicted land cover map (Figure 40), the predicted class boundaries are close to the true class boundaries, as in Random Forest. The confusion matrix (Table 14) is almost identical to Random Forest (Table 13), despite the vast difference in class probabilities.

The third classifier is Naïve Bayes. In the predicted land cover map (Figure 41), the river was consistently classified as river. But the forest-river boundary was classified as mining, part of the bare land was classified as river, the mining pond was classified as mining, and the area classified as mining includes the edge of the forest. The class probabilities show extreme

confidence for each class (Figure 41), which is consistent with the behaviour of Naïve Bayes when given a large sample size.

The second sensor to be tested is Sentinel-2, and the first test uses Random Forest. In Figure 42, the class labels have lines indicating the class boundaries in the highest resolution reference data, which at this resolution shows which pixels will be mixed pixels. As with Planetscope, Sentinel-2 with Random Forest produces a predicted land cover map where, in the mining site, the predicted class boundaries are close to the true class boundaries. The exception is that the path disappears in the predicted land cover map but is visible in the class probabilities (Figure 42). The second test uses Gradient Boosted Trees. The same pattern appears as the test with Planetscope (Figure 43). Naïve Bayes with Sentinel-2 produces very different results than Naïve Bayes with Planetscope. This time, Naïve Bayes produces a very accurate predicted land cover map with high certainty (Figure 44).

The third sensor to be tested is Landsat, and the first test uses Random Forest. In Figure 45, the lower resolution results in a large number of mixed pixels, and half of the path disappears, because the majority of the pixel is forest. The resulting predictions are very close to the true class labels. The second test uses Gradient Boosted Trees. In the predicted land cover map (Figure 46), the class boundaries are very close to the class labels, and again, the class probabilities (Figure 46) show low certainty, in sharp contrast with the high certainty of Random Forest. The third classifier is Naïve Bayes. With Landsat data, the predicted land cover map (Figure 47) classifies both the forest and the river as forest, but outlines the mining site well. The mining pond is mostly classified as forest except for one pixel, and there are some river pixels around the edge of the mining site. In the class probabilities (Figure 47), the probability of Forest is different for the river vs the forest. This again suggests that accurate predictions could be derived from the class probabilities using a different threshold, one other than 50%.

The purpose of this chapter is to compare the performance of three common classifiers and satellites in a spectrally simple area, and the main takeaway is that Random Forest is the best performing classifier. It produces accurate predicted land cover maps and probabilities with high confidence regardless of the data given to it. Naïve Bayes produces reasonably accurate predicted land cover maps in Sentinel-2 and Landsat, and an inaccurate predicted land cover map in Planetscope. The two differences between Planetscope and Sentinel-2 are that Gradient Boosted Trees produces accurate predictions but the probabilities have low certainty. Random Forest will therefore be the default classifier going forward. This is an observation that is not

mentioned in the existing literature on gradient boosted trees in remote sensing; Lawrence et al., 2004; Jun et al., 2021; Chirici et al., 2013 test gradient boosted trees in remote sensing classification problems, and find high prediction accuracy, but never mention class probabilities. The next comparison (Chapter 5) will test Random Forest using four SAR satellites and two multispectral satellites, in a more complex area.

Table 12: Summary of the differences between Google Earth Engine and Python.

	Google Earth Engine	Python
Data acquisition	Free and pre-processed data is built in More data can be uploaded	Data must be acquired manually Data must be pre-processed manually
Classification Metrics	Accuracy Confusion Matrices	Confusion Matrices Accuracy

	Consumer's Accuracy Producer's Accuracy	Precision Recall F1 Score Matthews Correlation Coefficient Jaccard Score Hinge Loss Log Loss
Sample Size Tuning	Must be done manually	A learning curve function is available
Feature Engineering	Must be done manually	Sequential Feature Selection Exhaustive Feature Selection
Hyperparameter Tuning	Must be done manually	Grid Search Cross Validation Random Search Cross Validation Bayes Search Cross Validation
Classifiers Available	CART Random Forest Gradient Boosted Trees Support Vector Machine K-Nearest Neighbours Naïve Bayes Minimum Distance	Decision Tree (Loh et al., 2014) Naïve Bayes K-Nearest Neighbours Radius Neighbours Nearest Centroid Logistic Regression Passive Aggressive Classifier Ridge Classifier Stochastic Gradient Descent Support Vector Machine Perceptron Multilayer Perceptron More advanced neural network architectures in TensorFlow

Figure 35: The location of the first set of tests in Guyana. This area is in the far west of the country, deep in the jungle near the border with Venezuela

Figure 36: The reference Skysat image data (top), and maps of the training data points (middle) and validation data points (bottom), superimposed over the class labels. Forest is green, river is blue, and mining is yellow. The purple dots are the locations of selected pixels.

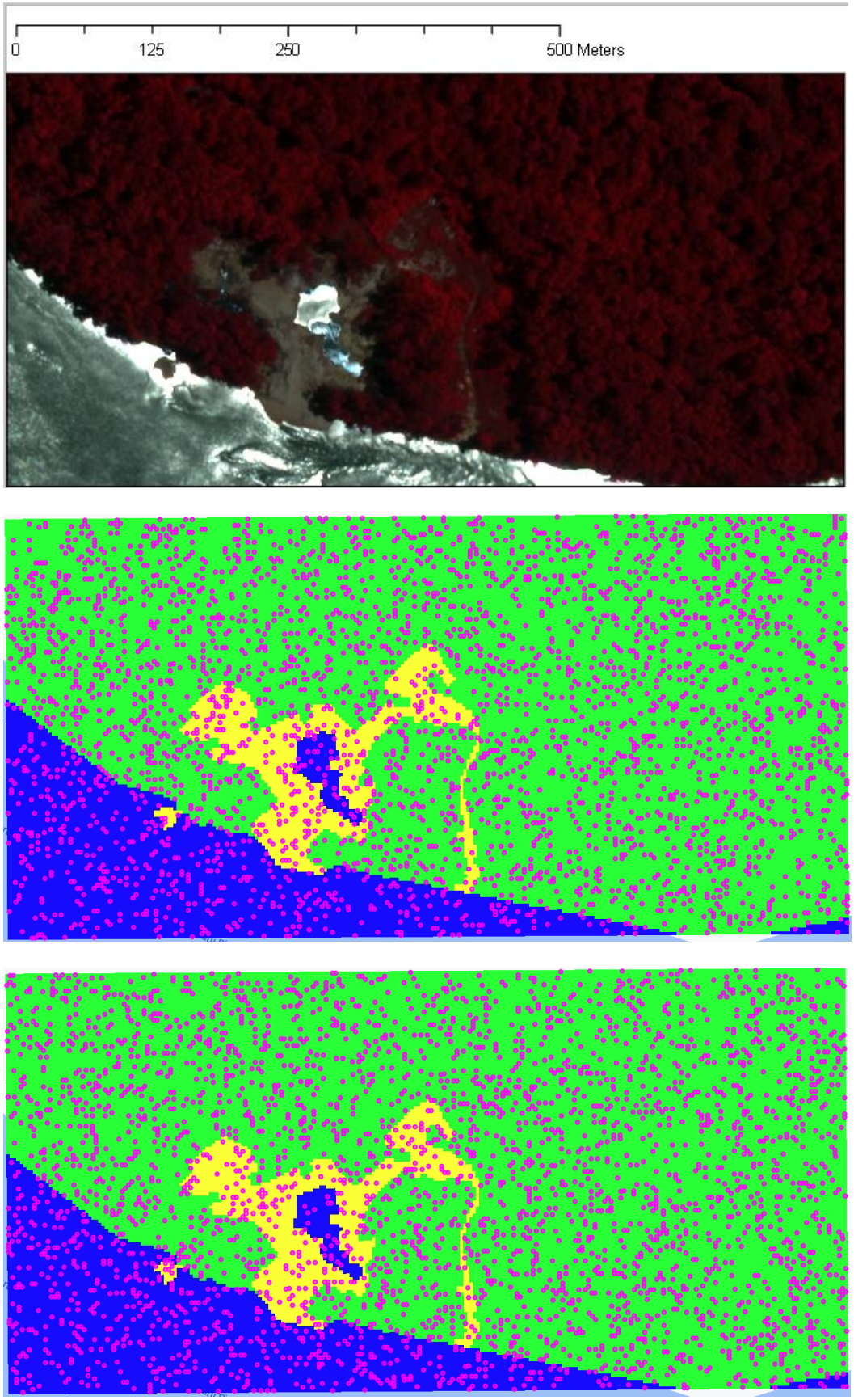


Figure 37: The Landsat source data. Top image: Red, Green and Blue bands (30m). Bottom image: Panchromatic band (15 m). The panchromatic band captures photons from the full visible light spectrum, and the multispectral bands capture photons from specific slices of the visible light spectrum.

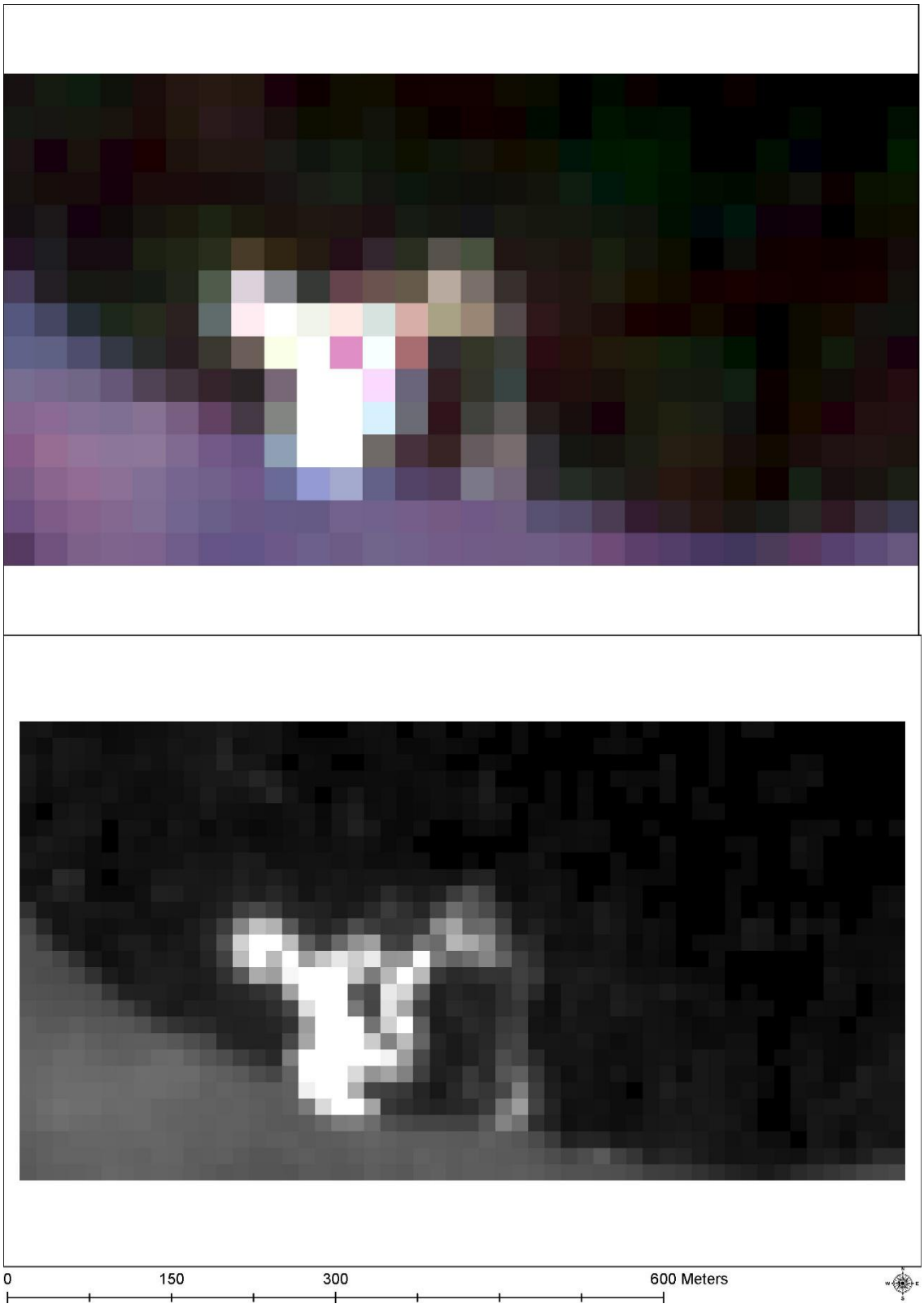


Figure 38:

Bottom left: The Enhanced Vegetation Index (EVI) values for each pixel. Dark green is the most vibrant vegetation, and the lightest green is no vegetation. It follows the forest/non-forest boundary very closely.

Bottom right: The results of clustering using X-Means Clustering. The forest and the river are in separate clusters. The bare surface is in another cluster, the mining pond is clustered with the river, and the forest-mining boundary is clustered with the water.

Top left: False colour composite image of Planetscope data used to calculate the EVI and clustering maps.

Top right: The class labels, as described in the legend on the bottom centre.

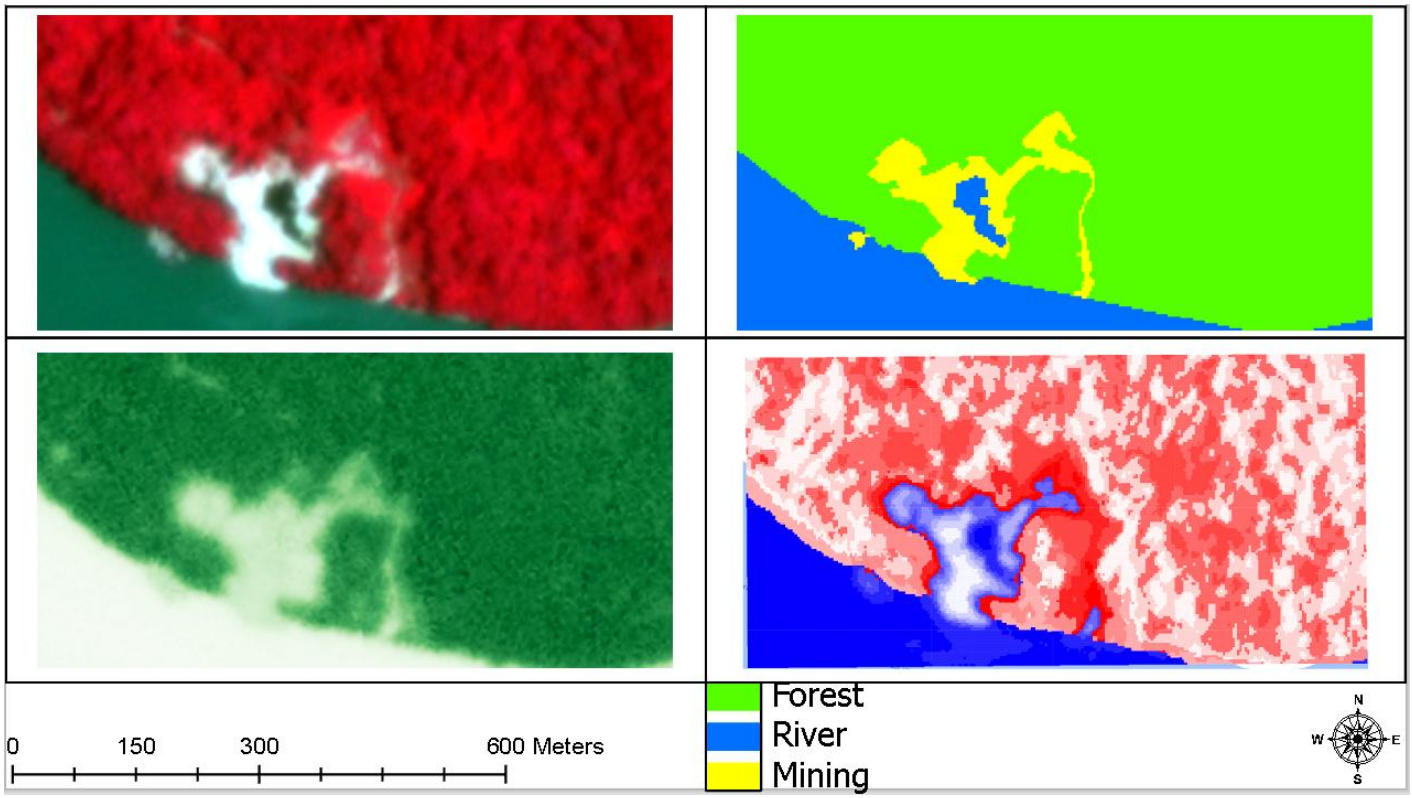


Table 13: Random Forest on Planetscope.

Accuracy: 95.99% Pixel Count: 3924		True			Consumer's Accuracy
		Forest	River	Mining	
Predicted	Forest	2470	4	34	98.48%
	River	22	539	15	93.58%
	Mining	39	18	153	72.86%
Producer's Accuracy		97.59%	96.08%	75.74%	

Figure 39: Planetscope classified using Random Forest. Data (top left), labels (middle left), predictions (bottom left), probability of Forest (top right), probability of River (middle right), probability of Mining (bottom right).

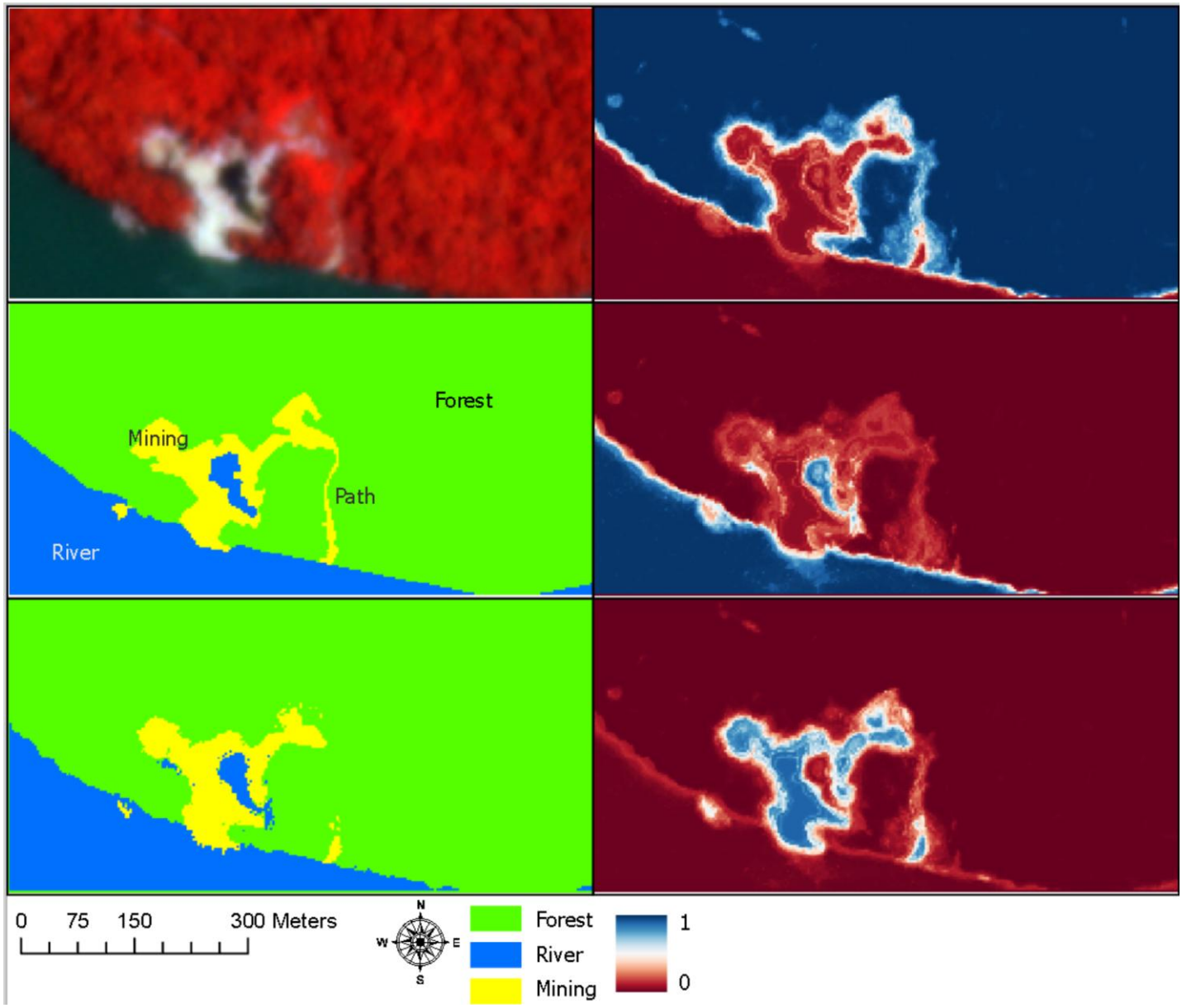


Table 14: Gradient Boosted Trees on Planetscope.

Accuracy: 96.21% Pixel Count: 3294		True			Consumer's Accuracy
		Forest	River	Mining	
Predicted	Forest	2462	4	42	98.17%
	River	19	544	13	94.44%
	Mining	30	17	163	77.62%
Producer's Accuracy		98.05%	96.28%	74.77%	

Figure 40: Planetscope classified using Gradient Boosted Trees. Data (top left), labels (middle left), predictions (bottom left), probability of Forest (top right), probability of River (middle right), probability of Mining (bottom right). The black line in the labels follows the class boundary to show mixed pixels.

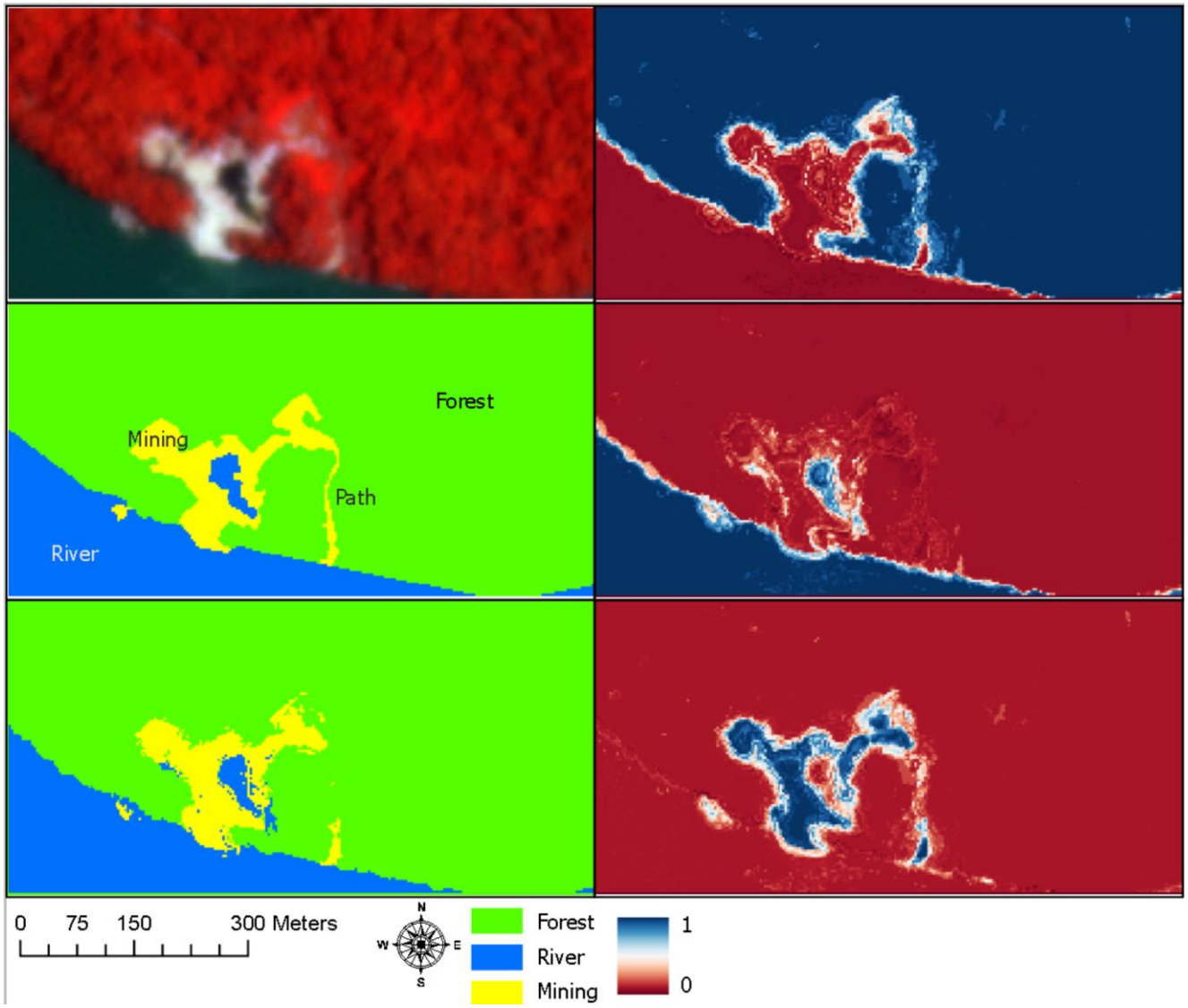


Table 15: Naive Bayes on Planetscope

Accuracy: 91.92% Pixel Count: 3294		True			Consumer's Accuracy	
		Forest	River	Mining		
Predicted	Forest	2403	1	104	95.81%	
	River	10	464	102	80.56%	
	Mining	21	28	161	76.67%	
Producer's Accuracy		98.73%	94.12%	43.87%		

Figure 41: Planetscope classified using Naïve Bayes. Data (top left), labels (middle left), predictions (bottom left), probability of Forest (top right), probability of River (middle right), probability of Mining (bottom right).

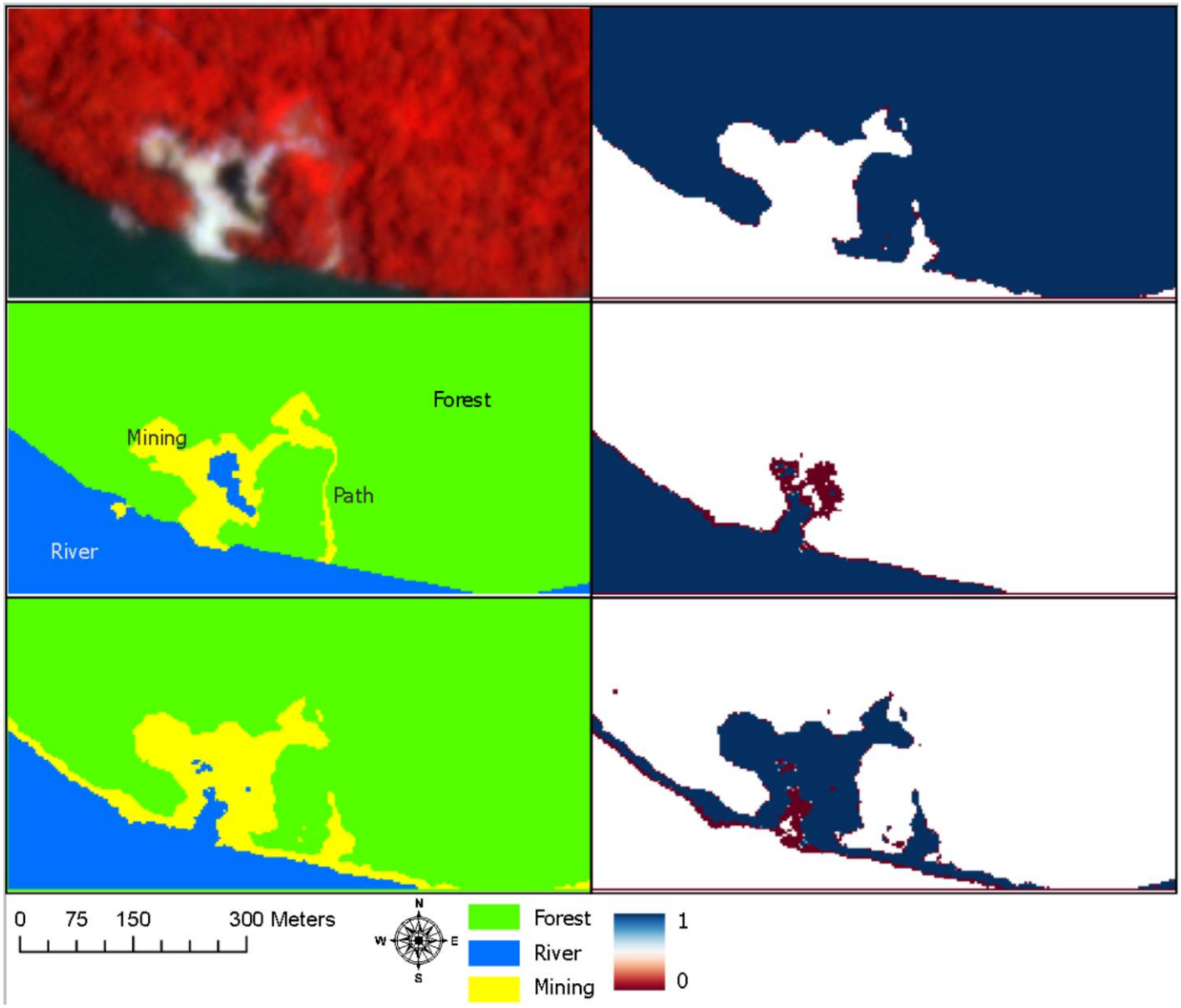


Table 16: Random Forest on Sentinel-2.

Accuracy: 95.74% Pixel Count: 305		True			Consumer's Accuracy
		Forest	River	Mining	
Predicted	Forest	219	0	7	96.90%
	River	3	54	0	94.74%
	Mining	3	0	19	86.36%
Producer's Accuracy		97.33%	100.00%	73.08%	

Figure 42: Sentinel-2 classified using Random Forest. Data (top left), labels (middle left), predictions (bottom left), probability of Forest (top right), probability of River (middle right), probability of Mining (bottom right).

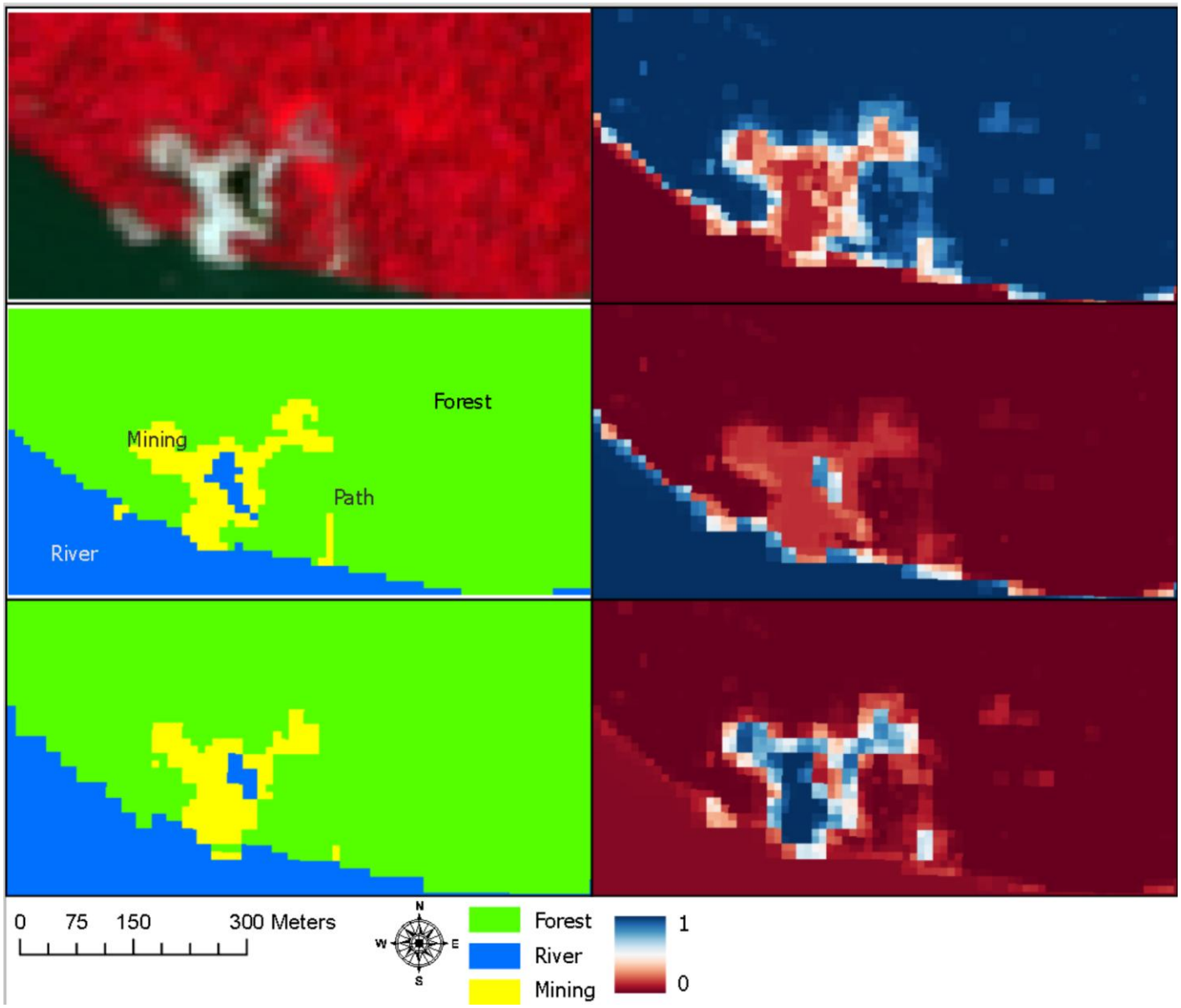


Table 17: Gradient Boosted Trees on Sentinel-2.

Accuracy: 95.08% Pixel Count: 305		True			Consumer's Accuracy
		Forest	River	Mining	
Predicted	Forest	212	0	4	95.58%
	River	5	51	1	96.49%
	Mining	1	0	21	86.36%
Producer's Accuracy		97.74%	94.83%	73.08%	

Figure 43: Sentinel-2 classified using Gradient Boosted Trees. Data (top left), labels (middle left), predictions (bottom left), probability of Forest (top right), probability of River (middle right), probability of Mining (bottom right).

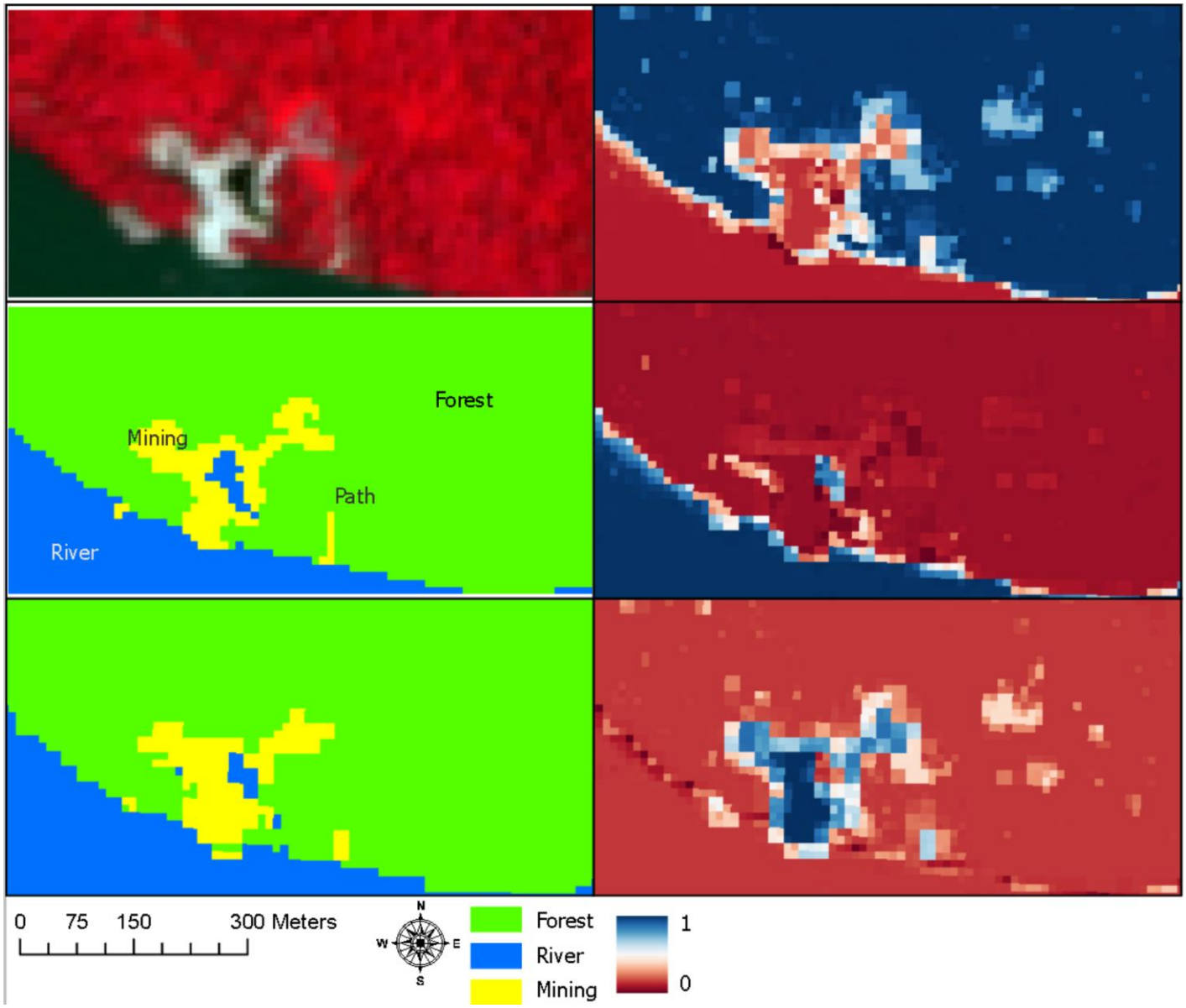


Table 18: Naive Bayes on Sentinel-2.

Accuracy: 93.11% Pixel Count: 305		True			Consumer's Accuracy
		Forest	River	Mining	
Predicted	Forest	212	0	14	93.81%
	River	5	51	1	96.49%
	Mining	1	0	21	94.45%
Producer's Accuracy		97.25%	100.00%	58.33%	

Figure 44: Sentinel-2 classified using Naïve Bayes. Data (top left), labels (middle left), predictions (bottom left), probability of Forest (top right), probability of River (middle right), probability of Mining (bottom right).

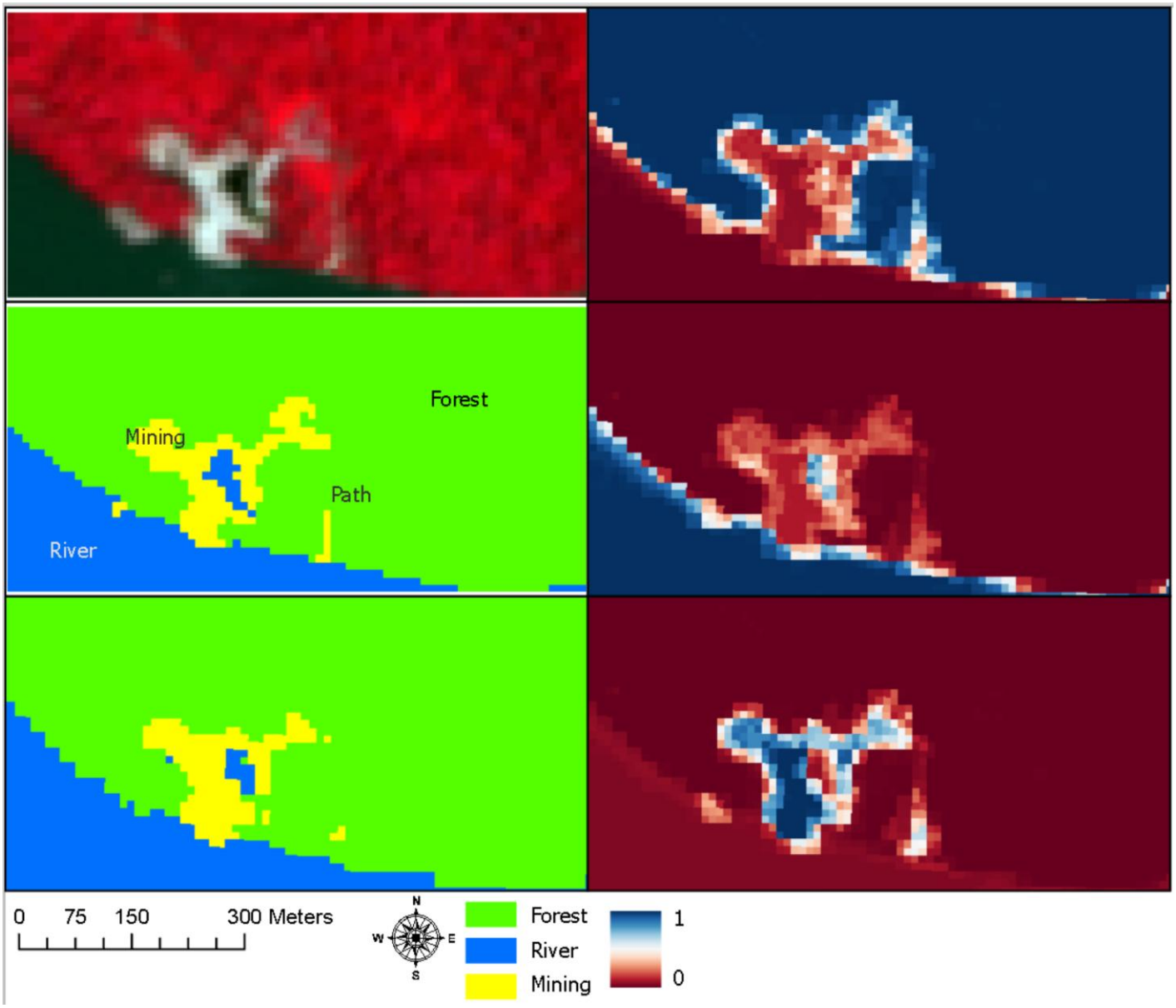


Table 19: Random Forest on Landsat.

Accuracy: 94.52% Pixel Count: 146		True			Consumer's Accuracy
		Forest	River	Mining	
Predicted	Forest	111	0	1	99.11%
	River	4	23	0	85.19%
	Mining	3	0	4	57.14%
Producer's Accuracy		94.07%	100.00%	80.00%	

Figure 45: Landsat classified using Random Forest. Data (top left), labels (middle left), predictions (bottom left), probability of Forest (top right), probability of River (middle right), probability of Mining (bottom right).

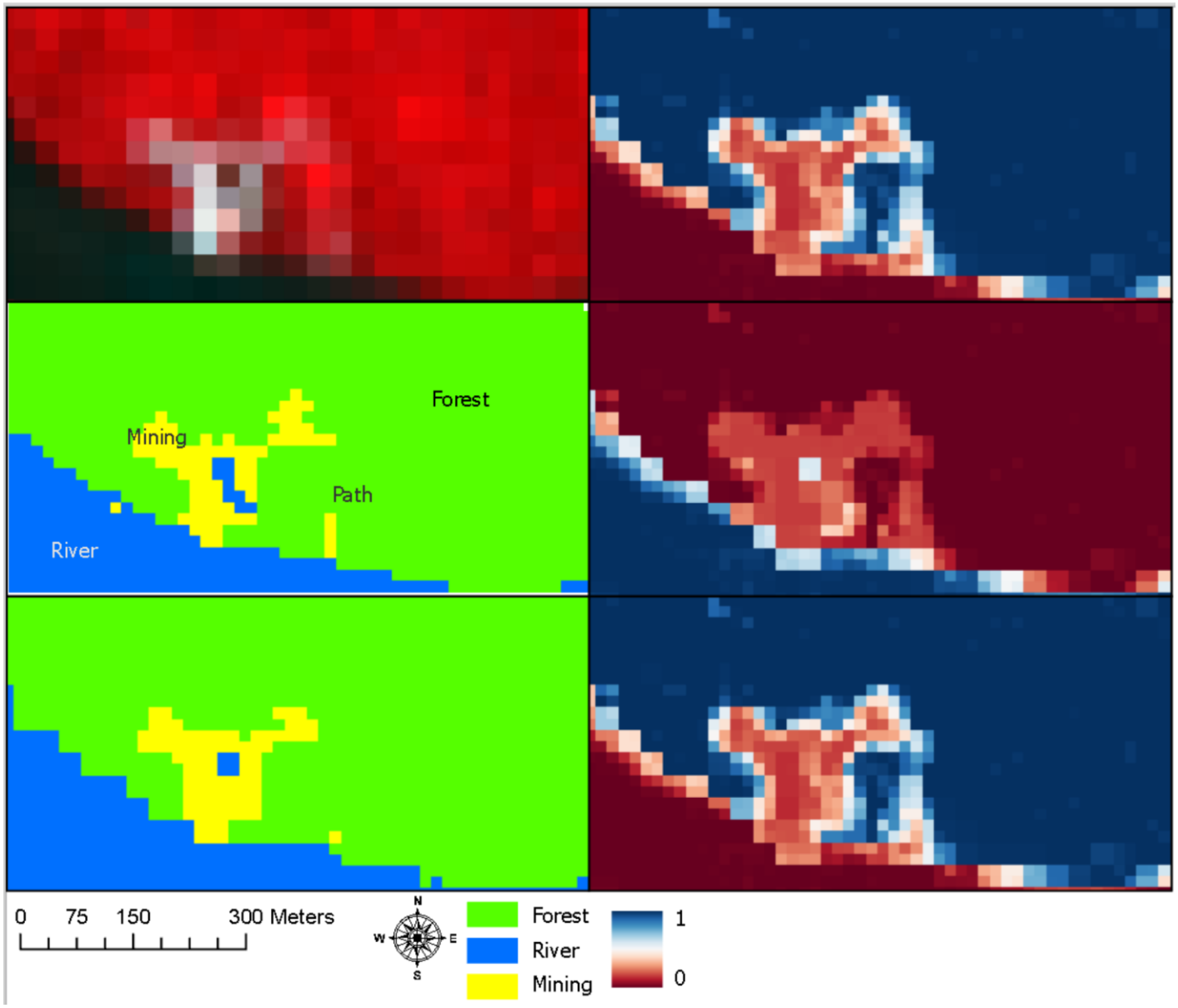


Table 20: Gradient Boosted Trees on Landsat.

Accuracy: 94.52% Pixel Count: 146		True			Consumer's Accuracy
		Forest	River	Mining	
Predicted	Forest	111	0	1	99.11%
	River	2	25	0	85.19%
	Mining	3	0	4	57.14%
Producer's Accuracy		94.07%	100.00%	80.00%	

Figure 46: Landsat classified using Gradient Boosted Trees. Data (top left), labels (middle left), predictions (bottom left), probability of Forest (top right), probability of River (middle right), probability of Mining (bottom right).

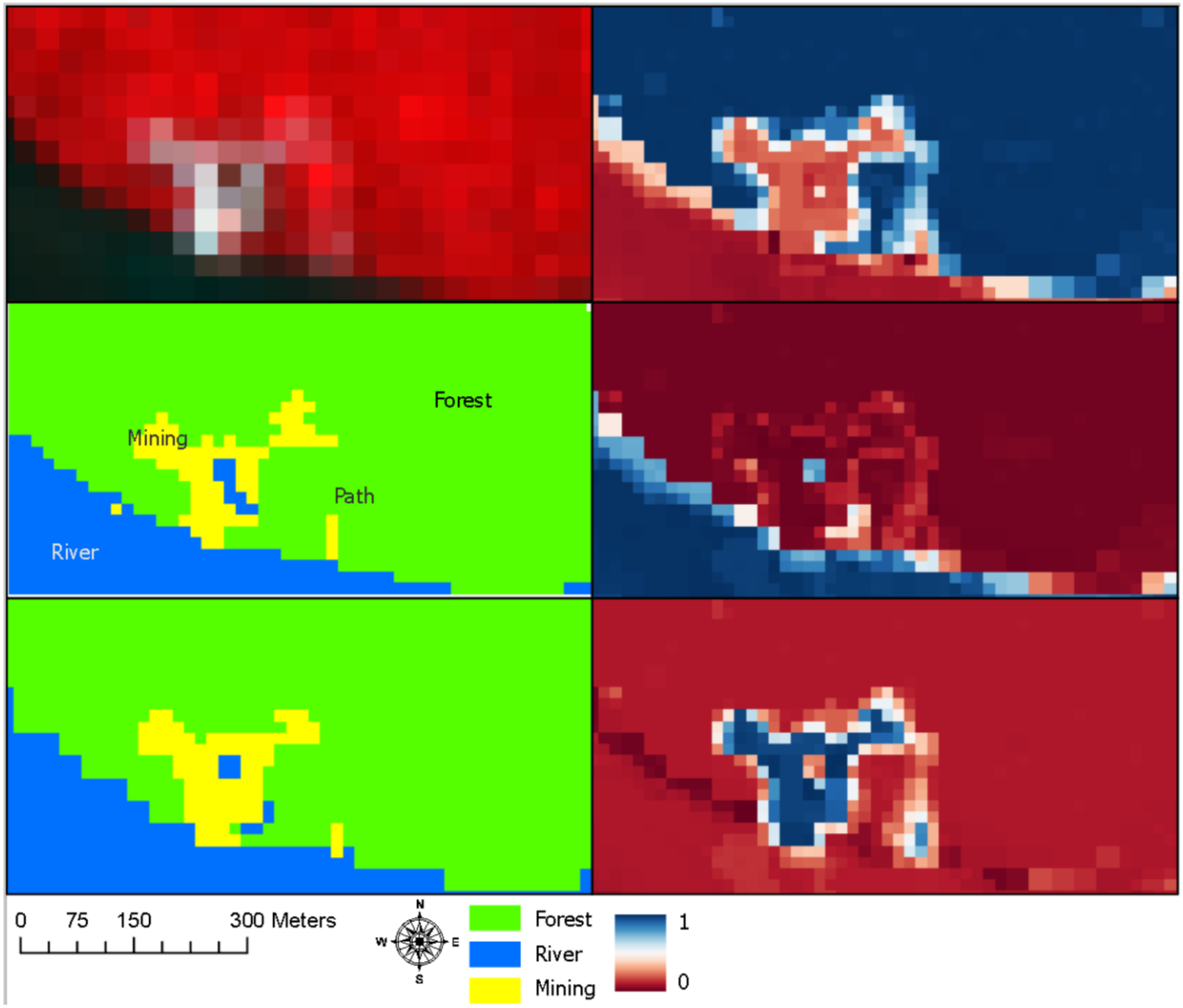
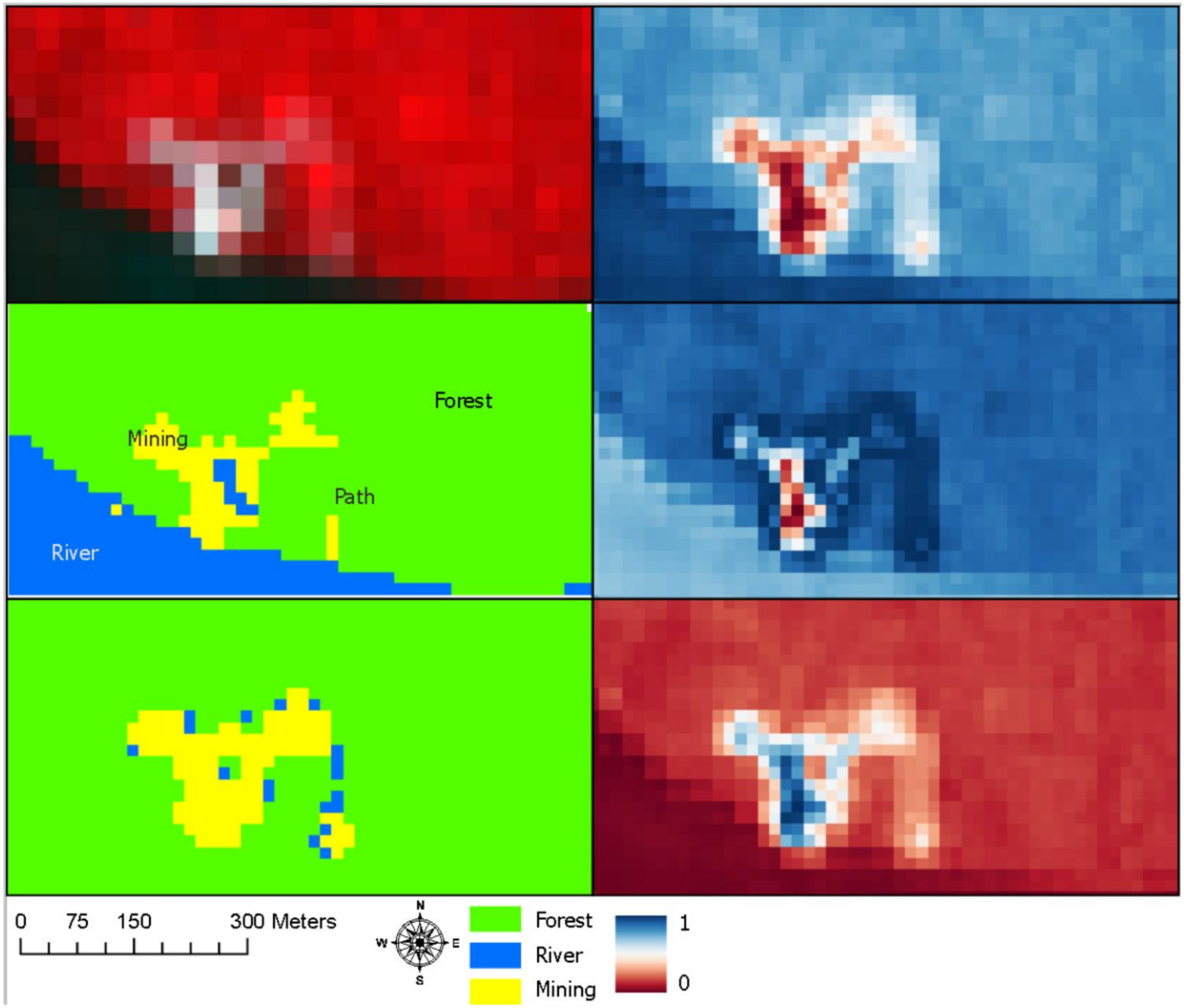


Table 21: Naive Bayes on Landsat.

Accuracy: 78.08% Pixel Count: 146		True			Consumer's Accuracy
		Forest	River	Mining	
Predicted	Forest	108	3	1	96.43%
	River	27	0	0	0.00%
	Mining	1	0	6	85.71%
Producer's Accuracy		79.41%	0.00%	85.71%	

Figure 47: Landsat classified using Naïve Bayes. Data (top left), labels (middle left), predictions (bottom left), probability of Forest (top right), probability of River (middle right), probability of Mining (bottom right).



CHAPTER 5: TESTING MORE SENSORS IN A NEW LOCATION

The study area in the previous chapter was one of the simplest Forest-River-Mining sites available, but real-world applications need to cope with more complexity than the simplest sites. The existing literature on remote sensing of forests tends to test classifiers in a single area with a simple set of land cover types. However, deforestation mapping of an entire country involves testing classifiers in a very heterogeneous set of land cover types. Rivers can be clear or turbid. Mining sites can be active (more bare surfaces), abandoned (more emphasis on new vegetation), or mixed (complex mix of bare surfaces, vegetation and mining ponds). This chapter and the next chapter will address this issue by testing sensors in two complex mining sites that both have a river passing by.

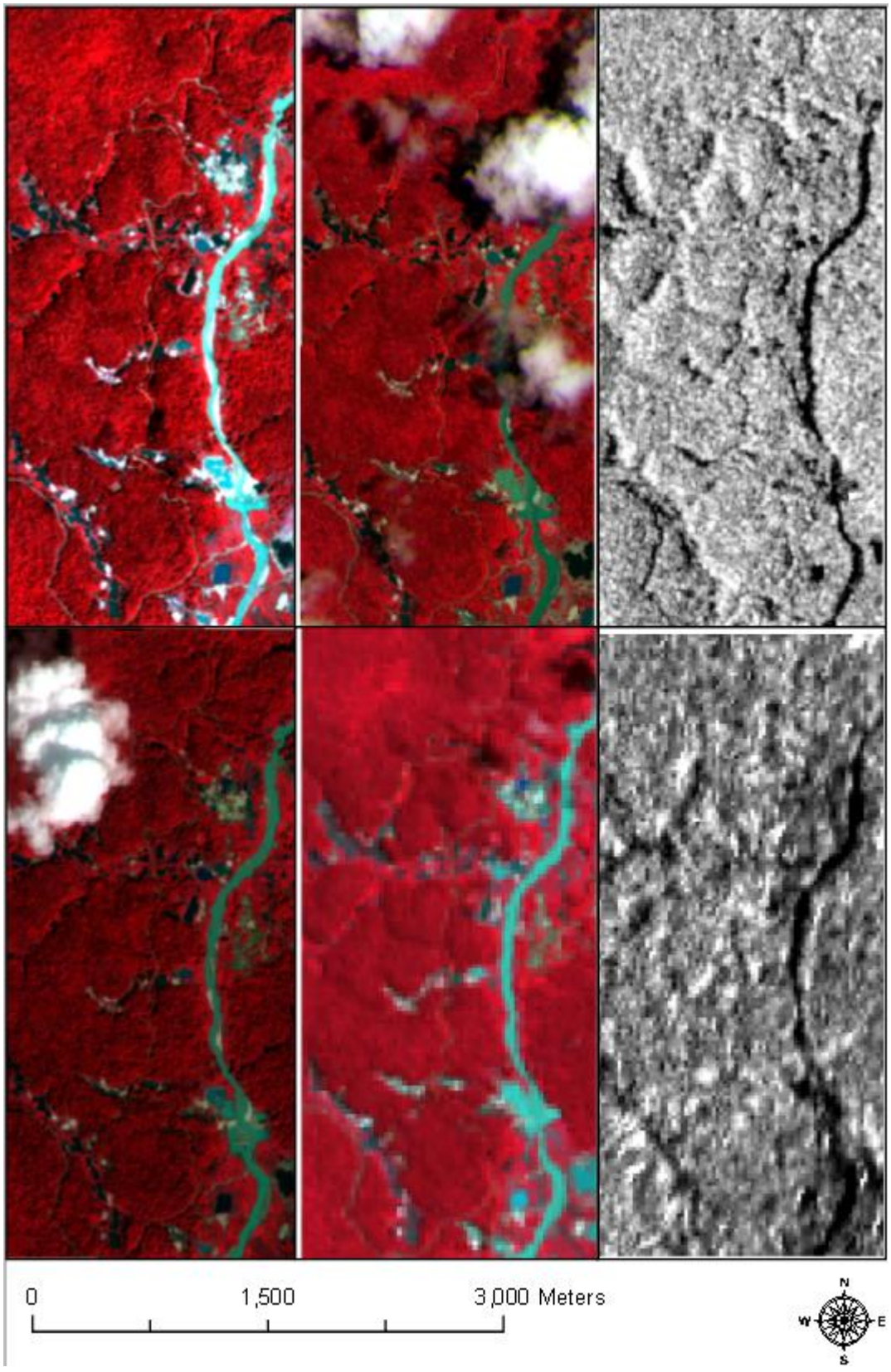
It would also be good to test a wider range of sensors. This chapter will accomplish that by comparing six sensors in a region with more complexity: a mining site beside a river that has ponds, bare surfaces and vegetation. This creates opportunities for overlap with other classes, that will test the capabilities of Random Forest with each data source. The location in Guyana is shown in Figure 47, and Figure 48 shows the multispectral data before clouds were removed. This location was chosen because it is the only site that has aerial photography of an area with forest, river, mining and roads, and all four multispectral sensors used in this study. The benefits of these sensors are:

- Planetscope is the highest resolution of the set, and has the highest revisit time.
- RapidEye is useful for historical work, because it provides five meter resolution images from 2008 to 2020.
- Sentinel-2 is the most commonly used sensor, is well calibrated, and has sufficient resolution for most remote sensing of forests work.
- Landsat is still frequently used in large-scale remote sensing of forests work.
- Sentinel-1 and ALOS-2 are SAR sensors, which penetrate clouds, and record very different information from the multispectral sensors. The ALOS-2 data is from a descending orbit, and the Sentinel-1 data is from an ascending orbit. Both sensors were right-looking.

The classifier is Random Forest, with the Max Nodes hyperparameter set to 24. This level of smoothing yielded the best accuracy. Otherwise, this is a continuation of the method in Chapter 4.

Figure 48: The location of the multispectral comparison in Guyana. The example image is Landsat.

Figure 49: The multispectral images before cloudy areas were clipped out. Top left: Planetscope. Top middle: RapidEye. Top right: Sentinel-1. Bottom left: Sentinel-2. Bottom middle: Landsat. Bottom right: ALOS-2.



Planetscope, Sentinel-2 and Landsat offer similar performance, and RapidEye has lower accuracy than the other multispectral sensors (Table 22). Of the SAR sensors, accuracy at classifying forest is similar to the multispectral sensors, but accuracy at classifying river and mining is much lower. Overall, in Planetscope (Figure 49) and Sentinel-2 (Figure 51), the predicted class boundaries were close to the true boundaries. When using RapidEye and Landsat, the predicted class boundaries are farther from the true boundaries than in Sentinel-2 and Planetscope.

Table 22: Summary of results for the multispectral comparison.

		Planetscope	RapidEye	Sentinel-2	Landsat	ALOS-2	Sentinel-1
Pixel Count		9822	2131	4060	3177	1208	7342
Consumer's Accuracy	Forest	94.69%	79.26%	92.01%	94.77%	91.77%	92.56%
	River	77.27%	59.83%	80.53%	78.95%	42.22%	22.60%
	Mining	72.57%	59.77%	80.67%	79.19%	42.63%	16.47%
Producer's Accuracy	Forest	92.16%	69.87%	89.33%	92.02%	81.44%	72.65%
	River	77.82%	71.43%	86.12%	85.71%	65.52%	51.97%
	Mining	79.42%	69.21%	83.71%	84.54%	60.45%	42.17%

When identifying the roads, Random Forest with Planetscope (Figures 49) is able to identify the majority of the road pixels. RapidEye (Figure 50) and Sentinel-2 (Figure 51) can identify a minority of road pixels, and Landsat (Figure 52) can identify almost no road pixels. Sentinel-1 (Figure 53) and ALOS-2 (Figure 54), the SAR satellites, are unable to identify roads at all.

When identifying the river, the Sentinel-2 was the best, finding almost all of the river, and the Planetscope and Landsat images also produced good river classifications. The RapidEye predictions were very good at the majority of the river, but the middle part of the river was missing a large part. ALOS-2 identifies about half of the river and is missing the other half. Sentinel-1 identifies the centre of the river but misses the edge.

When identifying the mining areas, the Sentinel-2 and Landsat predictions outline the mining areas very accurately. The Planetscope and RapidEye predictions accurately identify the majority of the mining areas, but miss large areas around the edges.

The SAR data here is much worse at identifying mining and river than the multispectral sensors (Table 22). However, the two SAR sensors are similar to Planetscope, Sentinel-2 and Landsat when identifying forest.

In the probability of river for Planetscope (Figure 49), Sentinel-2 (Figure 51), and Landsat (Figure 52), there is a darker red area that matches the forest in the true labels, and a lighter red area that matches the mining area in the true labels. In more numerical terms, there is a sharp gradient in class probabilities at the class boundary, and the two classes have distinct probabilities of river. By identifying the threshold value that matches the centre of this gradient, it would be possible to predict all three classes just from the probability of river.

In comparing the four multispectral sensors, the lowest performing sensor was RapidEye. Planetscope was the best at identifying roads, and Sentinel-2 and Landsat were both the best at outlining the mining areas. Planetscope's ability to find roads makes sense: at three meter spatial resolution, it is the only one that can resolve a feature as narrow as a road, and provide enough pure pixels for a classifier to identify it. This indicates that high spatial resolution is useful in some situations, but in many other situations, moderate resolution sensors (Sentinel-2 and even Landsat) are adequate for most deforestation mapping. The two best sensors at outlining the mining area have one feature in common that Planetscope and RapidEye lack: Shortwave Infrared capability. So far, it appears that SAR sensors perform much worse than multispectral sensors, meaning that they are not a viable alternative in cloudy areas.

At this point, all available multispectral sensors have been tested. Two SAR sensors have been tested, but data from two higher resolution SAR sensors is available from other parts of Guyana. The next chapter will compare results from four SAR sensors and three multispectral sensors in another location.

Some of the classification errors are where mining ponds were identified as River (ranges from 96/785 pixels or 12.2% for Sentinel-1 to 18/220 pixels or 8.19% for ALOS-2). However, the idea that is a misclassification, a deviant from the truth, requires some examination. The true labels were made with the policy definitions of land cover types in mind. For policy makers, what matters is where a mining site occurs in the forest. However, as far as ground features and physical processes are concerned, a mining site is not just a mining site. It is a collection of bare surfaces, mining ponds with varying chemical compositions and therefore varying spectral properties, and vegetation in varying stages of regrowth. The Random Forest classifier is picking up on this – the river classification includes similar mining ponds (which are also water), and the classifier identifies the denser vegetation within mining areas as forest. For the purposes of quantifying land cover types for policy purposes, the complexity of the mining area will result in errors. However, for the purposes of studying the physical and ecological properties of the area, the Random Forest classifier is improving on the labels, and identifying relevant features within the mining area. Therefore, classifier accuracy, how close the predictions are to the truth, depends on the context. The confusion matrices reflect the accuracy for policy purposes, but the usefulness to physical and ecological remote sensing is higher than the confusion matrices would suggest.

Three of the multispectral images have gaps in their coverage (Figures 49-51). In the case of RapidEye, the majority of the scene is missing due to the need to remove clouds. In many parts of the world, avoiding cloud cover is simply a matter of selecting data from a different date, and in a low-cloud area, it will be easy to find cloudless images that are close in time to each other. Guyana is one of the cloudiest places in the world, and this can be next to impossible. Images can be months apart, because cloud-free days were few and far between. SAR is often used as a solution to this problem, because microwaves pass through clouds. However, in this case, SAR classification accuracy was lower than the multispectral data, enough so that it isn't a suitable replacement.

SECTION 5.3 CONCLUSION AND NEXT STEPS

The take-aways from this chapter are:

- Overall, the multispectral sensors outperformed the SAR sensors.
- The ability to identify roads in SAR images is closely tied to pixel size.
- The two sensors that are best at finding mining are Sentinel-2 and Landsat: the two with Shortwave Infrared bands.
- The class probabilities show a gradient that follows the class boundary closely, even when the predictions don't. This suggests that variable thresholding could improve predictions.

The next step is to run this test again, in a different area. This test will find out whether the results are consistent across the country. Normally, when classifiers are tested in remote sensing of forests research, one study area is selected, which is very homogeneous, and therefore easy for the classifier to work with. However, nationwide deforestation mapping will require a classifier that can work with very heterogeneous data.

The next study site will also have a different set of satellites. The area is missing one multispectral sensor (RapidEye), but it does have two higher resolution SAR satellites: TerraSAR-X and Radarsat-2. The next study site will serve a second purpose: find out whether SAR performance improves with higher resolution.

Table 23: Confusion matrix for Planetscope classified with Random Forest.

Accuracy: 89.03% Pixels: 9822		True			Consumer's Accuracy
		Forest	River	Mining	
Predicted	Forest	6865	14	371	94.69%
	River	6	221	59	77.27%
	Mining	578	49	1659	72.57%
Producer's Accuracy		92.16%	77.82%	79.42%	

Figure 50: Planetscope classified using Random Forest. Top left: Data. Top middle: True labels. Top right: Predicted classes. Bottom left: Probability of forest. Bottom middle: Probability of river. Bottom right: Probability of mining.

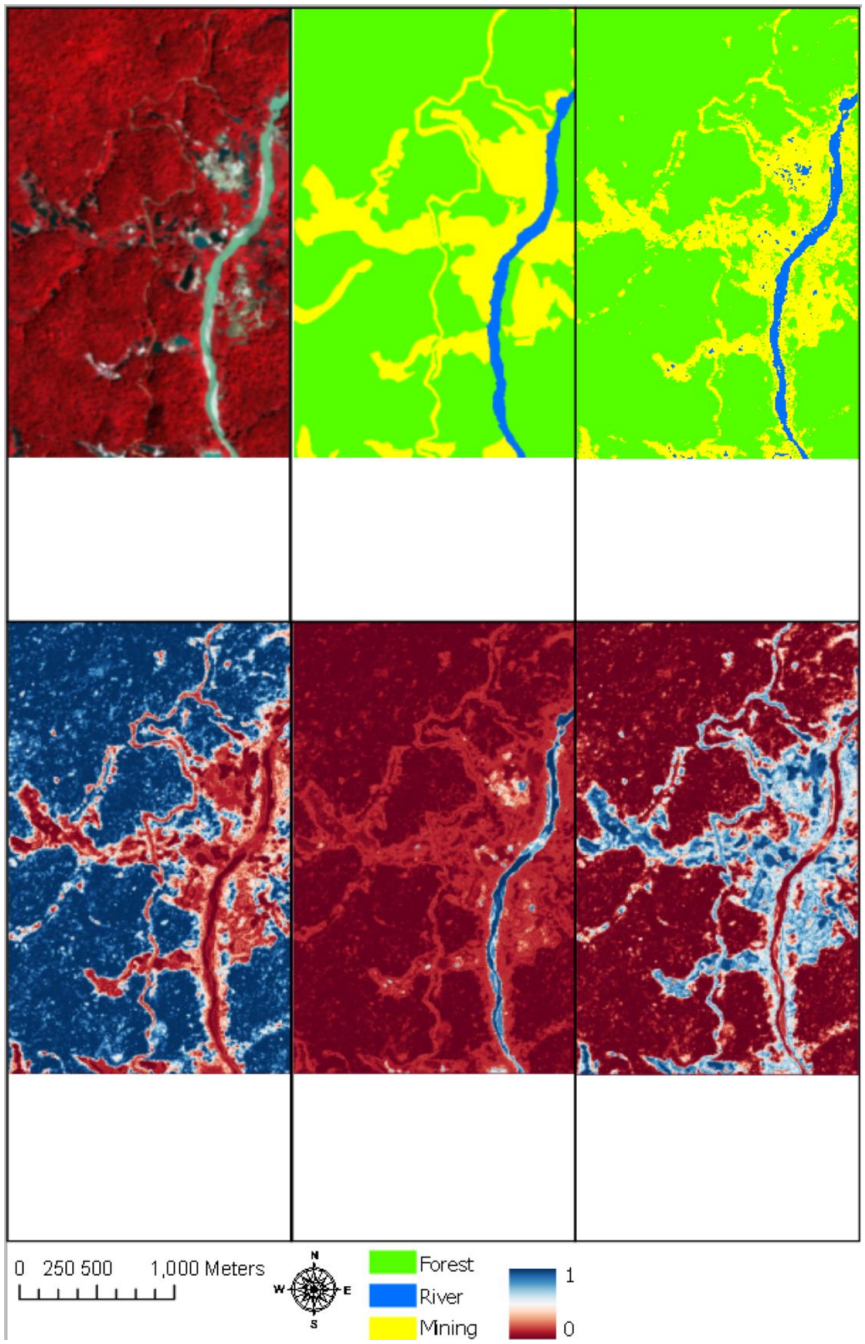


Table 24: Confusion matrix for RapidEye classified with Random Forest.

Accuracy: 69.68% Pixels: 2131		True			Consumer's Accuracy
		Forest	River	Mining	
Predicted	Forest	697	5	248	79.26%
	River	12	70	35	59.83%
	Mining	405	23	636	59.77%
Producer's Accuracy		69.87%	71.43%	69.21%	

Figure 51: RapidEye classified using Random Forest. Top left: Data. Top middle: True labels. Top right: Predicted classes. Bottom left: Probability of forest. Bottom middle: Probability of river. Bottom right: Probability of mining.

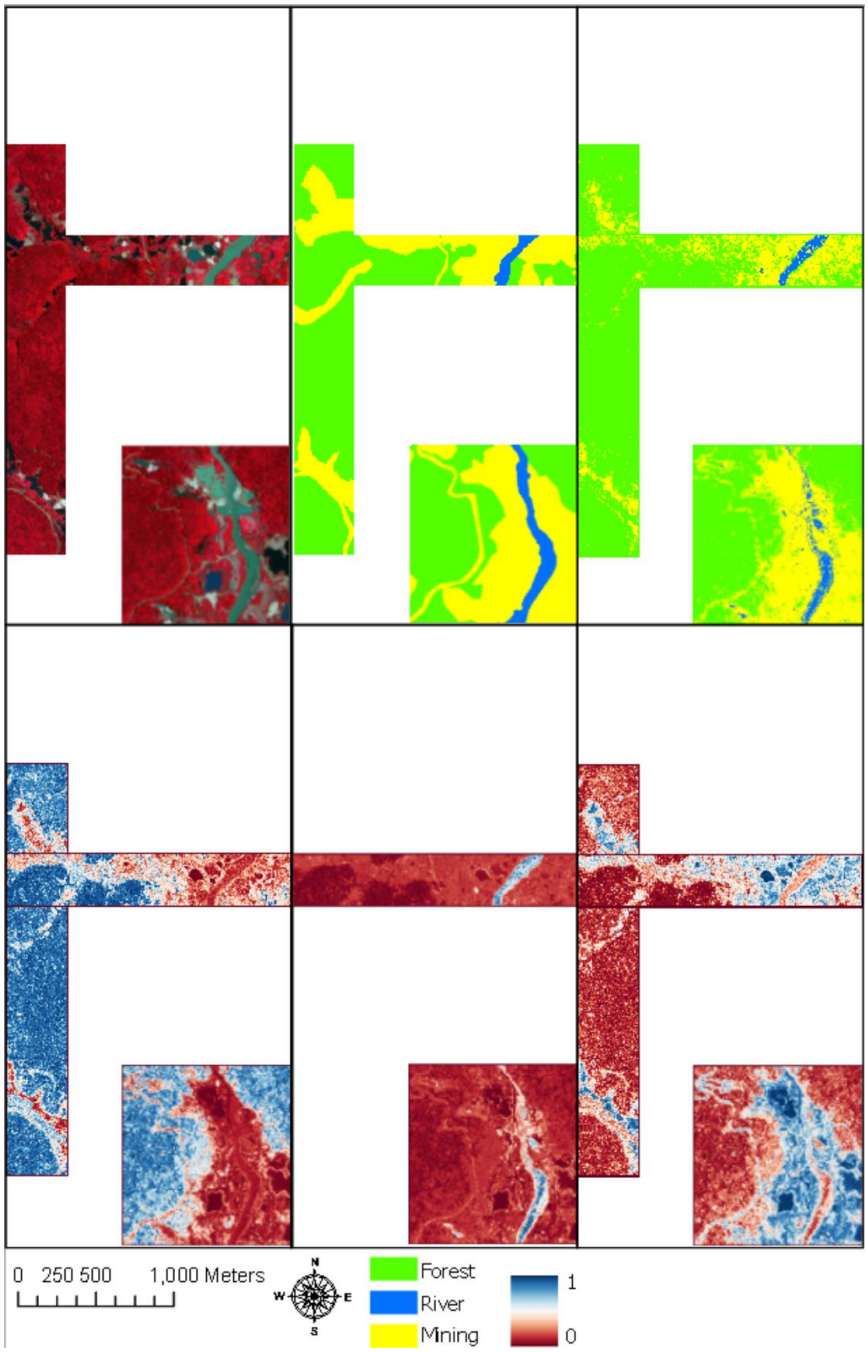


Table 25: Confusion matrix for Sentinel-2 classified with Random Forest.

Accuracy: 87.19% Pixels: 4060		True			Consumer's Accuracy
		Forest	River	Mining	
Predicted	Forest	2152	8	179	92.01%
	River	1	211	50	80.53%
	Mining	256	26	1177	80.67%
Producer's Accuracy		89.33%	86.12%	83.71%	

Figure 52: Sentinel-2 classified using Random Forest. Top left: Data. Top middle: True labels. Top right: Predicted classes. Bottom left: Probability of forest. Bottom middle: Probability of river. Bottom right: Probability of mining.

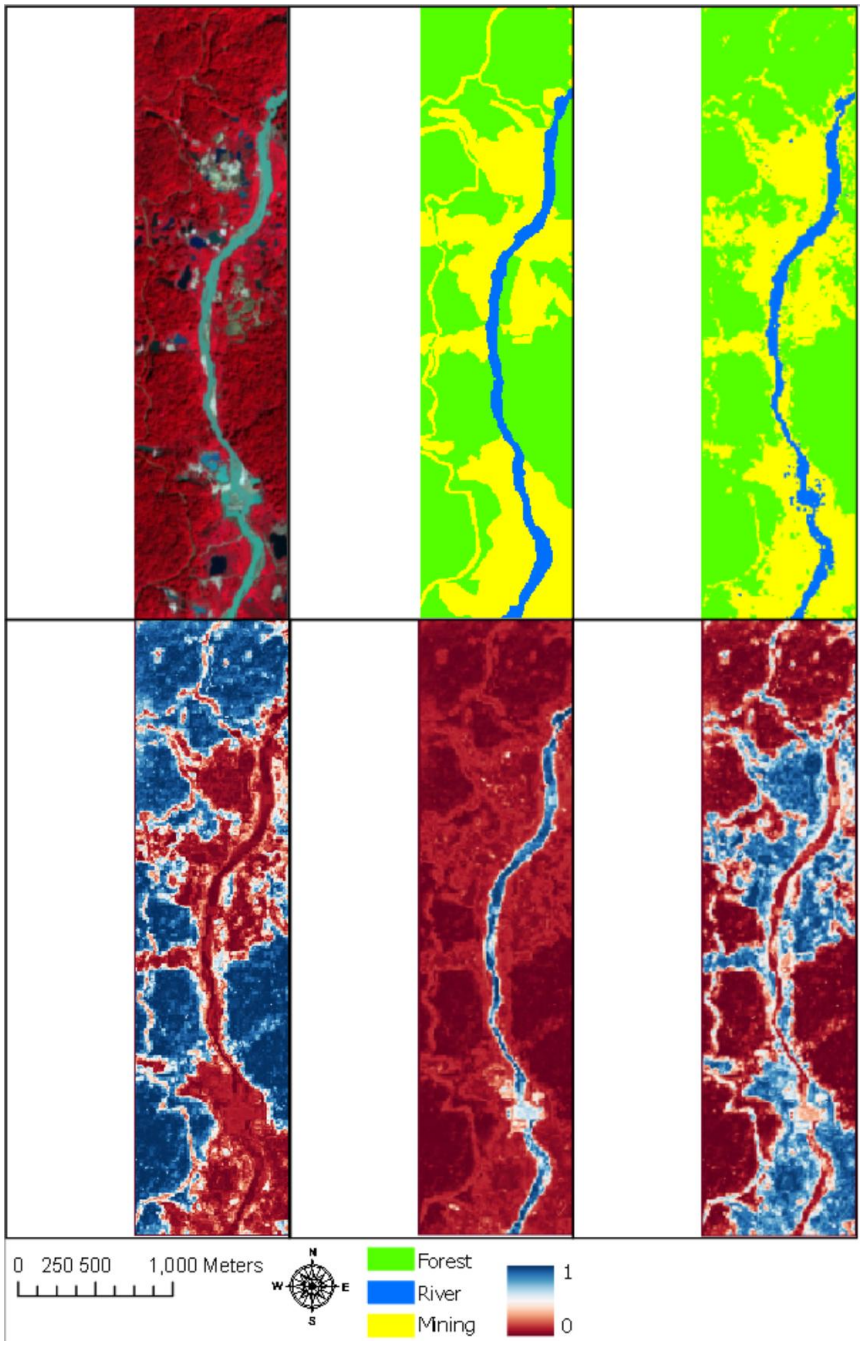


Table 26: Confusion matrix for Landsat classified with Random Forest.

Accuracy: 89.86% Pixels: 3177		True			Consumer's Accuracy
		Forest	River	Mining	
Predicted	Forest	2065	7	107	94.77%
	River	3	90	21	78.95%
	Mining	176	8	700	79.19%
Producer's Accuracy		92.02%	85.71%	84.54%	

Figure 53: Landsat classified using Random Forest. Top left: Data. Top middle: True labels. Top right: Predicted classes. Bottom left: Probability of forest. Bottom middle: Probability of river. Bottom right: Probability of mining.

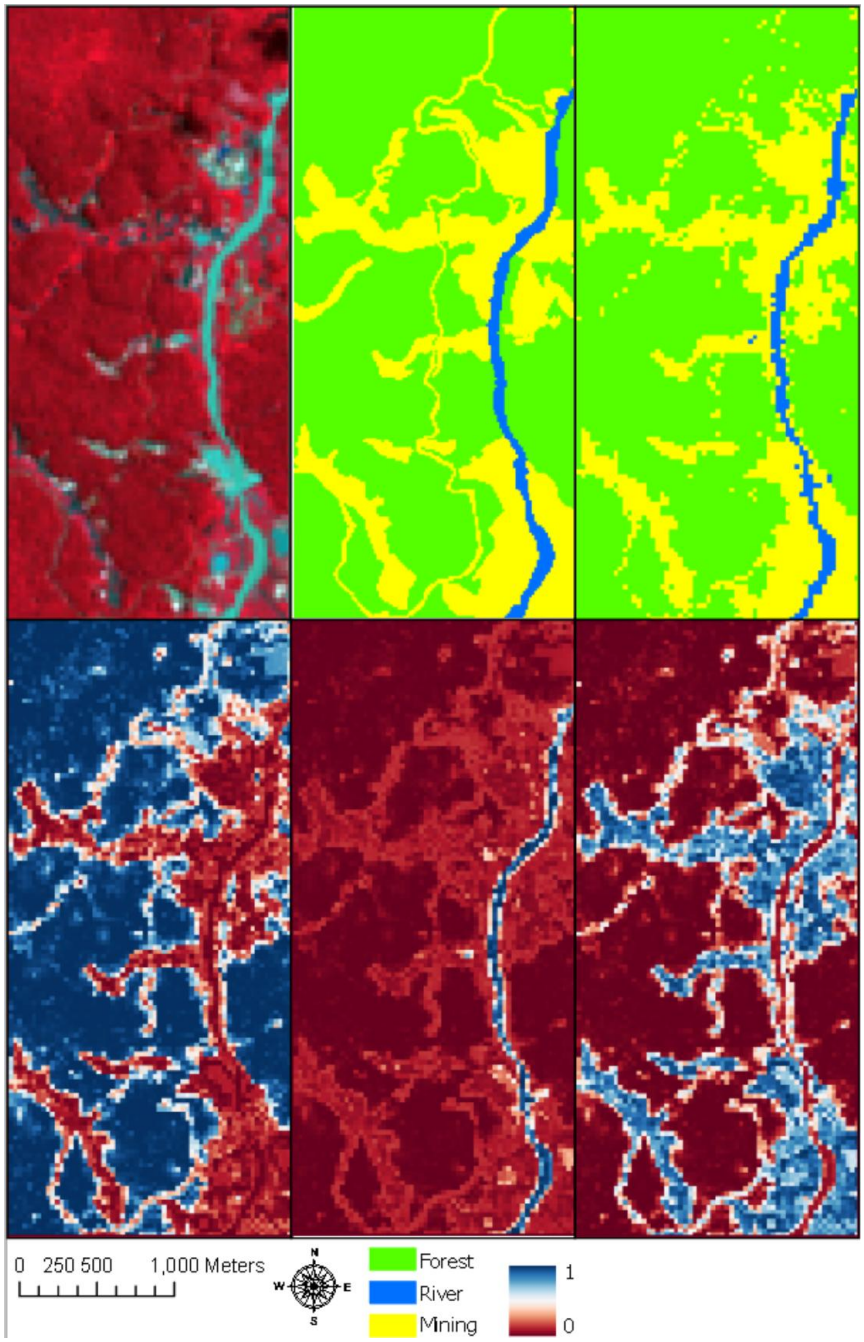


Table 27: Confusion matrix for Sentinel-1 classified with Random Forest.

Accuracy: 69.05% Pixels: 7342		True			Consumer's Accuracy
		Forest	River	Mining	
Predicted	Forest	4693	19	358	92.56%
	River	130	66	96	22.60%
	Mining	1637	42	331	16.47%
Producer's Accuracy		72.65%	51.97%	42.17%	

Figure 54: Sentinel-1 classified using Random Forest. Top left: Data. Top middle: True labels. Top right: Predicted classes. Bottom left: Probability of forest. Bottom middle: Probability of river. Bottom right: Probability of mining.

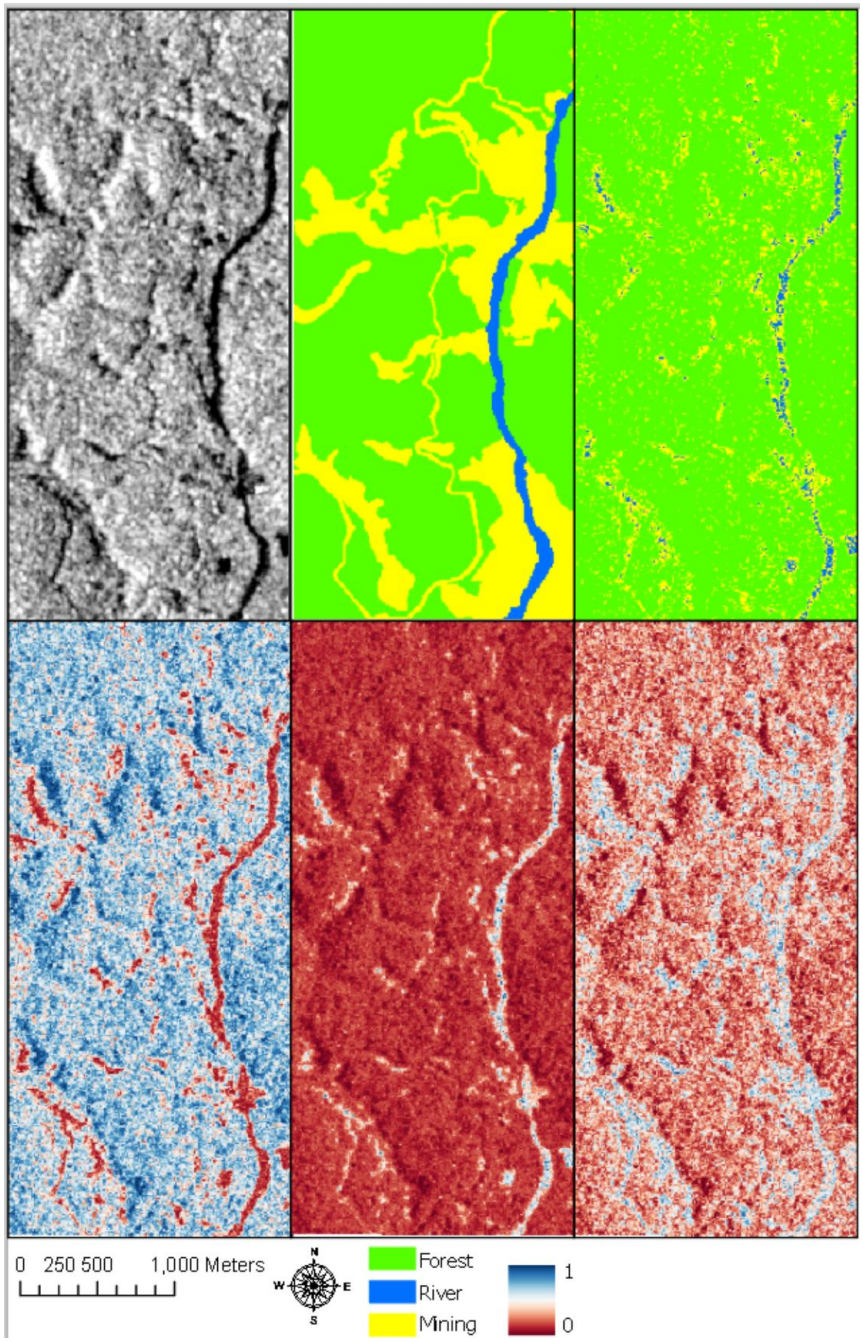
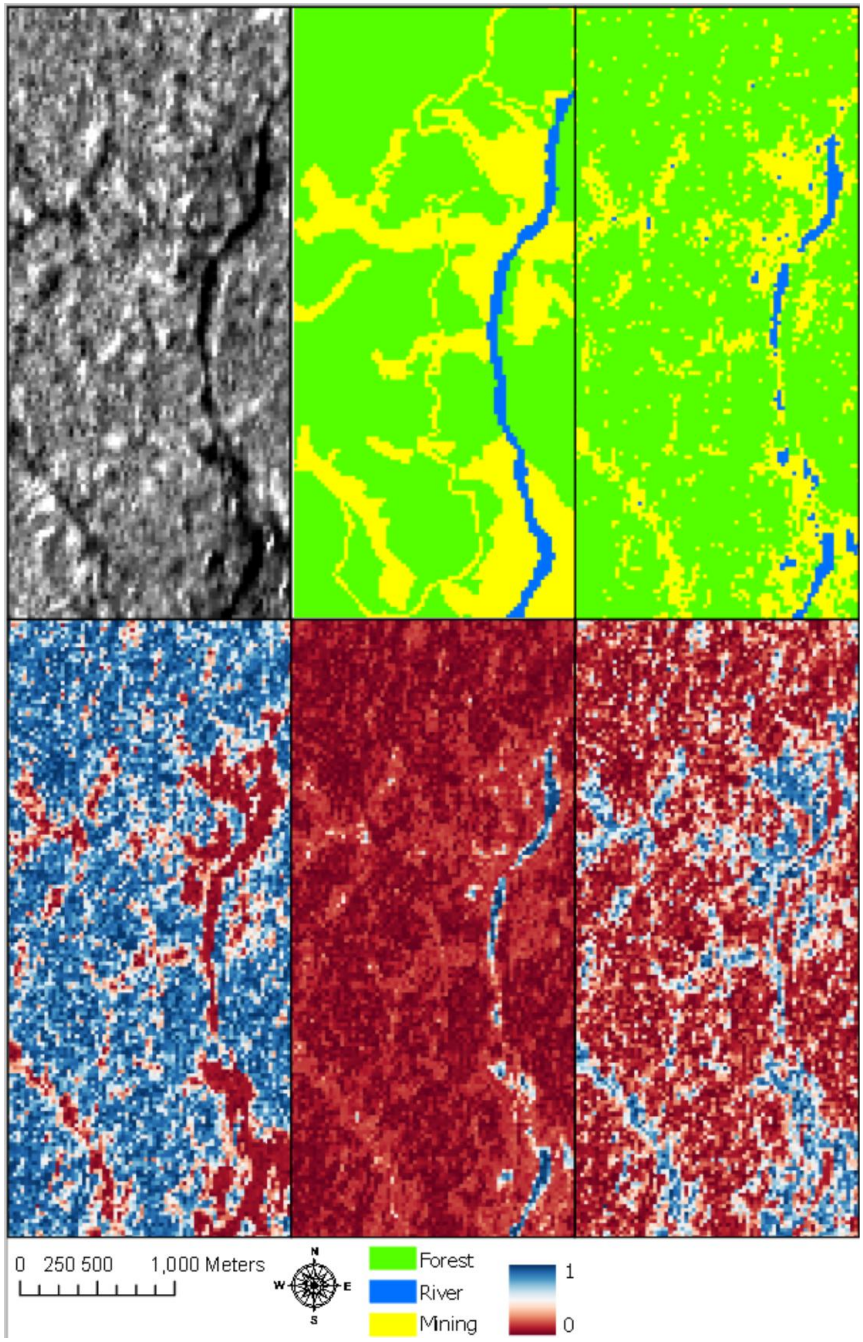


Table 28: Confusion matrix for ALOS-2 classified with Random Forest.

Accuracy: 77.24% Pixels: 1208		True			Consumer's Accuracy
		Forest	River	Mining	
Predicted	Forest	781	1	69	91.77%
	River	8	19	18	42.22%
	Mining	170	9	133	42.63%
Producer's Accuracy		81.44%	65.52%	60.45%	

Figure 55: ALOS-2 classified using Random Forest. Top left: Data. Top middle: True labels. Top right: Predicted classes. Bottom left: Probability of forest. Bottom middle: Probability of river. Bottom right: Probability of mining.



CHAPTER 6: SAR COMPARISON USING A FOREST-RIVER-MINING SITE

The previous two chapters have examined four multispectral sensors and two SAR sensors at two locations, which present different degrees of difficulty for classification. There is a third location with forest, river and mining in Guyana, located on the Cuyuni River (Figure 55), where data from four SAR sensors and three multispectral sensors are available. The four SAR sensors include both moderate and high spatial resolution data, with microwaves in both vertical and horizontal polarizations. The two SAR sensors in the last chapter (Sentinel-1 and ALOS-2) are available again, as are three of the multispectral sensors (Planetscope with the Dove Classic sensor, Sentinel-2, Landsat, but not RapidEye). This chapter will serve two purposes:

- To provide a more comprehensive assessment of the ability of SAR to map deforestation, at both moderate and high resolution.
- To compare with the results of Chapter 5 at a mining site on the Konawaruk River, an assessment of the performance of the same sensors in a different environment, this time on the Cuyuni River. This is important because nationwide deforestation mapping needs to be able to accurately classify the full diversity of land cover in Guyana.

Figure 55 shows the location of the area in Guyana. This is an area that includes the Cuyuni River and several mining areas that have a complex mix of vegetation and ponds. This area also has a mix of topography, including both flat areas and rapid changes in elevation. This means that this area will provide both high quality training data for a SAR test (low relief) and the ability to test SAR classifiers in both easy situations (low relief, simple land cover) and difficult situations (high relief, complex land cover types).

SECTION 6.1 CHOOSING THE SAR SATELLITES

The full discussion of selecting the SAR satellites can be found in Chapter 3. Here are the key points for this chapter, with a summary in Table 29:

- The ALOS-2 data is a 25 meter resolution ScanSAR image, because that is available for free on Google Earth Engine. The image is dual polarized HH/HV, and the microwaves are L-band, which penetrates to the forest floor. The satellite was in a descending orbit (top to bottom in the image), and the sensor is right-looking (the beam is oriented toward the left in the image). SAR shadows will be on the right side of a hill.

- The Sentinel-1 data is at 10 meter resolution (Interferometric Wide Swath) and is also freely available. The image is dual polarized VV/VH, and the microwaves . The satellite was in an ascending orbit (top to bottom in the image), and the sensor is right-looking (the beam is oriented toward the left in the image). SAR shadows will be on the left side of a hill.
- The Radarsat-2 data is a five meter resolution Fine beam mode image, in single polarization HH. It has been processed using radiometric calibration, terrain correction and speckle reduction using the Refined Lee algorithm. The satellite was in a descending orbit (top to bottom in the image), and the sensor is right-looking (the beam is oriented toward the left in the image). SAR shadows will be on the right side of a hill.
- The TerraSAR-X data is a 2.75 meter resolution StripMap image, in dual polarization VV/VH. It came with radiometric calibration and terrain correction applied. The satellite was in a descending orbit (top to bottom in the image), and the sensor is right-looking (the beam is oriented toward the left in the image). SAR shadows will be on the left side of a hill.
- The SAR images were collected in August 2017. The Planetscope data was collected on 5th September 2017, the Landsat data was collected on 12th April 2018, and the Sentinel-2 data was collected on 23rd September 2018, as these are the dates that offered a sufficient amount of cloud free data.

Table 29: A summary of the available synthetic aperture radar (SAR) satellites. X-band, C-band and L-band are the microwave frequency bands emitted. They differ in penetrating power: X-band (the shortest wavelengths) reflect off of the canopy, and L-band (the longest wavelength) reflects off of the forest floor. Polarizations are included: Horizontal or vertical.

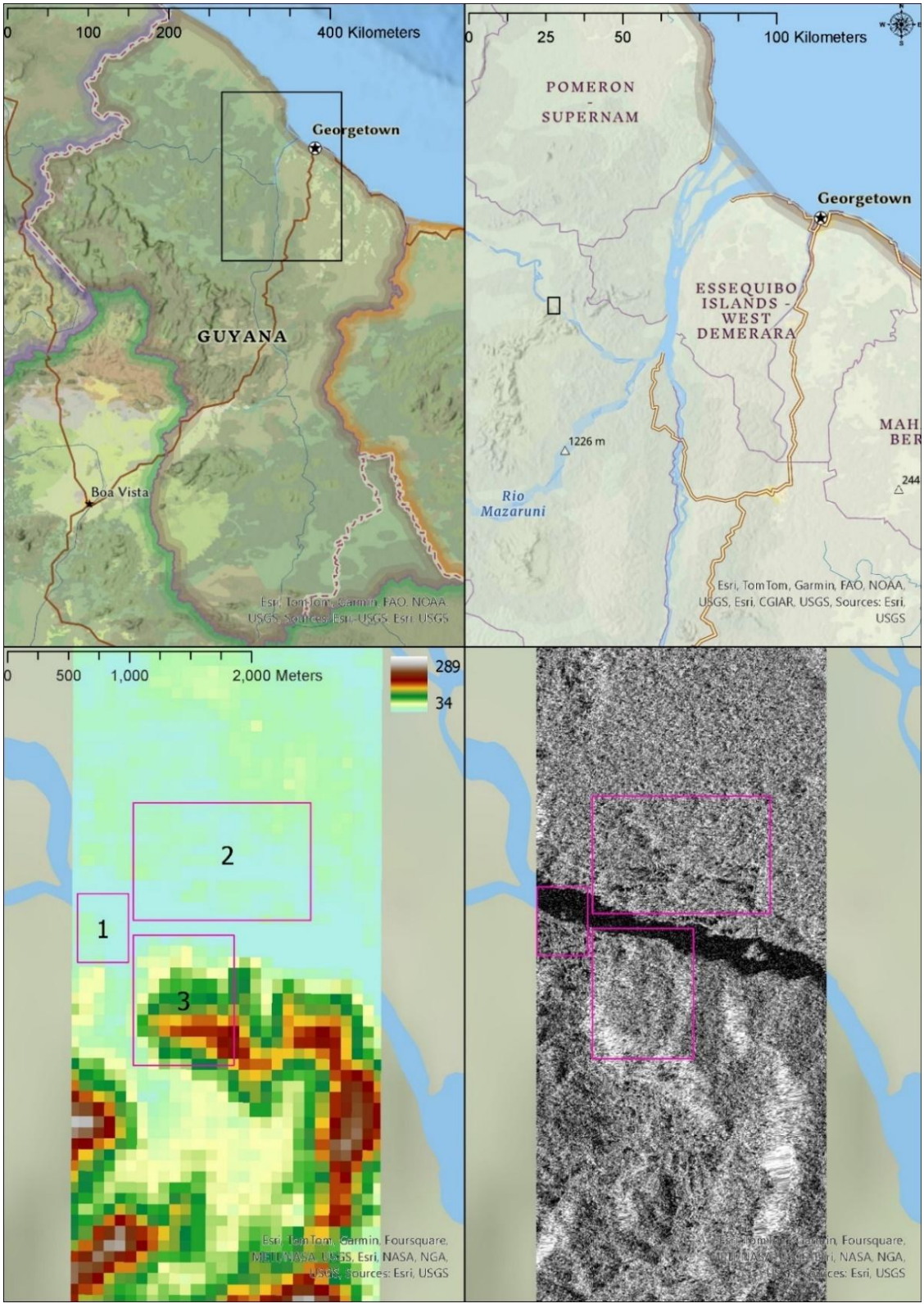
	X-band	C-band	L-band
2.75m	TerraSAR-X VV/VH		
5m		Radarsat-2 HH	
10m		Sentinel-1 VV/VH	
25m			ALOS-2 HH/HV

SECTION 6.2 COVERAGE OF THE TRAINING AND PREDICTION DATA

For multispectral and SAR data, there are two issues that determine what areas are covered by the training and validation data, and where the training samples are drawn from. They are: cloud cover and topography. For multispectral data, the cloud cover problem was solved by finding an image from the date with the least cloud cover, and then clipping out the clouds that were present. The training and validation samples were drawn from the cloud-free area that overlaps in all of the images. SAR penetrates clouds, which means that it may be a solution to this problem.

The second problem, steep topography, can limit the usefulness of SAR. Figure 42 shows a TerraSAR-X SAR image, along with an elevation map derived from Shuttle Radar Topography Mission (SRTM) data. The SRTM data shows that Area 1 and Area 2 are flat, with simple geometry, and Area 3 has high relief and complex geometry. In the SAR image, Area 3 has a large amount of white smearing, where that slope was in the shadow of the SAR signal, and pixel values had to be interpolated when terrain correction was done. The location of SAR shadow is determined by the satellite's orbit and the sensor's look direction. Depending on which way the sensor is pointed, one side of a hill or the other will be in the sensor's shadow. For this reason, two confusion matrices are being generated: one to measure performance in an easy area with little relief, and one to measure classifier performance in a difficult area with high relief.

Figure 56: The location of the study site in Guyana, with a TerraSAR-X example. The lower left image is a digital elevation model in meters. Area 1 is where the pixels for training were drawn from. Area 2 is where the first set of validation pixels were drawn from, and Area 3 is where the second set of validation pixels were drawn from. The whitest pixels (highest backscatter value) correspond with the left side of the hills in the DEM.



SECTION 6.3 RESULTS

These results use the same sensors and the same classifier (Random Forest) as Chapter 5, but these results are from a different location.

In this chapter, the classifier used for the test is Random Forest. The number of trees in the Random Forest was tuned manually, and it was found that reducing the number of trees (smoothing the model) reduces accuracy. It was also found that the orbit (ascending or descending in a polar orbit) either did not affect accuracy or did not change accuracy enough to solve any problems. Detailed results for both can be seen in Appendix E.

In this test, the confusion matrices and associated statistics were drawn from two sets of validation samples. One is from an area of low relief (Table 30), and the other is from an area of high relief (Table 31). Comparing SAR sensors within the flat part of the Cuyuni River mining site in different parts of a three-class problem, all sensors have high Consumer's Accuracy when classifying Forest (88% to 93%), and much lower Consumer's Accuracy when classifying River (27% to 81%) and Mining (12% to 34%). When classifying the hilly area, Consumer's Accuracy for Forest is 91% to 94%, for River it is 67% to 98%, and for Mining it is 8% to 19%. Overall, the pattern between the flat and hilly area is that the hilly area has slightly better classification of forest, and the classification of river and mining is much worse.

Comparing sensors, Sentinel-1 (C-band) is consistently the best, ALOS-2 (L-band) is second best, and Radarsat-2 (C-band) and TerraSAR-X (X-band) are much worse than either. Looking at polarization, the best and second worst sensors use dual VV/VH polarization, the second best uses dual HH/HV polarization, and the worst uses HH polarization. Looking at resolution, the two worst sensors are the two highest resolution sensors, the best sensor is in the middle, and the second-best sensor has the highest resolution. Looking at Figures 46-49, the distribution of predictions becomes noisier as the resolution increases. Note that this is after adjusting the Random Forest hyperparameters to smooth the model, by limiting the number of nodes in the trees. When unlimited nodes were allowed, the distribution of predictions was much noisier.

In comparison, the multispectral sensors showed higher performance at classifying Forest, and much higher performance at classifying River and Mining areas.

Table 30: Summary of confusion matrix results for the SAR comparison in the flat area, Area 2 in Figure 42, in the Cuyuni River mining site. Area 1 in Figure 42 was the training data. Accuracies are percentage out of the total pixel count.

		ALOS-2	Sentinel-1	Radarsat-2	TerraSAR-X	Planetscope	Sentinel-2	Landsat
Pixel size		25m	10m	5m	2.75m	3m	10m	15m
Pixel Count		2360	14 602	58 133	192 111	149 560	6139	914
Consumer's Accuracy	Forest	91.47	93.05	88.02	89.22	96.70	94.75	97.85
	River	65.33	80.68	26.71	35.45	91.24	97.87	68.34
	Mining	34.23	14.44	11.59	13.33	77.79	64.20	57.69
Producer's Accuracy	Forest	87.20	18.71	13.38	52.80	96.72	86.01	93.40
	River	55.06	74.49	63.14	78.74	89.11	78.25	97.14
	Mining	46.20	88.42	70.44	45.18	78.15	85.98	38.04

Table 31: Summary of confusion matrix results for the SAR comparison in the hilly area, Area 3 in Figure 42, in the Cuyuni River mining site. Area 1 in Figure 42 was the training data. Accuracies are percentage out of the total pixel count.

		ALOS-2	Sentinel-1	Radarsat-2	TerraSAR-X	Planetscope	Sentinel-2	Landsat
Pixel size		25m	10m	5m	2.75m	3m	10m	15m
Pixel Count		1620	10 283	40 497	326 005	112 587	10 154	709
Consumer's Accuracy	Forest	93.83	97.14	91.23	92.45	97.00	97.32	87.60
	River	90.01	98.44	67.14	76.60	98.16	96.54	53.41
	Mining	19.13	9.18	7.75	9.10	53.50	61.42	67.05
Producer's Accuracy	Forest	85.19	28.07	19.34	58.12	96.63	96.49	85.95
	River	76.34	88.74	88.01	80.91	80.82	95.80	26.14
	Mining	40.15	89.55	74.27	46.97	67.49	68.24	67.83

SAR RESULTS

The first SAR sensor to be tested is ALOS-2 (Table 35), at 25 meter resolution with HH/HV polarization. The predictions (Figure 59) show that the river is accurately outlined, and the mining ponds are mostly classified as river. However, a large minority of forest pixels are identified as mining, and a large minority of mining pixels are identified as forest. Class probabilities are in Figure 59, and show a high probability of mining in the mining area, and in many parts of the forest. Looking at the data images in Figure 59, there is partial overlap between the forest and mining pixel values – the darker forest pixels are as dark as the mining area.

The second SAR sensor to be tested is Sentinel-1 (Table 36). The results (Figure 60) are similar to ALOS-2 (Table 35). The river is clearly identified, and many of the mining ponds are identified as river. There is much confusion between forest and mining again, though this time the results are reversed: the majority of the forest area is classified as mining. In the data images (Figure 60), the bottom centre shows a bright area of smearing where the elevation changes quickly. Looking at the bottom centre of the predictions panel, this entire area is identified as forest. Looking above and to the right of that area, closer to the river and the right edge of the predictions panel, the higher backscatter (brighter in the image) forest pixels are classified as forest, and the lowest backscatter (darkest in the image) forest pixels are classified as mining. This indicates that there is a lot of variability in pixel values in a SAR image of a forest. The darker end of the spectrum overlaps with the mining area (which is majority dark), and areas that are influenced by the geometry of rough topography overlap with the brightest parts of a forest – remember that the training pixels were drawn from the mining area to the right, and the flat area north of the river, and the area of changing elevation was used only for predictions. This terrain effect actually makes the classifier more accurate in this area, because the entire area of changing elevation was classified as forest, and there is no mining or river in it, whereas forests with flatter topography are classified partly as forest and partly as mining.

The third sensor to be tested is Radarsat-2 (Table 37). This is corrected to σ_0 with terrain correction using Copernicus 30 m DEM. The predictions (Figure 61) show that the river is clearly outlined, and many mining ponds are identified as water, but the majority of both the mining site and the forest are identified as mining.

The fourth sensor to be tested is TerraSAR-X (Table 38). The predictions (Figure 62) reverse the Radarsat-2 predictions: the river is clearly outlined (though parts inside it are identified as

forest), and the mining site and forest are both largely identified as mining, with scattered pixels in both areas identified as mining. The class probabilities (Figure 62) indicate low certainty for forest and mining, and high certainty for water.

In the Google Earth Engine implementation, Random Forest has a Maximum Nodes hyperparameter. Values from 2 to 512 were tested, and 24 produced the highest accuracy scores, and was used for the results above. However, when unlimited nodes were allowed, the forest and mining predictions were reversed for Sentinel-1, Radarsat-2 and TerraSAR-X: the majority of pixels in both the mining and forest areas were classified as forest, and almost all of the water was identified as river.

MULTISPECTRAL RESULTS

In this comparison, the multispectral sensors consistently produced better results than the SAR sensors. Of the three multispectral sensors analysed (see Table 32 for Planetscope or Table 33 for Sentinel-2), Sentinel-2 is the most accurate for two classes (Mining and River), and Planetscope is the most accurate in Forest.

The first test will be Planetscope, the highest performing sensor. In this test (Figure 56), the entire mining site has been labelled as Mining. This is a spectrally complex area that includes many mining ponds and some patches of isolated forest new vegetation growth. Random Forest was still able to outline the river and the mining sites accurately, and it identifies some of the mining ponds as river. This shows that Random Forest can improve upon labels. The class probabilities (Figure 56) accurately follow the boundary between the mining site and the river, even though the line can't be seen by visual inspection of the data. Note that the two data images in Figure 56 show every band: the top image is Green-Red-Near Infrared, and the bottom image is Blue-Green-Red. First generation Planetscope sensors (Dove Classic) have no other bands available. This suggests that Random Forest, when looking at all four bands at once, can accurately identify land cover boundaries not visible to a human operator.

In Planetscope, the prediction image is "more accurate" because it accurately reflects features within a complex land cover class. However, this also means that the prediction at each pixel will be different from the label, which will reduce its accuracy in the confusion matrix. This may be an effect of Random Forest's ability to improve on labels.

The second test is another multispectral sensor, Sentinel-2 (Figure 57). This time, Random Forest shows a much greater ability to improve on class labels. The river and mining site are accurately outlined again, and more mining ponds were identified as water. The differences between Planetscope (Dove Classic sensor) and Sentinel-2 are the resolution (3 meters for Planetscope, 10 meters for Sentinel-2), and Sentinel-2 records more infrared bands. The class probabilities (Figure 57) show a difference in probability between mining surface and some of the mining ponds that were still classified as mining rather than river.

The third test is Landsat (Figure 58 and Table 34). Again, the river and the mining site predicted boundaries are very close to the true boundaries. One mining pond was identified as river, and large areas of vegetation in the southern end of the mining area were identified as forest.

SECTION 6.4 DISCUSSION

Overall, a clear pattern emerges. Both multispectral and SAR are very good at outlining the river. This is useful to know in hydrological research: SAR penetrates clouds, and so free-to-download SAR (Sentinel-1 or ALOS-2 ScanSAR) could be used for river and lake mapping with wall-to-wall coverage with 10 meter resolution. The problem is that Random Forest with Sentinel-1 and ALOS-2 identifies mining ponds as rivers, which would create a lot of false positives. This ties in with the discussion of accuracy in Chapter 5. The key point is that the classifier is capturing details not found in the true labels used for training and validation.

When classifying mining, the Planetscope (Figure 56) and Sentinel-2 (Figure 57) data both outline the mining site very accurately, and identify very little of the forest as mining. Almost none of the river is identified as mining. In the data, the river and the mining ponds are indistinguishable to the human eye, but the classifier was still able to separate them accurately. In the class probabilities, there is a visible difference in probability of river between mining and forest – as in the previous chapter, the gradient in probability between the two follows the true class boundary.

When Random Forest classifies mining using ALOS-2 data, the mining sites are mostly identified as either mining or river. However, there are a large number of false positives in the forest area. The backscatter variation in the forest partially overlaps with the mining area, in a way that not even Random Forest can correct for.

When classifying mining using Sentinel-1 data, part of the mining site is identified as mining, and part is identified as forest. The same pattern holds in Radarsat-2 and TerraSAR-X, with the proportion of pixels identified as mining increasing as the resolution increases.

It is also worth noting that the class probabilities follow the texture of the canopy.

Variable thresholding can be used to increase the number of mining sites detected, at the expense of false positives. Technicians can filter out false positives; false negatives will be difficult to find and correct.

SECTION 6.5 CONCLUSION

The last two comparisons have tested the same sensors and the same land cover type (Forest-River-Mining) at two locations, one on the Cuyuni River, and one on the Konawaruk River. The multispectral results were similar, but the SAR results were very different. At the Cuyuni River site, the SAR classification was very good at finding the river, and heavily biased toward classifying land pixels as mining. At the Konawaruk River site, the SAR classification missed about half of the river pixels, and was heavily biased toward identifying land pixels as forest. This indicates that study sites are not interchangeable, and the location has an impact on the classifier.

The last two comparisons have found that Random Forest is a high performing classifier in multispectral data, because of its ability to accurately outline class boundaries, and its ability to improve on labels, when given suitable data. Multispectral data consistently gives good results, even in this comparison, where the difficulty of the data was increased. SAR data is much worse than multispectral data in the same location. This outweighs the advantages of SAR (cloud penetration). The next chapter will look at the results of combining multispectral and SAR images.

Table 32: Confusion matrices for Planetscope. The top confusion matrix is from Area 2 in Figure 42, the flat area, and the bottom confusion matrix is from Area 3, the hilly area.

Accuracy: 93.92% Pixel Count: 149 560		True			Consumer's Accuracy
		Forest	River	Mining	
Predicted	Forest	119 932	45	4021	96.70%
	River	15	4533	539	91.24%
	Mining	4076	390	15 969	77.79%
Producer's Accuracy		96.72%	89.10%	78.15%	

Accuracy: 93.06% Pixel Count: 112 587		True			Consumer's Accuracy
		Forest	River	Mining	
Predicted	Forest	90 962	51	3120	97.00%
	River	225	8233	1729	98.16%
	Mining	2585	103	5579	53.50%
Producer's Accuracy		96.63%	80.82%	67.49%	

Figure 57: Planetscope classified using Random Forest. Data with Green-Red-Near Infrared (top left), labels (top middle), predictions (top right), probability of forest (bottom left), probability of river (bottom middle), probability of mining (bottom right). The areas below correspond with the areas in Figure 42.

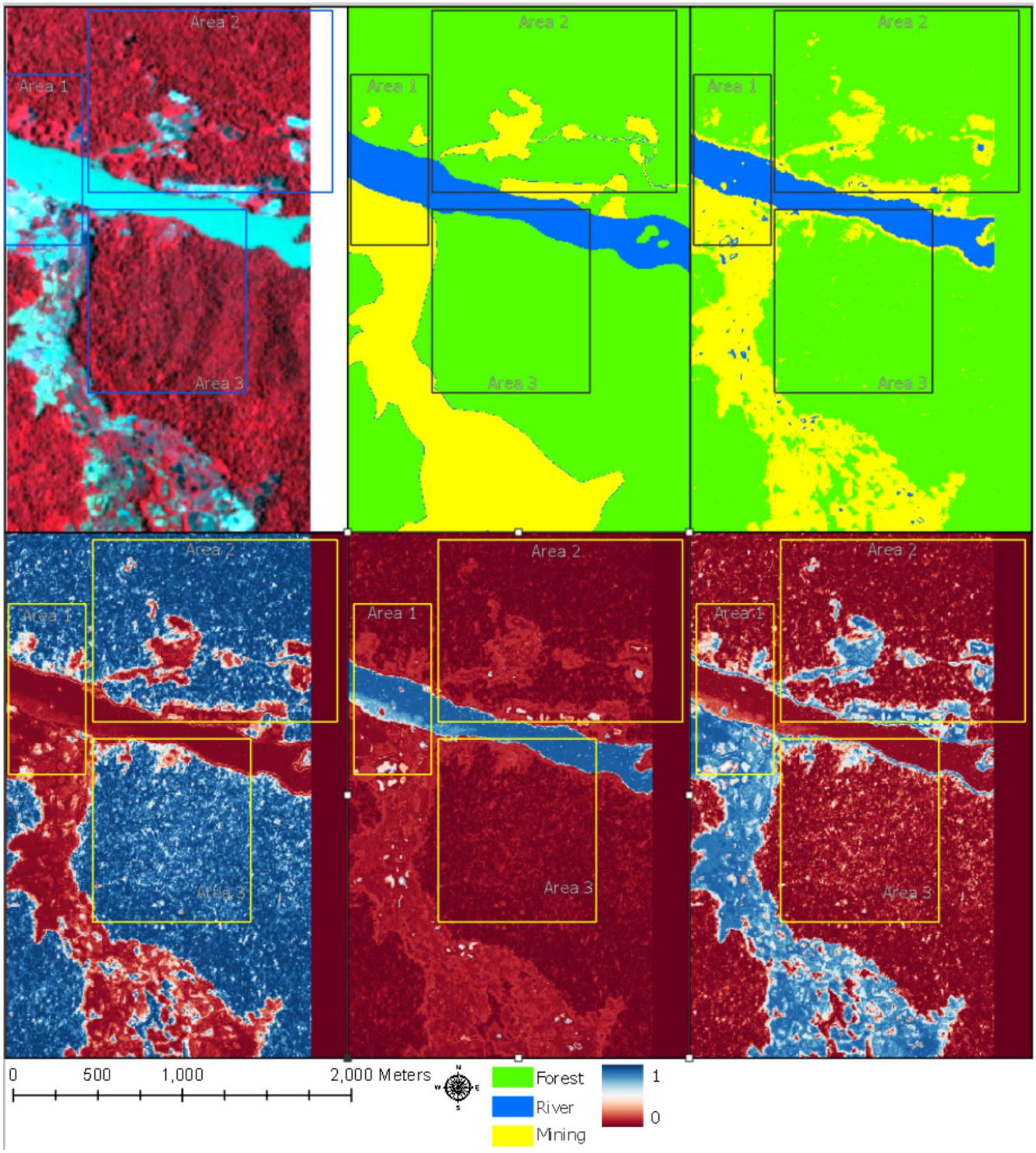


Table 33: Confusion matrices for Sentinel-2. The top confusion matrix is from Area 2 in Figure 42, the flat area, and the bottom confusion matrix is from Area 3, the hilly area.

Accuracy: 85.42% Pixel Count: 6139		True			Consumer's Accuracy
		Forest	River	Mining	
Predicted	Forest	3684	0	599	94.75%
	River	10	367	92	97.87%
	Mining	194	8	1239	64.20%
Producer's Accuracy		86.01%	78.25%	85.98%	

Accuracy: 94.43% Pixel Count: 10 154		True			Consumer's Accuracy
		Forest	River	Mining	
Predicted	Forest	8230	28	271	97.33%
	River	0	866	38	96.54%
	Mining	226	3	492	61.42%
Producer's Accuracy		96.49%	95.80%	68.24%	

Figure 58: Sentinel-2 classified using Random Forest. Data with Green-Red-Near Infrared (top left), labels (top middle), predictions (top right), probability of forest (bottom left), probability of river (bottom middle), probability of mining (bottom right). The areas below correspond with the areas in Figure 42.

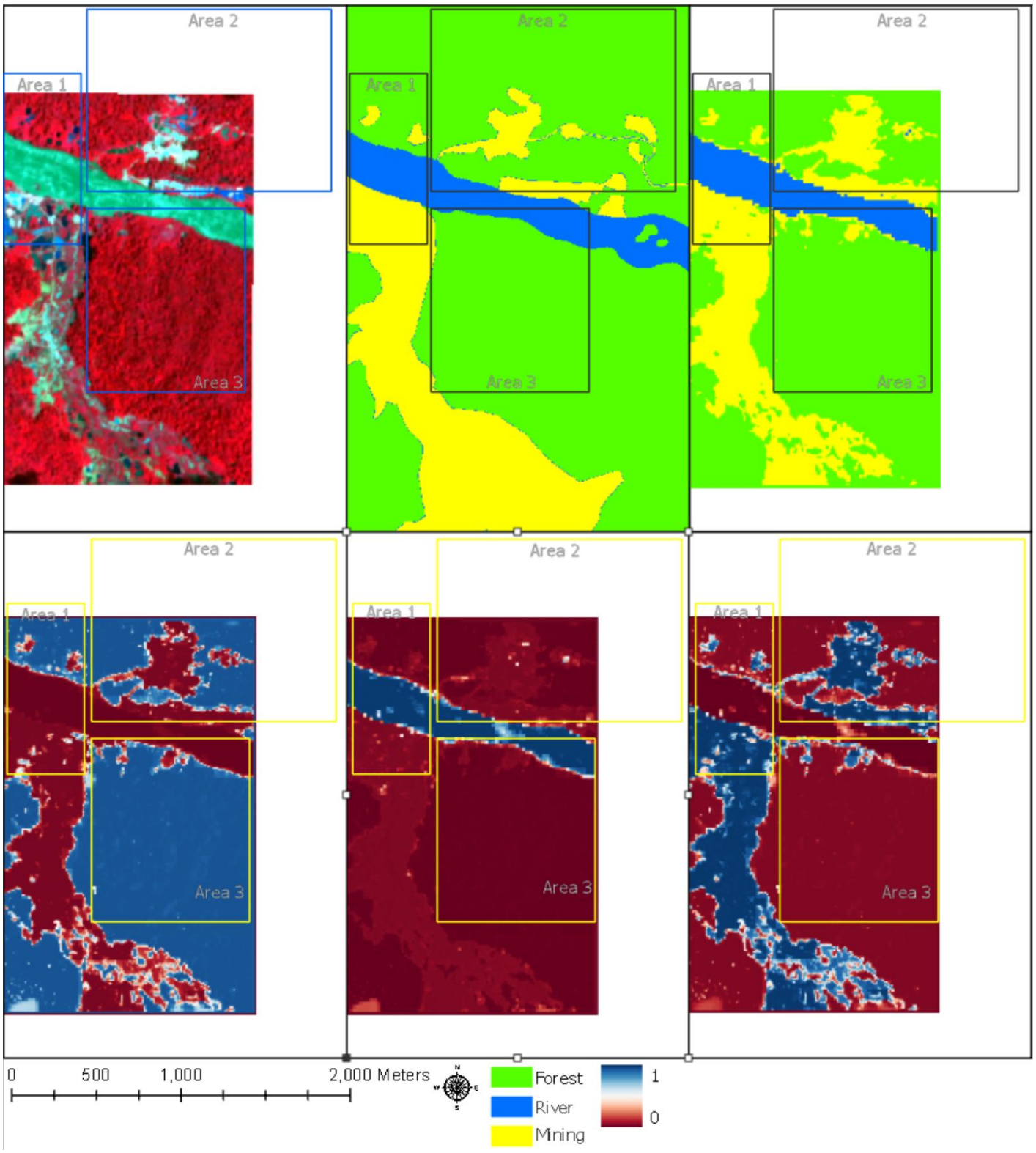


Table 34: Confusion matrices for Landsat. The top confusion matrix is from Area 2 in Figure 42, the flat area, and the bottom confusion matrix is from Area 3, the hilly area.

Accuracy: 88.40% Pixel Count: 914		True			Consumer's Accuracy
		Forest	River	Mining	
Predicted	Forest	637	0	14	97.85%
	River	20	136	43	68.34%
	Mining	25	4	35	57.69%
Producer's Accuracy		93.40%	97.14%	38.04%	

Accuracy: 75.88% Pixel Count: 709		True			Consumer's Accuracy
		Forest	River	Mining	
Predicted	Forest	318	0	45	87.60%
	River	14	47	27	53.41%
	Mining	85	0	173	67.05%
Producer's Accuracy		76.26%	100%	70.61%	

Figure 59: Landsat classified using Random Forest. Data showing Green-Red-Near Infrared (top left), labels (top middle), predictions (top right), probability of forest (bottom left), probability of river (bottom middle), probability of mining (bottom right). The areas below correspond with the areas in Figure 42.

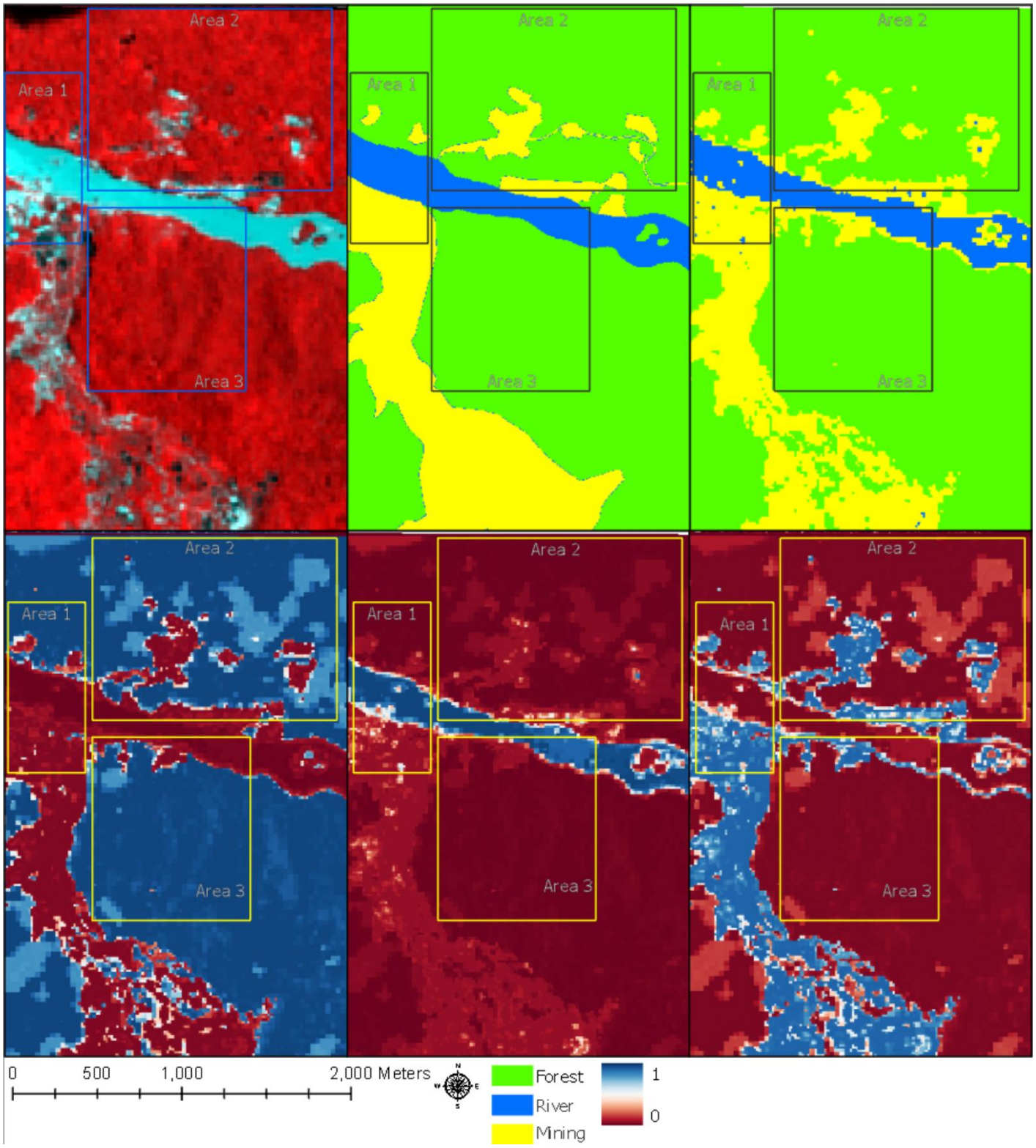


Table 35: Confusion matrices for ALOS-2. The top confusion matrix is from Area 2 in Figure 42, the flat area, and the bottom confusion matrix is from Area 3, the hilly area.

Accuracy: 80.72% Pixel Count: 2360		True			Consumer's Accuracy
		Forest	River	Mining	
Predicted	Forest	1716	4	248	87.20%
	River	19	49	21	55.06%
	Mining	141	22	140	46.20%
Producer's Accuracy		91.47%	65.33%	34.23%	

Accuracy: 80.80% Pixel Count: 1620		True			Consumer's Accuracy
		Forest	River	Mining	
Predicted	Forest	1156	4	197	85.19%
	River	4	100	27	76.33%
	Mining	72	7	53	40.15%
Producer's Accuracy		93.83%	90.09%	19.13%	

Figure 60: ALOS-2 classified using Random Forest. Data with HV polarization (top left), labels (top middle), predictions (top right), probability of forest (bottom left), probability of river (bottom middle), probability of mining (bottom right). The areas below correspond with the areas in Figure 42. The sensor is looking to the left.

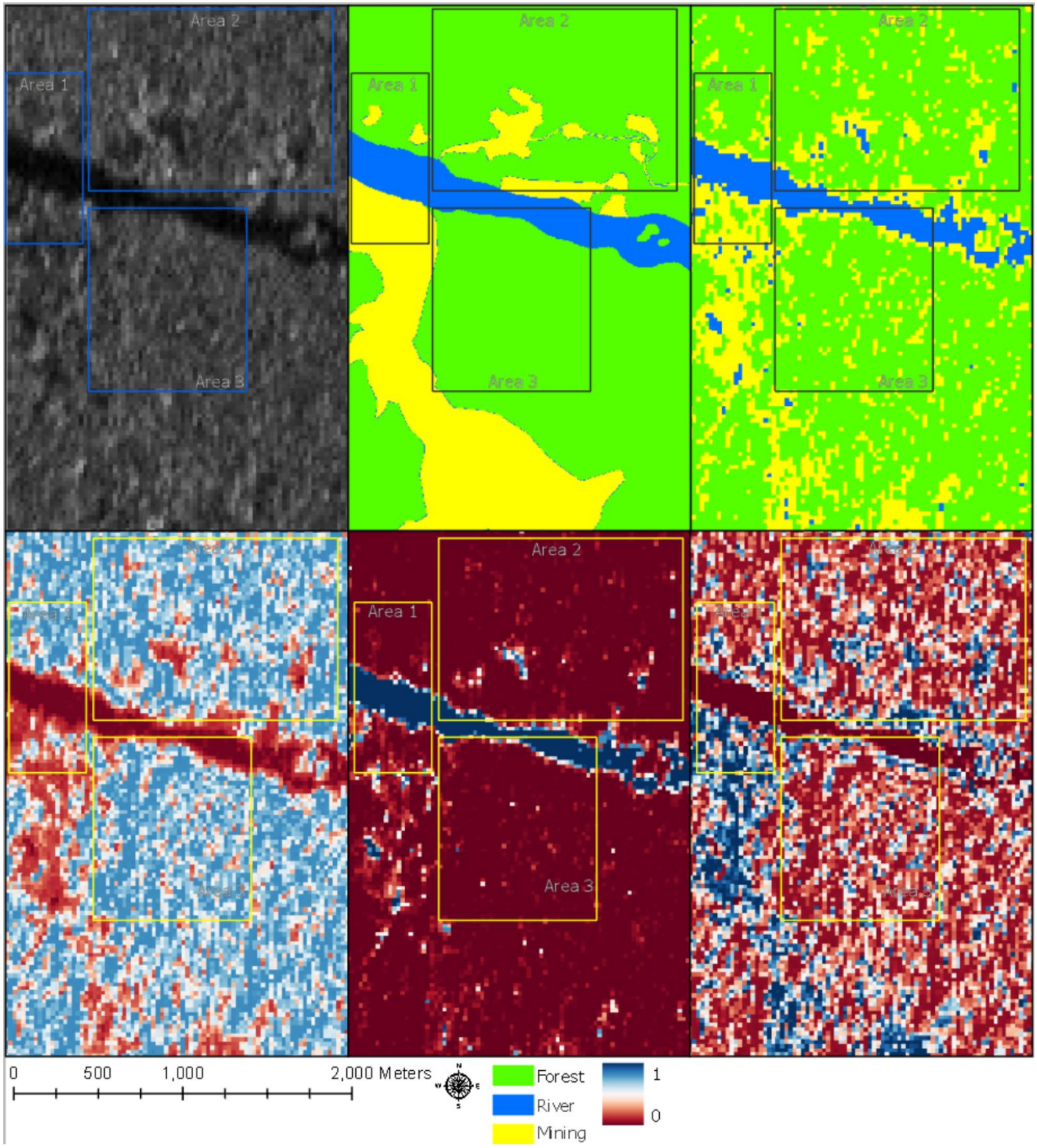


Table 36: Confusion matrices for Sentinel-1. The top confusion matrix is from Area 2 in Figure 42, the flat area, and the bottom confusion matrix is from Area 3, the hilly area.

Accuracy: 29.56% Pixel Count: 14 602		True			Consumer's Accuracy
		Forest	River	Mining	
Predicted	Forest	2291	26	9925	18.71%
	River	2	330	111	74.49%
	Mining	169	53	1695	88.42%
Producer's Accuracy		93.05%	80.68%	14.44%	

Accuracy: 38.17% Pixel Count: 10 283		True			Consumer's Accuracy
		Forest	River	Mining	
Predicted	Forest	2408	11	6161	28.07%
	River	0	883	112	88.74%
	Mining	71	3	634	89.55%
Producer's Accuracy		97.14%	98.44%	9.18%	

Figure 61: Sentinel-1 classified using Random Forest. Data with VV polarization (top left), labels (top middle), predictions (top right), probability of forest (bottom left), probability of river (bottom middle), probability of mining (bottom right). The areas below correspond with the areas in Figure 42. The sensor is looking to the right.

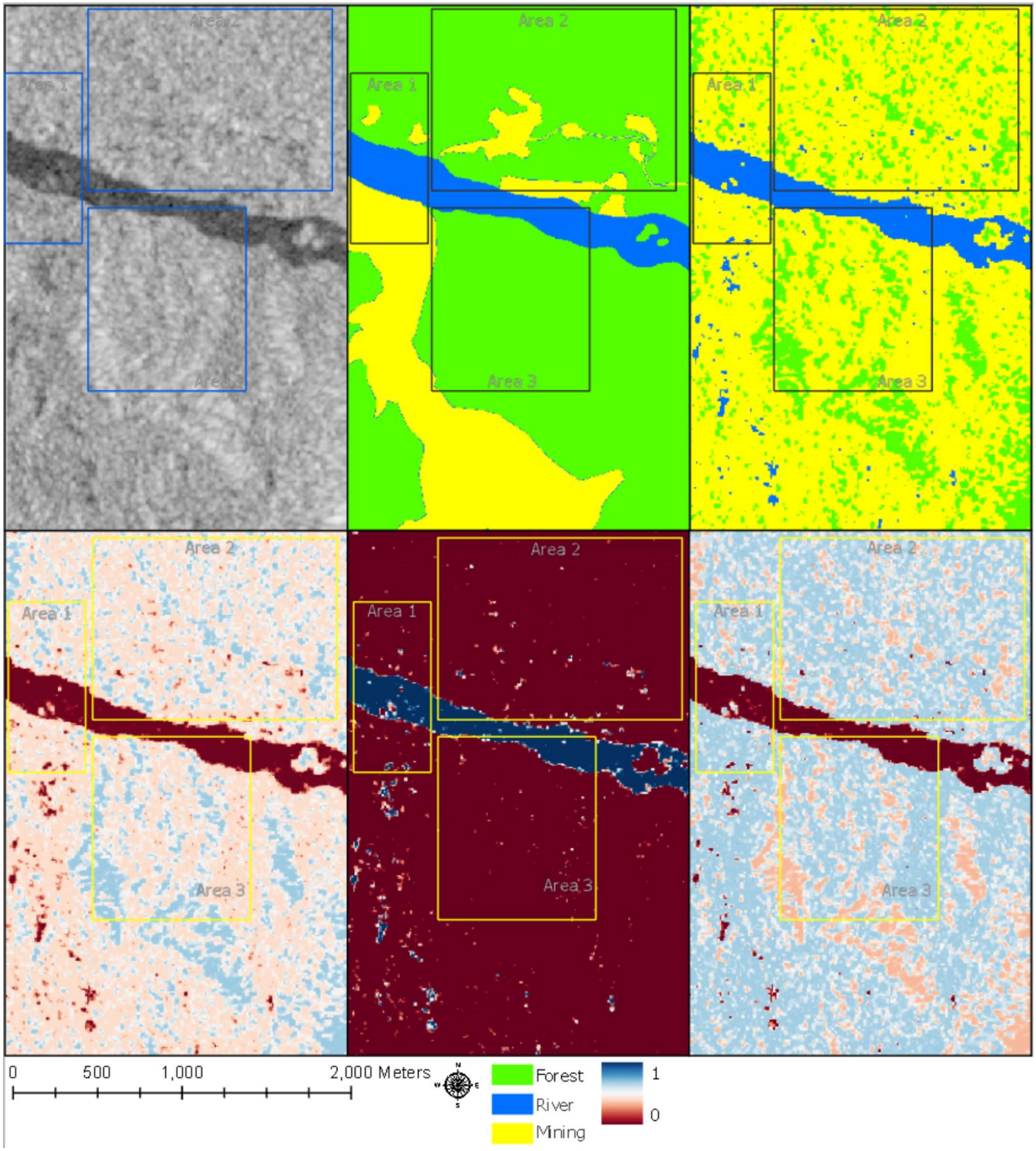


Table 37: Confusion matrices for Radarsat-2. The top confusion matrix is from Area 2 in Figure 42, the flat area, and the bottom confusion matrix is from Area 3, the hilly area.

Accuracy: 22.68% Pixel Count: 58 133		True			Consumer's Accuracy
		Forest	River	Mining	
Predicted	Forest	6474	2508	39 390	13.38%
	River	149	1451	698	61.14%
	Mining	732	1474	5257	70.44%
Producer's Accuracy		88.02%	26.71%	11.59%	

Accuracy: 29.76% Pixel Count: 40 497		True			Consumer's Accuracy
		Forest	River	Mining	
Predicted	Forest	6531	1448	25 783	19.34%
	River	41	3318	411	88.01%
	Mining	587	176	2202	74.27%
Producer's Accuracy		91.23%	67.14%	7.75%	

Figure 62: Radarsat-2 classified using Random Forest. Data with HH polarization (top left), labels (top middle), predictions (top right), probability of forest (bottom left), probability of river (bottom middle), probability of mining (bottom right) . The sensor is looking to the right.

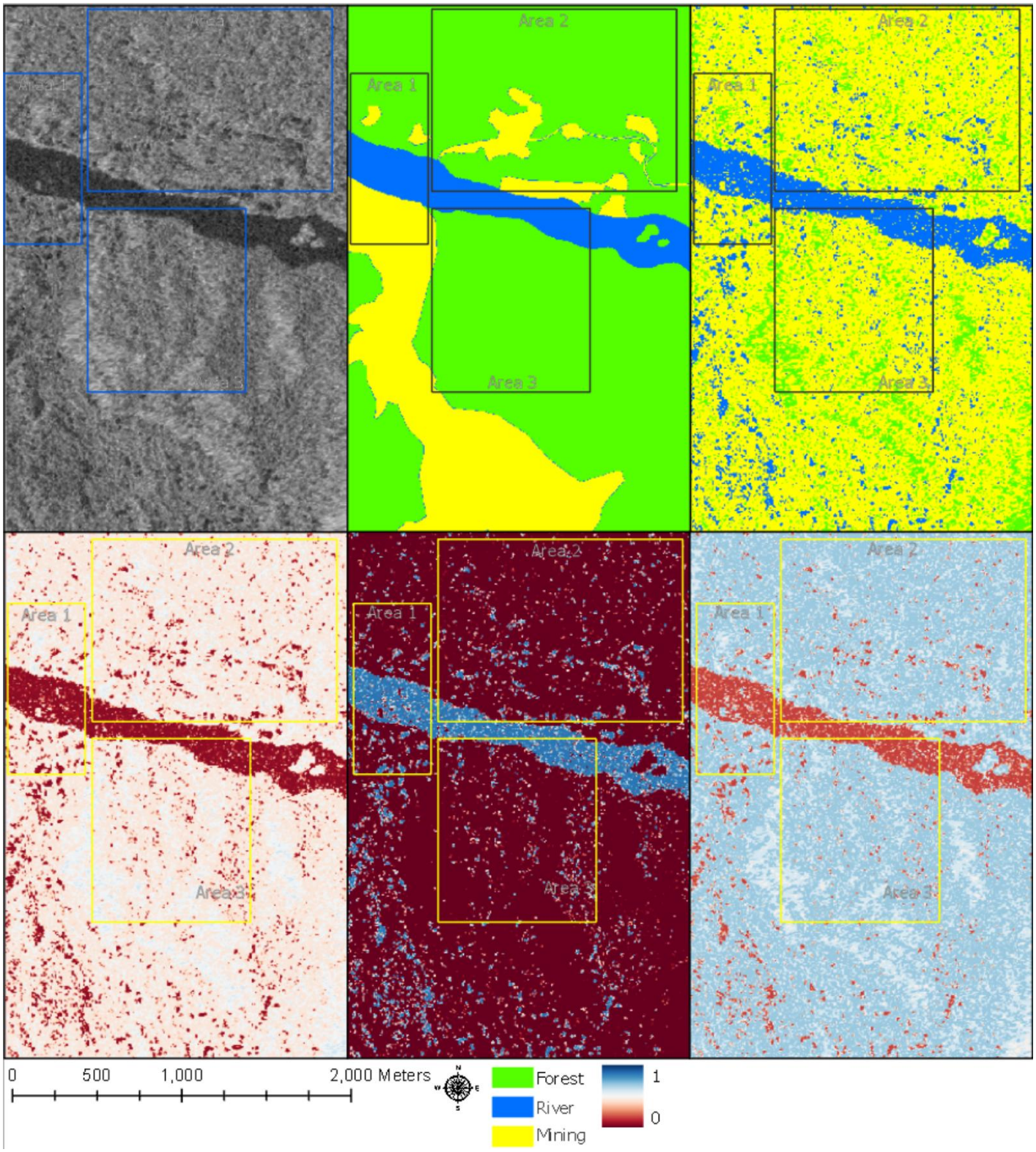
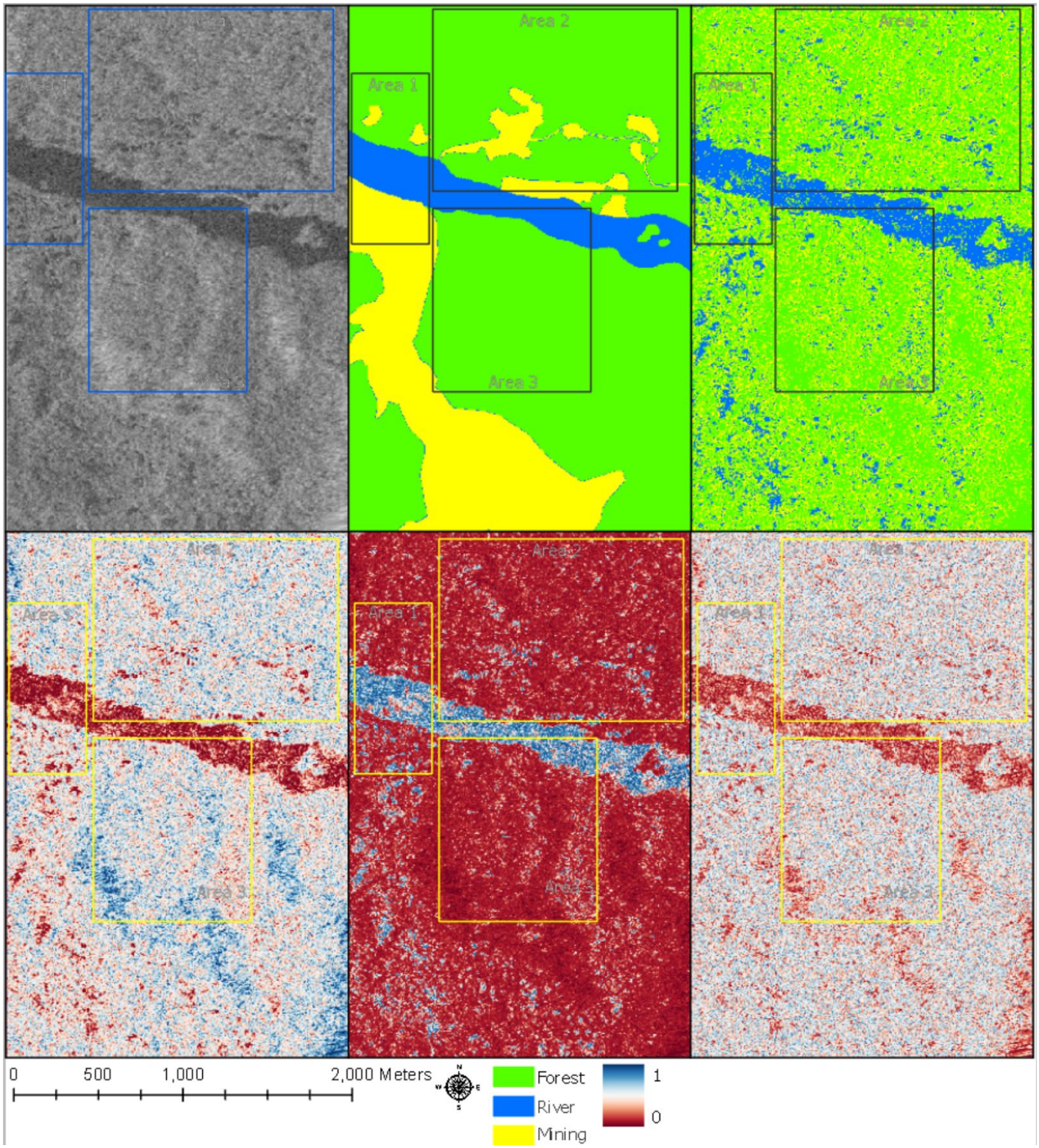


Table 38: Confusion matrices for TerraSAR-X. The left confusion matrix is from Area 2 in Figure 42, the flat area, and the right confusion matrix is from Area 3, the hilly area.

Accuracy: 73.73% Pixel Count: 192 111		True			Consumer's Accuracy
		Forest	River	Mining	
Predicted	Forest	133 136	9617	18 330	82.65%
	River	764	4636	617	77.05%
	Mining	16 084	5065	3862	15.44%
Producer's Accuracy		88.77%	24.00%	16.93%	

Accuracy: 80.10% Pixel Count: 326 005		True			Consumer's Accuracy
		Forest	River	Mining	
Predicted	Forest	96 127	5135	10 769	85.80%
	River	1050	9845	1129	81.88%
	Mining	7766	793	1280	13.01%
Producer's Accuracy		91.60%	62.42%	9.71%	

Figure 63: TerraSAR-X classified using Random Forest. Data with VV polarization (top left), labels (top middle), predictions (top right), probability of forest (bottom left), probability of river (bottom middle), probability of mining (bottom right). The sensor is looking to the right.



CHAPTER 7: RESULTS ON COMBINING MULTISPECTRAL AND SAR SENSORS

SECTION 7.1 INTRODUCTION

In the last two chapters, the SAR and multispectral images were strong and weak at different aspects of the classification. SAR (particularly ALOS-2) tended to be better at finding mining ponds and outlining the river than multispectral images. Multispectral images were much better at outlining the mining areas than SAR. But there is another possibility: to combine both data types into one. This chapter tests two methods of fusing data from different sensors.

There is some literature on this approach, though it isn't nearly as extensive as the literature on sensor fusion. Holmberg (2021) stacked Sentinel-1 and Sentinel-2 bands and found a small increase, though it was judged not to be worth the added computational requirements. Kuplich (2006) used both SAR and multispectral bands in a neural network, and class discrimination increased when both SAR and multispectral bands were used over SAR bands only.

There are five comparisons in this section, three at the Cuyuni River site (seen in Chapter 6) and two at the Konawaruk River site (seen in Chapter 5). The Cuyuni River site has data from Planetscope + TerraSAR-X, Sentinel-2 + Sentinel-1, and Landsat + ALOS-2. The Konawaruk River site has data from Sentinel-2 + Sentinel-1 and Landsat + ALOS-2.

SENSOR FUSION

The first method is the sensor fusion approach most commonly used in the remote sensing literature, where a three-band multispectral image and a one-band SAR image are fused into a three-band image that has information from both sources.

Here are the key concepts:

- In the RGB colour space, each colour is represented by three coordinates, that represent the mix of wavelengths in the colour: the recorded values of the Red, Green and Blue bands.
- A multispectral satellite sensor splits the incoming visible light into Red, Green and Blue fractions (sometimes more), and records their intensities as a vector of numbers at each pixel.
- There is a trade-off between spatial resolution and spectral resolution. When photons are drawn from fewer wavelengths, they need to be drawn from a larger spatial area to get

a clear signal. When photons are drawn from a wider range of wavelengths, they can be drawn from a smaller spatial area. Some multispectral sensors have both a panchromatic band and multispectral bands. The panchromatic band has lower spectral resolution (it draws from the entire visible light spectrum) and higher spatial resolution. The multispectral bands draw from the red, green and blue slices of the visible light spectrum, at a lower spatial resolution. Satellites that have a panchromatic band include Landsat, SPOT, Worldview and IKONOS. Satellites that do not have a panchromatic band include Sentinel-2, Planetscope and RapidEye.

- An image with the spatial resolution of the panchromatic band and the spectral resolution of the multispectral bands is a pansharpened image. It can be made by fusing the panchromatic band with the multispectral bands. Google Earth Engine has implemented the Hue-Saturation-Volume (HSV) transform for this.
- Hue can be thought of as the dominant wavelength or the colour in its purest form. Examples of hues are red, green or blue.
- Saturation is the tone of a hue, and can be adjusted by adding white or black.
- Volume is the brightness of a pixel.
- In the HSV colour space, each colour is represented by three coordinates: for the brightness, the primary colour, and the tone within the colour.
- Note that the panchromatic band has one value, the volume, with no spectral information. The multispectral bands do have colour information that can then become Hue and Saturation information.

There are two things to note, that can extend this idea:

1. Multispectral sensors have more than just visible light capability; they have at least Near Infrared capability. In remote sensing of forests, the Near Infrared band provides information on vegetation health and vitality, and in many situations, a Near Infrared – Red – Green image is more useful than a Red – Green – Blue image.
2. Computer displays are designed to display three channels. Normally, they are Red, Green and Blue, but the other bands in a multispectral image are pixel values of the same nature, and it's trivial to put the Near Infrared band into the monitor's Red channel, the Red band into the Green channel, and the Green band into the Blue channel.
3. Another kind of data that is of the same nature is the pixel values in a processed SAR image. Each pixel value is the intensity of the backscattered microwave photons. As

rasters – arrays of pixel values – SAR images and panchromatic bands are fundamentally of the same nature.

There is a decades-long body of literature (reviews can be found in Pohl and van Genderen, 1998; Wang et al., 2005; Pandit and Bhiwani, 2015; Ghassemian, 2016; Kulkarni and Rege, 2020) that shows that accuracy can be improved by combining data from multiple sources. In these reviews and the literature they review, the goal of sensor fusion is to combine images from different sources in a way that increases the quantity and quality of information available. A fused multispectral-SAR image has spatial, spectral, textural, and dielectric information in one image. Features not visible in either source image alone can be detected here. Pohl and van Genderen (1998) summarizes this as "1 + 1 = 3".

Sensor fusion is done using the HSV process implemented in Google Earth Engine, which includes these steps:

1. The multispectral image's Near Infrared, Red and Green bands are transformed to the HSV colour space. This combination provides more information about forests than the Red-Green-Blue combination.
2. A new image is made by stacking the Hue and Saturation bands with the SAR image in the intensity position.
3. The image is transformed to the RGB colour space.

STACKING IMAGES

That method of image fusion makes use of well-established methods in remote sensing, and it works within the constraints of the computer monitors used to visually interpret satellite images. However, in machine learning, classifiers are not limited to three bands. They can accept inputs with any number of bands. This opens up the possibility of testing the idea of fusing images by just stacking them. The inputs are an m-band multispectral image (4 for Planetscope, 10 for Sentinel-2, 8 for Landsat) and an n-band SAR image (2 for each). The output is an m + n band image (6 bands for Planetscope + TerraSAR-X, 12 bands for Sentinel-1 + Sentinel-2, 10 bands for Landsat + ALOS-2).

The name comes from NumPy's Stack method, a recent implementation of an old idea, which joins arrays along a specified axis. If the axis specified is the feature/channel/band axis, then a new array will be made that has the features of the source arrays. For example, RapidEye has five bands, for an array of dimension (height, width, 5), and Sentinel-1 has two bands, for a

dimension of (height, width, 2). Stacking them along axis 2 will produce an image that has dimension (height, width, 7). This differs from sensor fusion in that the exact pixel values of both source images are kept in the new image - in sensor fusion, the fused image has pixel values that are derived from but not identical to the source images. The images were stacked in Google Earth Engine, where they had already been projected to the same coordinate system.

Stacked images are made using the Add Bands method in Google Earth Engine. The first image is the multispectral image, the second image is the SAR image, and the output image has the multispectral bands and the SAR bands.

SECTION 7.2 RESULTS

In the confusion matrices in this section, there are results from four sensors:

- Stacked refers to an image that has all the bands of the source multispectral and SAR images intact.
- Fused refers to an image that was made through HSV fusion.
- After that, come a multispectral image alone and then SAR backscatter alone.

TESTING PLANETSCOPE + TERRASAR-X AT THE CUYUNI RIVER MINING SITE

When combining Planetscope with TerraSAR-X at the Cuyuni River mining site, the stacked and fused images both show little difference over Planetscope, but a great improvement over TerraSAR-X (Table 36). The Consumer's Accuracy for Forest in TerraSAR-X is 89%, and for Planetscope is 97%. For both fused and stacked images, the Consumer's Accuracy is 97% in both cases. When classifying River pixels, the Consumer's accuracy is 35% in TerraSAR-X and 91% in Planetscope, and becomes 86% in fused images and 91% in stacked images. When classifying Mining pixels, Consumer's Accuracy is 13% in TerraSAR-X and 78% in Planetscope, and 71% in fused images and 78% in stacked images. In effect, combining SAR with multispectral matches the performance of the multispectral, without adding anything.

When testing traditional sensor fusion, the predictions (Figure 63) show that the fused data outlines the river and the mining site as well as the multispectral source image, and much better than the SAR source image. Several mining ponds are identified as river, and large areas of vegetation inside the mining site are identified as forest. Some noise is present in the river.

When testing a stacked image, the noise in the river disappears, fewer mining ponds are identified as river, and larger areas of vegetation inside the mining site are identified as forest.

TESTING SENTINEL-2 + SENTINEL-1 AT THE CUYUNI RIVER MINING SITE

When combining Sentinel-2 with Sentinel-1 at the Cuyuni River mining site, the Consumer's Accuracy for mining in the flat area is 14% for Sentinel-1, 64% for Sentinel-2, 56% for fused images, and 64% for stacked images. In the hilly area, the Consumer's Accuracy for mining is 9% for Sentinel-1, 61% for Sentinel-2, 51% for fused images, and 54% for stacked images (Table 37). With this sensor combination, the combined images are actually worse than the source multispectral image, but much better than the source SAR image. This is different from the Planetscope-TerraSAR-X combination. There, combining images didn't change the accuracy. This is the first sign that the effect of combining images depends on the source satellites.

When testing traditional sensor fusion (Figure 64), the predictions outline all classes with good accuracy. Some small sections of the river are classified as land classes, and some mining ponds are classified as river. Most of the roads are clearly visible in the predictions.

When testing a stacked image, the predictions (Figure 64) outline the river almost perfectly, and outline the mining sites very accurately.

TESTING LANDSAT + ALOS-2 AT THE CUYUNI RIVER MINING SITE

Combining Landsat with ALOS-2 reports another pattern (Table 41). At the flat part of the Cuyuni River site, the Consumer's Accuracy for Mining is 34% for ALOS-2, 38% for Landsat, 64% when fused, and 73% when stacked. In the hilly area, there is a very different difference: 19% for ALOS-2, 63% for Landsat, 33% when fused, and 47% when stacked.

When testing traditional sensor fusion (Figure 65), much of the river is classified as mining. The riverbank is consistently misclassified as mining. Dense vegetation inside the mining areas is consistently classified as forest, and mining ponds are consistently classified as river. The upper half of the mining site is outlined accurately, but the majority of the bottom half is classified as forest. There are many small areas inside the forest that are classified as river.

When testing a stacked image, the predictions (Figure 65) follow the outlines of each class boundary very accurately. The only major deviation is where some mining ponds very close to the river are classified with the river.

There are also results from combined Sentinel-1 and Sentinel-2 images at the Konawaruk River mining site (Table 42). There, the Consumer's Accuracy for Sentinel-1 is 55%, for Sentinel-2 is 79%, for fused images is 77%, and for stacked images is 78%. This is consistent with the Planetscope-TerraSAR-X combination, and inconsistent with the Sentinel-1/Sentinel-2 combination at the Cuyuni River mining site.

At the Konawaruk River mining site, the difference is much smaller: 62% for ALOS-2, 78% for Landsat, 75% when fused, 80% when stacked.

When testing traditional sensor fusion (Figure 66), the predictions again outline the edges of the mining and river areas very accurately. This time, fewer mining ponds are classified as River than the Planetscope + TerraSAR-X combination, and more vegetation inside the mining areas is identified as forest.

When testing a stacked image, the predictions (Figure 66) outline the class boundaries very closely, and even capture many of the roads. In Figure 66, the stacked predictions improve greatly upon the fused predictions. The fused predictions carry over many of the errors found in the SAR image, and the stacked predictions share the accuracy of the multispectral predictions.

Table 39: Comparison of accuracy scores for Planetscope and TerraSAR-X at the Cuyuni River (same area as Chapter 6). The pixel size is 3 meters.

		Area 2 in Figure 42 (Flat Area)				Hilly Area			
		Stacked	Fused	Planetscope	TerraSAR-X	Stacked	Fused	Planetscope	TerraSAR-X
Consumer's Accuracy	Forest	96.70	96.68	96.70	89.22	97.01	98.11	97.00	92.45
	River	90.75	86.26	91.24	35.45	98.30	98.83	98.16	76.60
	Mining	77.80	70.77	77.79	13.33	53.77	51.04	53.50	9.10
Producer's Accuracy	Forest	96.76	95.36	96.72	52.80	96.68	94.95	96.63	58.12
	River	88.68	83.76	89.11	78.74	80.79	79.65	80.82	80.91
	Mining	77.95	77.14	78.15	45.18	67.65	81.95	67.49	46.97

Table 40: Comparison of accuracy scores for Sentinel-2 and Sentinel-1 at the Cuyuni River (same area as Chapter 6). The pixel size is 10 meters.

		Flat Area				Hilly Area			
		Stacked	Fused	Sentinel-2	Sentinel-1	Stacked	Fused	Sentinel-2	Sentinel-1
Consumer's Accuracy	Forest	94.96	91.56	94.75	93.05	97.36	97.94	97.32	97.14
	River	98.11	98.60	97.87	80.68	97.09	99.27	96.54	98.44
	Mining	64.31	55.72	64.20	14.44	60.56	48.51	61.42	9.18
Producer's Accuracy	Forest	86.15	81.81	86.01	18.71	96.35	94.06	96.49	28.07
	River	77.61	74.84	78.25	74.49	96.02	90.38	95.80	88.74
	Mining	86.40	77.72	85.98	88.42	68.79	76.70	68.24	89.55

Table 41: Comparison of accuracy scores for Landsat and ALOS-2 at the Cuyuni River (same area as Chapter 6). The pixel size is 15 meters.

		Flat Area				Hilly Area			
		Stacked	Fused	Landsat	ALOS-2	Stacked	Fused	Landsat	ALOS-2
Consumer's Accuracy	Forest	94.90	93.23	93.81	91.47	96.14	94.77	76.47	93.83
	River	97.04	31.49	100	65.33	99.72	48.10	100	90.01
	Mining	72.77	63.81	37.96	34.23	47.20	32.86	62.95	19.13
Producer's Accuracy	Forest	97.37	93.58	97.85	87.20	95.60	90.78	85.95	85.19
	River	42.71	25.26	63.82	55.06	77.17	41.30	26.14	76.34
	Mining	79.51	68.17	64.06	46.20	64.60	54.57	67.83	40.15

Table 42: Comparison of accuracy scores at the Konawaruk River (same as Chapter 5). The pixel sizes are 15 meters (Landsat + ALOS-2) and 10 meters (Sentinel).

		Landsat and ALOS-2				Sentinel-2 and Sentinel-1			
		Stacked	Fused	Landsat	ALOS-2	Stacked	Fused	Sentinel-2	Sentinel-1
Consumer's Accuracy	Forest	93.43	93.16	93.39	84.00	90.31	78.97	90.30	72.82
	River	94.06	80.39	94.06	84.27	83.08	77.07	79.81	52.04
	Mining	79.59	75.37	77.91	61.99	77.53	77.38	79.09	54.94
Producer's Accuracy	Forest	92.97	92.20	92.12	88.29	87.54	92.03	88.67	97.13
	River	66.67	35.10	70.91	35.89	70.08	67.26	74.05	26.04
	Mining	83.38	82.72	93.13	58.65	83.29	58.95	82.30	12.56

SECTION 7.3 DISCUSSION

Pohl and van Genderen, 1998 describes sensor fusion as "1 + 1 = 3", arguing that combining two sensors reveals details not visible in other one alone, and there is ample literature showing situations where sensor fusion does improve classification performance. In Adrian et al. (2021), fused SAR-multispectral data was found to have better accuracy at one study site in India, when classifying crop types. In Ahmad et al. (2024), the four study sites were cities of similar type. Sentinel-1 and Landsat data were stacked, and the resulting accuracy was 0.86 for

stacked, 0.84 for Landsat, and 0.67 for Sentinel-1. Overall, the results show that, in this context, the performance of fused SAR-multispectral data is not better than both sources, but between them. This shows that traditional sensor fusion methods won't result in much improvement in this context, despite what a lot of literature would suggest.

It should be noted that sensor fusion literature often discusses its usefulness in visual interpretation rather than machine learning – as an example, Liu et al. (2015) measures sensor fusion quality based on measures of visual interpretability. This explains why sensor fusion has been done this way – computer monitors are capable of outputting three-band RGB images, and so if you want the information of four or more bands, you need to reduce that to three bands.

However, machine learning is not limited to the three bands of a computer monitor. Classifiers are capable of working with vectors of any length, which opens up the option of creating a multimodal dataset by simply stacking the bands of multiple source images.

Overall, in this context, stacking images represents an improvement over traditional sensor fusion. The accuracy of the class boundaries is as good as the multispectral data, if not better. However, it doesn't represent a large improvement over the multispectral data. The SAR doesn't appear to be contributing very much to the final predictions.

SECTION 7.4 CONCLUSION

The take-away of this chapter is that combining images using sensor fusion or stacking does not improve classification performance. The effect varies: sometimes combined images perform as well as the multispectral source image, and sometimes they perform worse. Performance varies with both study site and sensor combination.

So far, all tests have been in mining areas that have a nearby river. The next chapter will test the performance of stacked Sentinel-1/Sentinel-2 data using Random Forest in a variety of land cover types.

Figure 64: Comparison of results from Planetscope, TerraSAR-X, fused data, and stacked data at the Cuyuni River mining site.

Top left: Data, fused TerraSAR-X + Planetscope. Top middle: Fused predictions.

Top right: Stacked predictions. Bottom left: True labels.

Bottom middle: TerraSAR-X predictions. Bottom right: Planetscope predictions.

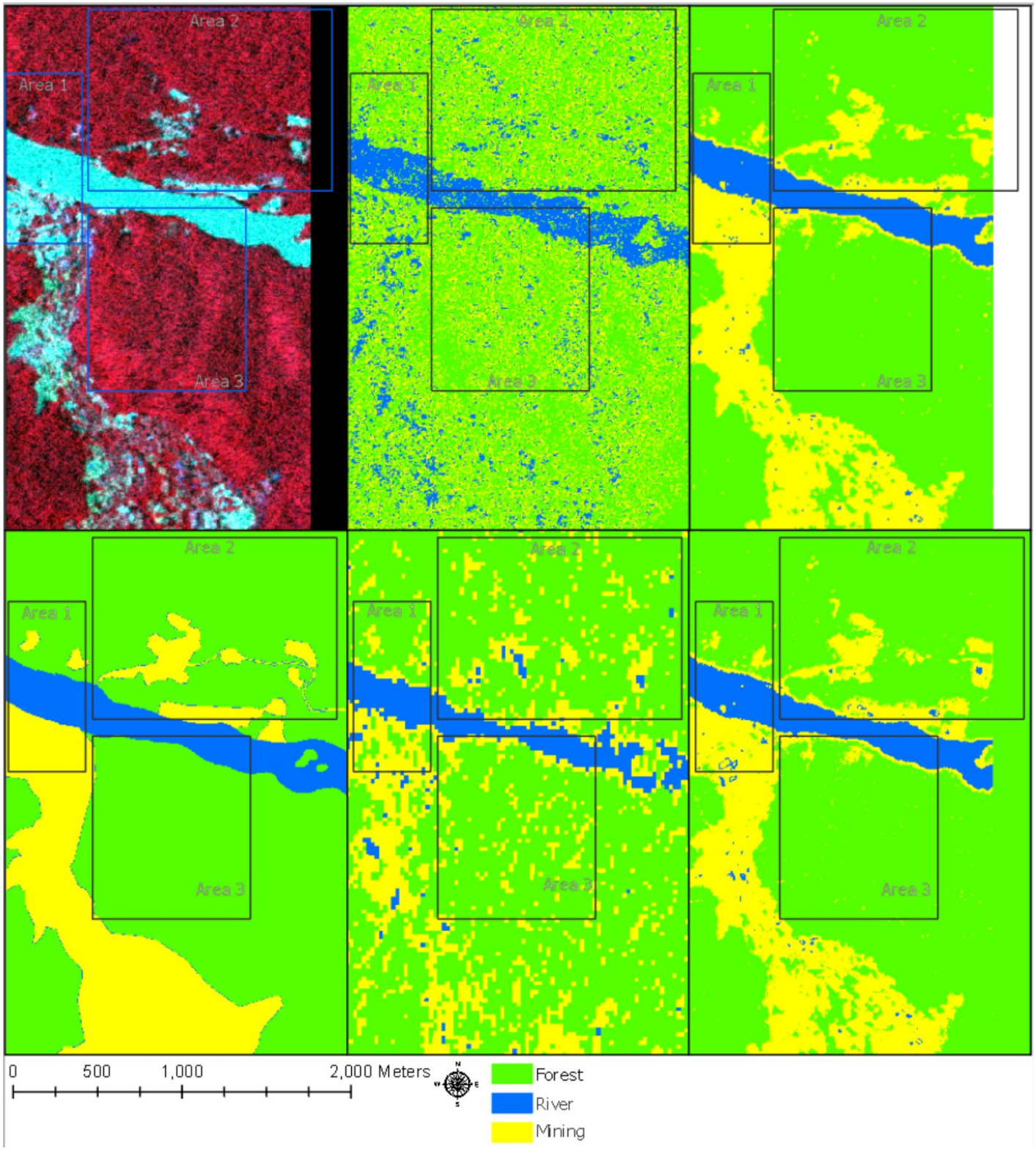


Figure 65: Comparison of results from Sentinel-2, Sentinel-1, fused data, and stacked data at the Cuyuni River mining site.

Top left: Data, fused Sentinel-1+ Sentinel-2. Top middle: Fused predictions.

Top right: Stacked predictions. Bottom left: True labels.

Bottom middle: Sentinel-1 predictions. Bottom right: Sentinel-2 predictions.

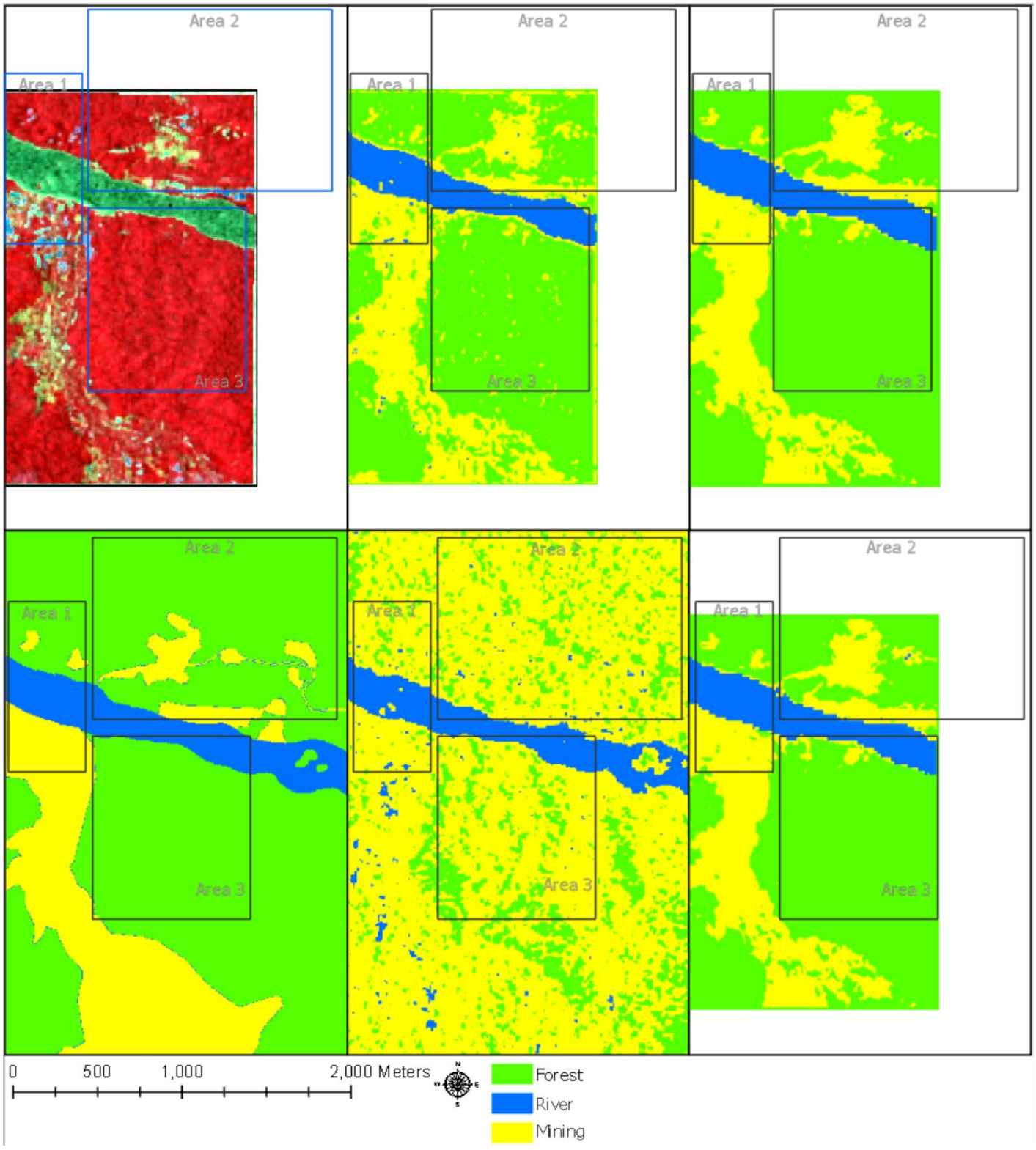


Figure 66: Comparison of results from Landsat, ALOS-2, fused data, and stacked data at the Cuyuni River mining site.

Top left: Data, fused ALOS-2 + Landsat. Top middle: Fused predictions.

Top right: Stacked predictions. Bottom left: True labels.

Bottom middle: ALOS-2 predictions. Bottom right: Landsat predictions.

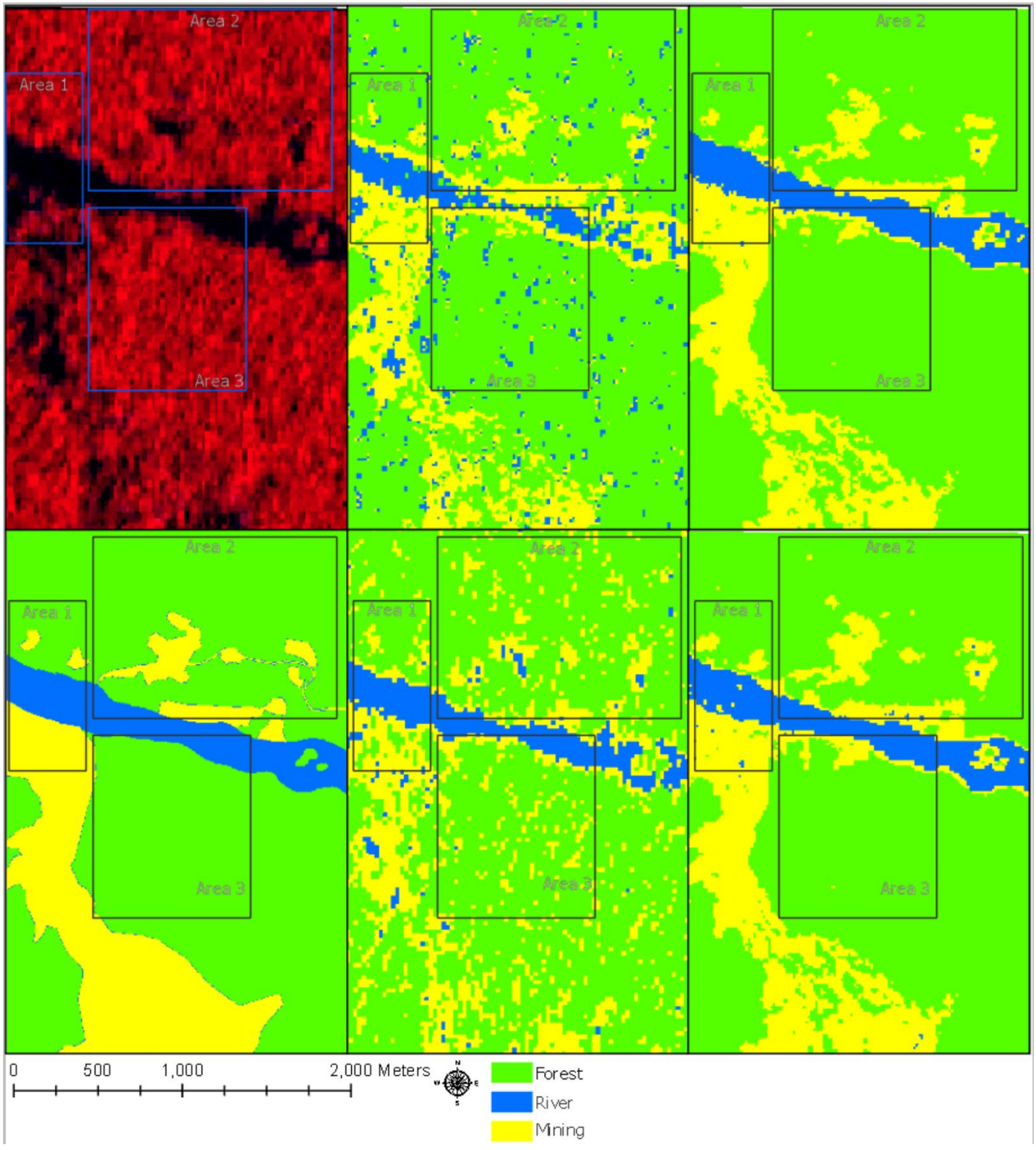


Figure 67: Comparison of results from Sentinel-2, Sentinel-1, fused data, and stacked data at the Konawaruk River mining site.

Top left: Data, fused Sentinel-1+ Sentinel-2. Bottom left: True labels.

Bottom middle: Sentinel-1 predictions. Bottom right: Sentinel-2 predictions.

Top middle: Fused predictions. Top right: Stacked predictions.

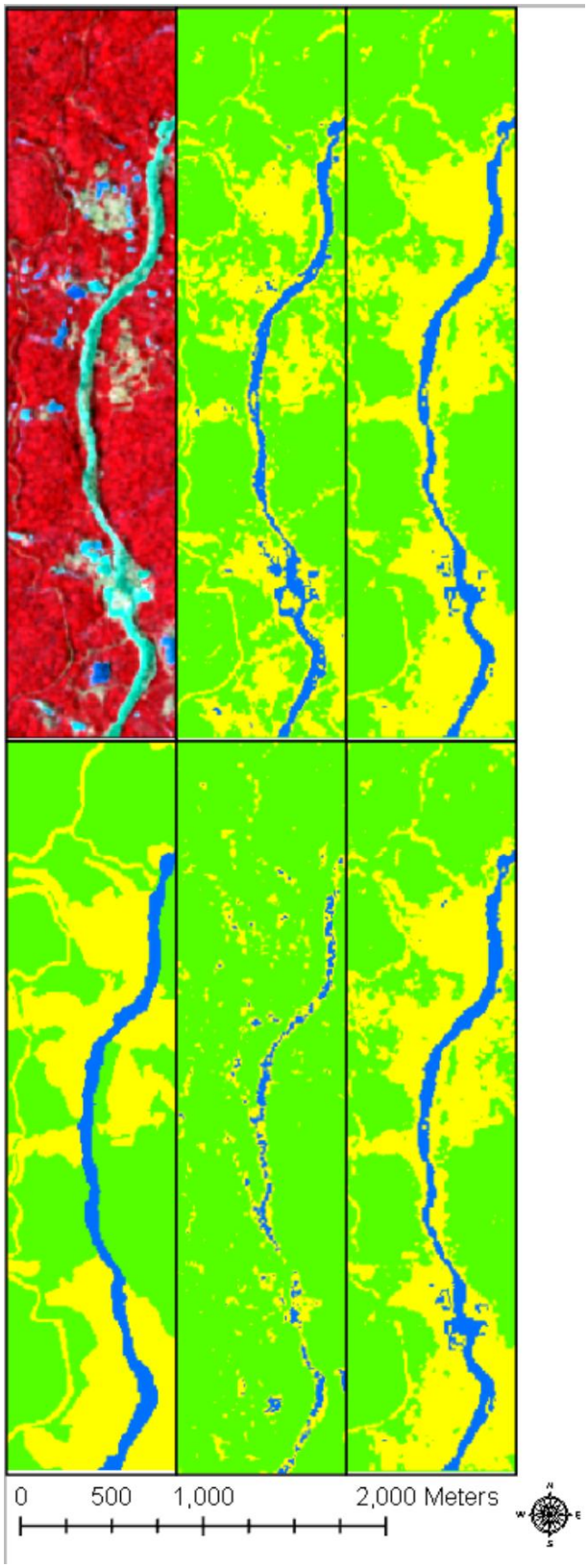
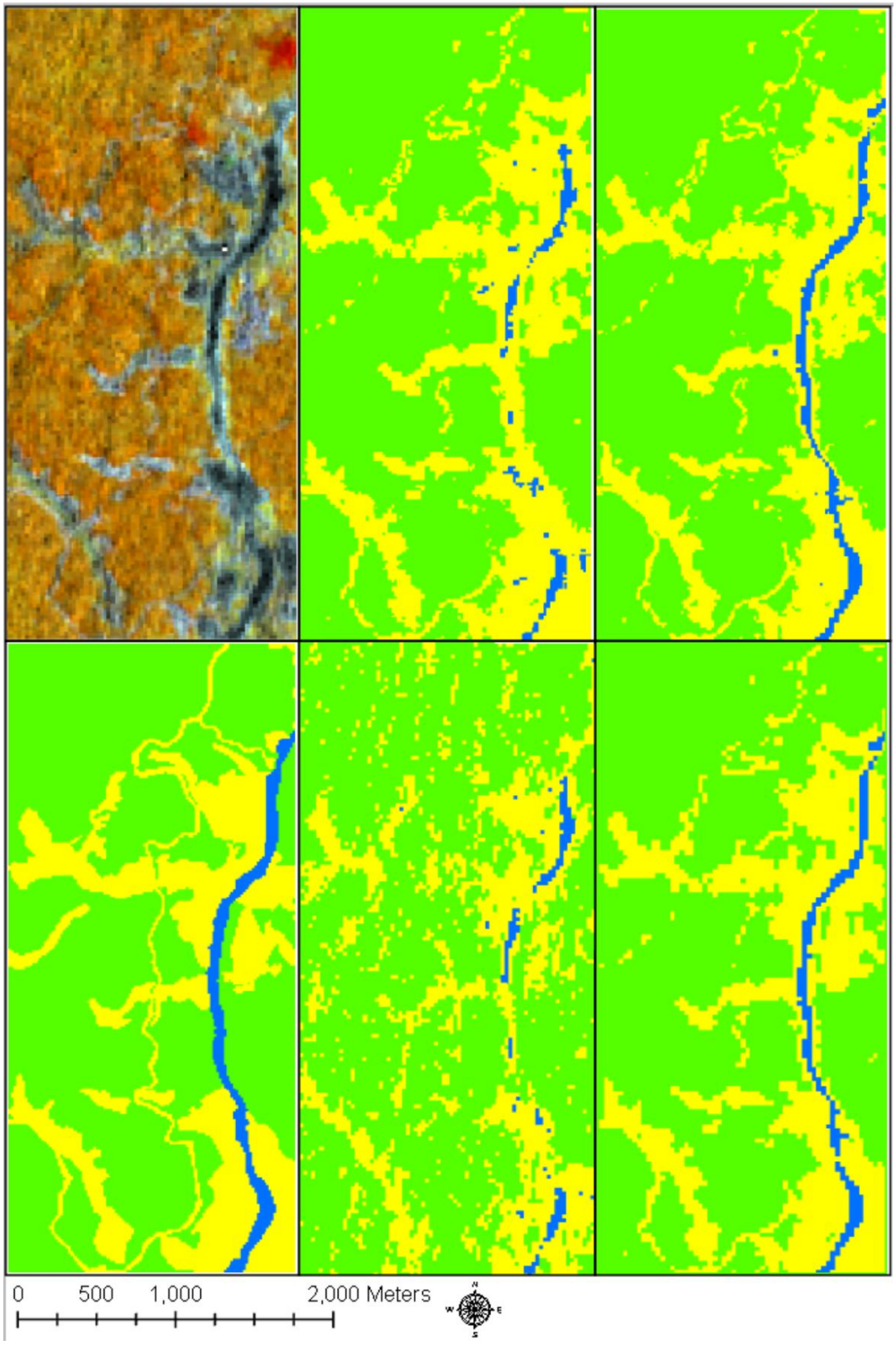


Figure 68: Comparison of results from Landsat, ALOS-2, fused data, and stacked data at the Konawaruk River mining site.

Top left: Data, fused ALOS-2 + Landsat. Top middle: Fused predictions.

Top right: Stacked predictions. Bottom left: True labels.

Bottom middle: ALOS-2 predictions. Bottom right: Landsat predictions.



CHAPTER 8: RESULTS IN KEY LAND COVER TYPES

SECTION 8.1 INTRODUCTION

Now that the optimal data and classification algorithm have been established in the previous chapters, this chapter will extend the study by using them to test classification performance in a variety of land cover types that better represents the variety of tasks in deforestation mapping in Guyana. Combinations of land cover types will include Forest and River, Forest and Mining, Forest and Road, Forest and Grassland, Forest, Settlement and Agriculture, and all of those classes at once.

In general, study locations were selected based on these reasons:

- The training data and the validation data need to have the chosen spectral characteristics, and they need to be similar. If the study sites have very different spectral characteristics, then that will introduce an extra variable, complicating the interpretation of the results.
- The deforested area, and forest around it, needs to be cloud free. There are no cloud-free images, and there are a few images that are completely clouded over. In selecting study sites, many potential areas were clouded over, which narrowed down the options to only a few. Study sites were chosen from among those that were clear above the deforested area.

Here is a description of the land cover types being tested:

- **Mining (Historic)**: The historic mining site is an abandoned mine, where there are still ponds, but the bare surfaces have new vegetation growing into them. This new vegetation is easy to distinguish visually from the surrounding forest. The two sites were chosen because they have the largest area of historic mining that has no bare surfaces.
- **Mining (Active)**: An active mine has ponds and bare surfaces, and can have some vegetation, but less than historic mining. The two sites were chosen because they have large bare surfaces and the least amount of new vegetation.
- **Mining (Mixed)**: A mixed mining site has both active mining and new vegetation growing in. In all three cases, the training and validation data were chosen from two locations that are spectrally similar to each other, to ensure that errors reflect the land cover type and not a spectral mismatch. The two sites were chosen because they are

spectrally similar to each other and they are the largest mining areas that have large amounts of both bare surfaces/mining ponds (active) and new vegetation (historic).

- **Roads**: In a physical sense, roads are deforestation – trees are cut down to clear the way for them. However, roads are ordinarily not counted as deforestation, because they are narrow enough to disappear in the medium resolution satellite images often used to measure and report deforestation. It would be good to test classification accuracy on roads anyway. The two road images chosen here are the widest roads available, to ensure that they will be at least one full pixel wide. This also means that most roads will be narrower.
- **Settlements and Agriculture**: Settlements have roads, paved surfaces, isolated trees, grass and buildings. Settlements and agriculture were tested together, because both are infrequent enough in Guyana that there are very few examples available in the reference data. The two sites chosen are the northern and southern halves of the same town, and this is the only pair of images that have settlements and agriculture and not mining or rivers.
- **Rivers**: As natural features, rivers are not an activity like the previous examples. Instead, rivers need to be classified when measuring the total non-forest area, and classifying rivers is useful to hydrological work.
- **Grassland**: Like rivers, grasslands are natural features that need to be classified to measure the non-forest area. Separating forest from grassland is very important in those parts of Guyana that are naturally savannah, and there are large areas of savannah in the interior uplands. The two study sites were chosen as the pair that is most spectrally similar.
- **All Classes**: The examples above separate two or three classes (one class is always Forest). The last test separates six of those classes at once (Forest, River, Mining, Road, Settlement, Agriculture). The training and validation data are taken from two locations in Linden, which has jungle, the Demerara River, an active bauxite mine, a town, and farm fields, all in close proximity to each other. As a High Forest Low Deforestation country, Guyana has little deforestation, and the majority of that is mining. Linden was chosen because there aren't any other areas that have this combination of deforestation drivers.

The sampling strategy in this chapter is different from the sampling strategy used in Chapters 4-7. In those chapters, because there was only one study site that met the criteria for each comparison, one image was used, and the pixels were split into training and validation samples.

In this comparison, each test will be done using two images from different study sites with the same land cover types.

One image will be the training image: the training data will consist of every pixel from that image. The other image will be the validation image: the confusion matrix, associated statistics, and predictions and class probabilities will all be calculated from that image. This means that the predictions and probabilities images will be based entirely on data that was not seen in training. The training and validation images will be from different locations, and they will be chosen from two sites that have similar spectral/backscatter characteristics. The two images will come from different times, though each image is from one date, which ensures that the training and validation data are consistent. Guyana is very cloudy, which greatly reduces the ground surface that is visible at a given point in time.

The locations of the study sites can be seen in Figure 68.

Figure 69: Location of the training and validation data for the tests in this chapter.

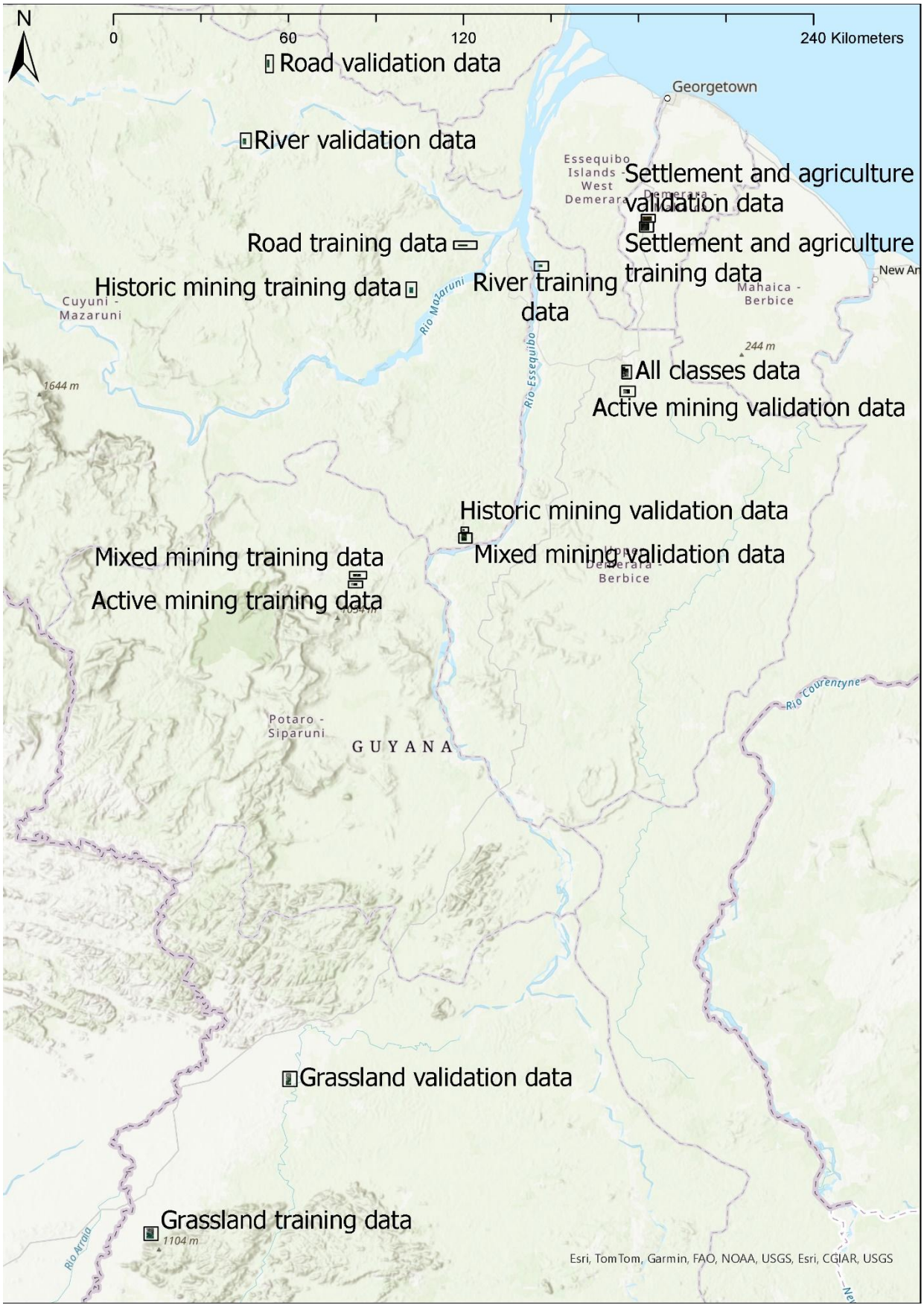


Table 43: Dates of the data for each study site.

Sedimented River	Training Image	7th December 2017
	Validation Image	18th October 2019
Clear River	Training Image	8th October 2017
	Validation Image	14th August 2018
Narrow River	Training Image	13th October 2019
	Validation Image	17th December 2017
Mixed River	Training Image	17th December 2017
	Validation Image	8th October 2017
Historic Mining	Training Image	30th June 2018
	Validation Image	23rd September 2017
Active Mining	Training Image	23rd September 2017
	Validation Image	23rd September 2017
Mixed Mining	Training Image	23rd September 2017
	Validation Image	20th June 2018
Roads	Training Image	6th April 2019
	Validation Image	18th October 2019
Grassland	Training Image	8th September 2019
	Validation Image	26th January 2019
Settlement and Agriculture	Training Image	3rd October 2017
	Validation Image	3rd October 2017
All Classes	Training Image	23rd September 2017
	Validation Image	23rd September 2017

RIVERS

Random Forest with stacked Sentinel-1 and Sentinel-2 data (as in Chapter 7) was tested in four river locations, which represent:

1. A river with sediment from nearby mining. The results can be seen in figure 70 in this chapter.
2. A clear river (low sediment). The results can be seen in figure 72 in this chapter.
3. A narrow river, one where there will be more mixed pixels. The results can be seen in figure 74 in this chapter.
4. A mix of river types, where the training data is a river with sediment, and the validation data is a clear river. The results can be seen in figure 76 in this chapter. This will test the classifier's ability to classify the full range of rivers in Guyana in wall-to-wall deforestation mapping.

The sedimented river results show that, for the most part, the boundaries in the predictions follow the true labels very closely (Figure 69 and Table 44). The exceptions are for very narrow rivers, and for an island that wasn't in the class labels. This is important, because the ability of Random Forest to correct labels means that it will produce an accurate land cover map even as the extent of the river channel or water level changes. In the class probabilities, in the bottom right corner, the probability of Forest is much lower where a small stream is located, indicating that class probabilities can find streams small enough to be dominated by mixed pixels. In the confusion matrix (Table 44), the Consumer's and Producer's Accuracy for Forest is near perfect (above 95%), but Consumer's Accuracy for River is 78% while Producer's Accuracy for River is 96%, which reflects the inability to accurately classify sections of the river that are one or two pixels (10-20 m) wide.

Classification of wide rivers with low sediment showed near perfect results. In the confusion matrix (Table 45), the Consumer's Accuracy and Producer's Accuracy are above 98.60% for all classes, and in the predictions (Figure 71), the class boundaries are very close to the true labels, with almost no class probabilities close to 50%.

Classification of narrow rivers showed far worse accuracy. In the confusion matrix (Table 46), the Consumer's Accuracy and Producer's Accuracy for Forest are both above 97%. But the Producer's Accuracy for River is 60%, and the Consumer's Accuracy is 39%. In looking at the

predictions (Figure 73), the predicted river follows the course of the real river perfectly, and there is no region of high uncertainty around it. However, the predicted river is narrower than the real river. In comparing the predictions with the validation data, two observations stand out. One is that the trees are casting shadows on the river, which change its spectral properties greatly. Another is that, because the river is so narrow, mixed pixels represent a greater proportion of the river.

Classification of a mix of river types has produced accurate results, if a different probability threshold is used. By default, the classifier outputs class probabilities, and in a binary problem, the pixel is predicted to be the class with a probability greater than 50%. The confusion matrix (Table 47) uses this 50% threshold, and finds that 99.94% of pixels (out of 60 376 total) have been classified as Forest. This is near-total inaccuracy. The predictions in Figure 63 show which pixels fall into three other thresholds: 60%, 80% and 95% probability of Forest. The only pixels with less than 60% probability of Forest are a small cluster located near the top of the image, in the river. The pixels that fall between 60% and 80% probability of Forest match the river almost perfectly, and the pixels with greater than 80% probability of Forest match the Forest almost perfectly.

MINING

The mining results (results in figures 65, 67, 69 in this chapter show that Random Forest tends to find vegetation inside the mining site and classify it as Forest, even when that vegetation is not in the class labels. The mixed mining site (Figures 81 and 82), and to a lesser extent the historic mining site (Figures 77 and 78), classified scattered areas of Forest as Mining. This follows the spectral variation in the forest canopy; some patches of forests are spectrally similar to vegetation in the mining site. Note that this does not occur in the active mining example (Figures 79 and 80), which has very little vegetation inside the mining site. Looking at the class probabilities, the certainty of Forest is high for historic mining and active mining, and low for mixed mining.

The confusion matrices offer three very different results. Table 48 (historic mining) has a Consumer's Accuracy of 47% for Mining. Table 49 (active mining) has a Consumer's Accuracy of 94% for Mining, and Table 50 (mixed mining) has a Consumer's Accuracy of 77% for Mining. Forest accuracy is consistently above 90%. There is no one accuracy for the Mining land cover type; it varies considerably from one kind of mining site to the next.

ROAD

The confusion matrix (Table 51) shows near perfect accuracy for Forest (above 98%), and a Consumer's Accuracy of 2% and Producer's Accuracy of 43% for Roads. These results are in the two areas with the widest roads available, to minimize the possibility of the roads being so narrow that they disappear.

The predictions (Figure 84) show that the choice of probability threshold makes the difference between an accurate classification and a false negative. If pixels with a probability of Road greater than 50% are classified as Road, then most of the road will be missed, only a few isolated sections. If the probability threshold is set to 10% probability of Road and 90% probability of Forest, then the predictions will accurately identify the extent of the road. The pixels with greater than 90% probability of Forest will accurately identify the extent of the Forest. These results make sense, considering the width of a road and the resolution of the data. With 10 meter pixels, a road will consist mostly of mixed pixels, and if the forest occupied the majority of the pixel, then that will skew its spectral properties in the direction of forest.

GRASSLAND

In Forest and Grassland predictions (results can be seen in figures 85 and 86 of this chapter), the predicted class boundaries closely follow the true class boundaries. Some areas labelled as Forest were classified as Grassland, where the vegetation is lower than the surrounding forest. In the confusion matrix (Table 52), the Producer's Accuracy for Forest is 79% and 96% for Grassland. The histograms (Table 71) show that the distributions for pixel values of forest and grassland have overlap, but they each have one peak, and their peaks are well separated.

SETTLEMENT AND AGRICULTURE

When classifying Settlement and Agriculture, the confusion matrix (Table 53) shows near perfect accuracy for Forest (Consumer's Accuracy of 97% and Producer's Accuracy of 96%), moderate accuracy in Settlements (Consumer's Accuracy of 65% and Producer's Accuracy of 54%), and almost no accuracy in Road and Agriculture (0% to 6%). The predictions (Figures 87 and 88) map the forest very close to the true Forest labels. The true Settlement areas are predicted accurately, but several areas labelled as Forest are classified as Settlement, where there is a clearing in the forest. The classifier missed all of the Road and Agriculture pixels. The class probabilities show high certainty for most classes.

ALL CLASSES

The final test site includes all land cover classes. In this test, the Forest, River, Mining, some of the Roads, and most of the Settlement were outlined very accurately (Figures 89 and 90). One of the Agriculture areas was identified partly as Forest, and partly as Settlement. In the confusion matrix (Table 54), all accuracies were above 85%, except for Agriculture, where the Consumer's Accuracy is 11%.

SECTION 8.3 DISCUSSION

One difference between this study and previous research is the variety in study sites. In Scarpa et al. (2018), there was just one study site. In Milodowski et al. (2017), the objective was to measure forest loss in areas where agriculture was expanding, and two study sites in different provinces in Brazil were used. This chapter tests eight study sites with different land cover types. They are:

- River classifications are very accurate when classifying wide rivers with similar spectral characteristics. Small streams disappear in the predictions, but the class probabilities indicate their location with areas of lower certainty. When classifying rivers with very different sediment composition and therefore spectral characteristics, classification was still very accurate when the probability threshold was adjusted.
- Mining classifications are variable in accuracy and certainty from one location to the next, depending on the spectral characteristics of each site. This includes the presence or absence of vegetation, the presence or absence of bare surfaces, and the spectral characteristics of mining ponds.
- Roads are too narrow to appear in predictions. However, class probabilities show areas of mixed pixels where the road is.
- Agriculture never appears in predictions or class probabilities.

The tests above show situations where the classifier misses most examples of the non-forest class with the default probability threshold, but outlines the class boundaries very accurately when a different threshold is used. Variable thresholds for class probabilities is a concept that has been extensively explored in the literature on other topics. Wu et al. (2018) tested a Fully Convolutional Network at land cover mapping, with and without adaptive probability thresholds. Measured using the Jaccard Index (also known as Intersection Over Union), accuracy when

classifying small vehicles went from 0.166 without adaptive thresholds to 0.238 with. When classifying roads, accuracy went from 0.863 without to 0.864 with.

This also indicates that classification accuracy depends heavily on location. This is a variable that will have to be addressed in further research on other aspects of machine learning in remote sensing of forests. In doing a classifier comparison, or a sensor comparison, or a study of classifier behaviour, or a study of feature engineering, sample size or hyperparameter tuning, the results will vary greatly between study sites, even within the sample land cover type. This means that site selection is a critically important decision, and there needs to be a strategy for selecting one site or many sites.

The knowledge needed is an understanding of how variation in the data causes and does not cause variation in classifier performance. This means understanding:

- The nature of the data from different areas, in statistical terms, to fully understand what the classifier is receiving. This means detailed descriptive statistics on spectral properties (multispectral and hyperspectral data), backscatter, polarimetry and interferometry (SAR), and 3D structure (Lidar and long-wavelength SAR).
- A detailed study of how different classifiers respond to data at different sites. This would go beyond accuracy comparison, and dive into the theory of statistical learning, and how each step of the algorithm is responding to the data at different sites. A summary of variations in classifier behaviour from one area to another would enable a researcher to choose a set of study sites that represent a large area of interest or would identify study areas that need to be classified by different models. This summary of variations can include sites that behave similarly and can be lumped together, sites that behave very differently and need to be analyzed separately, and exceptions to trends that need to be accounted for.
- The resulting workflow may include both supervised classification and clustering, along with vegetation indices, and different models for different data types, with a first step being segmenting the data based on the characteristics of different areas.

SECTION 8.4 CONCLUSION

This chapter shows that the same data and the same classifier in different settings get different results, even when the different settings belong to the same land cover type, and it shows that the classifier handles some land cover types very well, while struggling with others. It points

the way forward for research on machine learning in a REDD+ context. There are four follow-up studies that can be done based on the results of this chapter:

1. Research into variation in rivers, and how to build a workflow (not necessarily a single classifier) that can classify both clear and turbid rivers.
2. Research into variation in mines, and how to build a workflow (not necessarily a single classifier) that can classify the full variety in mining ponds, vegetation and bare surfaces.
3. Research into variable thresholding to improve predictions in mixed pixel situations, such as roads.
4. Research into the problem of classifying settlements and agriculture.
5. Further research into a strategy for creating class labels that reflect the physical and ecological processes that appear in the data, while still providing the information needed to meet policy requirements.

Figure 70: Data for rivers with sediment classification in Red, Green, Blue.

Top left: Training data in Sentinel-2 RGB.

Top right: Reference data from GeoVantage aerial photography.

Bottom left: Validation data in Sentinel-2 RGB.

Bottom right: Reference data from GeoVantage aerial photography.

Figure 71: Stacked Sentinel-1 + Sentinel-2 (as in Chapter 7) classified using Random Forest. Location has forest and river with sediment with no other land cover types.

Top left: True labels. Top right: Predictions.

Bottom Left: Probability of Forest. Bottom Right: Probability of River.

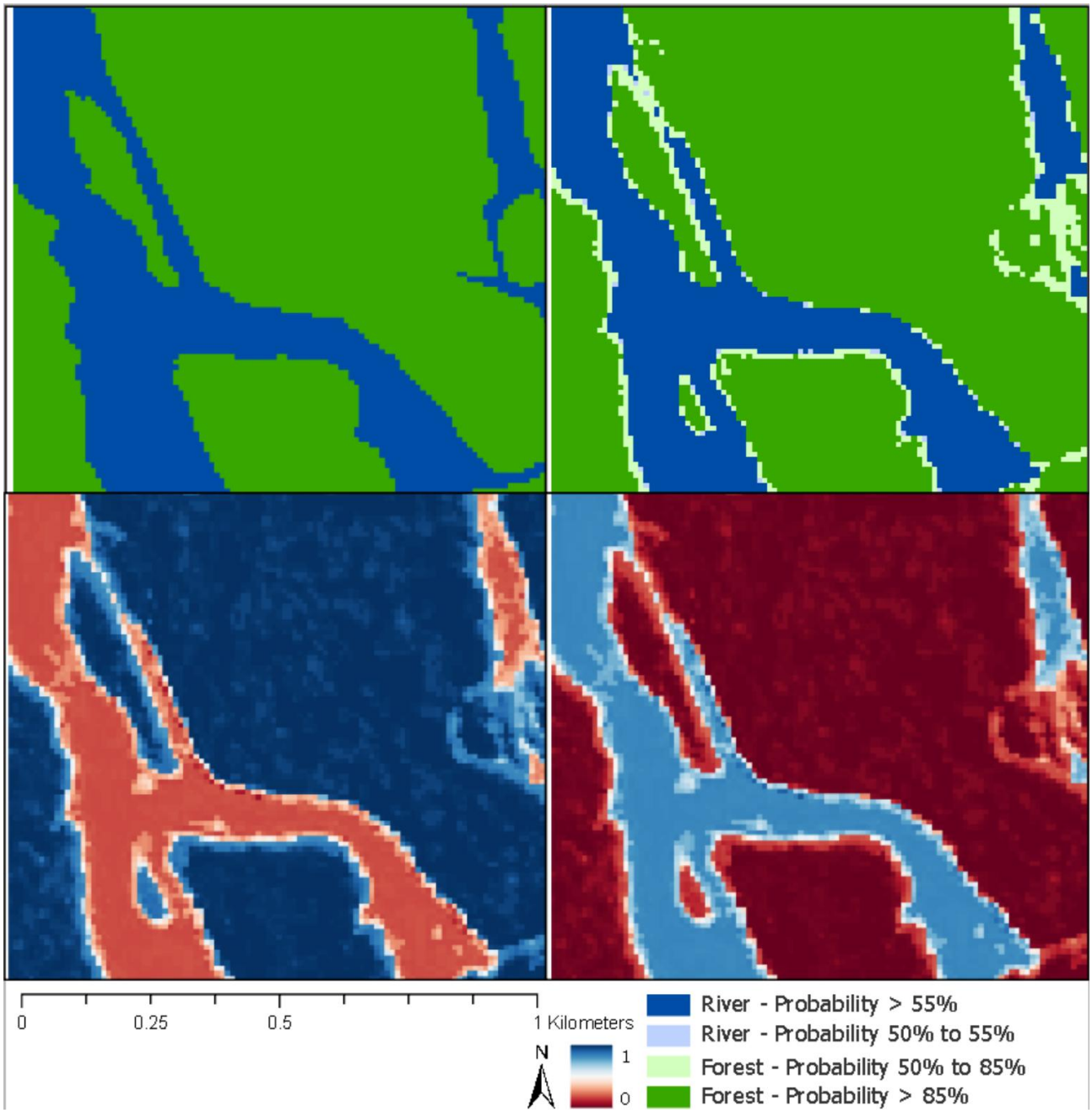


Table 44: Confusion matrix when classifying rivers with sediment.

Accuracy: 96.07% Pixel Count: 27 250		True		Consumer's Accuracy
		Forest	River	
Predicted	Forest	22 986	149	99.36%
	River	921	3194	77.62%
Producer's Accuracy		96.15%	95.55%	

Figure 72: Data for clear river classification in Red, Green, Blue.
Top: Training data in Sentinel-2 RGB.
Middle: Reference data from GeoVantage aerial photography.
Bottom left: Validation data in Sentinel-2 RGB.
Bottom right: Reference data from GeoVantage aerial photography.

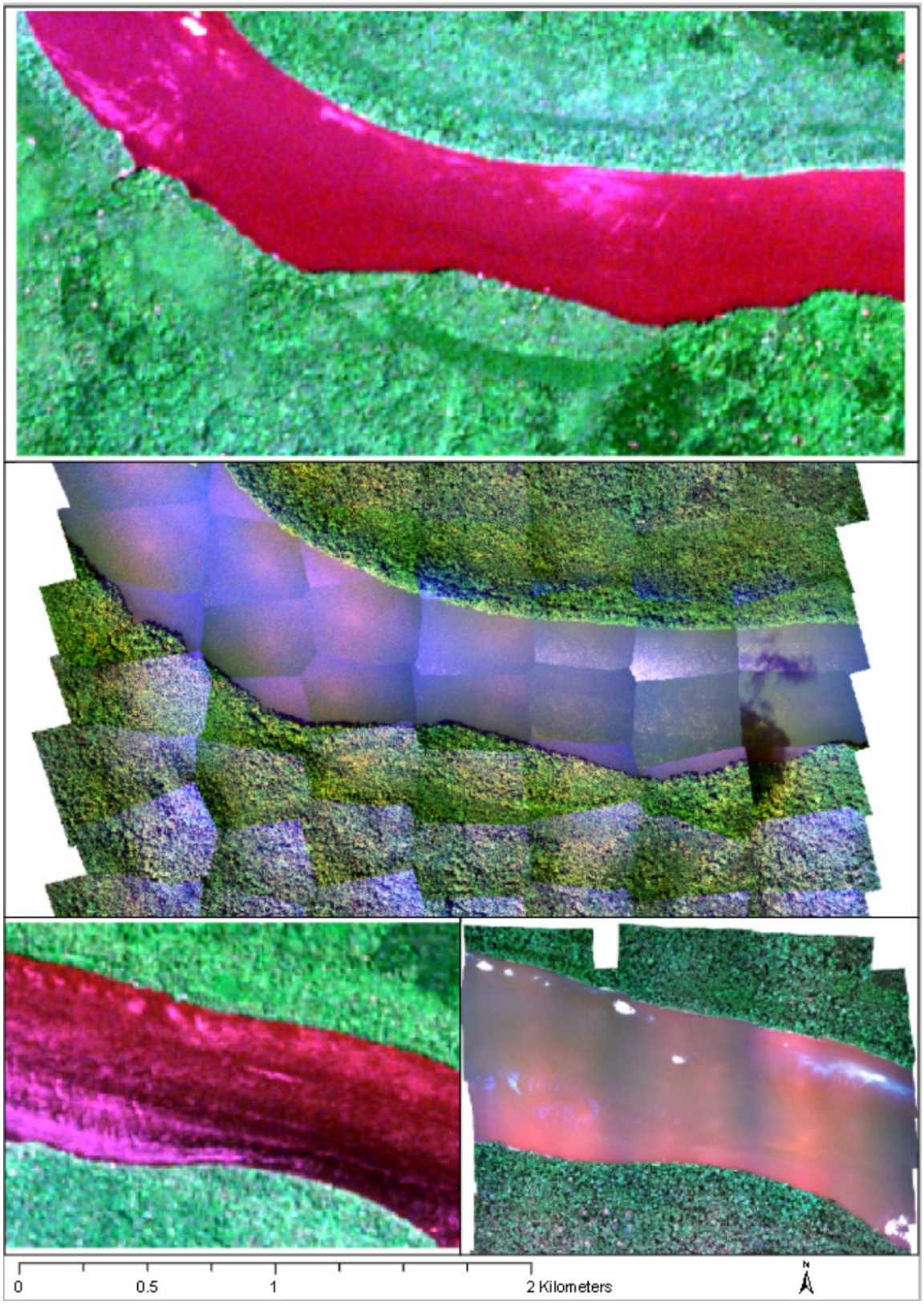


Table 45: Confusion matrix when classifying clear rivers.

Accuracy: 98.64% Pixel Count: 22987		True		Consumer's Accuracy
		Forest	River	
Predicted	Forest	10787	131	98.80%
	River	149	11920	98.77%
Producer's Accuracy		98.64%	98.91%	

Figure 73: Stacked Sentinel-1 + Sentinel-2 (as in Chapter 7) classified using Random Forest. Location has forest and clear river with no other land cover types.

Top left: True labels. Top right: Predictions.

Bottom Left: Probability of Forest. Bottom Right: Probability of River.

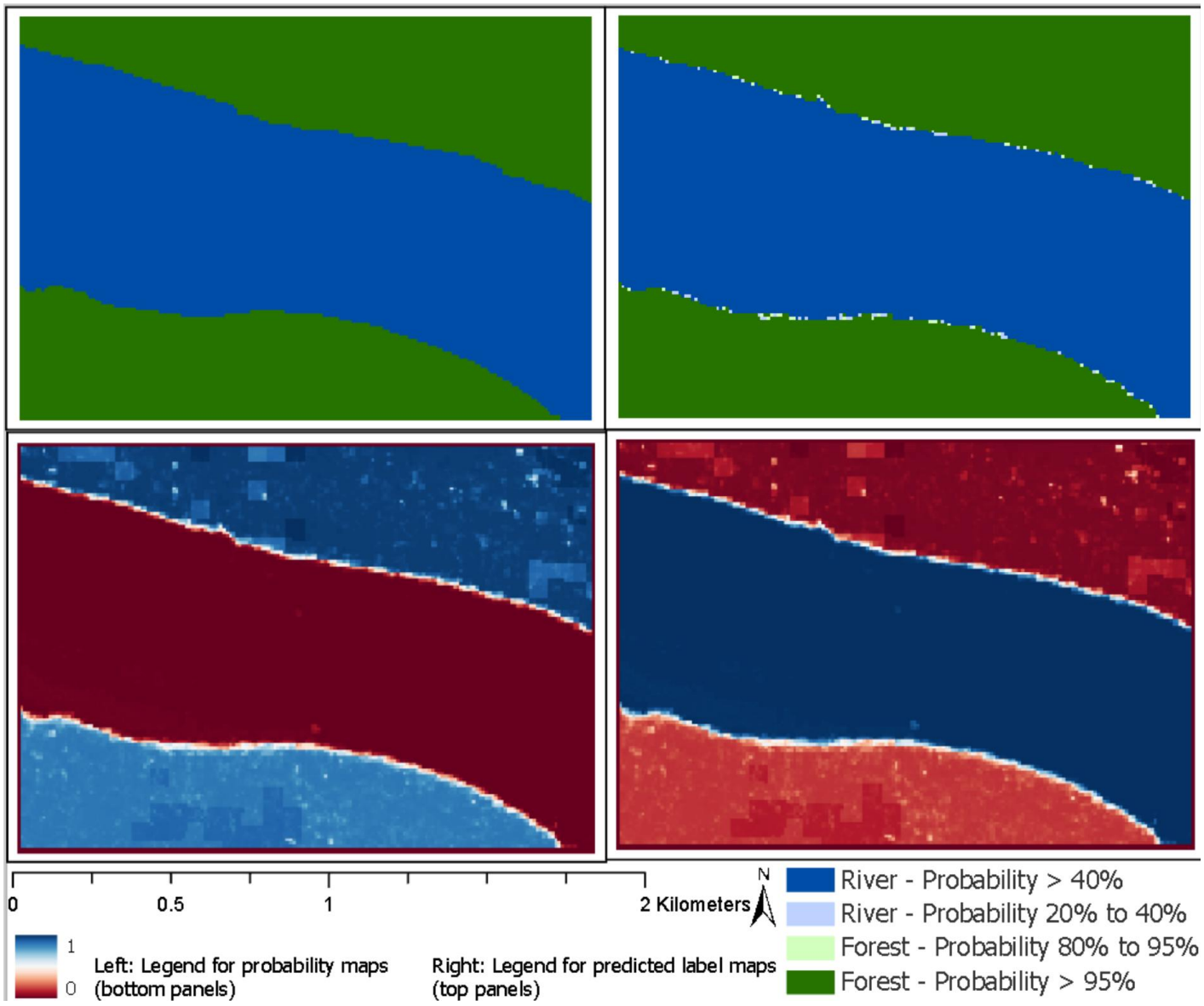


Figure 74: Data for narrow river classification in Red, Green, Blue.
Top left: Training data in Sentinel-2 RGB.
Top right: Reference data from GeoVantage aerial photography.
Bottom left: Validation data in Sentinel-2 RGB.
Bottom right: Reference data from GeoVantage aerial photography.

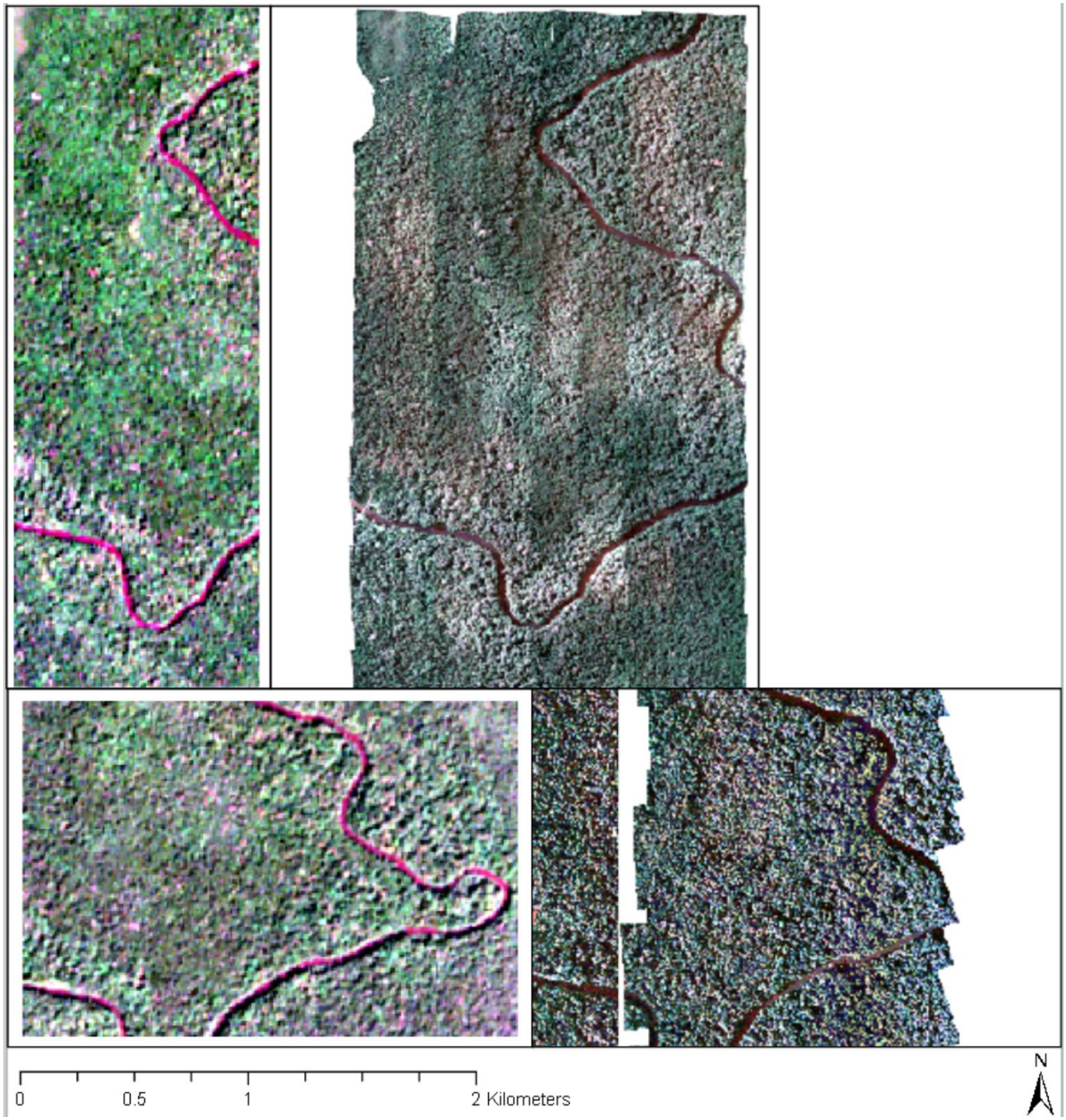


Table 46: Confusion matrix when classifying narrow rivers.

Accuracy: 96.61% Pixel Count: 31899		True		Consumer's Accuracy
		Forest	River	
Predicted	Forest	30334	325	98.94%
	River	755	485	39.11%
Producer's Accuracy		97.57%	59.88%	

Figure 75: Stacked Sentinel-1 + Sentinel-2 (as in Chapter 7) classified using Random Forest. Location has forest and narrow rivers with no other land cover types.

Top left: True labels. Top right: Predictions.

Bottom Left: Probability of Forest. Bottom Right: Probability of River.

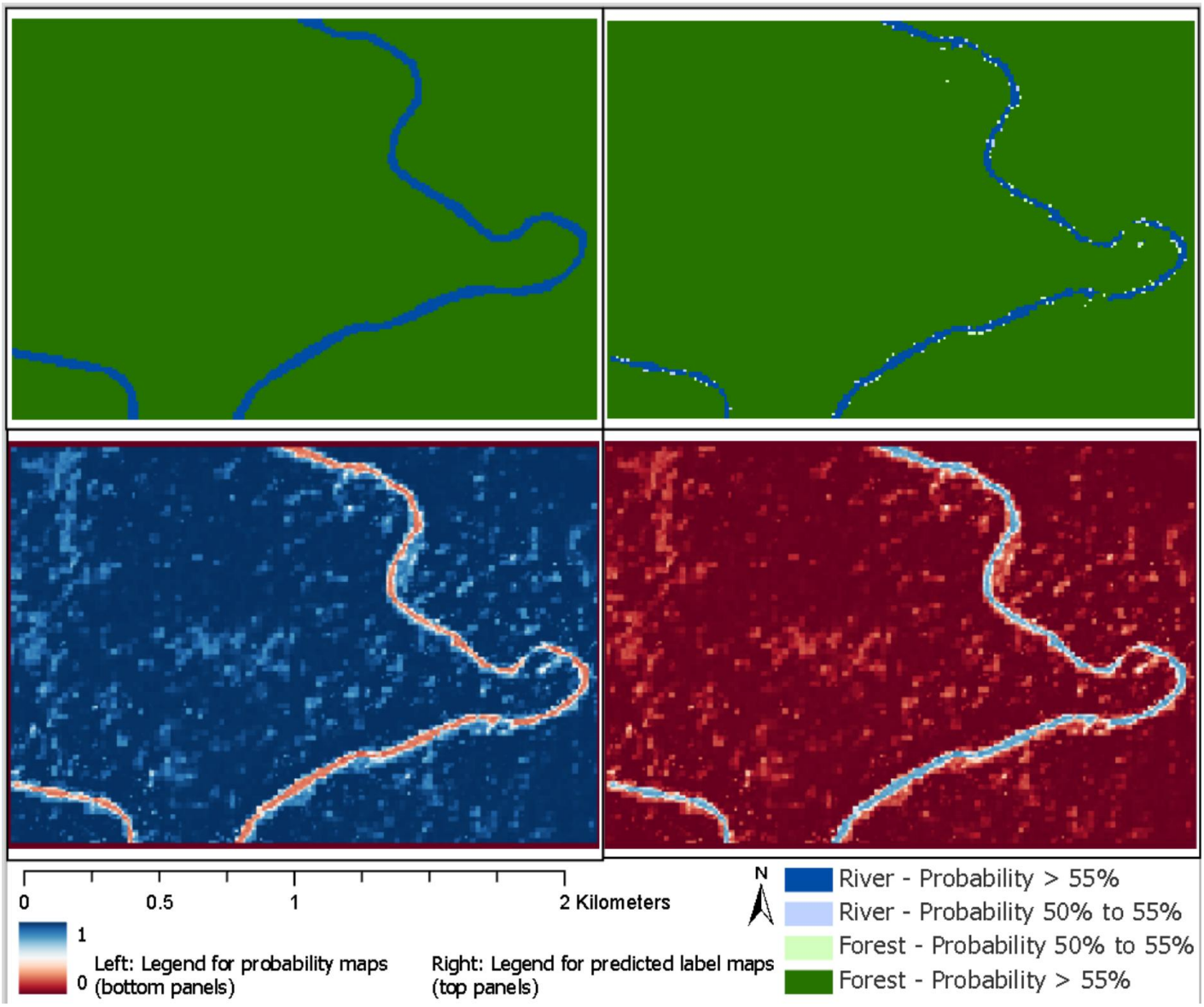


Figure 76: Data for mixed river classification in Red, Green, Blue.

Top left: Training data in Sentinel-2 RGB.

Top right: Reference data from GeoVantage aerial photography.

Middle: Validation data in Sentinel-2 RGB.

Bottom: Reference data from GeoVantage aerial photography.

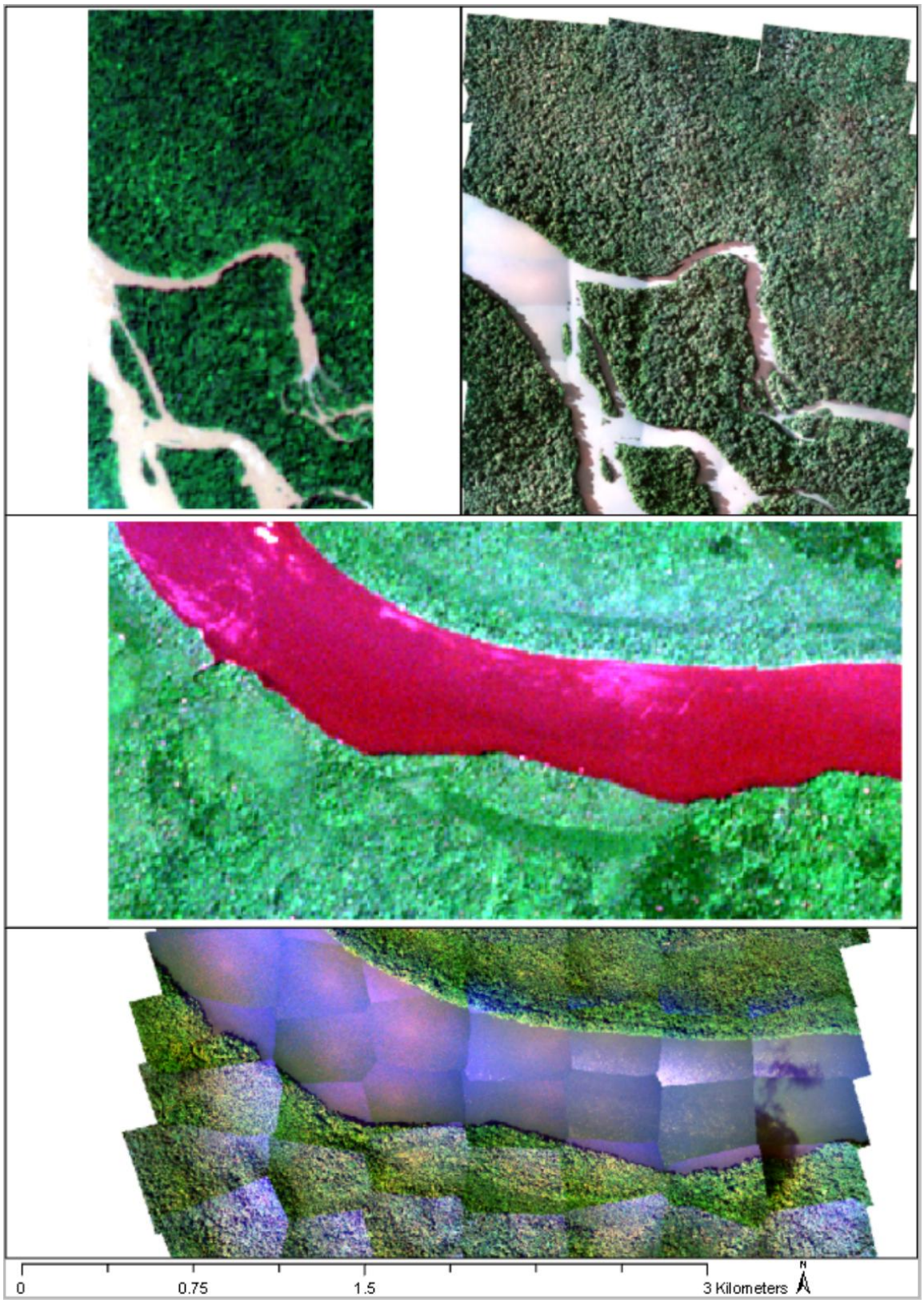


Table 47: Mixed River

Accuracy: 68.28% Pixel Count: 60376		True		Consumer's Accuracy
		Forest	River	
Predicted	Forest	41194	5	99.88%
	River	19144	33	0.17%
Producer's Accuracy		68.27%	86.84%	

Figure 77: Stacked Sentinel-1 + Sentinel-2 (as in Chapter 7) classified using Random Forest. Location has forest and a mix of clear and sedimented rivers with no other land cover types.

Top left: True labels. Top right: Predictions.

Bottom Left: Probability of Forest. Bottom Right: Probability of River.

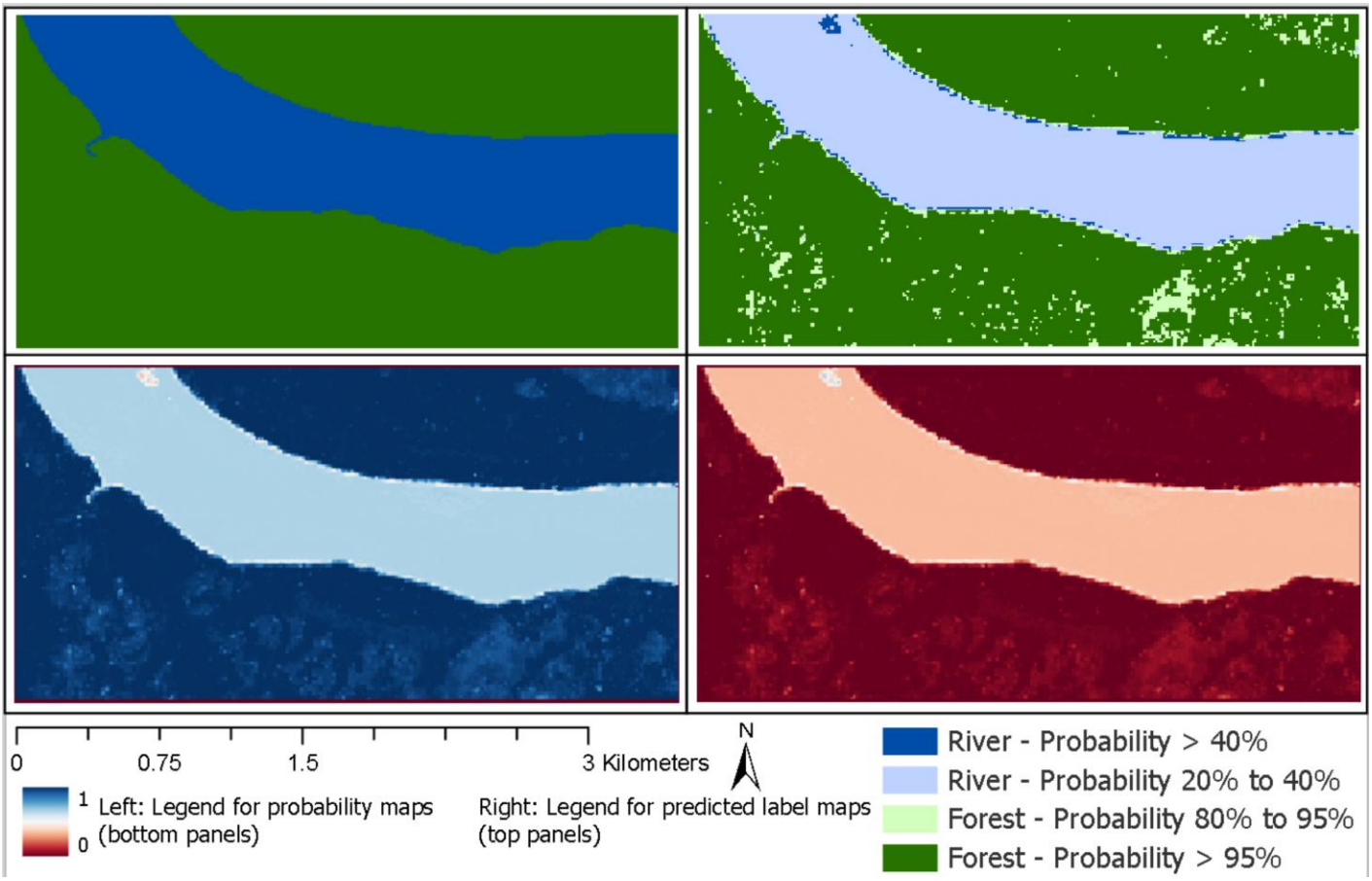


Figure 78: Data to classify forest and mining classes, in an example of historic (abandoned) mining, in Red, Green, Blue. Classification results are in Figure 59.

Top left: Training data in Sentinel-2 RGB.

Top right: Reference data from GeoVantage aerial photography.

Bottom left: Validation data in Sentinel-2 RGB.

Bottom right: Reference data from GeoVantage aerial photography.

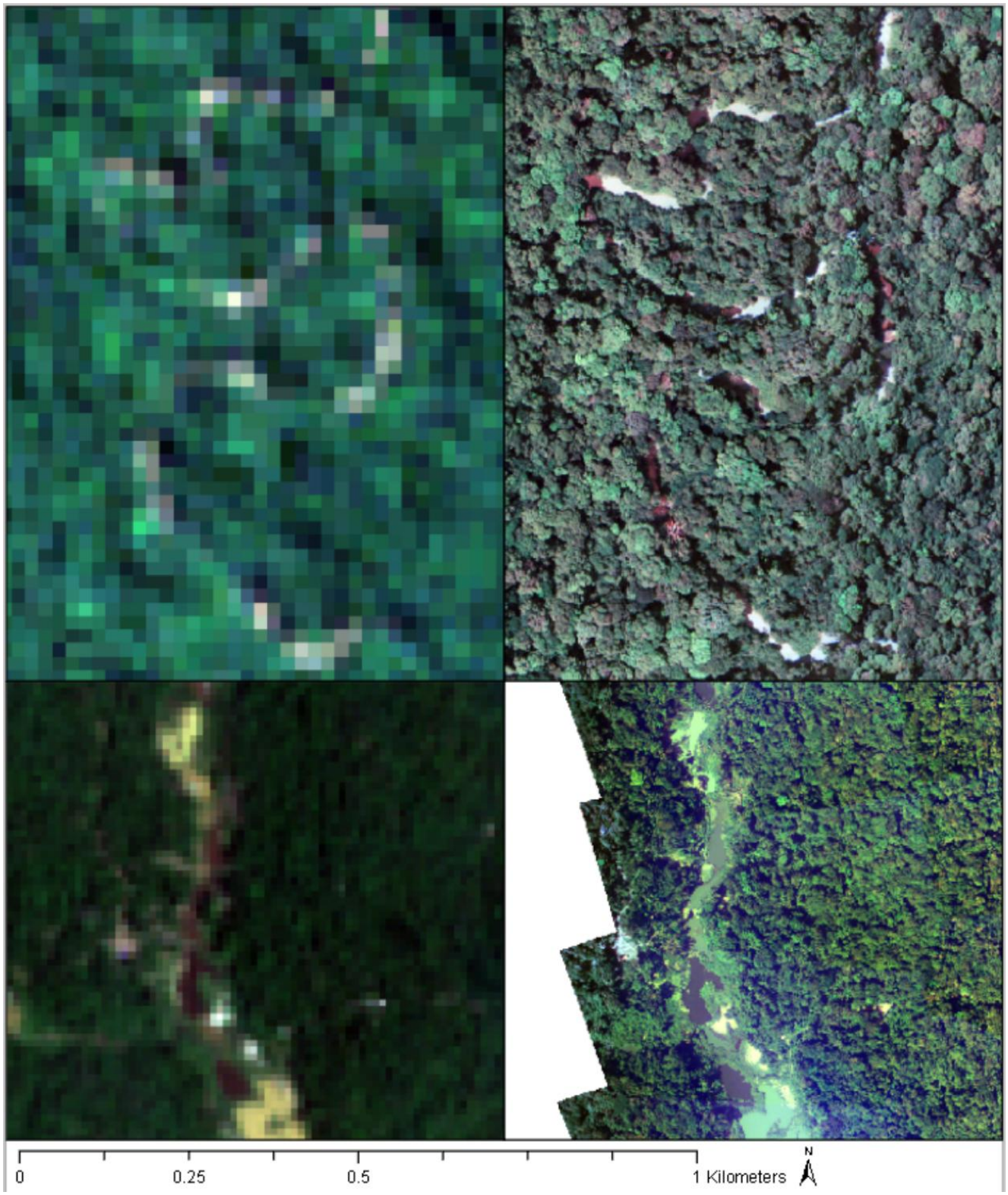


Table 48: Confusion matrix when classifying an historic mining site.

Accuracy: 91.04% Pixel Count: 5236		True		Consumer's Accuracy
		Forest	Mining	
Predicted	Forest	4397	60	98.65%
	Mining	409	370	47.50%
Producer's Accuracy		91.49%	86.05%	

Figure 79: Stacked Sentinel-1 + Sentinel-2 (as in Chapter 7) classified using Random Forest. Location has forest and historic mining with no other land cover types.

Top left: True labels. Top right: Predictions.

Bottom Left: Probability of Forest. Bottom Right: Probability of River.

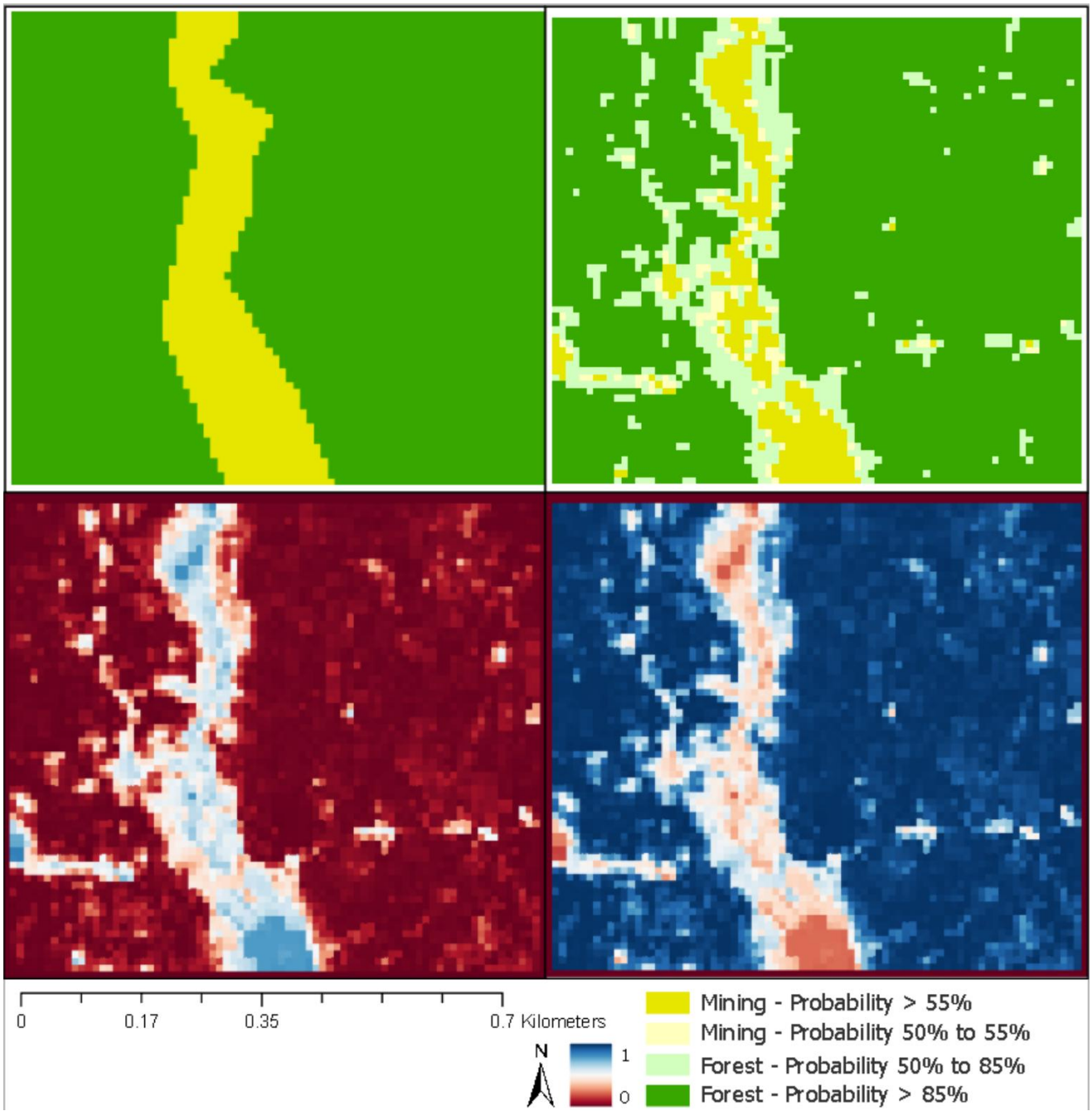


Figure 80: Data for active mining classification in Red, Green, Blue.

Top left: Training data in Sentinel-2 RGB. Location is a gold mine near Mahdia.

Top right: Validation data in Sentinel-2 RGB. Location is a bauxite mine near Linden.

Bottom left: Reference data from GeoVantage aerial photography.

Bottom right: Reference data from GeoVantage aerial photography.

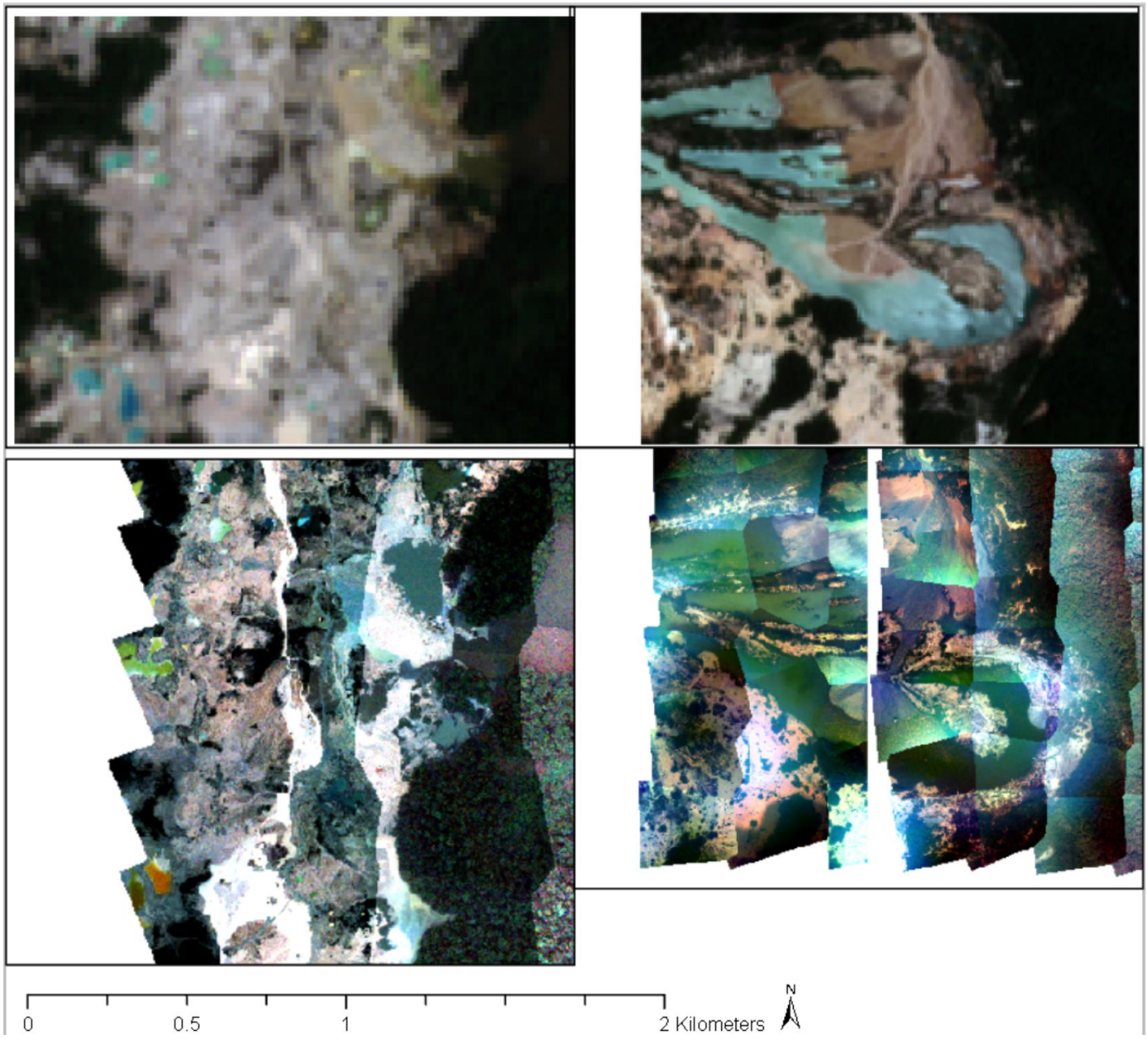


Figure 81: Stacked Sentinel-1 + Sentinel-2 (as in Chapter 7) classified using Random Forest at an active bauxite mine near Linden (see Figure 66). Location has forest and active mining with no other land cover types.

Top left: True labels.

Top right: Predictions.

Bottom Left: Probability of Forest.

Bottom Right: Probability of Mining.

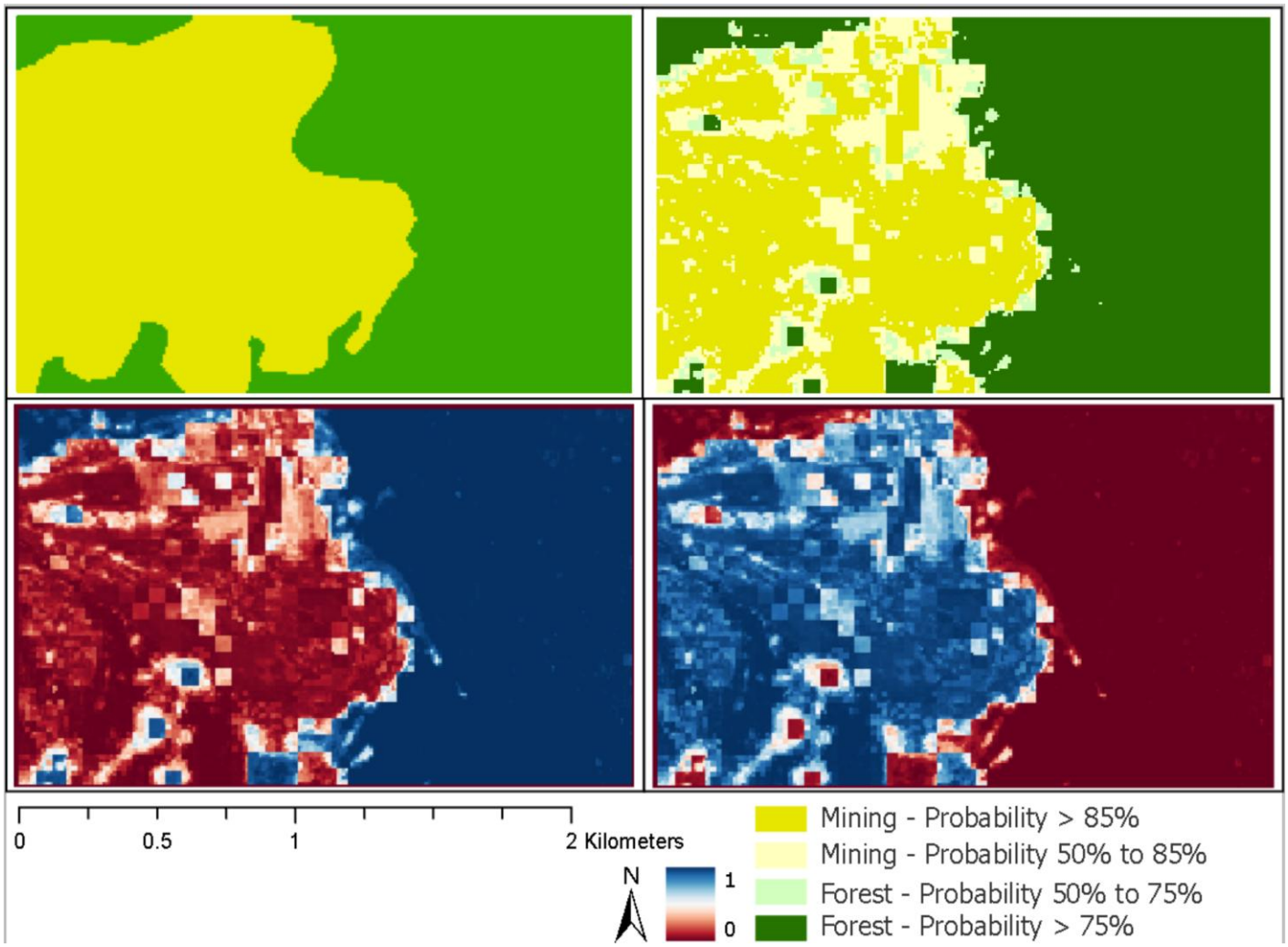


Table 49: Confusion matrix when classifying an active mining site.

Accuracy: 92.26% Pixel Count: 30 192		True		Consumer's Accuracy
		Forest	Mining	
Predicted	Forest	14 398	848	90.63%
	Mining	1489	13 457	94.07%
Producer's Accuracy		94.44%	90.04%	

Figure 82: Data for forest-mining classification in Red, Green, Blue.

Top left: Training data in Sentinel-2 RGB.

Top right: Reference data from GeoVantage aerial photography.

Bottom left: Validation data in Sentinel-2 RGB.

Bottom right: Reference data from GeoVantage aerial photography.

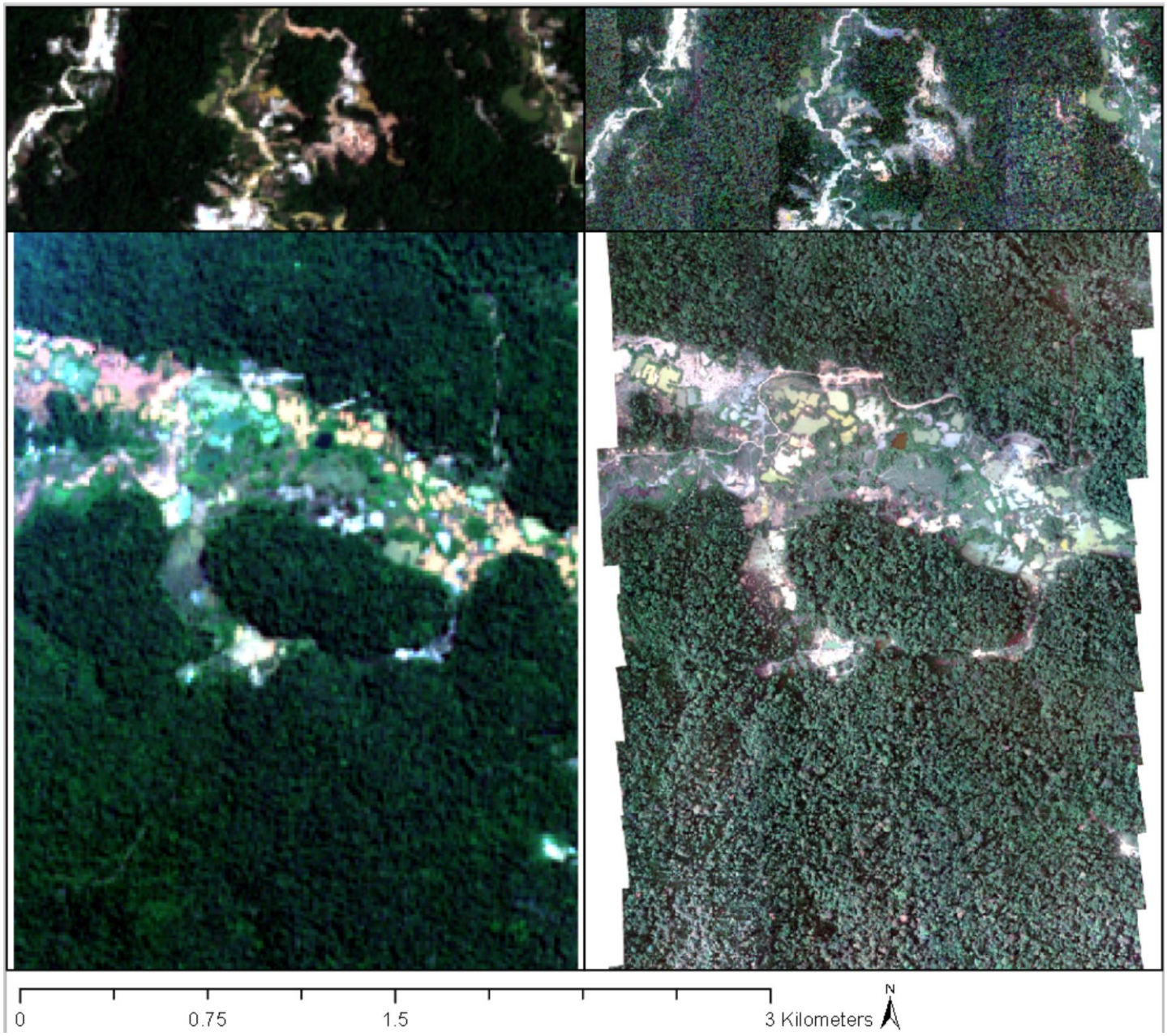


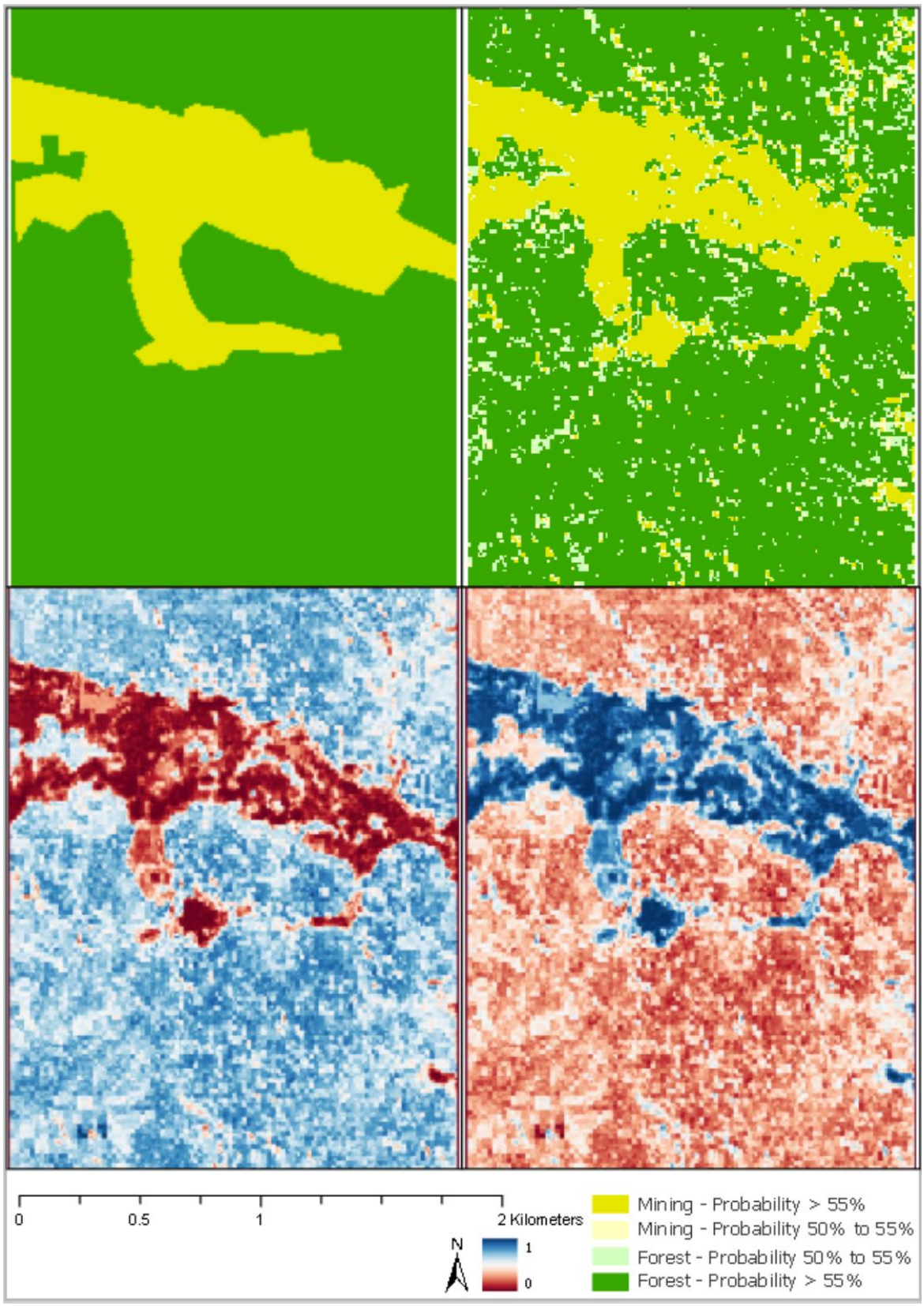
Table 50: Confusion matrix when classifying a mixed mining site.

Accuracy: 90.38% Pixel Count: 44 585		True		Consumer's Accuracy
		Forest	Mining	
Predicted	Forest	32 308	1896	94.46%
	Mining	2392	7989	76.96%
Producer's Accuracy		93.11%	80.82%	

Figure 83: Stacked Sentinel-1 + Sentinel-2 (as in Chapter 7) classified using Random Forest. Location has forest and active mining with no other land cover types.

Top left: True labels. Top right: Predictions.

Bottom Left: Probability of Forest. Bottom Right: Probability of Mining.



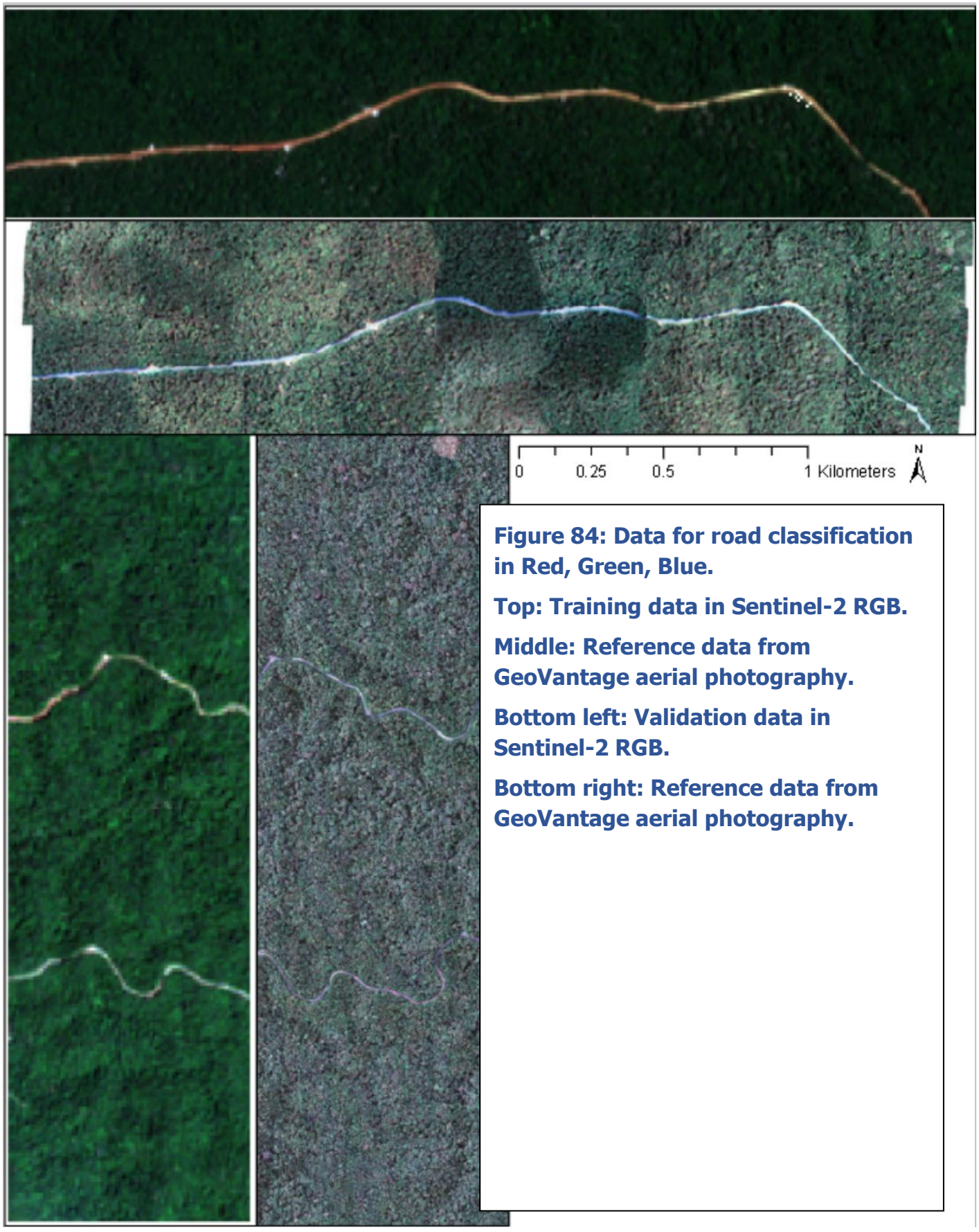


Figure 84: Data for road classification in Red, Green, Blue.

Top: Training data in Sentinel-2 RGB.

Middle: Reference data from GeoVantage aerial photography.

Bottom left: Validation data in Sentinel-2 RGB.

Bottom right: Reference data from GeoVantage aerial photography.

Figure 85: Stacked Sentinel-1 + Sentinel-2 (as in Chapter 7) classified using Random Forest. Location has forest and active mining with no other land cover types.

Top left: True labels.

Top right: Predictions.

Bottom Left: Probability of Forest.

Bottom Right: Probability of Mining.

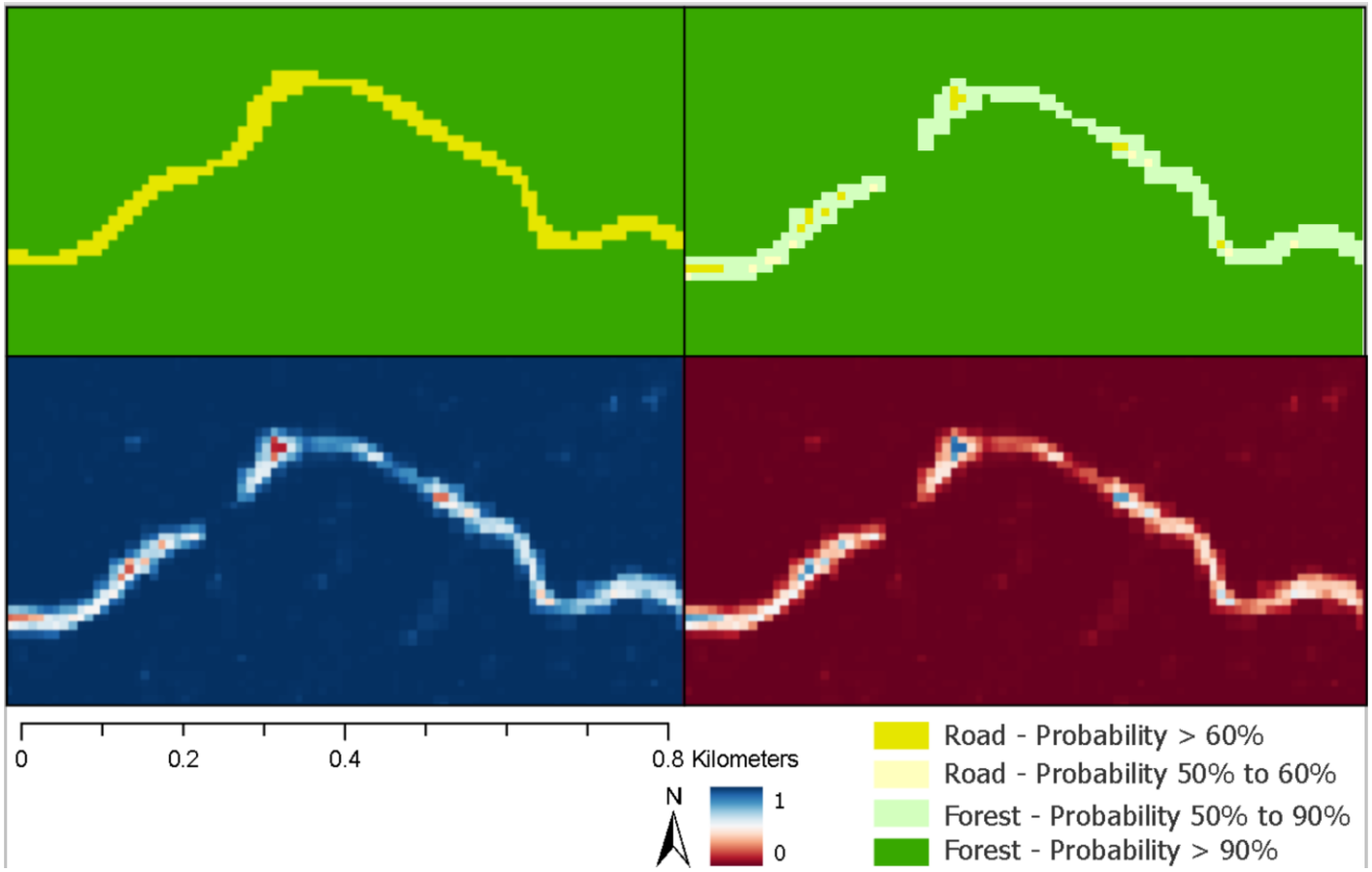


Table 51: Confusion matrix when classifying a road.

Accuracy: 98.09% Pixel Count: 22 932		True		Consumer's Accuracy
		Forest	Road	
Predicted	Forest	22 485	13	99.94%
	Road	424	10	2.30%
Producer's Accuracy		98.15%	43.48%	

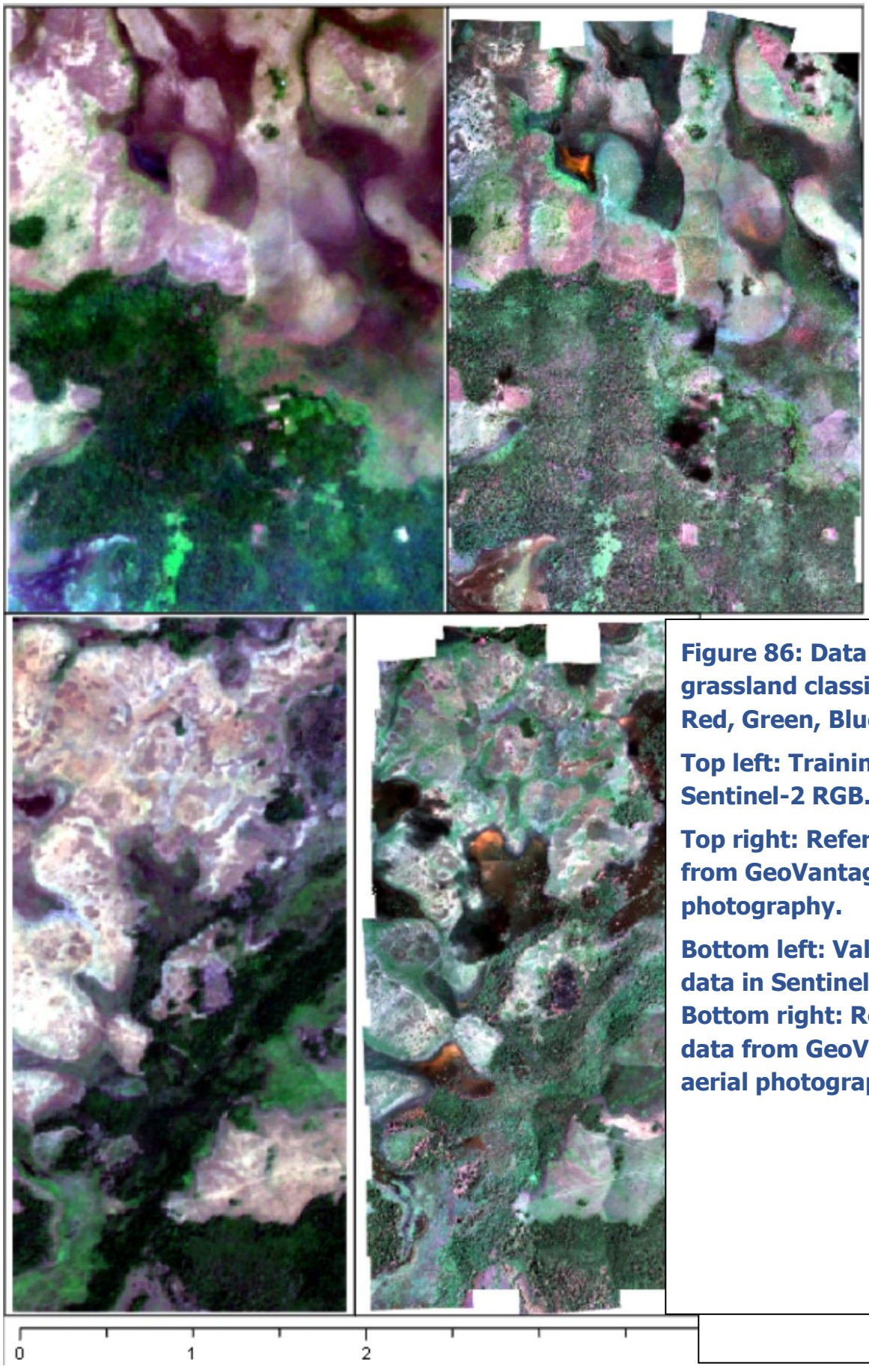


Figure 86: Data for grassland classification in Red, Green, Blue.

Top left: Training data in Sentinel-2 RGB.

Top right: Reference data from GeoVantage aerial photography.

Bottom left: Validation data in Sentinel-2 RGB.
Bottom right: Reference data from GeoVantage aerial photography.

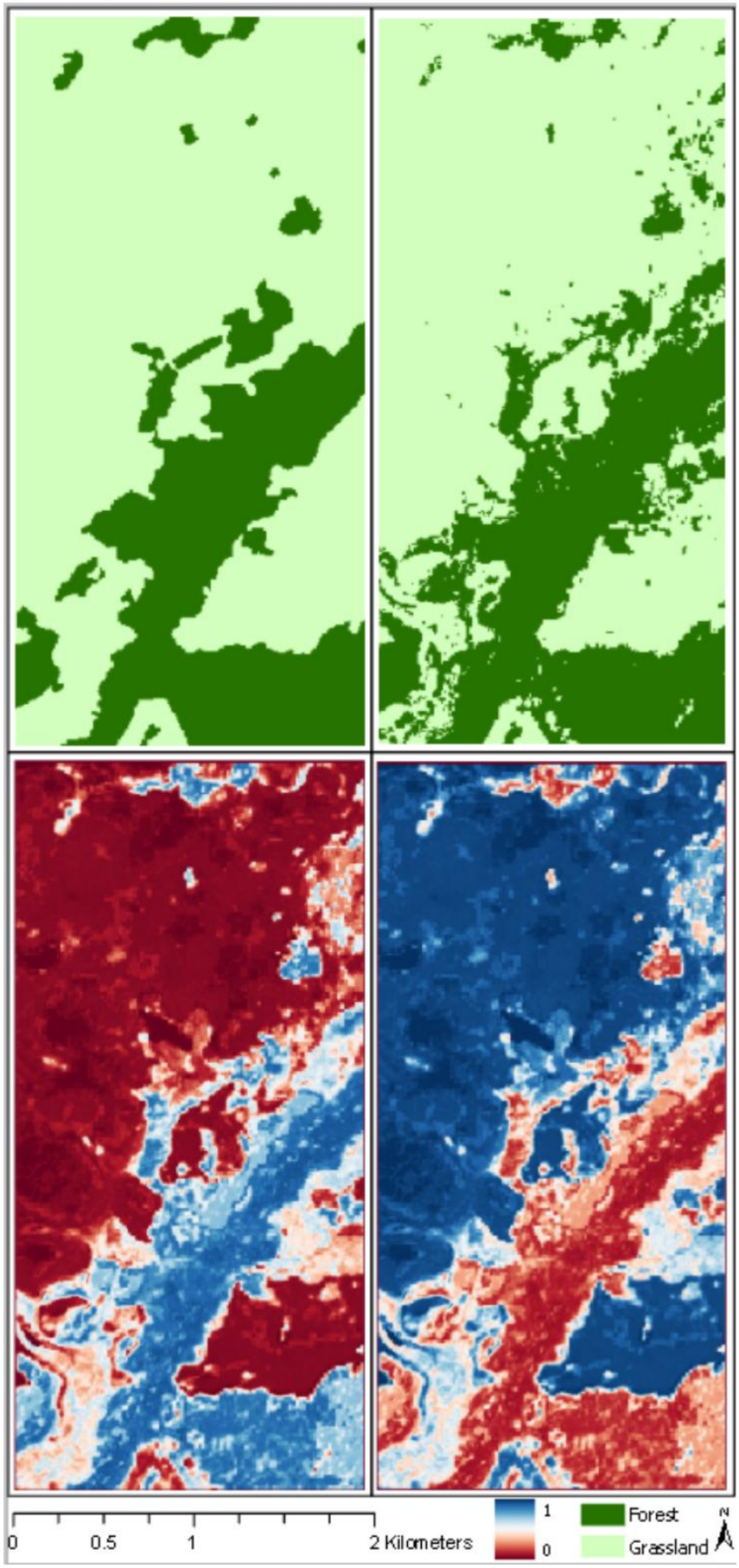


Figure 87: Stacked Sentinel-1 + Sentinel-2 (as in Chapter 7) classified using Random Forest. Location has forest and active mining with no other land cover types.

Top left: True labels.

Top right: Predictions.

Bottom Left: Probability of Forest.

Figure 88: Data for forest, road, settlement and agriculture.

Top left: Training data in Sentinel-2 RGB.

Top right: Reference data from GeoVantage aerial photography.

Bottom left: Validation data in Sentinel-2 RGB.

Bottom right: Reference data from GeoVantage aerial photography.

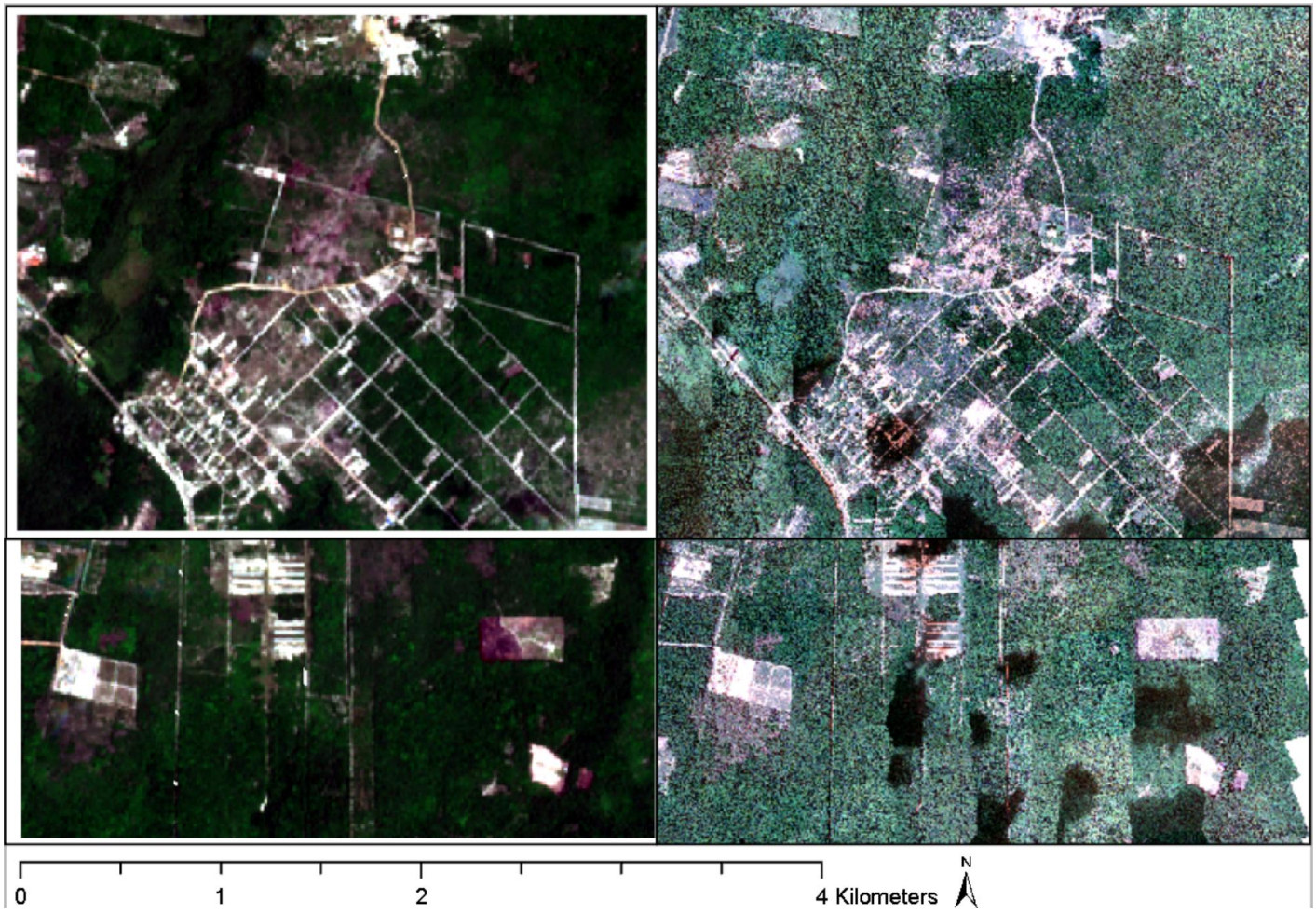


Table 52: Confusion matrix when classifying grassland.

Accuracy: 90.06% Pixel Count: 77 586		True		Consumer's Accuracy
		Forest	Grassland	
Predicted	Forest	20 361	2123	90.56%
	Grassland	5592	49 510	89.85%
Producer's Accuracy		78.45%	95.89%	

Figure 89: Stacked Sentinel-1 + Sentinel-2 (as in Chapter 7) classified using Random Forest. Location has forest and active mining with no other land cover types.

Top left: True labels. Top right: Predictions.

Bottom Left: Probability of Forest. Bottom Right: Probability of Mining.

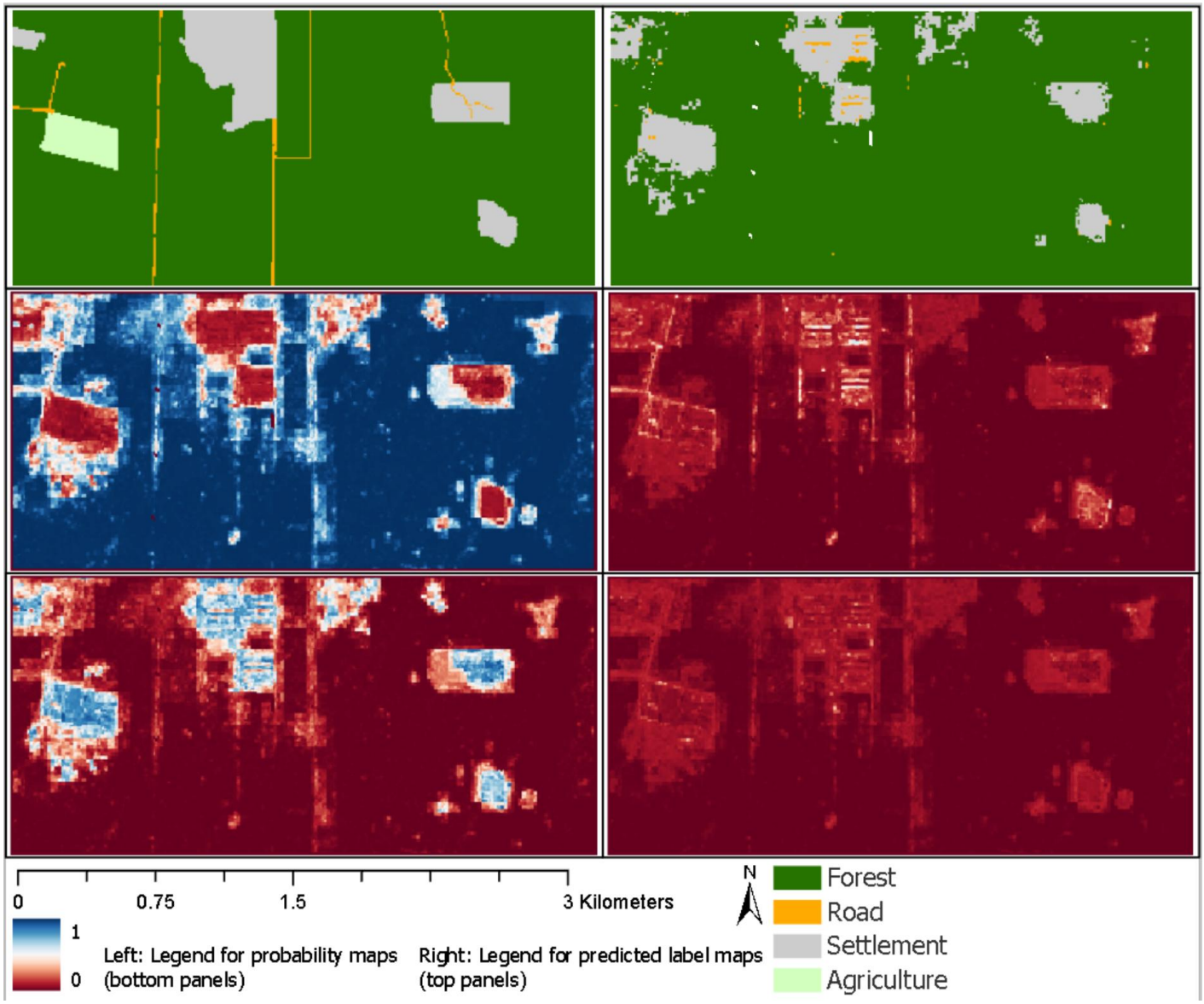


Table 53: Confusion matrix when classifying settlement and agriculture.

Accuracy: 91.46% Pixel Count: 47 574		True				Consumer's Accuracy
		Forest	Road	Settlement	Agriculture	
Predicted	Forest	40 933	27	1254	6	96.95%
	Road	468	9	45	0	1.72%
	Settlement	1227	127	2513	1	64.97%
	Agriculture	19	8	877	0	0.00%
Producer's Accuracy		95.98%	5.26%	53.59%	0.00%	

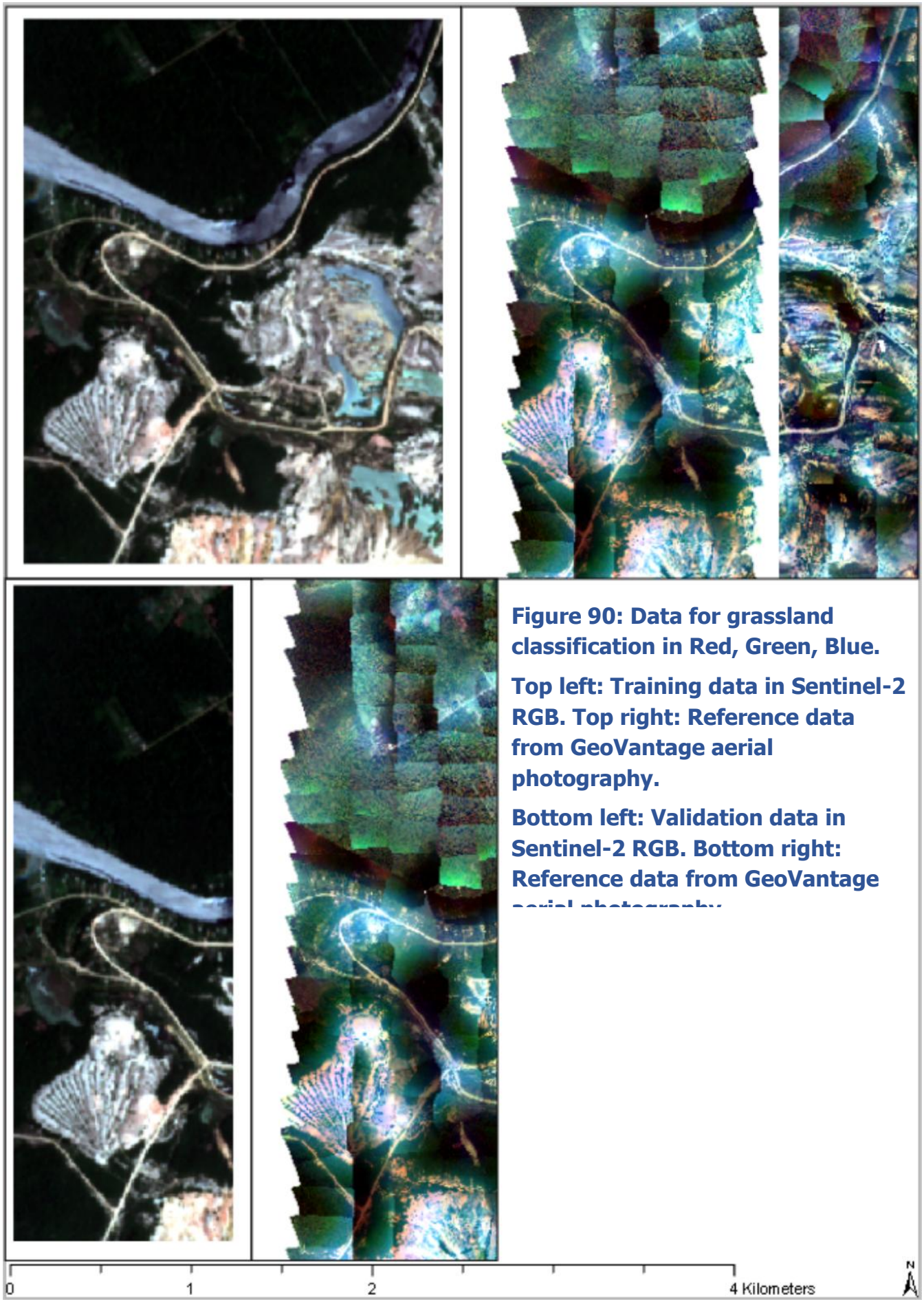


Figure 90: Data for grassland classification in Red, Green, Blue.

Top left: Training data in Sentinel-2 RGB. Top right: Reference data from GeoVantage aerial photography.

Bottom left: Validation data in Sentinel-2 RGB. Bottom right: Reference data from GeoVantage aerial photography

Figure 91: Stacked Sentinel-1 + Sentinel-2 (as in Chapter 7) classified using Random Forest. Location has forest and active mining with no other land cover types.

Top left: True labels. Top right: Predictions.

Bottom Left: Probability of Forest. Bottom Right: Probability of Mining.

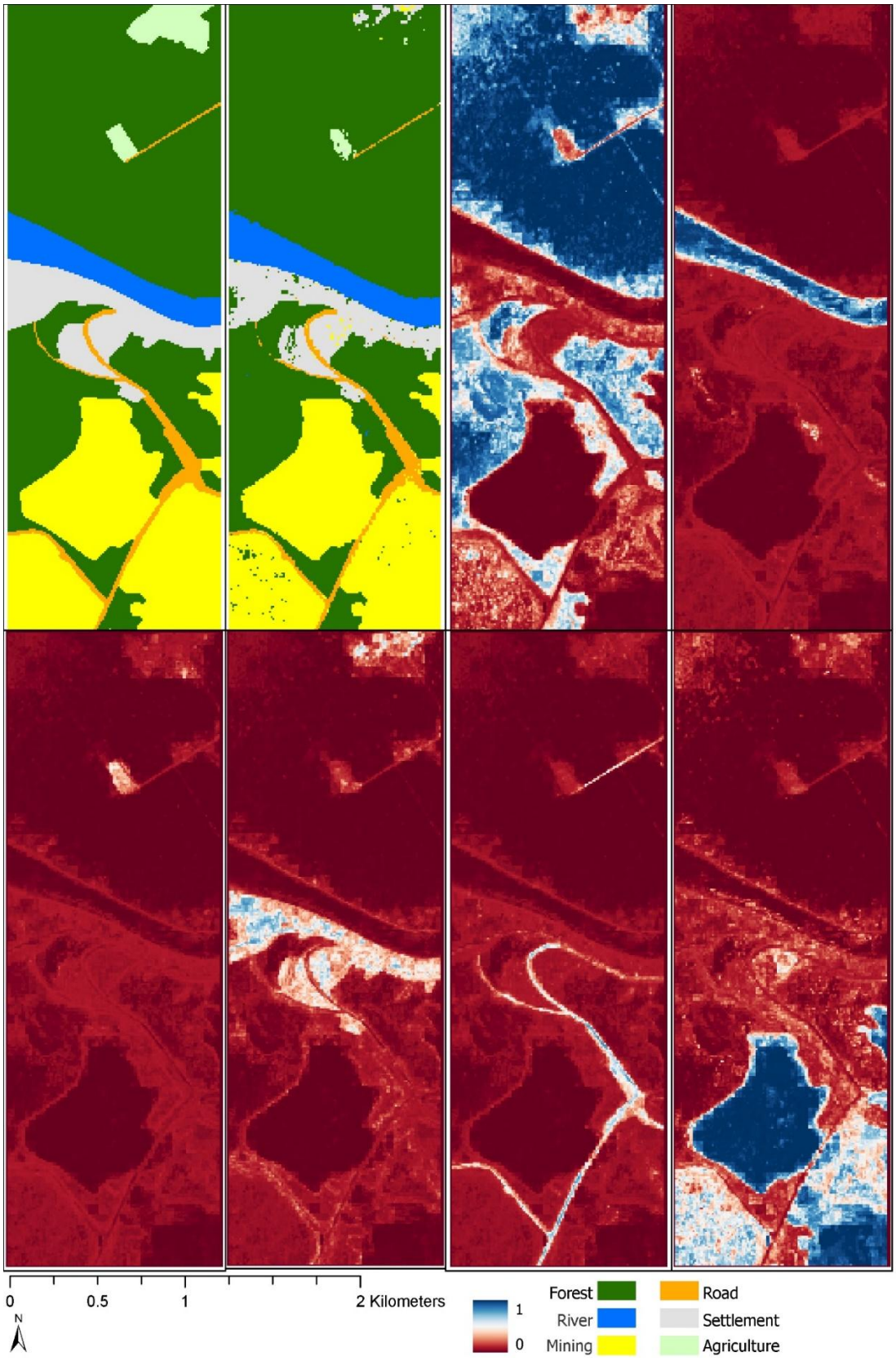


Table 54: Confusion matrix when classifying all classes.

Accuracy: 95.78% Pixel Count: 43 681		True						Consumer's Accuracy
		Forest	River	Mining	Road	Settlement	Agriculture	
Predicted	Forest	24 752	17	17	5	23	1	99.75%
	River	23	2120	0	0	1	0	98.88%
	Mining	141	0	10 583	2	0	0	98.67%
	Road	89	0	100	1176	0	13	85.34%
	Settlement	232	21	49	28	0	3068	90.29%
	Agriculture	864	0	26	1	192	137	11.23%
Producer's Accuracy		94.83%	98.24%	98.22%	97.03%	93.05%	99.28%	

CHAPTER 9: OVERALL DISCUSSION AND CONCLUSION

SECTION 9.1 CHAPTER REVIEW AND TYING IT ALL TOGETHER

Having finished presenting the results in detail in Chapters 4-8, this chapter will start by summarizing the results (Section 9.1), and then it will discuss the issues the results present and their place in the broader literature (Sections 9.2 to 9.6) before moving on to the conclusion (Section 9.7) and some comments about changes in technology over the last five years (Section 9.8). As stated at the end of Chapter 2, the overall aim of the study is to identify the classifiers, datasets and reference (labeled) data sets that can produce the accuracy needed to automate REDD+ measurement and monitoring. This was broken down into specific research questions:

1. How does land cover type mapping accuracy change with the type of classifier? This can start with single classifiers.
2. How does land cover type mapping accuracy change when using remote sensing data from different satellite sensors and combinations of sensors? This can start with individual sensors, and then move on to data fusion.
3. How does the complexity of the land cover type affect accuracy? This can be tested by first classifying two land cover types, then three and four land cover types.

The first question is the subject of Chapter 4. There, it was found that of all the classifiers tested, Random Forest gave the best overall performance using a location that represented forest, mining and river (water) land covers. When using Planetscope imagery from the first generation Dove Classic sensor, overall accuracies were 96% for Random Forest, 95% for Gradient Boosted Trees, and 80% for Naïve Bayes. The predicted land cover map for Random Forest most closely followed the true class boundaries, with fewer errors inside classes. Random Forest was able to accurately outline class boundaries and improve on class labels, and it was noticed that Enhanced Vegetation Index and clustering were useful aids in defining certain class boundaries. This indicates that EVI can assist in label creation by showing the outline of most areas of forest.

Chapters 5 and 6 address the first part of the second question, a comparison of multispectral and SAR satellite sensors most commonly used in remote sensing of forests. Where Chapter 4 compared multispectral sensors only, Chapter 5 tested the Random Forest classification algorithm on image data from seven multispectral and SAR satellite sensors at a location typical of large surface mines located on the Cuyuni River, in north-central Guyana. Overall, the

multispectral sensors outperformed the synthetic aperture radar (SAR) sensors. Classifications with Planetscope (with the Dove Classic sensor), Sentinel-2 and Landsat 8 data yielded overall accuracies from 81% to 90%, and the SAR satellites Sentinel-1 and ALOS-2 yielded accuracies of 69% and 77%. The ability to identify roads in SAR images is closely tied to pixel size. The two sensors that perform best classifying of mining sites are Sentinel-2 and Landsat 8. The class probabilities show a gradient that follows the class boundary closely, even when the predictions don't. This suggests that classification accuracies could be improved by tuning the probability threshold for predicting the class of a pixel.

In Chapter 6, a Random Forest classifier was tested at a riverside mine on the Konawaruk River, using data from seven satellites, a mix of multispectral and SAR sensors. SAR sensor classification accuracy ranges from 23% (Radarsat-2 in the hilly area) to 81% (ALOS-2 in both areas). Multispectral sensor classification accuracy was 85% to 94% for Planetscope (with the Dove Classic sensor) and Sentinel-2. When separating forest from river, classification using both multispectral and SAR sensors performed well, and also performed well at separating mining from river. With SAR data, Consumer's Accuracy ranged from 53% to 88%, and in multispectral, Consumer's Accuracy ranged from 53% to 98%. SAR sensors were much worse than multispectral sensors at separating forest from mining (23% to 81% Consumer's Accuracy for SAR sensors, 85%+ Consumer's Accuracy for multispectral sensors).

The second part of the second question, about combining multispectral and synthetic aperture radar (SAR) satellite images, is the subject of Chapter 7. In Chapter 7, two methods of combining data were tested, so that a Random Forest model can use the information from both sensor types at once. A dataset that fuses data from both Planetscope (multispectral) and TerraSAR-X (SAR) to classify forest had the same Consumer's Accuracy as using Planetscope alone (96.70% in both cases). When classifying forest with fused Landsat and ALOS-2 data, the Consumer's Accuracy went from 67% to 96%, but when classifying mining sites, the Consumer's Accuracy went from 63% to 47%. Sensor fusion does not improve classification accuracy over multispectral data, and both multispectral data and fused multispectral-SAR data outperform SAR data alone.

The third question, about the effect of the complexity of land cover types on accuracy, is the subject of Chapter 8. In that chapter, Random Forest with Sentinel-2 data was tested at a variety of locations, representing the typical land cover types associated with to tropical deforestation. Small rivers, close in width to the ten meter pixels used in Sentinel-1 and

Sentinel-2 data, are identified as forest in predictions, but visible in the class probability image, as the probability of forest values are consistently lower in the pixels labelled as rivers. Roads are largely classified as forest in predictions, but like small rivers, they are visible in the class probability image as a drop in the probability of forest values where the road is located. Tuning the probability threshold for prediction decisions results in very accurate predictions. When separating both settlements and agriculture, the classifier identifies forest with high accuracy (Consumer's Accuracy 95%), finds no agriculture pixels (Consumer's Accuracy 0%), and finds almost no road pixels (Consumer's Accuracy 1.72%).

These results show that variation within the same land cover type is an important consideration when using machine learning to map and measure deforestation extent. When a model is trained on spectrally homogeneous examples of a land cover type, it performs well on similar examples and poorly on spectrally different examples. Spectral variation within a class can also lead to overlap between classes. The results also provide detailed information on the spatial distribution of prediction errors and class probabilities within specific classes:

1. In forests, the classification errors often follow the spectral variation within the forest – darker or lighter parts of the canopy that spectrally overlap with another class.
2. In mining sites, errors follow the complex features within the mining site, which can include bare surfaces, new vegetation, and water bodies/ponds that can have many colours due to variability in depth and sediment content.
3. In roads, most of the pixels are erroneously predicted as forest, though the road is clearly visible in the class probability map ouputs. This problem can be solved by tuning the probability threshold for making predictions.

The problems in accurately classifying those three land cover types (forests, mining, roads) are about: 1) the complex nature of some land cover types, 2) the importance of spatial information, and 3) the importance of class probabilities to accurate prediction. Mines are a clear example of a complex land cover type – they have many different objects within them that have very different spectral (multispectral) and SAR backscatter properties, and these objects are often confused with other land cover types. Mining ponds can be confused with rivers, and the vegetation growing inside a mine (especially a recently abandoned mine) looks like intact forest. Rivers provide an example of complexity of a different sort, in that the sediment load varies across a country, and this also results in great spectral variation within the same land

cover type. Settlements are also a complex land cover type, with many objects within them (roads, buildings, grass, trees, bare ground).

The second and third points (about mining and roads) both highlight the importance of spatial information and high-resolution reference data to effective accuracy assessment, in their own ways (discussed further in Section 9.2). In this study, the understanding of the causes of classification errors comes from the spatial distribution and physical context of errors. This information is not provided by confusion matrices and the accuracy metrics derived from them, but from maps of predicted classes and class probabilities (discussed in detail in Section 9.3). In general, class probabilities provide information about uncertainty in its physical context. In the case of roads, the class probabilities reflect mixed pixels in such a way that tuning the class probabilities will result in accurate class predictions in situations that yield very low accuracy with the default probability threshold for predicting classes.

This understanding of the physical context and causes of classification errors depends on very detailed information about the study area in each test. This detailed information comes from comparing the predictions and class probabilities with high spatial resolution reference data. The end of Section 9.3 discusses the role of reference data in accuracy assessment.

The implication for practice is that a nationwide deforestation measurement strategy that uses machine learning must be designed for a heterogeneous dataset where there is large within-class variation, and where one model may not fit all locations. The clearest example is given early in Chapter 8, where a model trained on clear rivers performed poorly on sedimented rivers, and vice versa. Once the spectral variation between sites of the same type has been addressed, then detailed algorithm and sensor tests can be done, in a way that will allow a useful automated workflow to be developed.

SECTION 9.2 SPATIAL INFORMATION

Three issues were raised in Section 9.1, that will be discussed further in this chapter. They are:

- The heterogeneous nature of some of the land cover types.
- The challenge of the most difficult land cover types, and the limits of machine learning.
- The importance of probability threshold tuning to high accuracy.

The unifying thread of all those points is the importance of high-resolution spatial information. It's possible to discuss them because this study reported both confusion matrices and predicted land cover maps, and because every study site had very high resolution reference data to aid in accuracy assessment. All of these will be discussed in depth as this chapter goes on. First, some discussion of what is meant by spatial information.

In this study, the data points are pixels that represent discrete and spatially contiguous sections of a physical object. Each pixel has a location in space and time, and a context that consists of the physical object at that location, the surrounding objects, and the physical context of machine learning predictions at that location at that point in time. Information on the location and context of prediction errors and high uncertainty allows the reader to relate the model's performance to the physical phenomenon being studied. Maps of predicted land cover types show the spatial distribution and physical context of classification errors, and maps of class probabilities enable similar interpretation of uncertainty, and they allow the reader to understand how the model interacts with the data recording the physical objects being measured.

In this study, this was done by comparing predicted land cover maps and class probability maps with very high resolution remote sensing data, which in this study consisted of aerial photography where the pixels covered a ground area of less than one meter. That aerial photography shows details as small as individual shrubs and vehicles, which allows for maximum precision in creating class labels, and maximum information in evaluating predicted land cover types and class probabilities. The former point, about creating precise class labels, is critically important because accuracy measurement consists of comparing predictions with the correct label, and this is only valid if the correct labels are error-free. Even a small amount of error in the training data can lead to results that overestimate accuracy greatly (Foody, 2012).

SECTION 9.3 CONFUSION MATRICES AND SPATIAL INFORMATION

The standard practice for reporting classification accuracy in remote sensing of forests is a confusion matrix. It provides a numerical breakdown of which true classes were identified as which predicted classes. This can be summarized using a variety of derived metrics, such as overall accuracy, and per-class summaries in Consumer's/User's Accuracy and Producer's Accuracy. Confusion matrices and their associated statistics are very useful, because they indicate which classes the classifier can accurately identify, and which classes it struggles with, and which classes they're being confused with.

The problem with confusion matrices is that they provide no information about the spatial distribution and physical context of errors (a point also made in Comber et al., 2012), which means that they lack a lot of the information on spatial distribution and physical context needed to evaluate the causes of errors, and the usefulness of the predicted maps. This information is provided by a map of predicted land cover types and a map of the probability of each class at each pixel (every results chapter has many examples of these). When these prediction and probability maps are compared with high-resolution reference data for the location, the reader can see how the errors and uncertainty are distributed, and what physical features are associated with errors and high uncertainty.

In this study, Chapter 6 provides an example of the importance of spatial accuracy assessment: two confusion matrices were reported, for different regions in the area of interest. That study area had areas of flat topography and areas of very hilly topography, and SAR imagery responds very differently in flat vs hilly areas. There were separate confusion matrices for the flat and hilly areas. Confusion matrices, as normally used, are a global accuracy metric, but in remote sensing of forests, the number of misclassifications is not randomly or evenly distributed, but localized around phenomena that are difficult to classify in the chosen data-algorithm combination (Foody 2004 makes a similar point). In this case, variation in the landscape was the phenomenon that affected accuracy, and different parts of the landscape were assessed separately. The alternative, one confusion matrix for the whole area, would represent an average of errors that does not reflect the reality in each part of the landscape. This is another example showing why spatial information is necessary to accuracy assessment.

Another example of the benefits of spatial information to accuracy assessment is in Asner et al. (2009). That study tests a fully automated deforestation measurement software package called CLASlite, which classifies forest and non-forest. The results section of that study shows a

predicted land cover map that shows the proportion of each vegetation type at each pixel. It shows gradations in the amount of healthy vegetation across the study area, and it shows vegetation types relative to the river that passes through. There are also maps of the standard error at each pixel, which show high uncertainty near the river and in the areas of dead vegetation. In example from deforestation mapping Gabon (another REDD+ country), Sannier et al. (2014) and (2016) report predicted land cover maps that provide similar information about the patterns in the spatial distribution of deforestation, with similar benefits for understanding the results and the behaviour of the classifier. Together, this study, Sannier et al. (2014) and (2016), and Asner et al. show the value of predicted land cover maps and uncertainty maps in assessing accuracy and the causes of inaccuracy.

Espejo et al. (2018) provides a case study in the importance of uncertainty information in the form of maps of class probabilities. The figures in Chapter 5 to 8 of this dissertation have maps of the class probabilities of mining areas, and consistently find that there is variation in the probability of mining that correlates with the locations of bare surfaces, vegetation, and mining ponds. In Chapter 8, when classifying rivers and roads, it was observed that tuning the probability threshold increased prediction accuracy greatly. This was observed in the map of class probabilities at each pixel, and is not information that can be found in the confusion matrix. Further research can compare the class probabilities within mining sites, to determine whether the class probabilities associated with specific objects fall within thresholds that can be used in the classification workflow. Espejo et al. (2018) reports numerical results and discusses the difficulties of classifying landscape disturbances caused by mining, and the importance of identifying mining methods. Maps of predicted land cover types and maps of class probabilities would have provided information on what physical features are causing problems, which would point the way to a solution.

Joshi et al. (2015) is an example of a study whose results look similar to the results of this study when you only look at the confusion matrices, but important differences become apparent when predicted land cover maps are included. When assessing the predictions using overall accuracy, the SAR results in Chapters 5 and 6 are not too dissimilar from those reported by Joshi et al., which used ALOS data to map deforestation due to mining and reported 67% accuracy. Joshi et al. mention that there is a clear reduction in backscatter in gold mining areas as opposed to intact forest. A close look at the data for the ALOS-2 tests in Chapters 5-6 shows that there is a noticeable change in SAR backscatter in the mining areas versus the forest.

However, the prediction images in the figures in Chapters 5-6 show a detail that is not mentioned in Joshi et al.: there is a lot of variation in backscatter within both forest and mining classes, and the distribution of values overlaps enough that there are a large number of incorrectly classified pixels, resulting in a predicted land cover map where the location of the mining site is clearly visible, but there are many misclassified pixels within the bounds of the mining area. This noise will result in incorrect quantification of forest change, and potentially false positives when identifying deforestation events, and future research needs to quantify this uncertainty.

McRoberts et al. (2016) is a comparison of models for calculating change in forest area in Brazil, using a variety of maps as data. They found that the Global Forest Change map (McRoberts et al. cite Hansen et al. 2013, and Song et al. 2018 reports more recent research on the Global Forest Change data) resulted in the lowest accuracy, at 62% for the lowest forest/non-forest threshold and 66% for the highest threshold. The highest accuracy went to the Atlas 2008 map, with 82% accuracy. It is useful to have information on other options for data sources in identify change. However, the results indicate that, when using Global Forest Change data, a large minority of pixels are misclassified, and when using the Atlas 2008 data, 28% of pixels are misclassified. Where are these errors? The large number suggests that the errors are more than just a few scattered pixels, or a thin line along class boundaries where the pixels are mixed. But because only numerical metrics derived from confusion matrices are presented, there is no spatial information about how the errors are distributed. Are they random noise, are they concentrated around specific physical features, are the errors all along class boundaries? This determines whether these errors are something that a user can work around.

In conclusion, spatial information in the form of predicted land cover maps and maps of uncertainty are essential information in accuracy assessment. When maps of predictions are present, they indicate local variations in accuracy, the physical context of errors, and the impact that the errors would have on the usefulness of the product. When spatial information is not provided, the information available is much more limited, and key details are missing. The following sections will explore questions and issues that can be explored because this study has extensive spatial information.

An important aspect of spatial information is spatial resolution, the size of a pixel or the ground area covered by a pixel. The importance of spatial resolution becomes very clear when looking at the results in Chapters 5 and 6, particularly when comparing the details visible in Planetscope

and Landsat images. In the GeoVantage aerial reference data, the colour and boundaries of individual mining ponds and vegetation in a mining site are clearly visible, making for easy interpretation of the nature of the site and the causes of observed classifier behaviour. In the Landsat data, these important details are not visible in the larger pixels; the only thing that can be seen is that this is a mining site. Previous research on remote sensing of deforestation in Guyana has used this same high-quality reference data, and reported that "Detailed comparison of the Global Forest Change dataset reveals some errors, mainly associated with the 30 m pixel size", an observation that is possible because of the detailed information that high-quality reference data provides (Galiatsatos et al., 2020). This observation about the value of GeoVantage data is also true of other high-resolution data. Pickering et al. (2021) found that the higher resolution data provided information on spatial heterogeneity at small scales, and Pickering et al. (2019) reported that RapidEye data (5 meter pixels) showed thin ribbon-shaped mines not visible in Landsat images.

A non-Guyana example of the importance of high resolution reference data can be seen in Giljum et al. (2022), which measures the extent of mining-caused deforestation, and discusses in depth its environmental and human impact. Their forest data is from Global Forest Watch, and their mine location data is based on coordinates provided by the Metals and Mining Database, provided by Standard and Poors as part of their business intelligence services. The discussion of accuracy notes the need to quantify the indirectly deforested areas around mining, and the need for comprehensive data that can provide information on local circumstances, and methods that allow for insights beyond average effects. The results in this research provide a step in that direction, by assessing in detail the information and errors that come out of high-resolution satellite images of mining sites. Small details of ponds, active mining areas and vegetation regrowth could be seen, especially in higher resolution data. But this study highlights the need for something that Giljum et al. doesn't mention: validation of the forest/non-forest data. The classifications in Chapters 4-8 were not perfect: even the best workflows had at least some errors. Global forest change maps, and maps of deforestation drivers, need to come with uncertainty quantified, which Giljum et al. does not have.

SECTION 9.4 THE HETEROGENEOUS NATURE OF LAND COVER TYPES

Some land cover types are not homogeneous and uniform in the spectral properties of their surfaces. Mines have many surfaces within them (bare ground, ponds, vegetation) that are spectrally very different from each other. Settlements, with their mix of buildings, roads, grass,

trees and bare ground, have the same problem. Rivers vary greatly in their sediment load, and this produces great spectral variation. This section will explore that observation in depth.

The heterogeneous nature of the mining sites has come up once before in the literature. Espejo et al. (2018) states that:

> "While easily identifiable by visual inspection in both Landsat-based and high-resolution satellite data, landscape disturbances caused by ASGM [Artisanal Scale Gold Mining] are difficult to accurately classify using current methodologies on medium scale multispectral satellite data. This is due to the heterogeneous nature of the post-mining landscape, in which large areas of bare ground, water, and remnant or new vegetation are mixed in the same Landsat pixel".

In addition to summarizing their results, the quote above also serves as an excellent summary of the results in this research. They used data fusion consisting of Landsat images and geomorphological data, supplemented with Global Forest Change and CLASLite-derived maps. Chapters 5-8 of this study also used Landsat data, as well as other multispectral sensors, SAR sensors, and fused multispectral-SAR data, and all of them had the problem described in the quotation above.

Espejo et al. go on to say that "mining disturbance is heterogeneous in its spectral signature, but spatially contagious" [sic]. They suggest using CLASLite forest/non-forest maps as a spatial prior, and constraining the analysis to non-forest areas within that. At first glance, this seems like a sound idea. In this study, the model was able to consistently identify change, even when it couldn't identify the driver (such as the all-classes test in Chapter 8).

This study offers another approach: in Chapter 4, it was noticed that the Enhanced Vegetation Index (EVI) was very good at outlining the edge between the forest and non-forest areas. Overall, within carefully chosen study sites, forest/non-forest measurement is a solved problem. The Random Forest classifier used with an Enhanced Vegetation Index will reliably separate forests from non-forest land cover types. Random Forest is also very effective at separating rivers with similar sediment levels, and at classifying mining sites of similar type. This indicates that the problem is not unique to this research, and that there are potential solutions. One is a multi-step classification process, where EVI is used to separate forest from non-forest, and within the non-forest area, mining is classified.

However, there is a problem that is not discussed in Espejo et al.: whether using CLASLite, Global Forest Change, or EVI, the new vegetation that grows in a mining site could easily be classified as forest, something that was observed in the predicted land cover maps in Chapters 4-8. It would not be in the non-forest area that then gets analyzed to detect mining.

Settlements (analyzed in Chapter 8) are another spectrally complex area. In the predicted land cover types, significant portions of the settlement areas were classified as forest, and some areas inside the settlements were classified as roads. This shows that the model is unsuited to accurately measuring the extent of settlements, but it is useful for identifying the location of a settlement. The same is true of the mining results: the model can accurately identify the presence of mining, but due to confusion between forests and vegetation inside the mine, area measurements will be inaccurate.

The extensive literature on remote sensing of settlements offers another solution that was beyond the scope of a study of data fusion, but may be the solution to the problem of classifying both settlements and mines. It has been found that object-based classifications outperform pixel-based classifications when classifying settlements (Stow et al., 2007; Myint et al., 2011; Zhang et al., 2018), and object-based image analysis has been used to solve different problems in Guyana (Persaud et al., 2020). Before using object-based classification, though, this question needs to be answered: In this study, it was found that methods that work well in carefully chosen test areas perform poorly when used in nationwide datasets with high within-class variability. Object-based classification methods would need to be tested first to find out whether they are a solution to that problem, or would be affected by that problem.

SECTION 9.5 CLASSIFYING AGRICULTURE

Regarding the results on agriculture, it should be noted that there was no test of forest vs agriculture, without other classes present. The analysis was constrained to areas that have aerial photography as reference data and have no cloud cover in one of the cloudiest countries on Earth. In addition, mining is the majority driver of deforestation in Guyana, with agriculture representing a minor component. This limits the number of suitable examples that can be used for testing.

In Chapter 8, there were two tests of the model's ability to find agriculture, and the model was able to identify the presence of change, but not the driver of change. There is a wide body of literature on remote sensing of agriculture (for reviews, see Khanal et al., 2020 and Weiss et al,

2020), covering vegetation extent, crop condition, nutrients and water, weed identification, and more. Here, the key issue is that methods that work well in small spatial scale studies that use data that is carefully chosen for simplicity do not scale well to nationwide deforestation measurement. A study that illustrates this well is Doggart et al. (2020). They map deforestation due to agriculture in Tanzania using ALOS-2 data, and an independent accuracy assessment reported 98% agreement between predictions and reference data. At first glance, this indicates a much higher accuracy than the results in Chapter 8 of this study. Doggart et al. excluded areas identified as wetlands, water or flooded cropland in the 2010 NAFORMA Land Use Land Cover map, and within that, intact forest and deforestation were separated using a threshold. There are two differences between Doggart et al. and this study that are worth noting:

- Doggart et al. is a simpler classification problem, where other complex land cover types are not included.
- Ancillary GIS data (land cover data) is being used to narrow the range of data that will be sent to the model, simplifying the data to be classified, and eliminating potential sources of confusion.

Therefore, classification accuracy can be increased by simplifying the data and the labels. This works when testing machine learning within the context of a small, carefully chosen area, with little within-class variation. In the context of country-wide quantification of all deforestation drivers, that must include all the variation that occurs within land cover types. Using existing land cover maps and other government GIS data has one problem: it only contains deforestation events already known to the government, and remote sensing has value in detecting recent and illegal activity that has not already been recorded. The results in Doggart et al. reflect a context where the problem can be simplified easily, and the results in Chapter 8 better reflect the challenge of machine learning in the more complex context of nationwide REDD+ measurement.

SECTION 9.6 TUNING THE PROBABILITY THRESHOLD

In Chapter 8, the accuracy of the model when identifying roads depended heavily on the probability threshold used to make predictions. This section will explain probability threshold tuning and its implications for classifying mixed pixels.

The output of most classification algorithms is class probabilities. In the case of decision trees, the class probability represents the percentage of pixels in a node that belong to that class. In

an ensemble of decision trees, the class probability is the average probability across all trees in the ensemble. By default, classes are predicted by selecting the highest probability class. In a binary classification problem, this is the class that has greater than 50% probability of membership.

In Chapter 8, when using the default probability threshold for predictions, almost all road pixels were identified as Forest. This is because the probability of forest was as high as 90% in most pixels that were labelled as road. However, the distributions of probability of forest values for forest and road were distinct. When the probability threshold was moved to 90% probability of forest, then pixels with a 10% probability of road or lower were predicted to be road, and the prediction accuracy was very high.

There has been other research on using probability threshold tuning to improve accuracy in remote sensing of forests. Botelho Jr. et al. (2022) mapped roads in the Brazilian Amazon using Sentinel-2 data and a Convolutional Neural Network. They used a probability threshold of 0.2 to decide which class each pixel belongs to, and high prediction accuracy resulted. This lends credence to the observation that probability threshold tuning is important for accurate road detection.

Tuning the probability threshold has been found to improve classification accuracy in other contexts as well. Li et al. (2023) found that, in agricultural remote sensing, tuning the probability threshold improved classification accuracy. Komarov and Buehner (2018) increased ice and open water detection accuracy by tuning the probability threshold. When they used a class probability threshold of 0.95, the fraction of misclassified ice pixels decreased from 0.98 to 0.24, and the fraction of misclassified water pixels decreased from 0.35 to 0.09. Bangira et al. (2019) found success in using threshold tuning to accurately classify optically complex water bodies, which could be applicable to the heterogeneous classes problem discussed earlier. Hancock et al. (2022) applied threshold tuning to imbalanced data to improve classification accuracy in that situation.

The results of this study and others show that this has the potential to become standard practice in remote sensing, including developing a multiclass tuning implementation. There are two things to note about implementation.

First, the ability to tune the probability threshold for binary classifications has been implemented in Scikit-Learn version 1.5 (released in May 2024). Scikit-Learn does not have a

version for tuning multiclass probability thresholds. One option is in Wu et al. (2018), which used a Convolutional Neural Network with an adaptive threshold that maximizes the Jaccard Similarity Score. Wang et al. (2020) tuned the probability threshold of a Convolutional Neural Network by optimizing the accuracy score.

Second, this study shows a simpler approach to mapping roads than previous research. Where Botelho Jr. et al. used a computationally expensive Convolutional Neural Network, this study was able to classify roads with high accuracy using Sentinel-2 data and the relatively computationally inexpensive Random Forest algorithm available in Google Earth Engine.

SECTION 9.7 KEY OUTCOMES

The most important outcome is that country-wide deforestation mapping requires a strategy for the variation in data from one location to the next. The same classifier with the same data classifying the same land cover type yielded very different results. Rivers vary in their sediment composition (Chapter 8), and mining sites vary in their vegetation cover, bare surfaces and mining ponds (Chapters 4-7). Section 9.6 ended with the point that high classification accuracy often occurs in situations that are made simple by being restricted to a small area, and nationwide deforestation needs to be designed for the complexity of nationwide data. This means that classification at carefully chosen specific sites is easy in most cases, with greater than 98% accuracy in some cases, but it prevents accurate wall-to-wall classification with a single model.

Location is one variable among several interacting variables, and the others include sensor and the classification algorithm. Multispectral sensors consistently outperform SAR sensors at identifying forest and mining (Chapters 5-6), though both multispectral and SAR tend to be very accurate at identifying rivers (Chapters 4-8), especially when an optimal probability threshold is chosen (Chapter 8). When multispectral and SAR backscatter data were combined (Chapter 7), the result is that sensor fusion in this context did not produce the improved results that a lot of remote sensing literature reports.

However, even after selecting the optimal sensor and classifier in one site, the performance varied when the same sensor/classifier combination was used across study sites (Chapter 8). Separating forest from river using a wide and clear river produced near-perfect accuracy, while training on sedimented river and testing on clear river produced near-complete inaccuracy. There was also variation in classification outcomes between different types of mining sites.

Other outcomes include some options for improving classification accuracy. The Enhanced Vegetation Index follows the Forest/Non-forest boundary very closely, which opens up the possibility of automatic generation of accurate labels. It was also found that the class probabilities accurately outline class boundaries even when they're invisible in predictions. This occurs when mixed pixels are an issue. This indicates that prediction accuracy can be increased in mixed pixel situations by adjusting the probability threshold.

The way forward is therefore to study the performance of classifiers and sensors in different environments, and come to understand how performance varies, and what can be done to classify large and heterogeneous areas. There is now a commitment to ending global deforestation and forest degradation by 2030 (five years after the submission of this dissertation), but deforestation will continue to occur, if not legally then illegally. It is vitally important that all countries have the ability to identify illegal deforestation events quickly and at low cost, so that they can respond to them right away. This research takes a large step toward that by identifying the problem that has prevented machine learning from automating deforestation measurement so far.

The next and final section will comment on upcoming Earth observation satellite missions and the data that they will provide to follow-up research.

SECTION 9.8 A NOTE ABOUT SENSORS GOING FORWARD

Planetscope SuperDove is now available. In late 2019, almost all of the data available was from the Dove Classic sensor, with some Dove R data just beginning to be released. The revisit time is near daily and improves as new cubesats are added to the constellation. Planet provides free pan-tropical basemaps, and provides free data to participating governments via the Norway Climate and Forests Initiative. The trade-off is that the radiometric consistency is lower than Sentinel-2 (Tu et al., 2022), and the geolocation quality is inconsistent (Semple et al., 2023). Planet is also expanding its very high resolution capabilities with the new Pelican sensor, which records 30 cm pixels. Planet is also expanding into space-based hyperspectral remote sensing with the Tanager mission, which offers 5 nm continuous bands covering the visible light, near infrared and shortwave infrared spectra, and it has announced its own automatic multispectral-radar data fusion platform with Sentinel-1.

Global **E**cosystem **D**ynamics **I**nvestigation (GEDI) Lidar data has come and gone, and come again. It began collecting data in March 2019, and was stowed in March 2023, though the

return of the GEDI mission has been confirmed from 2024 to 2030. At first glance, space-based Lidar appears to be tremendously useful in Guyana, where a nationwide airborne or drone-based Lidar survey would be prohibitively expensive. However, due to the length of time needed to collect enough returns for data of sufficient coverage and resolution, and the fact that the sensor has been stowed, GEDI data is unsuitable for detection of yearly change going forward, and its usefulness in historical research is very limited.

Sentinel-1A failed, leaving just one functioning Sentinel-1 SAR satellite instead of two. In 2019, Sentinel-1 was providing adequate coverage of Guyana for testing and supplementary use to multispectral data. Until the new satellites Sentinel-1C and Sentinel-1D are launched, Sentinel-1 coverage is insufficient for use in deforestation measurement, and the amount of data available for testing is limited.

This may change in 2025, when the Biomass and NISAR satellites launch. The Biomass mission will provide forest biomass and carbon stock data that may be highly applicable to deforestation quantification, and NISAR will provide quad-polarized L-band and S-band SAR data. These longer wavelength radar sensors will provide deep canopy penetration, to the ground in some cases, providing information on forest structure and biomass that will complement the shorter wavelength radar sensors that reflect off of the canopy. Biomass and NISAR are ESA and NASA-ISRO missions, respectively, which means that they will be made available at not cost, and will likely be available through Google Earth Engine. Minh et al. (2021) report results from Gabon that indicate that the vertical structure information of P-band SAR increases classification accuracy by 33%. Santos et al. (2000), da Costa Freitas (2008) and Rosa et al. (2017) have all reported benefits from using airborne P-band SAR to map deforestation in the Brazilian Amazon. In addition, the Sentinel-1C satellite was launched on 5th December 2024. This opens up the possibility of exploring the potential for a new kind of sensor fusion: combining SAR data at four wavelengths and four penetrating depths: C-band Sentinel-1, S-band and L-band NISAR data, and P-band Biomass data.

REFERENCES

Adam, Elhadi; Onisimo Mutanga; John Odindi; Elfatih M. Abdel-Rahman

2012 Land-use/cover classification in a heterogeneous coastal landscape using RapidEye imagery: evaluating the performance of random forest and support vector machines classifiers. *International Journal of Remote Sensing* 35(10): 3440-3458

Amani, Meisam; Arsalan Ghorbanian; Seyed Ali Ahmadi; Mohammad Kakooei; Armin Moghimi; S. Mohammad Mirmazloumi

2020 Google Earth Engine Cloud Computing Platform for Remote Sensing Big Data Applications: A Comprehensive Review. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 13: 5326-5350

ART

2021 ART Approves TREES Documents Submitted by Guyana. Last accessed 18th June 2025.

2022 ART Issues World's First Jurisdictional Forstry TREES Carbon Credits to Guyana. Last accessed 18th June 2025.

Asner, Gregory Paul; David E. Knapp; Aravindh Balaji; Guayana Paez-Acosta

2009 Automated mapping of tropical deforestation and forest degradation: CLASlite. *Journal of Applied Remote Sensing* 3(1)

Asner, Gregory Paul; William Llactayo; Raul Tupayachi; Ernesto Ráez Luna

2013 Elevated rates of gold mining in the Amazon revealedthrough high-resolution monitoring. *Proceedings of the National Academy of Sciences* 110(46): 18454-18459

Atkinson, P. M.; A. R. L. Tatnall

2010 Introduction Neural networks in remote sensing. *International Journal of Remote Sensing* 18(4): 699-709

Aznar-Sánchez, José A.; Luis J. Belmonte-Ureña; María J. López-Serrano; Juan F. Velasco-Muñoz

2018 Forest Ecosystem Services: An Analysis of Worldwide Research. *Forests* 9(8), 453

Bangira, Tsitsi; Silva Maria Alfieri; Massimo Menenti; Adriaan van Niekerk

2019 Comparing Thresholding with Machine Learning Classifiers for Mapping Complex Water *Remote Sensing* 11(11): 1351

Belgiu, Mariana; Lucian Drăguţ

2016 Random forest in remote sensing: A review of applications and future directions. *ISPRS Journal of Photogrammetry and Remote Sensing*

Bergstra, James and Yoshua Bengio

2012 Random Search for Hyper-Parameter Optimization. *Journal of Machine Learning Research* 13: 281-305

Bholanath, P.; K. Cort

2015 National Scale Monitoring, Reporting and Verification of Deforestation and Forest. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* XL-7/W3

Bonan, Gordon B.

2008 Forests and Climate Change: Forcings, Feedbacks, and the Climate Benefits of Forests. *Science* 320: 1444

Botelho Jr., Jonas; Stefany C. P. Costa; Julia G. Ribeiro; Carlos M. Souza, Jr.

2022 Mapping Roads in the Brazilian Amazon with Artificial Intelligence and Sentinel-2. *Remote Sensing* 14(15): 3625

Bovolo, Isabella; Thomas Wagner; Geoff Parkin; David Hein-Griggs; Ryan Pereira; Richard Jones

2018 The Guiana Shield rainforests—overlooked guardians of South American climate. *Environmental Research Letters* 13(7)

Brandon, Katrina

2014 Ecosystem Services from Tropical Forests: Review of Current Science. *Centre for Global Development Working Paper No. 380*

Breiman, Leo

1996 Bagging Predictors. *Machine Learning* 24: 123-140

1999 Pasting Small Votes for Classification in Large Databases and On-Line. *Machine Learning* 36: 85-103

2001 Random Forest. *Machine Learning* 45:5-32

Briem, G. J.; J. A. Benediktsson; J. R. Sveinsson

2022 Multiple classifiers applied to multisource remote sensing data. *IEEE Transactions on Geoscience and Remote Sensing* 40(10): 2291-2299

Broich, M.; Stephen V. Stehman, Matthew C. Hansen, Peter Potapov, Yosio E. Shimabukuro

2009. A comparison of sampling designs for estimating deforestation from Landsat imagery: A case study of the Brazilian Legal Amazon. *Remote Sensing of Environment*, 113(11): 2448-2454.

Cannon, Robert L.; Jitendra V. Dave; James C. Bezdek; Mohan M. Trivedi

1986 Segmentation of a Thematic Mapper Image Using the Fuzzy c-Means Clusterng Algorthm. *IEEE Transactions on Geoscience and Remote Sensing* 24(3): 400-408

Chen, Lin; Yeqiao Wang; Chunying Rena; Bai Zhang; Zongming Wang

2019 Assessment of multi-wavelength SAR and multispectral instrument data for forest aboveground biomass mapping using random forest kriging. *Forest Ecology and Management* 447: 12-25

Clements, F. E.

1905 Research methods in ecology. University Publishing Company, Lincoln, Nebraska

da Costa Freitas, Corina; Luciana de Souza Soler; Sidnei João Siqueira Sant'Anna; Luciano Vieira Dutra; João Roberto dos Santos; José Claudio Mura; Antonio Henrique Correia

2008 Land Use and Land Cover Mapping in the Brazilian Amazon Using Polarimetric Airborne P-Band SAR Data. IEEE Transactions on Geoscience and Remote Sensing 46(10): 2956-2970

Doggart, Nike; Theron Morgan-Brown; Emmanuel Lyimo; Boniface Mbilinyi; Charles K. Meshack; Susannah M. Sallu; Dominick V. Spracklen

2020 Agriculture is the main driver of deforestation in Tanzania. *Environmental Research Letters* 15

Douzas, Georgios; Fernando Bacao; Joao Fonseca; Manvel Khudinyan

2019 Imbalanced Learning in Land Cover Classification: Improving Minority Classes' Prediction Accuracy Using the Geometric SMOTE Algorithm. *Remote Sensing* 11(24): 3040

Espejo, Jorge Caballero; Max Messinger; Francisco Román-Dañobeytia; Cesar Ascorra; Luis E. Fernandez; Miles Silman

2018 Deforestation and Forest Degradation Due to Gold Mining in the Peruvian Amazon: A 34-Year Perspective. *Remote Sensing* 10(12): 1903

Ferreira, Matheus Pinheiro; Fabien Hubert Wagner; Luiz E.O.C. Aragão; Yosio Edemir Shimabukuro; Carlos Roberto de Souza Filho

2019 Tree species classification in tropical forests using visible to shortwave infrared WorldView-3 images and texture analysis. *ISPRS Journal of Photogrammetry and Remote Sensing* 149: 119-131

Foody, Giles M.

1992 A fuzzy sets approach to the representation of vegetation continua from remotely sensed data : an example from lowland health. *Photogrammetric engineering and remote sensing* 58(2): 221-225

1994 Fuzzy modelling of vegetation from remotely sensed imagery. *Ecological Modelling* 85: 3-12

2009 The impact of imperfect ground reference data on the accuracy of land cover change estimation. *International Journal of Remote Sensing* 30(12): 3275-3281

2010 Assessing the accuracy of land cover change with imperfect ground reference data. *Remote Sensing of Environment* 114(10): 2271-2285

2011 Impacts of imperfect reference data on the apparent accuracy of species presence–absence models and their predictions. *Global Ecology and Biogeography* 20(3): 498-508

2013 Ground reference data error and the mis-estimation of the area of land cover change as a function of its abundance. *Remote Sensing Letters* 4(8): 783-792

Foody, Giles M.; Mahesh Pal; Duccio Rocchini; Carol X. Garzon-Lopez; Lucy Bastin

2016 The Sensitivity of Mapping Methods to Reference Data Quality: Training Supervised Image Classifications with Imperfect Reference Data. *International Journal of Geoinformation* 5(11): 199

Forest Carbon Partnership

2023 FCPF Country Participants. Last accessed 18th June 2025.

Fortin, Marie-Josée; Pierre Drapeau

1995 Delineation of Ecological Boundaries: Comparison of Approaches and Significance Tests. *Oikos* 72(3): 323-332

Fortin, Marie-Josée; R. J. Olson; S. Ferson; L. Iverson; C. Hunsaker, G. Edwards; D. Levine; K. Butera and V. Klemas

2000 Issues related to the detection of boundaries. *Landscape Ecology* 15: 453-466

Freund, Yoav; Robert E. Schapire

1997 A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences* 55(1):119-139

Friedman, Jerome H.

1999 Stochastic Gradient Boosting. *Computational Statistics & Data Analysis* 38(4):367-378

Galiatsatos, Nikolaos; Daniel N. M. Donoghue; Pete Watt; Pradeepa Bholanath; Jeffrey Pickering; Matthew C. Hansen; Abu R. J. Mahmood

2020 An Assessment of Global Forest Change Datasets for National Forest Monitoring and Reporting *Remote Sensing* 12(11): 1790

Gebhardt, Steffen

n.d. MAD-MEX: Landsat Classification – Methodology. Last accessed 18th June 2025.

Gebhardt, Steffen; Thilo Wehrmann; Miguel Angel Muñoz Ruiz; Pedro Maeda; Jesse Bishop; Matthias Schramm; Rene Kopeinig; Oliver Cartus; Josef Kellndorfer; Rainer Ressl; Lucio Andrés Santos; Michael Schmidt

2014 MAD-MEX: Automatic Wall-to-Wall Land Cover Monitoring for the Mexican REDD-MRV Program Using All Landsat Data. *Remote Sensing* 6(5): 3923-3943

Gentine, Pierre; Julia K Green; Marceau Guérin; Vincent Humphrey; Sonia I Seneviratne; Yao Zhang; Sha Zhou

2019 Coupling between the terrestrial carbon and water cycles—a review. *Environmental Research Letters* 14

Georganos, Stefanos, Tais Grippa, Assane Niang Gadiaga, Catherine Linard, Moritz Lennert, Sabine Vanhuysse, Nicholus Mboga, Eléonore Wolff, Stamatis Kalogirou

2019 Geographical random forests: a spatial extension of the random forest algorithm to address spatial heterogeneity in remote sensing and population modelling. *Geocarto International* 36(2): 121-136

Ghassemian, Hassan

2016 A review of remote sensing image fusion methods. *Information Fusion* 32: 75-89

Giljum, Stefan, Victor Maus, Nikolas Kuschnig, Sebastian Luckeneder, Michael Tost, Laura J Sonter, Anthony J Bebbington

2022 A pantropical assessment of deforestation caused by industrial mining. *Proceedings of the National Academy of Sciences* 119(38)

Gislason, Pall Oskar, Jon Atli Benediktsson, Johannes R. Sveinsson

2006 Random Forests for land cover classification. *Pattern Recognition Letters* 27(4): 294-300

GOFC-GOLD

2016 A sourcebook of methods and procedures for monitoring and reporting anthropogenic greenhouse gas emissions and removals associated with deforestation, gains and losses of carbon stocks in forests remaining forests, and forestation. GOFC-GOLD Report version COP22-1, (GOFC-GOLD Land Cover Project Office, Wageningen University, The Netherlands)

Gopal, S.; C. Woodcock

1996 Remote sensing of forest change using artificial neural networks. *IEEE Transactions on Geoscience and Remote Sensing* 34(2): 398-404

Graves, Sarah J.; Gregory P. Asner; Roberta E. Martin; Christopher B. Anderson; Matthew S. Colgan; Leila Kalantari; Stephanie A. Bohlman

2016 Tree Species Abundance Predictions in a Tropical Agricultural Landscape with a Supervised Classification Model and Imbalanced Data. *Remote Sensing* 8(2): 161

Grohnfeldt, Claas; Michael Schmitt; Xiaoxiang Zhu

2018 A Conditional Generative Adversarial Network to Fuse Sar And Multispectral Optical Data For Cloud Removal From Sentinel-2 Images. 2018 IEEE International Geoscience and Remote Sensing Symposium, 22-27 July 2018, Valencia, Spain

Guyana Forestry Commission

2019 Guyana REDD+ Monitoring Reporting and Verification Systems (MRVS) Year 8 Summary Report.

2020 Guyana MRV Support – Mid Term Evaluation, Final Report

Hancock, John; Justin M. Johnson; Taghi M. Khoshgoftaar

2022 A Comparative Approach to Threshold Optimization for Classifying Imbalanced Data *2022 IEEE 8th International Conference on Collaboration and Internet Computing (CIC)* Atlanta: 14-16 December 2022

Hansen, Matthew C.; Peter V. Potapov; Rebecca Moore; Matt Hancher; Svetlana A. Turubanova; Alexandra Tyukavina; David Thau; Stephen V. Stehman; Scott J. Goetz; Thomas R. Loveland; Anil Kommareddy; Alexey Egorov; Louise Chini; Christopher O. Justice; John R. G. Townshend

2013 High-resolution global maps of 21st-century forest cover change. *Science* 342(6160): 850-853

He, Haibo; Edwardo A. Garcia

2009 Learning from Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering* 21(9): 1263-1284

Healey, Sean P.; Warren B. Cohen; Zhiqiang Yang; C. Kenneth Brewer; Evan B. Brooks; Noel Gorelick; Alexander J. Hernandez; Chengquan Huang; M. Joseph Hughes; Robert E. Kennedy; Thomas R. Loveland; Gretchen G. Moisen; Todd A. Schroeder; Stephen V. Stehman; James E. Vogelmann; Curtis E. Woodcock; Limin Yang; Zhe Zhu

2018 Mapping forest change using stacked generalization: An ensemble approach. *Remote Sensing of Environment* 204: 717-728

Heinimann, Andreas; Ole Mertz; Steve Frolking; Andreas Egelund Christensen; Kaspar Hurni; Fernando Sedano; Louise Parsons Chini; Ritvik Sahajpal; Matthew Hansen; George Hurtt

2017 A global view of shifting cultivation: Recent, current, and future extent. *PLOS One*

Hermosilla, Txomin; Michael A. Wulder; Joanne C. White; Nicholas C. Coops and Geordie W. Hobart

2015a An integrated Landsat time series protocol for change detection and generation of annual gap-free surface reflectance composites. *Remote Sensing of Environment* 158: 220-234

2015b Regional detection, characterization, and attribution of annual forest change from 1984 to 2012 using Landsat-derived time-series metrics. *Remote Sensing of Environment* 170: 121-132

Hirschmugl, Manuela; Carina Sobe; Janik Deutscher; Mathias Schardt

2018 Combined Use of Optical and Synthetic Aperture Radar Data for REDD+ Applications in Malawi. *Land* 7(4): 116

Holmberg, Andreas

2021 Combination analysis of multispectral and radar satellite data. Master's thesis, Department of Physics, Umeå University.

Houghton, R.

2012 Carbon emissions and the drivers of deforestation and forest degradation in the tropics. *Current Opinion in Environmental Sustainability* 4(6): 597-603

Houghton, Richard A. and Alexander A. Nassikas

2017 Negative emissions from stopping deforestation and forest degradation, globally. *Global Change Biology* 24(1): 350-359

Houghton, Richard A; Jo I House; Julia Pongratz; Guido R Van Der Werf; Ruth S Defries; Mathew C Hansen; Corinne Le Quéré; Navin Ramankutty

2012 Carbon emissions from land use and land-cover change. *Biogeosciences* 9(12): 5125-5142

Huang, C.; L. S. Davis; J. R. G. Townshend

2010 An assessment of support vector machines for land cover classification. *International Journal of Remote Sensing* 23(4): 725-749

Inkpen, Rob; Graham Wilson

2013 *Science, Philosophy and Physical Geography*. 2nd ed. Routledge: London.

Intergovernmental Panel on Climate Change

2015 Carbon and Other Biogeochemical Cycles. Chapter 6 in the Fifth Assessment Report: The Physical Science Basis.

James, Gareth; Daniela Witten; Trevor Hastie; Robert Tibshirani

2023 An Introduction to Statistical Learning with Applications in R. Springer, New York.

Jensen, J. R.; Fang Qiu; Minhe Ji

1998 Predictive modelling of coniferous forest age using statistical and artificial neural network approaches applied to remote sensor data. *International Journal of Remote Sensing* 20(14): 2805-2822

Johnston, C. A.; J. Bonde

1989 Quantitative analysis of ecotones using a geographic information system. *Photogrammetric Engineering and Remote Sensing*, 55: 1643-1647

Joshi, Netti; Edward T. A. Mitchard; Natalia Woo; Jorge Torres; Julian Moll-Rocek; Andrea Ehammer; Murray Collins; Martin R. Jepsen; Rasmus Fensholt

2015 Mapping dynamics of deforestation and forest degradation in tropical forests using radar satellite data. *Environmental Research Letters* 10(3)

Kattenborn, Teja; Jens Leitloff; Felix Schiefer; Stefan Hinz

2021 Review on Convolutional Neural Networks (CNN) in vegetation remote sensing. *ISPRS Journal of Photogrammetry and Remote Sensing* 173: 24-49

Kennedy, R. E.; Z. Yang; W. B. Cohen

2010 Detecting trends in forest disturbance and recovery using yearly Landsat time series: 1. LandTrendr—temporal segmentation algorithms. *Remote Sensing of Environment* 114: 2897-2910

Khanal, Sami; Kushal KC; John P. Fulton; Scott Shearer; Erdal Ozkan

2020 Remote Sensing in Agriculture—Accomplishments, Limitations, and Opportunities *Remote Sensing* 12(22): 3783

Kokaly, Raymond F; Gregory P Asner; Scott V Ollinger; Mary E Martin; Carol A Wessman

2009 Characterizing canopy biochemistry from imaging spectroscopy and its application to ecosystem studies. *Remote sensing of environment* 113: S78-S91

Kulkarni, Samadhan C.; Priti P. Rege

2020 Pixel level fusion techniques for SAR and optical images: A review. *Information Fusion* 59: 13-29.

Kuplich, Tatiana A.

2006 Classifying regenerating forest stages in Amazônia using remotely sensed images and a neural network. *Forest Ecology and Management* 234:1-3, 1-9

Kustiyo, O. Roswintiarti; A. Tjahjaningsiha; R. Dewantia; S. Furbyb; J. Wallace

2015 Annual Forest Monitoring As Part of the Indonesia's National Carbon Monitoring System. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* XL-7/W3

Lawrence, Rick L.; Christopher J. Moran

2015 The AmericaView classification methods accuracy comparison project: A rigorous approach for model selection. *Remote Sensing of Environment* 170: 115-120

LEAF Coalition

n.d. Tackling Deforestation and Building Sustainable Futures. Last Accessed 18th June 2025.

Leckie, D. G.

1990 Synergism of Synthetic Aperture Radar and Visible/Infrared Data for Forest Type Discrimination. *Photogrammetric Engineering and Remote Sensing* 56(9): 1237-1246

Li, Haijun; Xiao-Peng Song; Matthew C. Hansen; Inbal Becker-Reshef; Bernard Adusei; Jeffrey Pickering; Li Wang; Lei Wang; Zhengyang Lin; Viviana Zalles; Peter Potapov; Stephen V. Stehman; Chris Justice

2023 Development of a 10-m resolution maize and soybean map over China: Matching satellite-based crop classification with sample-based area estimation. *Remote Sensing of Environment* 294

Linhui, Li; Jing Weipeng; Wang Huihui

1990 Extracting the Forest Type From Remote Sensing Images by Random Forest. *IEEE Sensors Journal* 21(16): 17447-17454

Lippitt, Christopher D.; Rogan, John; Li, Zhe; Eastman, J. Ronald; Jones, Trevor G.

2008 Mapping Selective Logging in Mixed Deciduous Forest. *Photogrammetric Engineering & Remote Sensing* 10: 1201-1211

Louppe, Gilles; Pierre Guerts

2012 Ensembles on Random Patches. Machine Learning and Knowledge Discovery in Databases 346-361

Lynch, Jim; Mark Maslin; Heiko Balzter; Martin Sweeting

2013 Choose satellites to monitor deforestation. *Nature* 496: 293–294

Ma, Yue, et al.

2020 Remote Sensing of Turbidity for Lakes in Northeast China Using Sentinel-2 Images With Machine Learning Algorithms. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 14: 9132-9146

Malhi, Yadvinder and John Grace

2000 Tropical Forests and Atmospheric Carbon Dioxide. *TREE* 15(8): 332-337.

Mackey, Brendan; Cyril F. Kormos; Heather Keith; William R. Moomaw; Richard A. Houghton; Russell A. Mittermeier; David Hole; Sonia Hugh

2020 Understanding the importance of primary tropical forest protection as a mitigation strategy. Mitigation and Adaptation Strategies for Global Change, 25: 763-787

Mas, Jean-François; Stéphane Couturier; Jaime Paneque-Gálvez; Margaret Skutsch; Azucena Pérez-Vega; Miguel Angel Castillo-Santiago; Gerardo Bocco

2016 Comment on Gebhardt et al. MAD-MEX: Automatic Wall-to-Wall Land Cover Monitoring for the Mexican REDD-MRV Program Using All Landsat Data. *Remote Sensing* 8(7): 533

Maxwell, Aaron E.; M. P. Strager; T. A. Warner; N. P. Zegre; C. B. Yuill

2013 Comparison of NAIP orthophotography and RapidEye satellite imagery for mapping of mining and mine reclamation. *GIScience & Remote Sensing* 51(3): 301-320

Maxwell, Aaron E.; Timothy A. Warner; Michael P. Strager; Mahesh Pal

2014 Combining RapidEye Satellite Imagery and Lidar for Mapping of Mining and Mine Reclamation. *Photogrammetric Engineering & Remote Sensing* 2: 179-189

Maxwell, Aaron E.; T.A. Warner; M.P. Strager; J.F. Conley; A.L. Sharp

2015 Assessing machine-learning algorithms and image- and lidar-derived variables for GEOBIA classification of mining and mine reclamation. *International Journal of Remote Sensing* 36(4): 954-978

Maxwell, Aaron E.; Timothy A. Warner; Fang Fang

2018 Implementation of machine-learning classification in remote sensing: an applied review. *International Journal of Remote Sensing* 39(9): 2784-2817

Mhatre, Apurva; Navin Kumar Mudaliar; Mahadevan Narayanan; Aaditya Gurav; Ajun Nair and Akash Nair

2020 Using Deep Learning on Satellite Images to Identify Deforestation/Afforestation. International Conference On Computational Vision and Bio Inspired Computing 1078-1084

Milodowski, D. T.; E. T. A. Mitchard; M. Williams

2017 Forest loss maps from regional satellite monitoring systematically underestimate deforestation in two rapidly changing parts of the Amazon. *Environmental Research Letters* 12(9)

Minh, Dinh Ho Tong; Yen-Nhi Ngo; Thu Trang Lê

2021 Potential of P-Band SAR Tomography in Forest Type Classification. *Remote Sensing* 13(4): 696

Morley, Peter J.; Daniel N.M. Donoghue; Jan-Chang Chen; Alistair S. Jump

2018 Integrating remote sensing and demography for more efficient and effective assessment of changing mountain forest distribution. *Ecological Informatics* 43: 106-115.

2019 Quantifying structural diversity to better estimate change at mountain forest margins. *Remote Sensing of Environment* 223: 291-306

Mountrakis, Giorgos; Jungho Im; Caesar Ogole

2011 Support vector machines in remote sensing: A review. *ISPRS Journal of Photogrammetry and Remote Sensing* 66(3): 247-259

Myint, Sowe W.; Patricia Gober; Anthony Brazel; Susanne Grossman-Clarke; Qihao Weng

2011 Per-pixel vs. object-based classification of urban land cover extraction using high spatial resolution imagery *Remote Sensing of Environment* 115(5): 1145-1161

Nunes, Leonel J. R.; Catarina I. R. Meireles; Carlos J. Pinto Gomes; Nuno M. C. Almeida Ribeiro

2019 Forest Management and Climate Change Mitigation: A Review on Carbon Cycle Flow Models for the Sustainability of Resources. *Sustainability* 11(19): 5276

Pal, M.

2003 Random forest classifier for remote sensing classification. *International Journal of Remote Sensing* 26(1): 217-222

Pal, Mahesh; Paul M. Mather

2003 An assessment of the effectiveness of decision tree methods for land cover classification. *Remote Sensing of Environment* 86(4): 554-565

Pan, Yude et al.

2011 A Large and Persistent Carbon Sink in the World's Forests. *Science* 333: 988-993

Pandit, Vaibhav R.; R. J. Bhiwani

2015 Image Fusion in Remote Sensing Applications: A Review. *International Journal of Computer Applications* 120(10): 22-32

Persaud, Haimwant; Rehana Thomas; Pradeepa Bholanath; Towana Smartt; Pete Watt

2020 Object-Based Image Analysis Approach to Determine the Fallow Periods for Shifting Cultivation in Indigenous Communities in Guyana *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*

Pickering, Jeffrey; Stephen V. Stehman; Alexandra Tyukavina; Peter Potapov; Pete Watt; Samuel M. Jantz; Pradeepa Bholanath; Matthew C. Hansen

2019 Quantifying the trade-off between cost and precision in estimating area of forest loss and degradation using probability sampling in Guyana. *Remote Sensing of Environment* 221: 122-135

Pickering, Jeffrey; Alexandra Tyukavina; Ahmad Khan; Peter Potapov; Bernard Adusei; Matthew C. Hansen; André Lima

2021 Using Multi-Resolution Satellite Data to Quantify Land Dynamics: Applications of PlanetScope Imagery for Cropland and Tree-Cover Loss Area Estimation *Remote Sensing* 13(11): 2191

Pohl, Christine; J. L. Van Genderen

1995 Image fusion of microwave and optical remote sensing data for topographic map updating in the tropics. Proceedings Volume 2579, Image and Signal Processing for Remote Sensing II

1998 Review article Multisensor image fusion in remote sensing: Concepts, methods and applications. *International Journal of Remote Sensing* 19(5): 823-854

Psistaki, Kyriaki; Georgios Tsantopoulos; Anastasia K. Paschalidou

2024 An Overview of the Role of Forests in Climate Change Mitigation. *Sustainability* 16(14): 6089

Quarmby, N. A.; J. R. G. Townshend; J. J. Settle; K. H. White; M. Milnes; T. L. Hindle

1992 Linear mixture modelling applied to AVHRR data for crop area estimation. *International Journal of Remote Sensing* 13(3): 415-425

Rajasugunasekar, D.; Avdhesh Kumar Patel; Khumanthem Babina Devi; Akhilesh Singh; Panneer Selvan; Anup Chandra

2023 An Integrative Review for the Role of Forests in Combating Climate Change and Promoting Sustainable Development. *International Journal of Environment and Climate Change* 13(11)

Ranson, K. J.; G. Sun; V. I. Kharuk; K. Kovacs

2004 Assessing tundra–taiga boundary with multi-sensor satellite data. *Remote Sensing of Environment* 93(3): 283-295

Reiche, Johannes; Sytze De Bruin; Dirk Hoekman; Jan Verbesselt; Martin Herold

2015 A Bayesian Approach to Combine Landsat and ALOS PALSAR Time Series for Near Real-Time Deforestation Detection. *Remote Sensing* 7(5), 4973-4996.

Reichstein, Markus; Nuno Carvalhais

2019 Aspects of Forest Biomass in the Earth System: Its Role and Major Unknowns. *Surveys in Geophysics* 40: 693-707

Rodriguez-Galiano, V.F.; B. Ghimire; J. Rogan; M. Chica-Olmo; J.P. Rigol-Sanchez

2012 An assessment of the effectiveness of a random forest classifier for land-cover classification. *ISPRS Journal of Photogrammetry and Remote Sensing* 67: 93-104

Rogan, John; Janet Franklin; Doug Stow; Jennifer Miller; Curtis Woodcock; Dar Roberts

2008 Mapping land-cover modifications over large areas: A comparison of machine learning algorithms. *Remote Sensing of Environment* 5:2272-2283

Romijn, Erika; John Herbert Ainembabazi; Arief Wijaya; Martin Herold; Arild Angelsen; Louis Verchot; Daniel Murdiyarso

2013 Exploring different forest definitions and their impact on developing REDD+ reference emission levels: A case study for Indonesia. *Environmental Science & Policy* 33: 249-259

Rosa, Rafael A. S.; David Fernandes; Thiago L. M. Barreto; Christian Wimmer; João B. Nogueira

2017 Deforestation detection in Amazon rainforest with multitemporal X-band and p-band sar images using cross-coherences and superpixels. 2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), 23-28 July 2017

Rosenblatt, F.

1958 The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review* 65(6): 386–408

Sannier, Christophe; Ronald E. McRoberts; Louis-Vincent Fichet; Etienne Massard K. Makaga

2014 Using the regression estimator with Landsat data to estimate proportion forest cover and net proportion deforestation in Gabon. *Remote Sensing of Environment* 151: 138-148

Sannier, Christophe; Ronald E. McRoberts; Louis-Vincent Fichet

2016 Suitability of Global Forest Change data to report forest cover estimates at national level in Gabon. *Remote Sensing of Environment* 173: 326-338

Santos, João R; Corina C. Freitas; Luciana S. Araujo; Luciano V. Dutra; José C. Mura; Fábio F. Gama; Luciana S. Soler; Sidnei J.S. Sant'Anna

2003 Airborne P-band SAR applied to the aboveground biomass studies in the Brazilian tropical rainforest. *Remote Sensing of Environment* 87(4): 482-493

Scarpa, Giuseppe; Massimiliano Gargiulo; Antonio Mazza; Raffaele Gaetano

2018 A CNN-Based Fusion Method for Feature Extraction from Sentinel Data. *Remote Sensing* 10(2): 236

Schaffer, Cullen

1993 Selecting a classification method by cross-validation. *Machine Learning* 13: 135-143

Semple, Alana; Bin Tan; Guoqing Lin

2023 SuperDove Geometric Quality Assessment Summary. National Aeronautics and Space Administration.

Serpico, Sebastiano Bruno; Paolo Pellegretti; Lorenzo Bruzzone

1994 Feature selection for remote-sensing data classification. *Image and Signal Processing for Remote Sensing* 2315

Sesnie, Steven E.; Bryan Finegan; Paul E. Gessler; Sirpa Thessler; Zayra Ramos Bendana; Alistair M. S. Smith

2010 The multispectral separability of Costa Rican rainforest types with support vector machines and Random Forest decision trees. *International Journal of Remote Sensing* 31(11): 2885-2909

Sicre, C. Marais; R. Fieuzal; F. Baup

2020 Contribution of multispectral (optical and radar) satellite images to the classification of agricultural surfaces. *International Journal of Applied Earth Observation and Geoinformation* 84

Skidmore, A.K.; B.J. Turner; W. Brinkhof; E. Knowles

1997 Performance of a Neural Network: Mapping Forests Using GIS and Remotely Sensed Data. *Photogramrnetric Engineering & Remote Sensing* 63(5): 501-514

Song, Xiao-Peng; Matthew C. Hansen; Stephen V. Stehman; Peter V. Potapov; Alexandra Tyukavina; Eric F. Vermote; John R. Townshend

2018 Global land change from 1982 to 2016. *Nature* 560: 639-648

Souza Jr., Carlos; Laurel Firestone; Luciano Moreira Silva; Dar Roberts

2003 Mapping secondary tropical forest and forest age from SPOT HRV data. *International Journal of Remote Sensing* 20(18): 3625-3640.

Souza Jr., Carlos M.; Dar A. Roberts; Mark A. Cochrane

2005 Combining spectral and spatial information to map canopy damage from selective logging and forest fires. *Remote Sensing of Environment* 98(2-3): 329-343.

Stehman, Stephen V.; Giles M. Foody

2019 Key issues in rigorous accuracy assessment of land cover products. *Remote Sensing of Environment* 231

Stow, D.; A. Lopez; C. Lippitt; S. Hinton; J. Weeks

2007 Object - based classification of residential land use within Accra, Ghana based on QuickBird satellite data *International Journal of Remote Sensing* 28(22): 5167-5173

Thomas, I. L.; N. P. Ching; V. M. Benning; J. A. D'Aguanno

1987 A review of multi-channel indices of class separability. *International Journal of Remote Sensing*. 8(3): 331-350

Trisasongko, B. H.

2010 The use of polarimetric SAR data for forest disturbance monitoring. *Sensing and Imaging: An International Journal* 11(1-13)

Tu, Yu-Hsuan; Kasper Johansen; Bruno Aragon; Marcel M. El Hajj; Matthew F. McCabe

2022 The radiometric accuracy of the 8-band multi-spectral surface reflectance from the planet SuperDove constellation. *International Journal of Applied Earth Observation and Geoinformation* 114

UNFCCC

n.d. Bali Road Map Intro. Last accessed 18th June 2025.

UN-REDD Programme Collaborative Online Workspace

2020 ABOUT REDD+. Accessed 24th September 2020.

USGS

n.d. Landsat Enhanced Vegetation Index. Last accessed 18th June 2025.

Van der Werf, G. R.; D. C. Morton; R. S. DeFries; J. G. J. Olivier; P. S. Kasibhatla; R. B. Jackson; G. J. Collatz; J. T. Randerson

2009 CO_2 emissions from forest loss. *Nature Geoscience* 2(11): 737-738

Vieira, Ima Célia G.; Arlete Silvade Almeida; Eric A. Davidson; Thomas A. Stone; Cláudio J. Reis de Carvalho; José Benito Guerrero

2003 Classifying successional forests using Landsat spectral properties and ecological characteristics in eastern Amazônia. *Remote Sensing of Environment* 87(4): 470-481

Wang, Benjamin X.; Nathalie Japkowicz

2009 Boosting support vector machines for imbalanced data sets. *Knowledge and Information Systems* 25: 1-20

Wang, Yunxia; Guy Ziv; Marcos Adami; Edward Mitchard; Sarah A. Batterman; Wolfgang Buermann; Beatriz Schwantes Marimon; Ben Hur Marimon Junior; Simone Matias Reis; Domingos Rodrigues; David Galbraith

2018 Mapping tropical disturbed forests using multi-decadal 30 m optical satellite imagery. *Remote Sensing of Environment* 221: 474-488

Wang, Zhijun; D. Ziou; C. Armenakis; D. Li; Qingquan Li

2005 A comparative analysis of image fusion methods. *IEEE Transactions on Geoscience and Remote Sensing* 43(6): 1391-1402

Weiss, M.; F. Jacob; G. Duveiller

2020 Remote sensing for agricultural applications: A meta-review *Remote Sensing of Environment* 236

Williams, David P.; Vincent Myers; Miranda Schatten Silvious

2009 Mine Classification With Imbalanced Data. *IEEE Geoscience and Remote Sensing Letters* 6(3): 528-532

Wolpert, David H.

1992 Stacked generalization. *Neural Networks* 5(2): 241-259

Wood, T. F.; G. M. Foody

1989 Analysis and representation of vegetation continua from Landsat Thematic Mapper data for lowland heaths. *International Journal of Remote Sensing* 10(1): 181-191

Wu, Zhihuan; Yongming Gao, Lei Li, Junshi Xue, Yuntao Li

2018 Semantic segmentation of high-resolution remote sensing images using fully convolutional network with adaptive threshold. *Connection Science* 31(2)

Yang, Li and Abdallah Shami

2020 On hyperparameter optimization of machine learning algorithms: Theory and practice. *Neurocomputing* 415(20): 295-316

Zhang, Caiyun; Zhixiao Xie

2013 Object-based Vegetation Mapping in the Kissimmee River Watershed Using HyMap Data and Machine Learning Techniques. *Wetland* 33: 233-244

Zhang, Ce; Isabel Sargent; Xin Pan; Huapeng Li; Andy Gardiner; Jonathan Hare; Peter M. Atkinson

2018 An object-based convolutional neural network (OCNN) for urban land use classification *Remote Sensing of Environment* 216: 57-70

Zhang, Yuzhen; Jingjing Liu; Wenjuan Shen

2022 A Review of Ensemble Learning Algorithms Used in Remote Sensing Applications. *Applied Sciences* 12(17): 8654

Zhao, Junfang; Dongsheng Liu; Yujie Zhu; Huiwen Peng; Hongfei Xie

2022 A review of forest carbon cycle models on spatiotemporal scales. *Journal of Cleaner Production* 339

Zhao, Qingxia; Shichuan Yua; Fei Zhao; Linghong Tian; Zhong Zhao

2019 Comparison of machine learning algorithms for forest parameter estimations and application for forest quality assessments. *Forest Ecology and Management* 434: 224-234

Zhou, Zheng; Change Zheng; Xiaodong Liu; Ye Tian; Xiaoyi Chen; Xuexue Chen; Zixun Dong

2023 A Dynamic Effective Class Balanced Approach for Remote Sensing Imagery Semantic Segmentation of Imbalanced Data. *Remote Sensing* 15(7): 1768

APPENDIX A: PLATFORM COMPARISON

The key result of this research is that the core problem of nationwide deforestation mapping is the variety in the data, and the main body of the dissertation argued that point. Along the way, many other observations about classifier and sensor behaviour were made. This appendix will report results on a wide range of classifiers in Python, feature selection, choice of metrics, sample size, and class balance.

COMPARING A BROADER RANGE OF CLASSIFIERS

TRANSITIONING FROM GOOGLE EARTH ENGINE TO PYTHON

So far, all tests have been in Google Earth Engine (GEE). But there are others that can do machine learning on satellite images. The Python programming language is the most common language in data science today, and it is capable of classifying satellite images. The workflow was constructed using these modules:

- RasterIO opens GeoTIFF images and converts them to NumPy arrays.
- Scikit-Learn tunes hyperparameters, classifies the data, and assesses accuracy using a wide range of metrics.
- RasterIO saves predictions and class probabilities as GeoTIFF files for viewing in ArcGIS.

The data is a test of a section of the Cuyuni River (Figure 91), upstream from the Cuyuni River mining site in Chapter 6. The training data and the validation data are from different but adjacent areas, which will ensure that there are no spectral differences in the water.

METHODOLOGY: GOOGLE EARTH ENGINE VS PYTHON

The Python workflow follows these steps:

1. Using the RasterIO module, open GeoTIFF data and label images and convert them to 3D NumPy arrays (width, height, bands).
2. Reshape the data to a 2D array (samples, bands), and reshape the labels to a vector (samples,).
3. Delete no data entries from the array.

4. Tune hyperparameters. Bayes Search Cross-Validation (using the Scikit-Optimize module) is efficient at searching large search spaces. Grid Search Cross Validation can be used for small search spaces.
5. Fit the classifier using tuned hyperparameters.
6. Calculate confusion matrix and associated statistics.
7. Create a predictions array (samples, features) from the validation image.
8. Reshape the predictions image from (samples, 1) to (height, width, 1), and then to (1, height, width), and then save the resulting image using RasterIO.
9. Create class probability arrays (samples, number of classes) from the validation image.
10. Reshape the class probabilities image from (samples, number of classes) to (height, width, number of classes), and then to (number of classes, height, width), and then save the resulting image using RasterIO.

The Google Earth Engine workflow follows these steps:

1. The training and validation data for the classifier is feature collections created from images. At each pixel in the training image, a point feature is created with the pixel's values, and a feature collection of points is made for each pixel in the validation image.
2. The classifier (Random Forest) is trained using the training feature collection.
3. The trained classifier is applied to the validation feature collection, and a confusion matrix is calculated from that.
4. A prediction image and a class probabilities image are made by classifying the validation image.

SUMMARY OF RESULTS

A summary of Consumer's Accuracy and Producer's Accuracy is provided in Table 55.

- CART (GEE) and Decision Tree (Python): In Consumer's Accuracy, Forest is slightly lower, and River is much higher. In Producer's Accuracy, River is lower and Forest is much higher.
- Random Forest: In Consumer's Accuracy, River is lower for Python than GEE. Accuracies are otherwise similar.
- Gradient Boosted Trees: Python accuracy is much higher – in Google Earth Engine, all pixels were assigned to the majority class.

- Naïve Bayes: In Consumer's Accuracy, River is much lower in Python. In Producer's Accuracy, River is much higher in Python.
- Minimum Distance (GEE) and Nearest Centroid (Python): Python accuracy is consistently slightly lower.
- K-Nearest Neighbours: In Consumer's Accuracy, River is higher in Python. In Producer's Accuracy, River is lower in Python.
- Support Vector Machines: Accuracy in Google Earth Engine is high, and very low in Python. The performance of Support Vector Machines is an outlier among the Python classifiers; the others have much higher performance. The specific implementation used here is LIBSVM in both Python and Google Earth Engine. In Python, several hyperparameters were tuned (C, kernel, degree, gamma, $coef_0$), and the same hyperparameters were used in Google Earth Engine. In Table 56, two implementations were tested in Python: LIBLINEAR and LIBSVM. LIBLINEAR showed similar results to the other classifiers.

The Python workflow was built from scratch, and the process is fully understood. The full details of the GEE workflow are not available in the documentation. What is known is that the training data is input as point features, but the documentation does not say whether the classifier converts that to a 1D array (the NumPy approach) or processes point features natively. This is the first question that needs to be answered.

Some of the classifiers are identical in implementation between platforms, but most are not. Support Vector Machine uses the libsvm implementation in both cases. The Minimum Distance classifier in Google Earth Engine does not have a Python equivalent with the same name, so Nearest Centroid was used as the closest equivalent. What is also known is the difference in hyperparameter tuning. The Google Earth Engine implementations have fewer hyperparameters that can be tuned, compared with their Python counterparts. For hyperparameter tuning, Google Earth Engine only allows for manual testing of different hyperparameter values. Python includes Grid Search CV (in Scikit-Learn) and Bayes Search CV (in Scikit-Optimize).

CONCLUSION

The original goal of this research was to test the benefits of sensor fusion for classification accuracy in deforestation mapping in Guyana. It has been found that there are other, more important issues that need to be addressed. The most important problem is that classifier

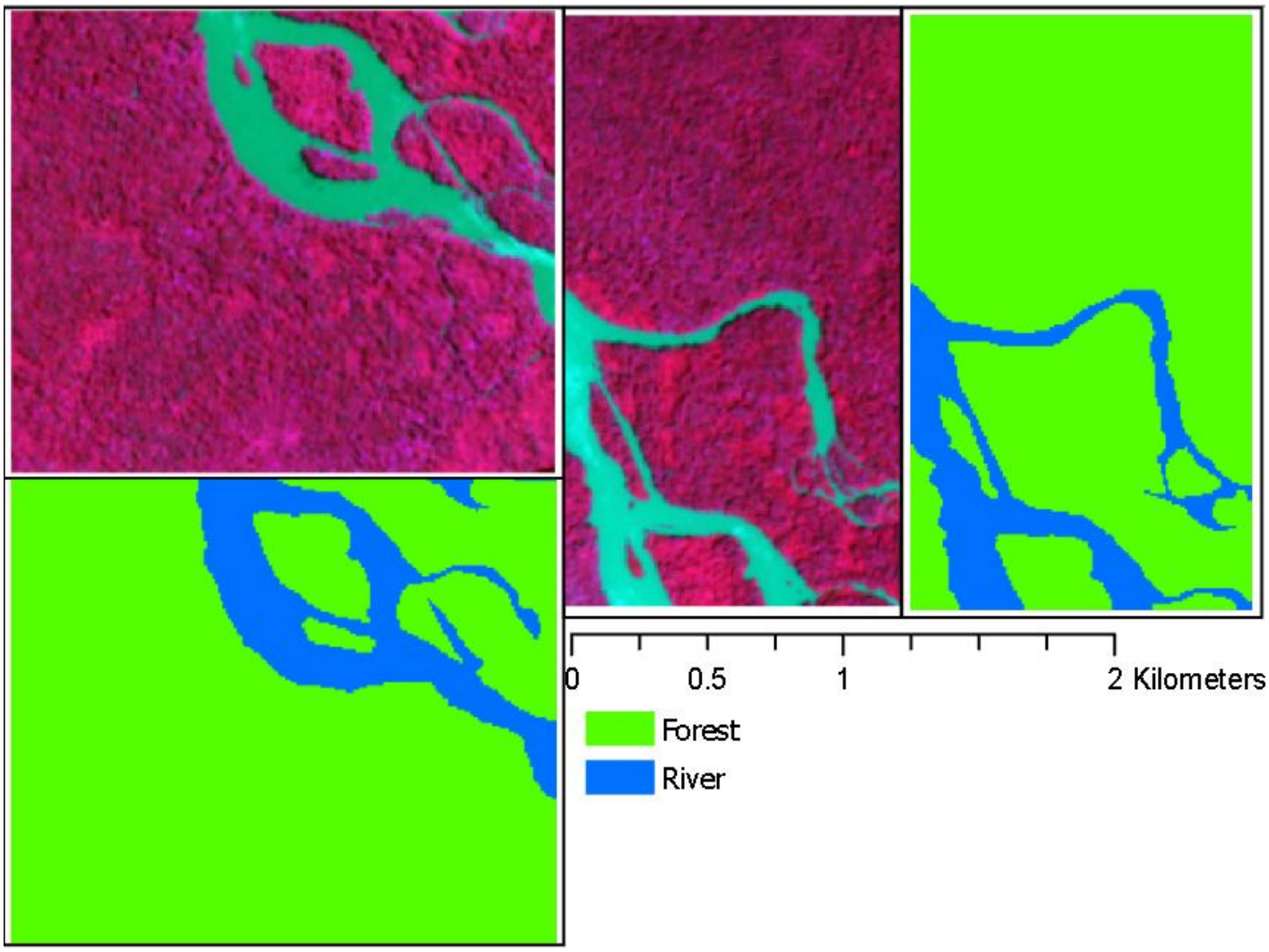
behaviour varies greatly from one location to another. This appendix has also found that the platform also matters – results will change depending on whether you are using Google Earth Engine or Python. Understanding why these differences occur, and how they vary from one study site to the next, will allow a researcher to choose the best platform for the task.

Table 55: Accuracy of near-equivalent classifiers in Google Earth Engine and Python.

		Consumer's Accuracy		Producer's Accuracy	
		Forest	River	Forest	River
GEE	CART	98.16	39.63	76.65	91.43
Python	Decision Tree	96.73	81.24	96.67	81.54
GEE	Random Forest	98.20	88.03	97.96	89.29
Python	Random Forest	98.13	75.48	95.74	87.79
GEE	Gradient Boosted Trees	85.64	0.00	100.00	0.00
Python	Gradient Boosted Trees	99.07	71.49	95.13	93.22
GEE	Naïve Bayes	96.50	93.22	98.92	78.57
Python	Naïve Bayes	98.92	72.69	95.32	92.29
GEE	Minimum Distance	96.50	92.44	98.92	78.57
Python	Nearest Centroid	92.85	86.10	97.41	68.16
GEE	K-Nearest Neighbours	97.60	86.62	97.72	87.86
Python	K-Nearest Neighbours	98.98	71.98	95.21	92.59
GEE	Support Vector Machine (LIBSVM)	98.23	85.83		
Python	Support Vector Machine (LIBSVM)	51.98	5.01	75.47	1.82

Figure 92: The location of the training and validation data in Guyana.

Figure 93: Top left: Training data in Sentinel-2. Bottom left: True labels for training data. Middle: Validation data. Right: True labels for validation data.



CLASSIFIERS SPECIFIC TO SCIKIT-LEARN

Scikit-Learn has a wide range of classifiers, including many not found in Google Earth Engine. While doing this research, other Scikit-Learn classifiers were tested as well. Here are some observations:

- The best performing classifier is Naïve Bayes.
- Soft Voting is the second best classifier. Soft Voting fits each of the base classifiers and calculates the class probabilities. Predictions are made based on the argmax of the probabilities at each pixel. Soft Voting outperforms Hard Voting, which reports the majority predictions at each pixel.
- Radius Neighbours Classifier assigned all pixels to Forest – all Forest probabilities were above 50%.
- The LIBSVM implementation of Support Vector Machines in Python gives very poor performance, and this only happens in Python.
- The Consumer's Accuracy for the other classifiers is no lower than 70% for River (Stacking Classifier with passthrough) and 91% for Forest (Support Vector Machine using the LIBLINEAR implementation). The Producer's Accuracy is no lower than 87% for Forest (Nearest Centroid) and 50% for River (for Passive Aggressive Classifier).

Table 56: Accuracy of Scikit-Learn classifiers, sorted in ascending order of Forest accuracy.

	Consumer's Accuracy		Producer's Accuracy	
	Forest	River	Forest	River
Naïve Bayes	98.92	72.69	95.32	92.29
Soft Voting	97.52	79.32	96.37	85.07
K-Neighbours Classifier	98.98	71.98	95.21	92.59
Multilayer Perceptron	99.16	70.64	95.00	93.74
Stacking Classifiers	99.28	69.96	94.89	94.55
Stacking Classifiers with Passthrough	99.23	70.11	94.91	94.22
Decision Tree	96.73	81.24	96.67	81.54
Hard Voting	96.68	91.34	96.68	81.32
Perceptron	99.63	64.88	94.10	96.88
Nearest Centroid	92.85	86.10	87.41	68.16
Stochastic Gradient Descent	82.25	86.44	97.45	66.47
Logistic Regression	91.69	86.63	97.47	64.97
Linear Support Vector Machine (LIBLINEAR)	91.36	86.85	97.50	61.14
Ridge Classifier	86.58	88.63	97.72	54.01
Passive Aggressive Classifier	93.85	89.40	97.80	49.61
Radius Neighbours Classifier	99.47	0.00	84.86	4.84
Support Vector Machine (LIBSVM)	51.98	5.01	75.47	1.82

Table 57: Accuracy of bagged and boosted ensembles of decision trees.

	Consumer's Accuracy		Producer's Accuracy	
	Forest	River	Forest	River
Decision Tree	96.73	81.24	96.67	81.54
Bagged Decision Tree	98.13	75.48	95.74	87.79
Bagged Extra Randomized Trees	51.73	22.79	79.01	7.75
Random Forest	98.13	75.48	95.74	87.79
Boosted Decision Trees	98.60	75.12	95.70	90.51
Boosted Extra Randomized Trees	45.00	26.55	77.48	7.91
Gradient Boosted Trees	99.07	71.49	95.13	93.22
Histogram Gradient Boosted Trees	43.86	26.81	77.10	8.84

Table 58: Accuracy scores for bagging and boosting of Scikit-Learn classifiers.	Consumer's Accuracy		Producer's Accuracy	
	Forest	River	Forest	River
Naïve Bayes	98.92	72.69	95.32	92.29
Bagged Naïve Bayes	98.97	72.25	95.25	92.59
Boosted Naïve Bayes	98.92	72.69	95.32	92.29
K-Neighbours Classifier	98.98	71.98	95.21	92.59
Bagged K-Neighbours Classifier	98.97	72.30	95.26	92.59
Radius Neighbours Classifier	99.47	0.00	84.86	4.84
Bagged Radius Neighbours Classifier	98.98	71.69	95.16	92.56
Nearest Centroid	92.85	86.10	87.41	68.16
Bagged Nearest Centroid	92.82	86.10	97.41	68.10
Logistic Regression	91.69	86.63	97.47	64.97
Bagged Logistic Regression	91.64	86.71	97.48	64.86
Boosted Logistic Regression	97.55	79.27	96.36	85.21
Passive Aggressive Classifier	93.85	89.40	97.80	49.61
Bagged Passive Aggressive Classifier	97.18	80.27	96.51	93.51
Ridge Classifier	86.58	88.63	97.72	54.01
Bagged Ridge Classifier	95.81	88.89	97.75	52.71
Boosted Ridge Classifier	98.56	75.63	95.79	90.36
Stochastic Gradient Descent	82.25	86.44	97.45	66.47
Bagged Stochastic Gradient Descent	92.03	86.56	97.47	65.89
Boosted Stochastic Gradient Descent	100.00	0.00	84.89	0.00
Perceptron	99.63	64.88	94.10	96.88
Bagged Perceptron	93.75	85.73	97.30	70.83
Boosted Perceptron	78.39	90.72	97.94	42.75
Linear Support Vector Machine	91.36	86.85	97.50	61.14

Bagged Linear Support Vector Machine	91.30	86.90	97.51	63.98
Boosted Linear Support Vector Machine	99.49	67.68	94.54	95.97
Support Vector Machine	51.98	5.01	75.47	1.82
Bagged Support Vector Machine	66.18	92.78	98.10	32.79
Multilayer Perceptron	99.16	70.64	95.00	93.74
Bagged Multilayer Perceptron	98.78	74.00	95.53	91.52

NEURAL NETWORKS IN KERAS

The previous tests used the classifiers found in Scikit-Learn. This section is a preliminary test of three simple deep learning architectures in Keras. The architectures are:

- A multilayer perceptron with three hidden Dense layers and two Dropout layers. The number of units, and the fraction of units dropped, are tuned using Bayesian Optimization.
- A multilayer perceptron with three hidden Dense Flipout layers that use Bayesian weights. These come from the Tensorflow-Probability module. There are two Dropout layers again.
- A Convolutional Neural Network based on the Unet architecture.

The workflow follows these steps:

1. Using RasterIO, open the GeoTIFF data and label files and convert them to NumPy arrays.
2. Using NumPy, expand them from (height, width, bands) to (batches, height, width, bands).
3. Optimize hyperparameters using Bayesian optimization.
4. Fit the final model on the training data for a maximum of 20 000 epochs, with early stopping implemented.
5. Calculate predictions and confusion matrices using the validationd data.

For consistency, the same data was used for the Google Earth Engine test, the Scikit-Learn classifier test, and the Keras deep learning test (Figures 100 and 101).

All neural networks using Keras have 98% Consumer's Accuracy when classifying Forest (Table 60). Consumer's Accuracy for River is much higher for Multilayer Perceptrons with Bayesian weights than regular weights (16% vs 2%) and Convolutional Neural Networks based on the Unet architecture (16% vs 0%).

Table 59: Accuracy for three neural network architectures built using Keras.

	Consumer's Accuracy		Producer's Accuracy	
	Forest	River	Forest	River
Multilayer Perceptron	98.05	1.77	11.32	87.69
Bayesian Multilayer Perceptron	98.26	15.70	99.70	3.11
Convolutional Neural Network	98.21	0.00	100.00	0.00

APPENDIX B: CHOICE OF METRIC FOR COMPARING PERFORMANCE

The goal of this research is to find the workflow that produces the most useful land cover map as a prediction, meaning high accuracy relative to a ground truth. For industry applications, outputs so inaccurate that they have no information need to be discarded. Outputs that do provide information need to be sorted to quickly find the best one. For science, the cause of variation needs to be identified. This section will identify the classification metric that accomplishes that.

- Definition of Accuracy: Correct pixels / All pixels.
- Definition of Precision: True Positives / True Positives + False Positives. In the multiclass case, Recall is calculated for each class and averaged.
- Definition of Recall: True Positives / True Positives + False Negatives. In the multiclass case, Recall is calculated for each class and averaged.
- Definition of Matthews Correlation Coefficient:
- Definition of Jaccard Score: Size of the intersection of two classes divided by the size of the union of two classes. In the multiclass case, this is calculated for each class and averaged.
- Definition of Confusion Matrix: A square matrix of size n x n, where n is equal to the number of classes.
- Definition of Consumer's Accuracy: For each row, divide the number of pixels in the correct column by the total number of pixels in the row.
- Definition of Producer's Accuracy: For each row, divide the number of pixels in the correct row by the total number of pixels in the column.
- Definition of Log Loss: For each pixel, with a true label in the range 0 to 1, the log loss in the binary case is the negative log-likelihood of the classifier. This can be generalized to the multiclass case with a matrix of probability estimates for each label.

There is a substantial amount of literature arguing that the Matthews Correlation Coefficient is the best option. Chicco and Jurman (2020) argue that Matthews Correlation Coefficient should be used instead of Precision and Recall by all scientists because it is not affected by class imbalance or class swapping. Yao and Shepperd (2020) argue that F1 Score should be deprecated in favour of the Matthews Correlation Coefficient. Gonzalez-Ramirez et al. (2021) analyzed seven classification metrics and found that Matthews Correlation Coefficient is the most robust. Jurman et al. (2012) found that "MCC is a good compromise among discriminancy,

consistency and coherent behaviors with varying number of classes, unbalanced datasets, and randomization".

Contrary to those opinions, Foody (2023) argues that Matthews Correlation Coefficient is sensitive to class imbalance and the quality of the reference data, and that it is not necessarily the final word on classification accuracy assessment. MCC can produce a high score on a poor classifier if errors in the reference data correlate with errors in the classifier, and that the quality of the reference data has a major impact on classification metrics (Foody, 2023)

Because of this, the choice of classification metrics cannot be chosen based on what the literature usually recommends. Instead, the objective of this research must be made clear, and the performance of each metric must be assessed based on how well it accomplishes the objectives of this research.

Interpretation of Confusion Matrix: The rows and columns represent the true labels and predicted labels. In data science in general, which one represented which is chosen by the researcher and labelled, and in remote sensing, the convention is to have the rows represent true labels and the columns represent predicted labels. In this research, the remote sensing convention is followed. The diagonal of the matrix is the correct predictions, and the upper and lower triangular matrices around it are the incorrect predictions, with the true class and falsely predicted class indicated.

Usefulness of Confusion Matrices: The goal of this research is to understand the ability of different classifiers and sensors to classify deforestation drivers, and a confusion matrix provides the most detailed class-by-class breakdown of the errors.

Interpretation of Accuracy: In terms of a confusion matrix, the accuracy is the number of pixels in the diagonal divided by the number of pixels in the confusion matrix. Accuracy is strongly influenced by class imbalance. A classifier where the majority class is predicted very accurately, and the minority class is predicted very inaccurately, will have high accuracy because the majority of pixels were classified accurately.

Usefulness of Accuracy: Accuracy is a simple way of summarizing the overall performance of a classifier, and accuracy scores are easy to compare. Accuracy scores do not provide details of which classes the classifier handled well and which it didn't, and accuracy is sensitive to unbalanced data.

Interpretation of Consumer's Accuracy: This indicates how many predictions in each class were correct.

Interpretation of Producer's Accuracy: This indicates how many examples of each class were found.

Usefulness of Producer's and Consumer's Accuracy: These are two ways of summarizing the per-class performance of a classifier, and they can be easily added to a confusion matrix as an extra row and column. In this way, they show the number of predictions by class that were correct, and the number of samples of each class that were found, with the confusion matrix entries that they're summarizing.

Interpretation of Precision: In the binary case, Precision is the percentage of all pixels identified as the positive class that really were the positive class. In the multiclass case, this is the average of that for all classes. A weighted option is available for imbalanced data.

Usefulness of Precision: In the binary case, Precision provides information on the classifier's ability to avoid false positives. In the multiclass case, this is averaged, and so the per-class information provided by Consumer's Accuracy is summarized, but the details are lost. The average can be weighted, to reduce sensitivity to unbalanced data.

Interpretation of Recall: In the binary case, Recall is the percentage of the positive class that was predicted. In the multiclass case, this is the average of that for all classes. A weighted option is available for imbalanced data.

Usefulness of Recall: In the binary case, Recall provides information on the classifier's ability to avoid false negatives. In the multiclass case, this is averaged, and so the per-class information provided by Producer's Accuracy is summarized, but the details are lost. The average can be weighted, to reduce sensitivity to unbalanced data.

Interpretation of Matthews Correlation Coefficient: This is a correlation coefficient where +1 is a perfect classification, 0 is a completely random prediction, and -1 is a perfectly inverse prediction. As the score moves from 0 to 1, is becomes increasingly less random and better than chance. It takes into account all aspects of the confusion matrix.

Usefulness of Matthews Correlation Coefficient: In the binary case, MCC uses all four of True Positives, True Negatives, False Positives and False Negatives. At first glance, this makes it the most complete summary of a confusion matrix, using information found in neither Precision nor

Recall. MCC is useful for comparing the randomness of two classifications – the one that is closer to 0 is the more random of the two. Like Precision and Recall, MCC also loses per-class information.

Interpretation of Jaccard Score: In the binary case, Jaccard Score represents the number of correct predictions as the intersection of the predicted and true labels (the number of correct predictions) and the number of total predictions as the union of predictions and true labels. This makes it an alternative way of calculating accuracy.

Usefulness of Jaccard Score: The difference between Jaccard Score and Accuracy is the calculation of the percentage of correct predictions, and this difference does not provide any new information.

Interpretation of Log Loss: The log loss is a description of the uncertainty of class probabilities, by calculating the distance between the predicted probability and the label (which is given a value of 1).

Usefulness of Log Loss: Log Loss is a useful complement to confusion matrices, because it provides information on the uncertainty of the class probabilities, information that is not found in metrics based on predictions.

In summary, the metrics that provide the necessary information are the standard set in remote sensing: a confusion matrix combined with Consumer's Accuracy, Producer's Accuracy and Accuracy. Log Loss is a useful complement, by providing a summary of uncertainty. The other metrics do not provide any useful information not provided by the metrics described above.

APPENDIX C: FEATURE ENGINEERING

While developing the Python workflow, some observations were made about feature selection. Table 60 below shows accuracy scores when classifying with only one band from Sentinel-1 or Sentinel-2. The range of accuracy scores is very small – from 79.88% (VH polarization) to 81.11% (Red Edge 3). The most important observation is that all bands are nearly equally useful – there is no one band that will produce a terrible classification, and there is no band that will produce a perfect classification, either.

Table 61 below shows accuracy scores using all bands except for the one specified. Here, performance is nearly identical – every accuracy score can be rounded to either 86.34% or 86.35%. Once again, removing features does not make any meaningful difference.

Table 60: Accuracy scores when using only one band from Sentinel-1 and Sentinel-2 data.

Accuracy	Band
0.8111	Red Edge 3
0.8106	Shortwave Infrared 2
0.8101	Coastal
0.8095	VV
0.8081	Green
0.8080	Shortwave Infrared 1
0.8065	Near Infrared
0.8065	Red
0.8040	Red Edge 2
0.8030	Blue
0.8020	Red Edge 1
0.7988	VH

Table 61: Accuracy scores when each band is omitted.

Accuracy	Band Combination
0.86348	Without Shortwave Infrared 2
0.86345	Without Red Edge 2
0.86345	Without Green
0.86342	Without VV
0.86342	Without Shortwave Infrared 1
0.86342	Without Red
0.86342	Without Coastal
0.86342	All Bands
0.86339	Without VH
0.86339	Without Near Infrared
0.86339	Without Red Edge 3
0.86339	Without Red Edge 1
0.86336	Without Blue

APPENDIX D: CLASS BALANCE

Class imbalance is a major issue, and there is no perfect solution, only trade-offs. Preserving the class proportions of the original data will lead to a classifier that favours the majority class, and undersampling gives the classifier a dataset that's different from the prediction data (Salmon et al., 2015).

Class balance is required to train good models (Belgiu and Drăguţ 2016), but class imbalance is a recurring problem in remote sensing. The class imbalance seen here is consistent with previous research (Graves et al., 2016). This is a result of a classifier seeking to minimize an overall error rate (Maxwell et al., 2018).

The remote sensing of forests literature has several approaches to deal with imbalanced data (Douzas et al., 2019):

- Introduce a cost matrix that applies higher misclassification costs for the minority classes.
- Modify the algorithm to reinforce learning of the minority classes. An exampe of this is Infinitely Imbalanced Logistic Regression (Williams et al., 2009), which is designed so that, as the size of the majority class approaches infinity, the coefficients are minimally affected.
- Sampling strategies to rebalance the class distribution, by:
    - Undersampling to reduce the size of the larger classes.
    - Oversampling to create artificial data for the smaller classes.
    - Hybrid approaches that combine oversampling and undersampling.

Examples of sampling strategies include dynamic weighting (Zhou et al., 2023), the Synthetic Minority Oversampling Technique (SMOTE) (Douzas et al., 2019, Mallah et al., 2022),

Undersampling and oversampling come with their own problems (He and Garcia, 2009). Undersampling reduces the number of points sampled in the larger classes, which means that key information about the larger classes may not be included. Oversampling duplicates points in the smaller classes, which can lead to overfitting.

Douzas et al., 2019 tests a variation of SMOTE using land cover classification of Landsat data. The results are reported based on Accuracy, F1 Score, and G-Mean, using Decision Trees,

Random Forest, Gradient Boosted Trees, and K-Nearest Neighbours. Accuracy was always highest when using no oversampling. When using F1 Score, accuracy was always higher when using oversampling, by 2% to 10%.

Maxwell et al., 2018 argues that we should consider the opposite approach: change the classifiers to meet the needs of the data. One of the questions they ask is: "What kind of assumptions will make imbalanced learning algorithms work better compared to learning from the original distributions?". That suggests that there is much potential for further research into designing classifier implementations that are insensitive to class imbalance.

Google Earth Engine offers two sampling options:

- Random sampling, where all pixels are randomly selected.
- Stratified sampling, where a specified number of pixels per class are randomly selected.
- Sampling by region, where pixels within a bounding box are randomly selected.

Scikit-Learn offers these sampling options:

- The simplest approach to sampling is to randomly sample the data so that every pixel has an equal chance of being selected.
- Scikit-Learn also offers the option of random sampling where the proportion of classes in the sample matches the proportion of classes in the original dataset.
- NumPy offers the tools needed to implement undersampling as well.

Tables 62-65 show results from four tests of the effect of class balance. The classifications separate forest from grassland (the same data as in Chapter 8). In Google Earth Engine, the results show that balancing the class proportions produces a small increase in Consumer's Accuracy for Grassland increases from 78% to 79%, and the Producer's Accuracy for Forest decreases from 96% to 84%. Balanced classes here actually have lower performance. In the Scikit-Learn test on the same data, undersampling increased the overall accuracy (72% to 88%), but the Producer's Accuracy for Grassland went from 89% to 13%, and the Consumer's Accuracy for Forest went from 98% to 88%. Despite the increase in overall accuracy, the distribution of errors has changed such that balanced classes do not solve any problems.

Table 62: Confusion matrices using all pixels, with no resampling or weighting, in Google Earth Engine.

Accuracy: 96.07% Pixel Count: 34170		True		Consumer's Accuracy
		Forest	Grassland	
Predicted	Forest	22986	149	99.36%
	Grassland	921	3194	77.62%
Producer's Accuracy		96.15%	95.54%	

Table 63: Stratified Sampling (equal number of pixels sampled per class), in Google Earth Engine.

Accuracy: 89.77% Pixel Count: 9328		True		Consumer's Accuracy
		Forest	Grassland	
Predicted	Forest	4629	35	99.25%
	Grassland	863	3253	79.03%
Producer's Accuracy		84.29%	98.94%	

Table 64: Confusion matrix usng all pixels, with no resampling or weighting, in Python.

Accuracy: 71.76% Pixel Count: 9328		True		Consumer's Accuracy
		Forest	Grassland	
Predicted	Forest	20370	515	97.53%
	Grassland	9136	4149	31.23%
Producer's Accuracy		69.04%	88.96%	

Table 65: Undersampling (equal pixels per class) in Python.

Accuracy: 88.05% Pixel Count: 34170		True		Consumer's Accuracy
		Forest	Grassland	
Predicted	Forest	29477	4053	87.91%
	Grassland	29	611	95.47%
Producer's Accuracy		99.90%	13.10%	

APPENDIX E: SAR CLASSIFICATIONS

This appendix elaborates on the radar comparison in Chapter 5, by testing the effects of these variables:

- Radar sensor: There are four sensors tested, and they are ALOS-2, Sentinel-1, Radarsat-2, and TerraSAR-X.
- Satellite orbit: This can be ascending or descending. These satellites travel in polar orbits, and the orbit listed indicates whether they are heading from the south pole to the north pole, or from the north pole to the south pole. Radar sensors look to the side as they travel, which means that opposite orbits mean opposite look directions. This changes the location of the radar shadows that can affect classification results.
- Model smoothing. Random Forest models can be smoothed by controlling the number of trees. Two values are tested here: 24 trees, or unlimited trees.

The take-aways are:

- Model smoothing makes the results worse. In all cases, better classifications were done with unlimited trees than 24 trees.
- In ALOS-2, changing the look direction changed the predictions, but never produced predictions that are close to the true labels. In TerraSAR-X, changing the look direction does not change the results.
- In Radarsat-2, model smoothing does not change the predictions.

Table 66: ALOS-2, Ascending Orbit, Random Forest with 24 Trees

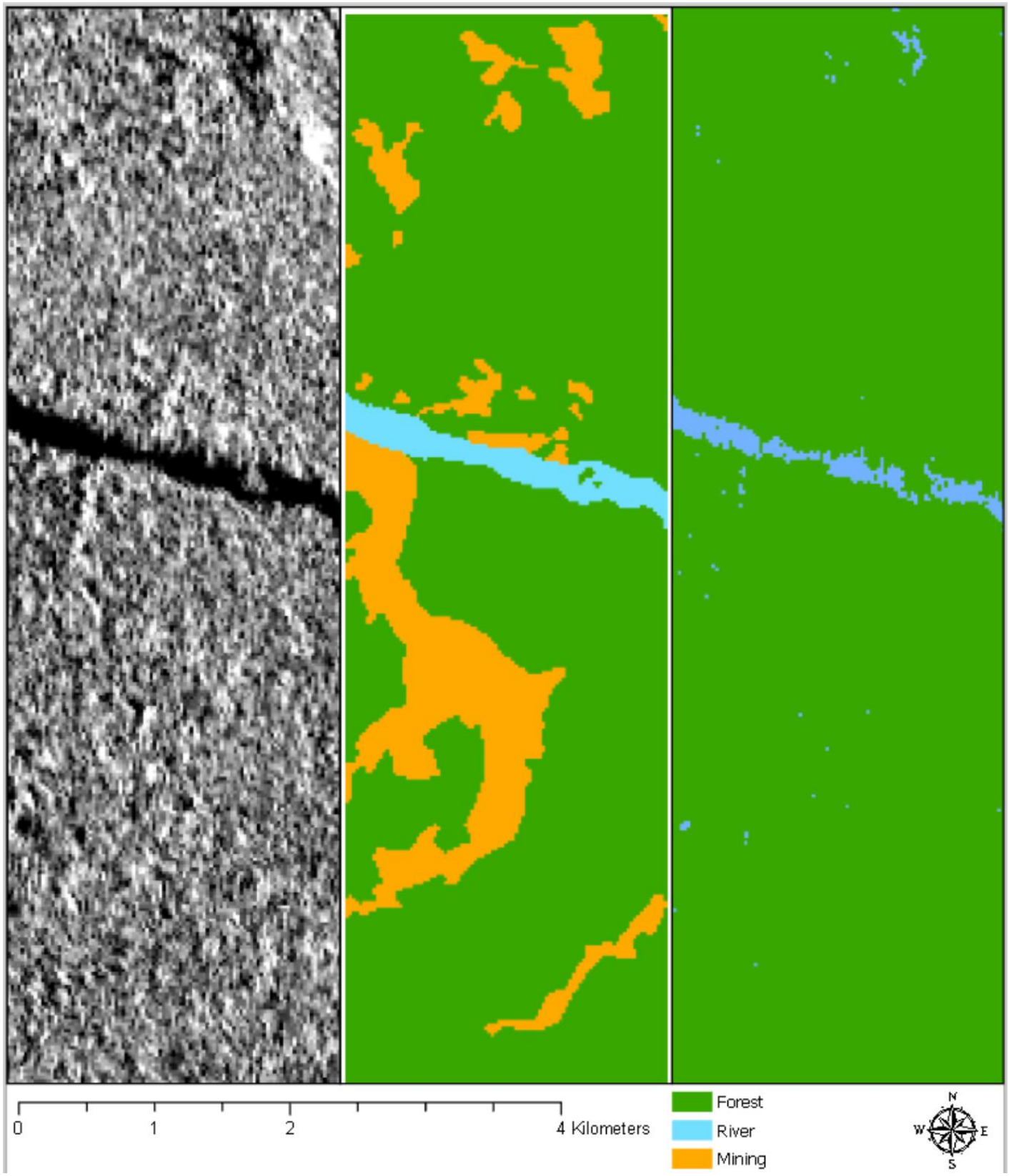


Table 67: ALOS-2, Ascending Orbit, Random Forest with Unlimited Trees

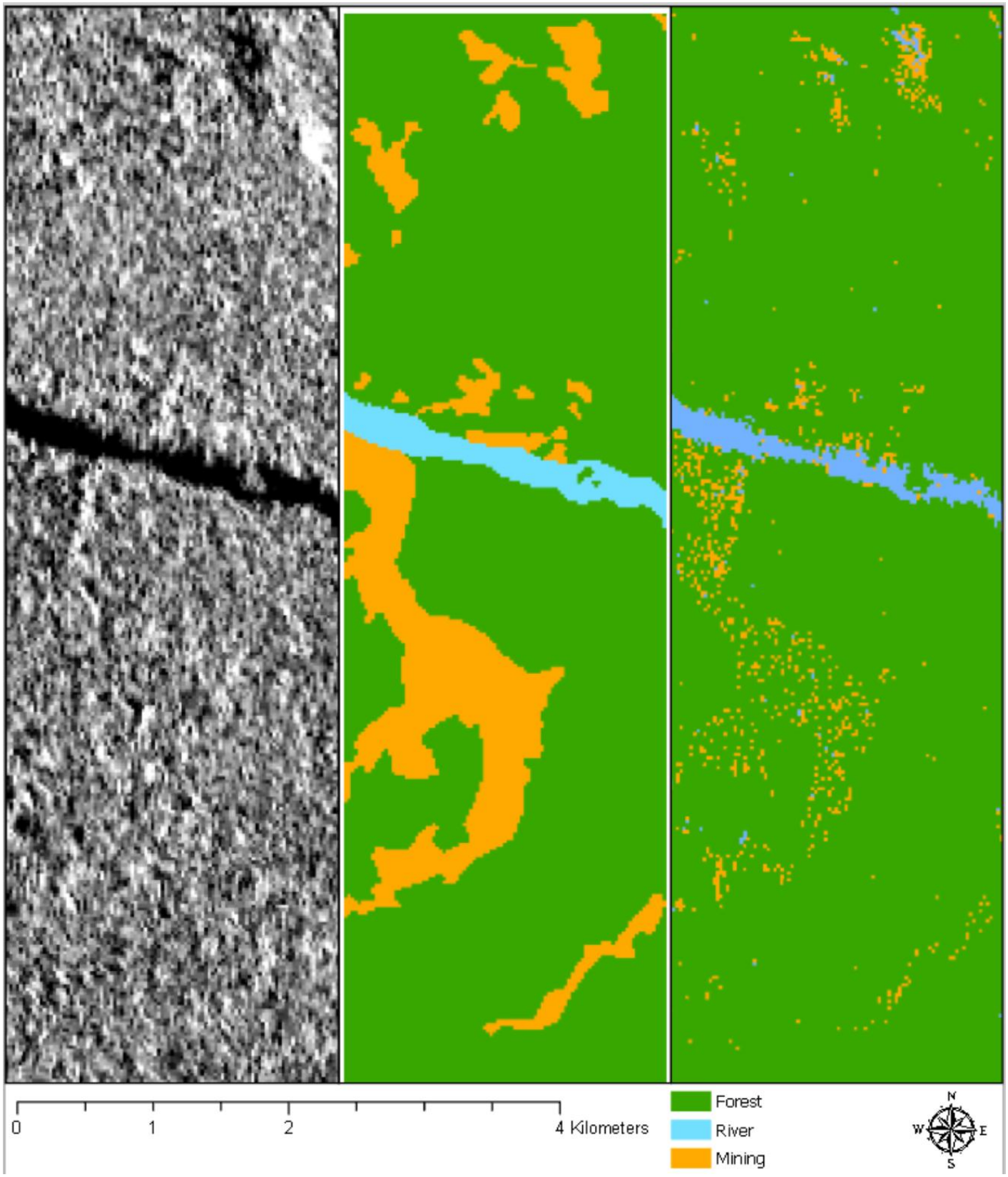


Table 68: ALOS-2, Descending Orbit, Random Forest with 24 Trees

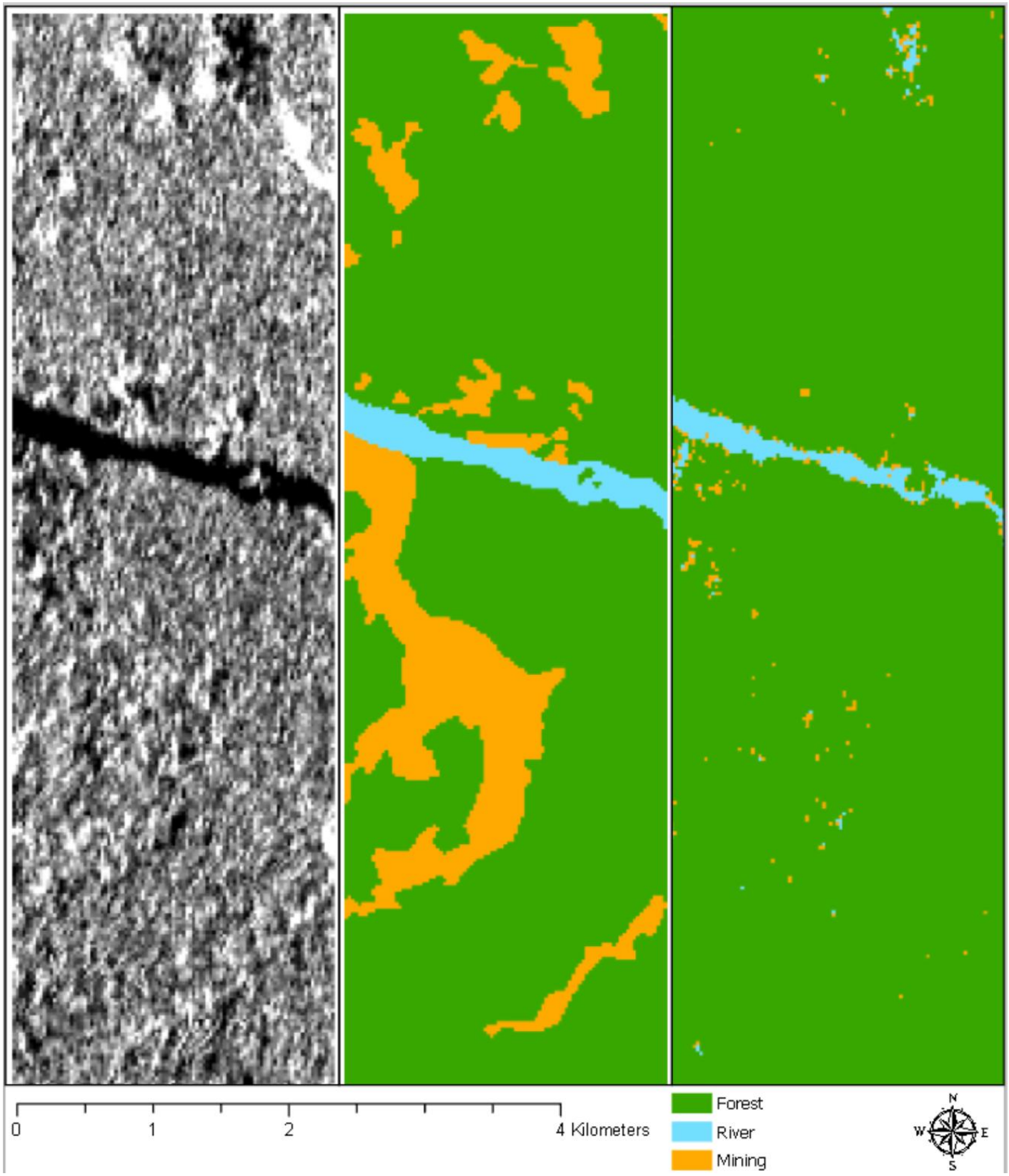


Table 69: ALOS-2, Descending Orbit, Random Forest with Unlimited Trees

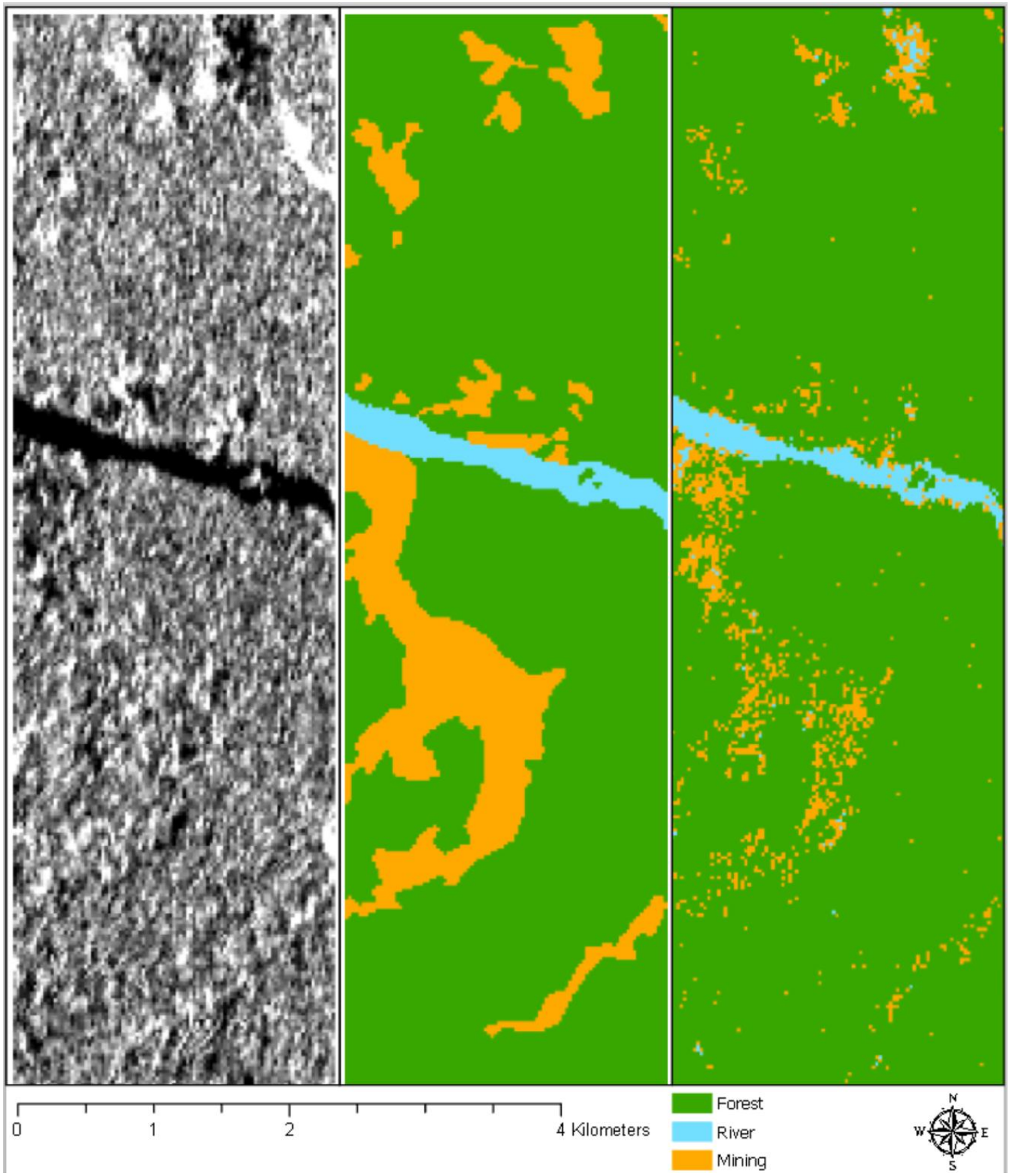


Table 70: Sentinel-1, Ascending Orbit, Random Forest with 24 Trees

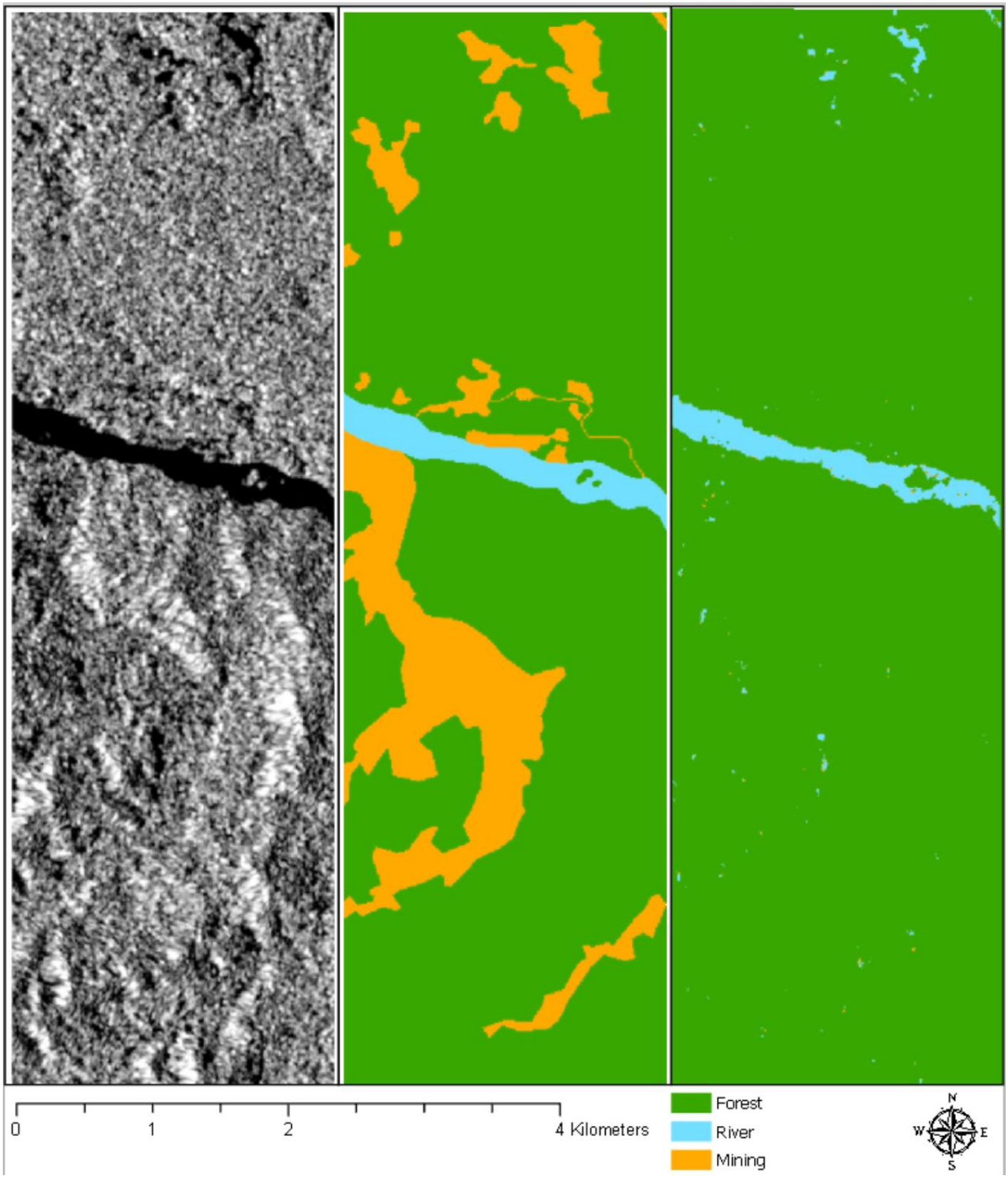


Table 71: Sentinel-1, Ascending Orbit, Random Forest with Unlimited Trees

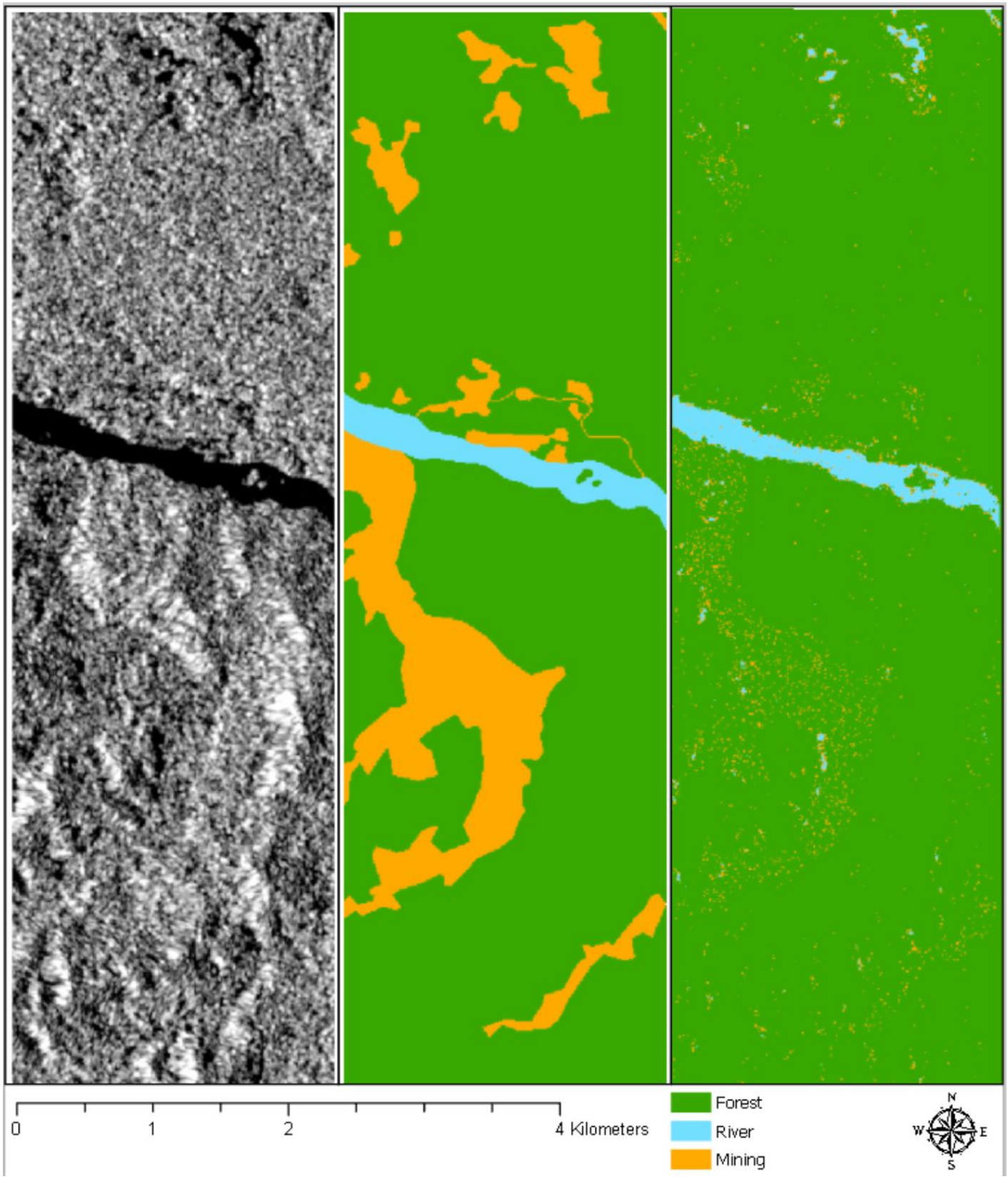


Table 72: Radarsat-2, Ascending Orbit, Random Forest with 24 Trees

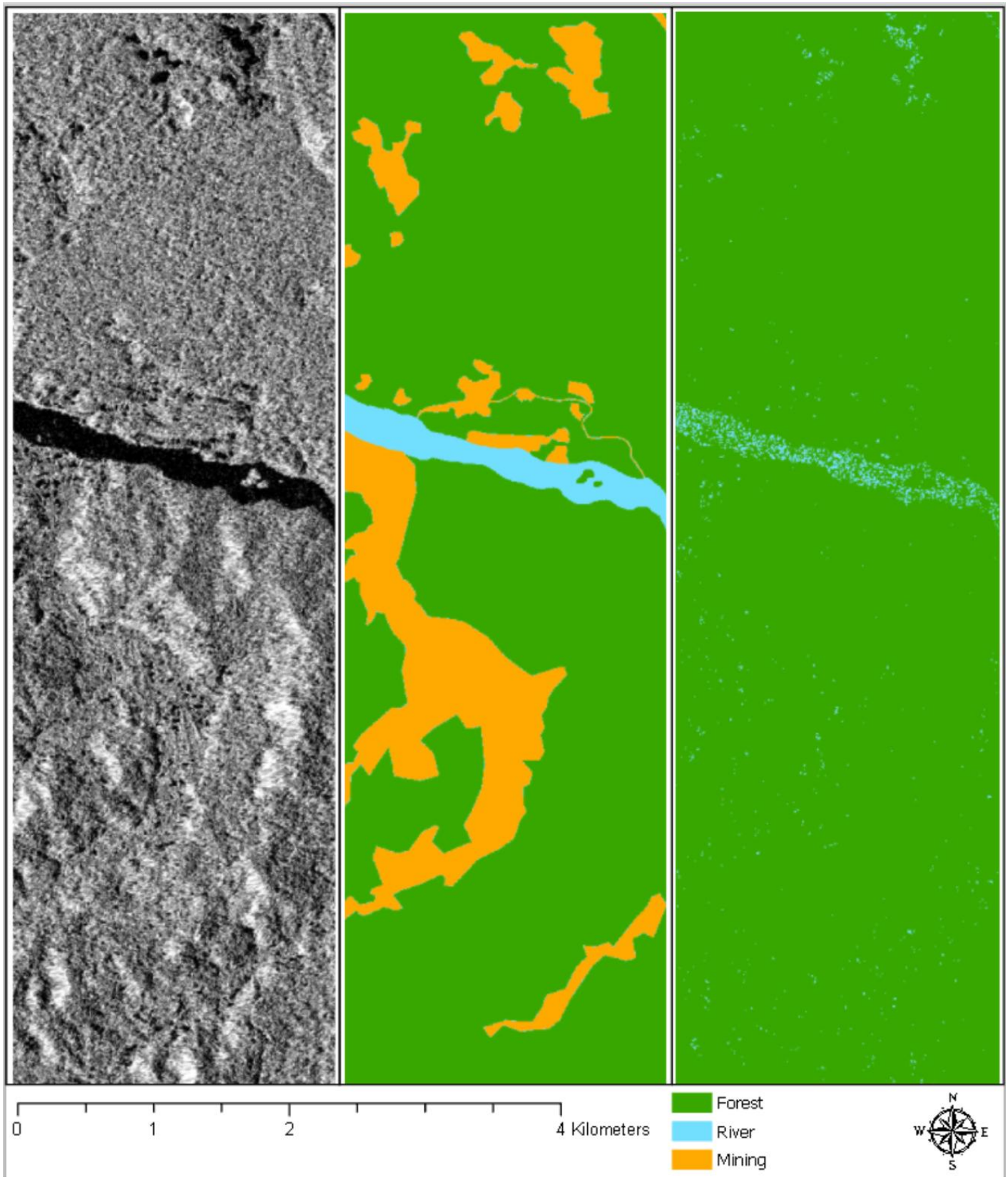


Table 73: Radarsat-2, Ascending Orbit, Random Forest with Unlimited Trees

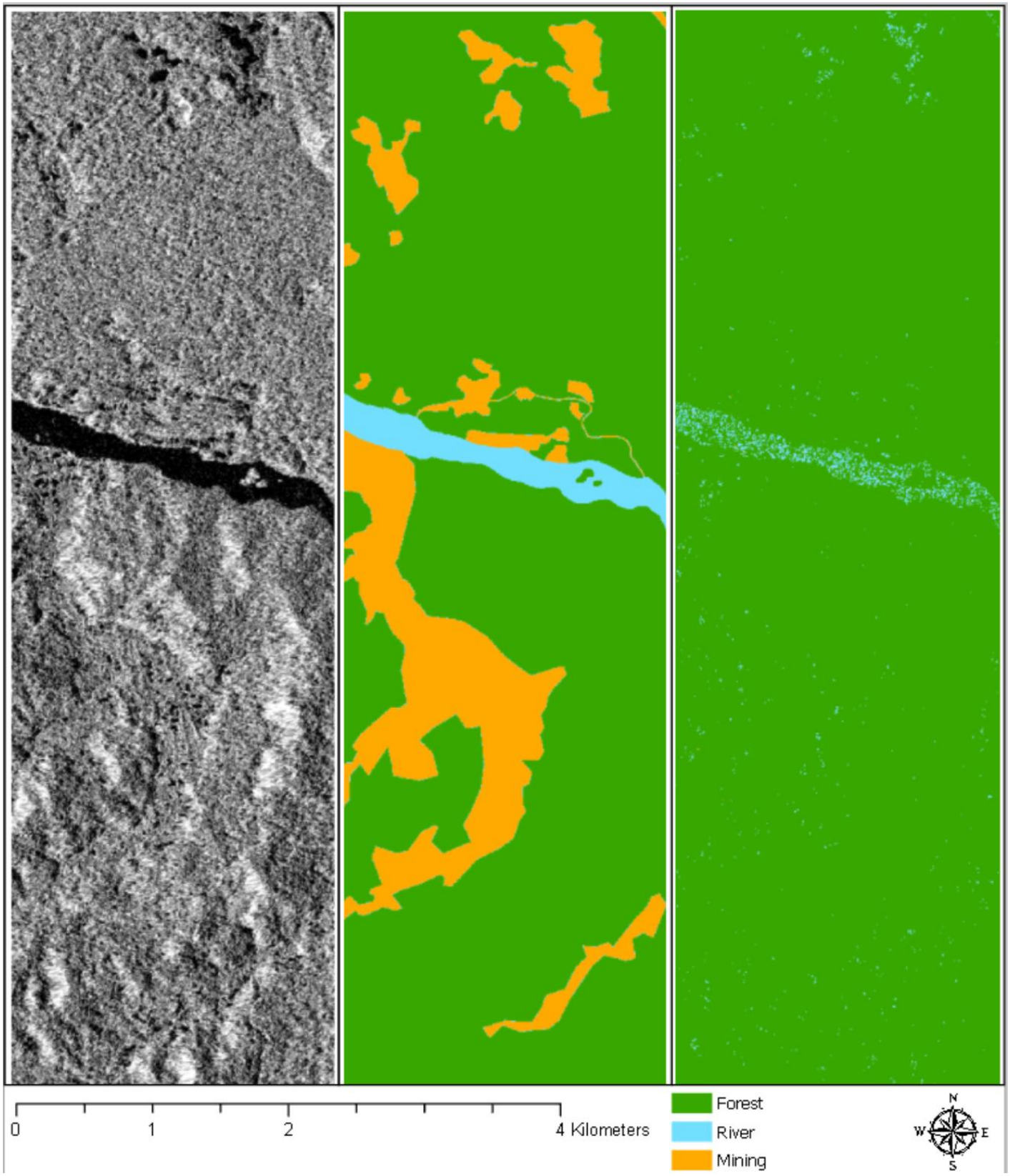


Table 74: TerraSAR-X, Ascending Orbit, Random Forest with 24 Trees

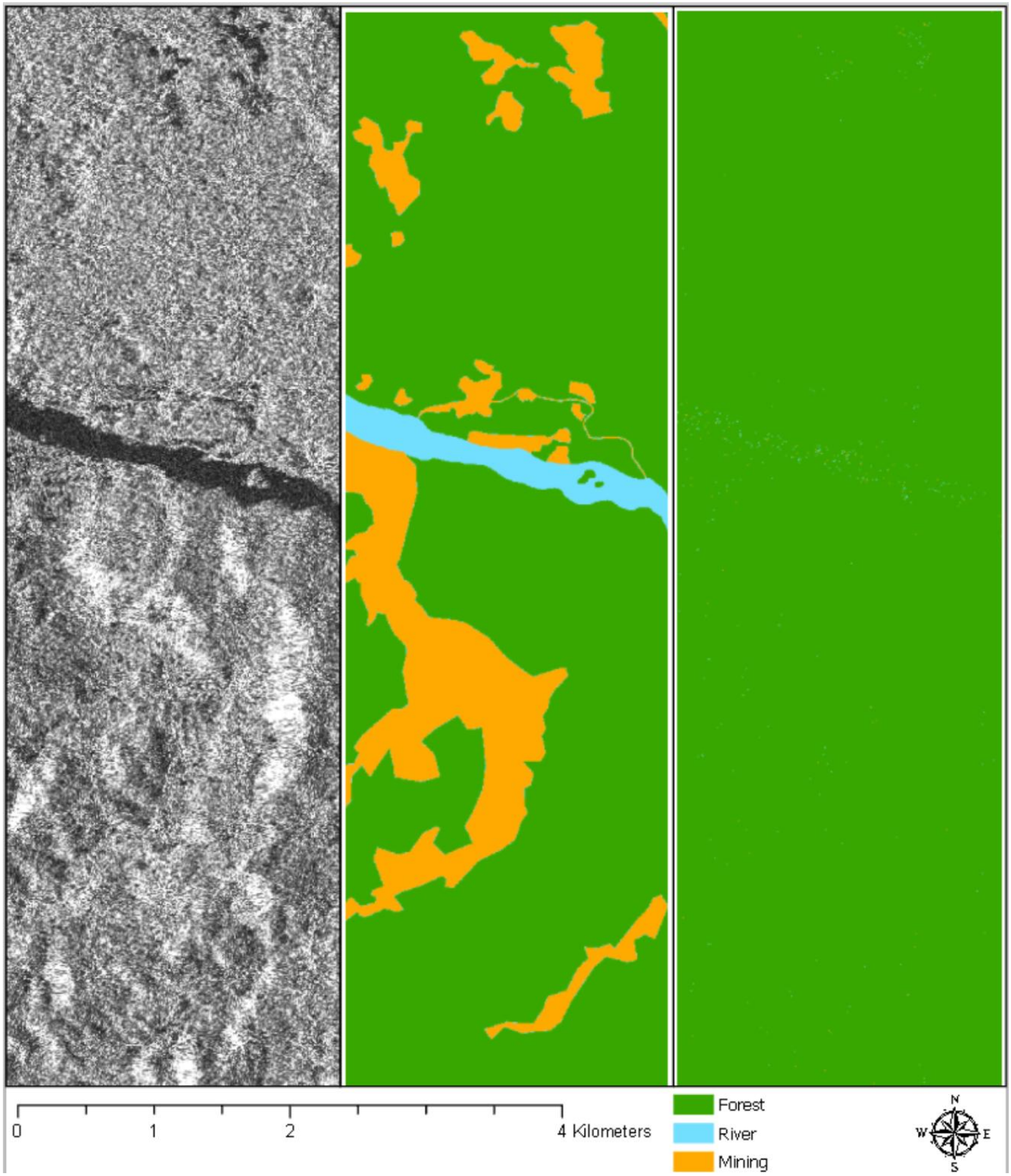


Table 75: TerraSAR-X, Ascending Orbit, Random Forest with Unlimited Trees

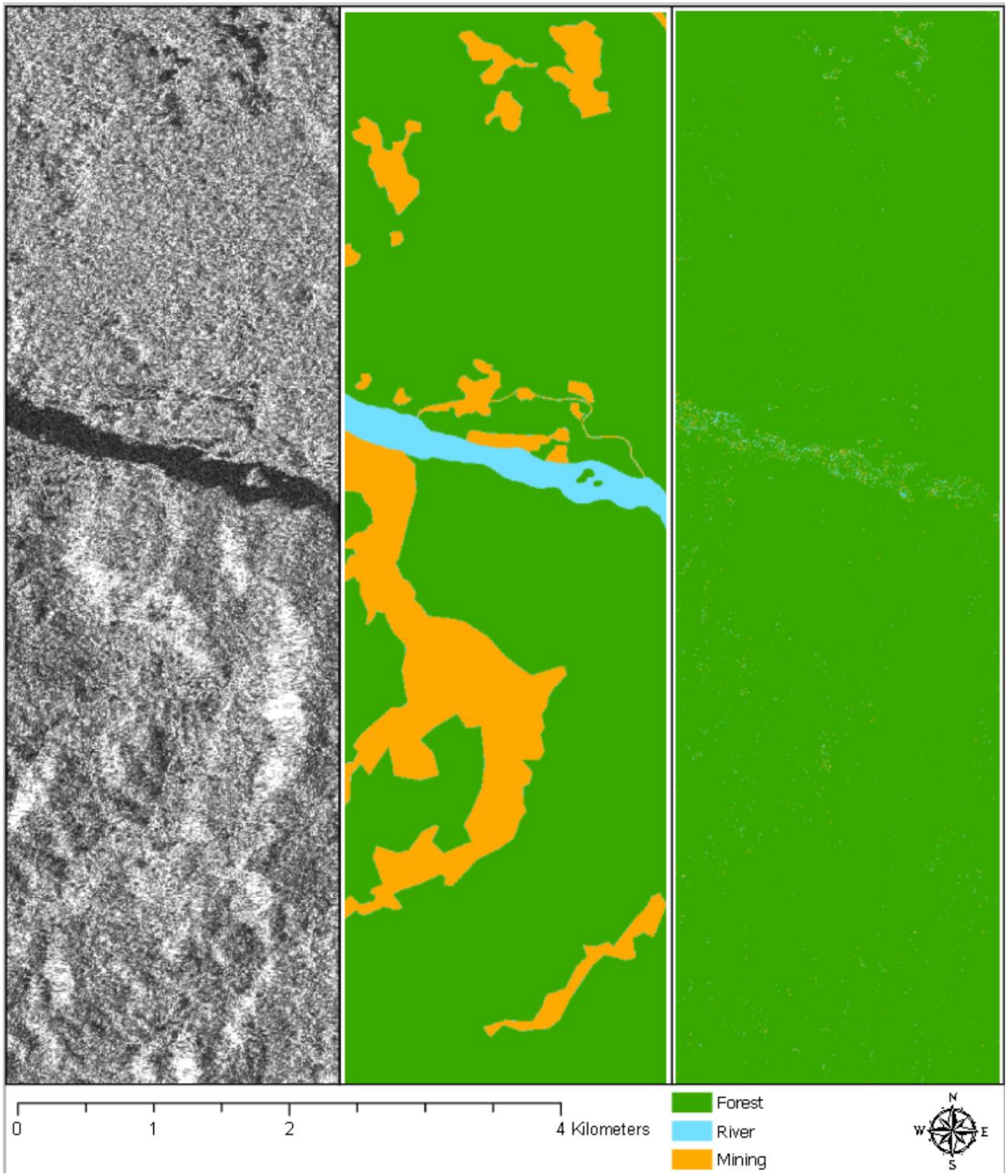


Table 76: TerraSAR-X, Descending Orbit, Random Forest with 24 Trees

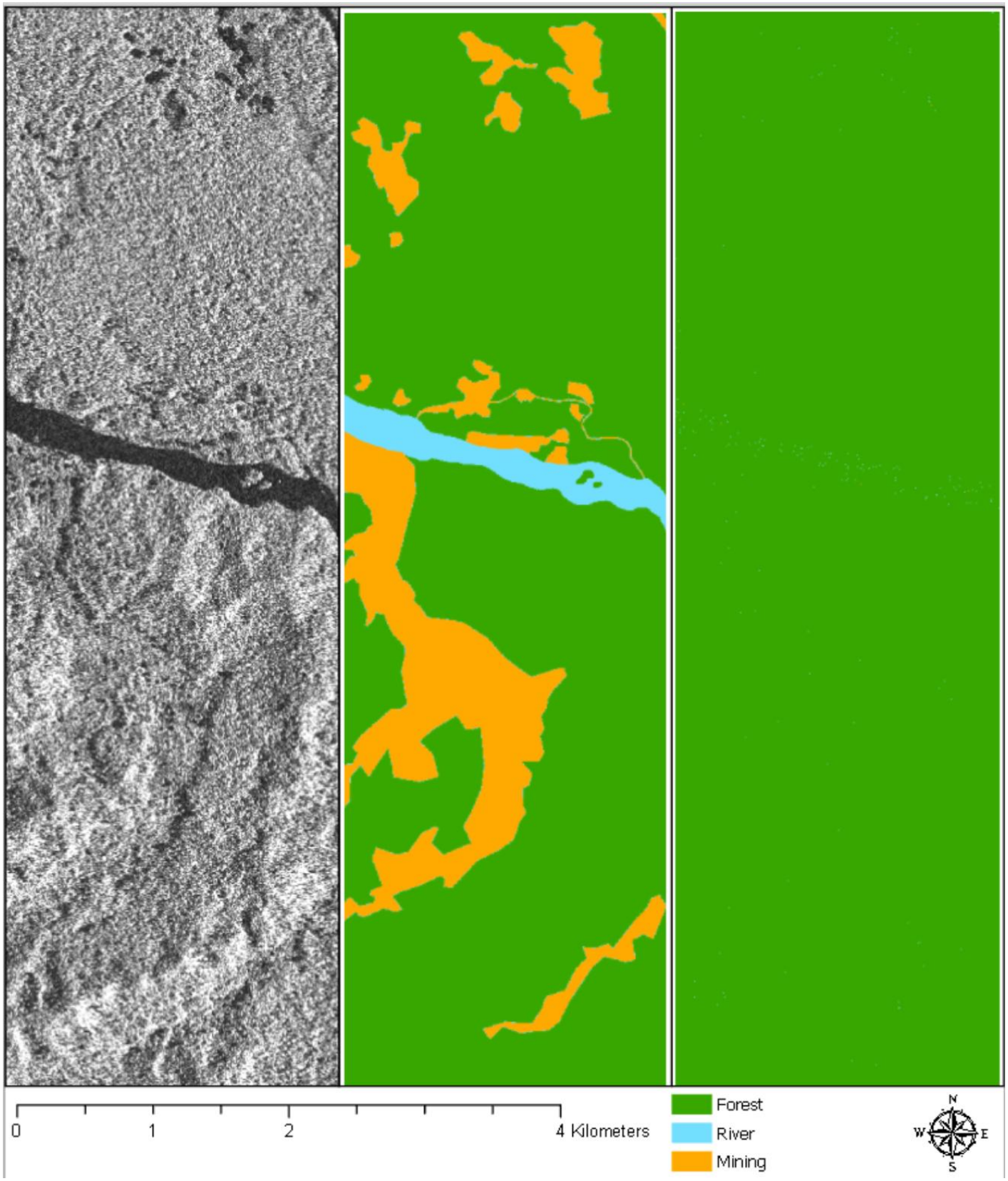


Table 77: TerraSAR-X, Descending Orbit, Random Forest with Unlimited Trees