Universiteit Leiden
The Netherlands

Evaluating the Performance of Machine Learning Models in Remote Sensing for Sustainable Development Goals: A Meta-Analysis

Nina Maria Leach

Thesis advisor: Dr. J. Burger[1]
Thesis advisor: Dr. J. Klingwort[1]
Thesis advisor: Prof. Dr. Mark de Rooij[2]

Defended on July 7, 2025

MASTER THESIS

STATISTICS AND DATA SCIENCE

UNIVERSITEIT LEIDEN

[1]Department of Research and Development, Statistics Netherlands (CBS), CBS-weg 11, PO Box 4481, 6401 CZ Heerlen, the Netherlands.
[2]Department of Methodology and Statistics, Leiden University, Leiden, The Netherlands

Foreword

The aim of this research was to assess whether study features can explain variations in results across studies. To my knowledge, this is the first study to apply weighted meta-analysis techniques to examining the performance of machine learning models in remote sensing. While I intended to adhere to PRISMA guidelines, the exploratory nature of the topic—and my own learning journey—meant pre-registration was ultimately not conducted, and with data extraction carried out solely by myself, there is a degree of subjectivity and potential for error there. But I'm giving away spoilers for the discussion, so I'll stop myself there!

I have succeeded in building this manuscript using Quarto, with minimal stylistic adjustments after rendering. The entire code for this project is available on GitHub, and an HTML version with integrated code chunks can be accessed on GitHub Pages website. The data processing and paper selection analysis scripts are all available on the GitHub-hosted site under the appendix (more details about the file organisation are available on the GitHub page). I have also integrated the Leiden University master thesis cover format into the Quarto book, so if any future student would like to reuse it please do! I toyed with the idea of creating a Quarto book template but that is a project for another day.

One last thing, in the discussion of this thesis I suggest that journals should begin requesting data submissions alongside the manuscript to enable active, ongoing meta-analyses. In support of this idea, I developed a small pilot website to demonstrate how such a system might work. If anyone feels inclined to add to the dataset, there are instructions on how to do that there.

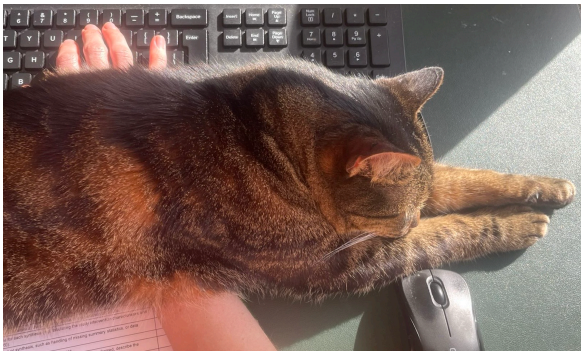
To the reader: thank you for taking the time to read my thesis— there's still time to stop reading, and I won't know any different! If you've made it this far —hi supervisors, independent reader (and mum⁉)— I would like to apologize in advance for continuing the convention of inconsistent notation across meta-analysis research. I can only hope I have been consistent within my own work as its best not to change notations μ-dstream.

Acknowledgements

I am very grateful to everyone that has helped me through the process and completion of this thesis. First and foremost, my thesis supervisors at the CBS, Dr. Joep Burger and Dr. Jonas Klingwort. I would like to thank (and berate) them for giving me the freedom of choosing my topic. What a journey it led us through! I never thought I would refer to back and forth articles on the misconceptions of double arcsine back-transformations as drama, but here I am, waiting for the next episode (no spoilers). In all seriousness, I truly appreciate the time and effort they dedicated and the detailed and feedback they provided. I would also like to extend my thanks to my internal supervisor, Prof. Mark de Rooij, for his support and feedback throughout the process, and to my independent reader, Prof. E.M.L. Dusseldorp for stepping, in short notice.

Perhaps unconventionally, I would like to acknowledge Dr. Wolfgang Viechtbauer— the statistician behind the *Metafor* package. I aspire to write documentation like his when I "grow up".

A sincere thank you goes to my friends and family for their unwavering encouragement through all my self-doubt. Special thanks to Paul, Anka and (my favorite uncle) Mike for their proofreading and feedback, and to Bo for the continuous support and for pretending to be interested in the drama of double arcsines. Lastly, a special thank you to Capo for almost never leaving my side and reminding me of the importance of taking breaks, which, of course, are essential because he should have my undivided attention.



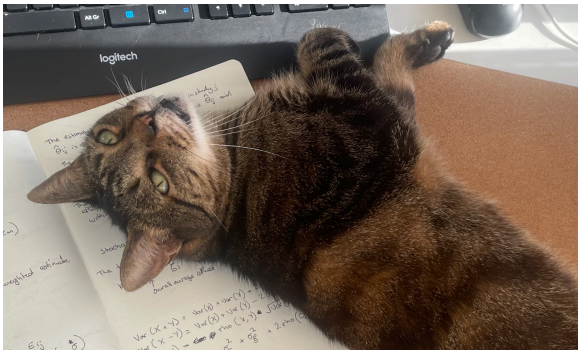


Table of contents

Abstract

Objective: This meta-analysis aims to evaluate machine learning methods in remote sensing applications for monitoring Sustainable Development Goals (SDGs). Specifically, it aims to (1) estimate the average performance (summary effect size); (2) determine the degree of heterogeneity within and across studies; (3) assess whether specific study features influence model performance; and (4) compare the sample-weighted and unweighted estimate summary effect.

Methods: The meta-analysis used the PRISMA guidelines. A search was performed across multiple academic databases to identify peer-reviewed studies that applied machine learning models to remote sensing data for SDG monitoring. A random sample of 200 relevant studies was selected for abstract screening, which was reduced to $n = 20$ studies with $k = 86$ effect sizes for the analysis. To estimate the overall accuracy of machine learning models both a three-level random-effects model and an unweighted model were used.

Results: The average overall accuracy of the unweighted model, $\hat{\mu}_{\text{unweighted}} = 0.90$ (95% CI [0.85; 0.94]), which is not substantially different from the weighted model, $\hat{\mu}_{\text{weighted}} = 0.89$ (CI 95% [0.85, 0.94]. The weighted models found substantial heterogeneity between results. Unsurprisingly, the proportion of the majority class was identified as the most important factor affecting the overall accuracy, followed by the inclusion of ancillary data. However, machine learning model group (i.e., neural networks, tree-based models) or SDG goal did not have a significant effect on the reported overall accuracy.

Conclusion: This study demonstrates the high variability model performance in remote sensing applications. As well as the impact class imbalance has on the reported overall accuracy. These findings suggest the need for precise metrics to assess model performance, particularly in imbalanced datasets. Future research should examine a broader range of performance metrics and explore additional study features to explore further what features affect the outcomes. In addition, the robustness of the random-effects meta-analysis methods application to this field should be further examined.

Table of Notation

Notation	Definition	Section
rc	Index for the rows and columns of a matrix	Chapter 2, Chapter 3
m_{rc}	Confusion matrix: the number of instances where the actual class is r and the predicted class is c. Where r is the row index, representing the actual class (reference) and c is the column index, representing the predicted class.	Chapter 2
m	The total number of classified instances on which the $i-$th effect size in the $j-$th study is based.	"
s	Number of correctly classified instances.	"
n	Total number of primary studies used in this study. In Figure 3.3 n is the number of studies in each section.	Chapter 3
j	The study index: $j = 1, ..., n$	"
k_j	Total number of effect sizes in the $j-$th study	"
i	The effect size index within a study, $i = 1, ..., k_j$	"
k	Total number of effect sizes gathered.	"
$\theta_{ij}, \widehat{\theta}_{ij}$	True and estimated effect size (overall accuracy) from the $i-$th effect size the $j-$th study.	"
κ_j	Average effect size in study j.	"
μ	Summary effect size: average effect size in the population.	"
$\widehat{\mu}_{\text{unweighted}}$	Unweighted estimate- where the unit of analysis are the numbers inclued in this study.	
$\widehat{\mu}_{\text{weighted}}$	Weighted estimate where the unit of analysis is the sample sizes of the primary studies .	
v_{ij}	Variance of the $i-$th effect size in study j.	"
σ^2_{level2}	Within-study variance	"
σ^2_{level3}	Between-study variance	"

Introduction

In 2015, all United Nations member states adopted the Sustainable Development Goals (SDGs) to address global challenges such as climate change, environmental degradation, poverty, and inequality (UN DESA, 2023; UN-GGIM:Europe, 2019). This international plan outlines 17 global goals to achieve a better and more sustainable future (UN DESA, 2023; UN-GGIM:Europe, 2019; United Nations, 2024). Having passed the midpoint of the SDGs' timeline with significant setbacks, the critical role of timely and high-quality data has never been more apparent (UN DESA, 2023; United Nations, 2024). These data are vital to identifying challenges, formulating evidence-based solutions, monitoring the implementation of solutions, and making essential course corrections (UN-GGIM:Europe, 2019). However, despite this necessity for high-quality data, traditional monitoring approaches, such as household- or field-level surveys (ground-acquired data), remain the primary source of data collection for key indicators of SDGs by National Statistical Institutes (NSIs) (Burke et al., 2021; UN-GGIM:Europe, 2019). These methods are expensive and time-consuming to conduct (Burke et al., 2021). As a result, the frequency of ground-acquired data varies significantly around the world; for example, the most recent agricultural census for 24% of the world's countries was more than 15 years ago (Burke et al., 2021). Recognizing this challenge, both the United Nations SDG Report (2023, p. 49) and the Global Working Group on Big Data for Official Statistics underscore the importance of innovative methodology and data sources, including remote sensing and machine learning, to enhance the monitoring and implementation of the SDGs (UN-GGIM:Europe, 2019; United Nations, 2017).

Remote sensing — data collected from a distance via satellite, aircraft, or drones — offers a cost-effective approach for monitoring wide-ranging geographic areas (Khatami et al., 2016; Maso et al., 2023; UN-GGIM:Europe, 2019; Zhao et al., 2022). Remote sensing imagery has been limited to agricultural and socioeconomic applications for decades (Burke et al., 2021; Lavallin & Downs, 2021; Y. Zhang et al., 2022). For instance, the Laboratory for Applications of Remote Sensing (LARS) has utilized satellite data and machine learning methods for crop identification since the 1960s (Holloway & Mengersen, 2018). However, in recent years, there has been a considerable increase in the spatial, spectral, and temporal resolution of remote sensing data, alongside a significant increase in free sensor data and computational power for complex data analysis (Burke et al., 2021; Thapa et al., 2023; Y. Zhang et al., 2022). The magnitude of possible applications and increased availability of remote sensing data have rapidly increased the number of published research papers in this field (Burke et al., 2021; Khatami et al., 2016). Earth observation satellites alone can measure 42% of the SDG targets (Y. Zhang et al.,

2022).

Despite the increased research and availability the uptake of remote sensing data by NSIs has been slow. However, many NSIs are now capitalizing on the potential of using new and consistent data sources and methodologies to support and inform official statistics (United Nations, 2017). These can be generated by combining geospatial information, remote sensing, and other big data sources, allowing for the filling of data gaps, providing information where no measurements were previously made, and improving the temporal and spatial resolutions of data (e.g., daily updates on crop area and yield statistics). This paradigm shift from traditional statistical methods—such as counting and measuring by humans—towards estimation from sensors, simulation, and modelling, presents challenges, and requires convincing, statistically sound results, rigorous validation, and a significant shift in resources within institutions to adapt to the higher spatial and temporal resolutions necessary to address emerging policy questions (United Nations, 2017).

Given the wide variety of methodologies and contexts in previous studies, a critical question arises: *What factors influence the performance of machine learning models using remote sensing data for SDG monitoring?* A meta-analysis statistically combines the body of evidence on a specific topic, aiming to produce unbiased summaries of evidence (Iliescu et al., 2022). There are many potential methods to choose from to combine results. One choice that is made when conducting a meta-analysis is whether to use the study's sample size to weigh the result of each study (sample-weighted estimate) or an unweighted approach, which treats all results equally, disregarding sample size (J. A. Hall & Rosenthal, 2018). The current standard in meta-analysis research is to use the sample-weighted estimate (J. A. Hall & Rosenthal, 2018). However, the previous meta-analyses investigating the performance of machine learning models on remote sensing data have exclusively relied on unweighted approaches. While these studies have found that certain models, such as Support Vector Machines (SVM) and deep learning methods, often outperform traditional classifiers, the magnitude of these differences can vary across applications. For example, Khatami et al. (2016) selected studies with more than one model, and by making pairwise comparisons they concluded that SVM consistently outperformed other classification models. However, these meta-analyses relied on unweighted approaches, potentially overlooking if these variations in results are due to differences in sample sizes, which could affect the reliability and precision of the findings, as larger studies generally provide more accurate estimates.

Therefore, this study seeks to address the question of how machine learning models perform when applied to remote sensing data for SDG monitoring. By conducting a meta-analysis on peer-reviewed research articles in this domain, the study aims to; (1) estimate the average performance (summary effect size), (2) determine the degree of heterogeneity within and across studies, (3) assess whether specific study features influence model performance, and (4) compare the sample-weighted and unweighted estimate summary effect.

Background

This chapter provides an overview of the concepts and methodologies analysed in this research. It provides a brief introduction to remote sensing, machine learning techniques, and their applications in land cover mapping and SDG monitoring.

2.1 Remote Sensing

In the broadest sense, remote sensing involves acquiring information about an object or phenomenon without direct contact (Campbell & Wynne, 2011). More specifically, remote sensing refers to gathering data about land or water surfaces using sensors mounted on aerial or satellite platforms that record electromagnetic radiation reflected or emitted from the Earth's surface (Campbell & Wynne, 2011, p. 6). The origins of remote sensing lie with the development of photography in the 19th century, with the earliest aerial or Earth Observation photographs taken with cameras mounted on balloons, kites, pigeons, and aeroplanes. (Burke et al., 2021; Campbell & Wynne, 2011, p. 7). The first mass use of remote sensing was during World War I with aerial photography. The modern era of satellite-based remote sensing started with the launch of Landsat 1 in 1972, the first satellite specifically designed for Earth Observation (Campbell & Wynne, 2011, p. 15). Today, remote sensing technology enables frequent and systematic collection of data about the Earth's surface with global coverage, revolutionizing our ability to monitor and analyze the Earth's surface (Burke et al., 2021; NASA, 2019). As of May 2023, roughly 1039 active nonmilitary Earth Observation satellites are in orbit; 51% were launched in 2020 (UCS, 2021).

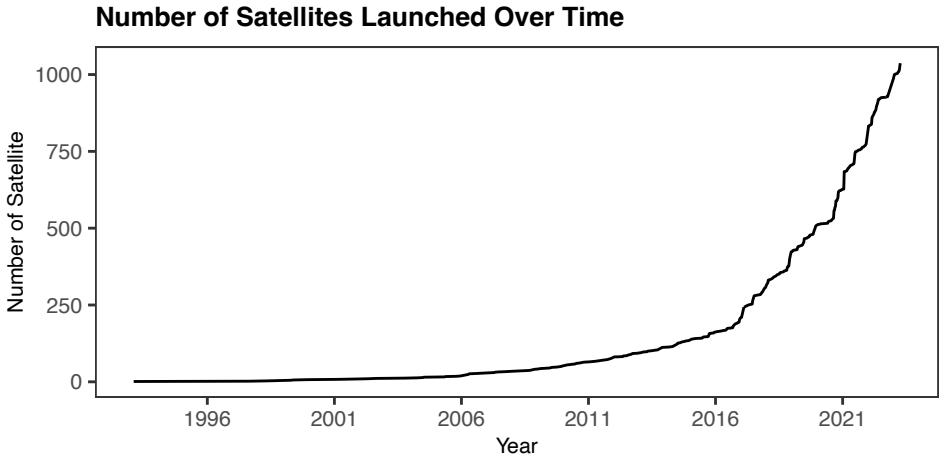


Figure 2.1: Number of active satellites by date of launch — data acquired from UCS (2021).

Sensors on remote sensing devices such as satellites measure electromagnetic radiation reflected by objects on the Earth's surface. This is done in two different ways: passive and active. Passive sensors rely on natural energy sources, like sunlight, to record incident energy reflected off the Earth's surface. While active sensors generate their own energy, which is emitted and then measured as it reflects back from the Earth's surface (NASA, 2019).

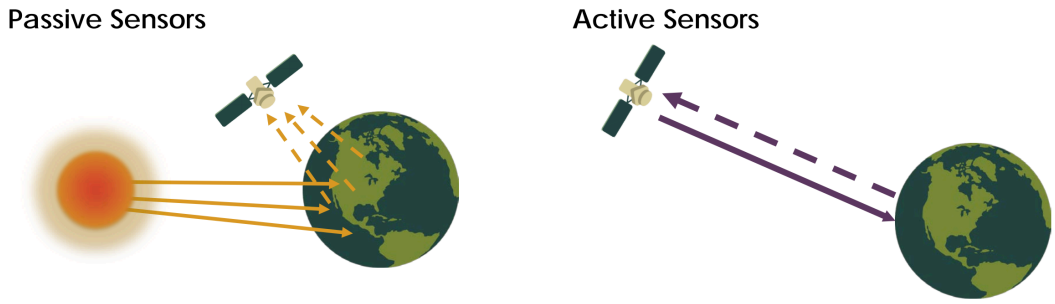


Figure 2.2: Illustration of a passive sensor and an active sensor —source: NASA (2019).

Components of the Earth's surface have different spectral signatures — i.e., reflect, absorb, or transmit energy in different amounts and wavelengths (Campbell & Wynne, 2011). Remote sensing devices have several sensors that measure specific ranges of wavelengths in the electromagnetic spectrum; these are referred to as spectral bands; e.g., visible light, infrared, or ultraviolet radiation (NASA, 2019; SEOS, 2014). By capturing information from particular bands, the spectral signatures of surfaces can be used to identify objects on the ground. Figure 2.3 illustrates the differences between the spectral signatures of soil, green vegetation, and water across various wavelengths. The grey bands in the figure represent the specific spectral bands on the Landsat TM satellite (SEOS, 2014). The distinct reflectance properties of each material within these bands enable the differentiation of surface materials, making it possible to identify different land cover types. This information can be used directly for classification, or it can be combined into indices—such as the Normalized Difference Vegetation Index (NDVI)—to enhance the detection of specific features like vegetation health and coverage (Campbell & Wynne, 2011; NASA, 2019). The $NDVI$ uses red light and near-infrared (NIR) to distinguish green vegetation. Higher $NDVI$ values indicate green vegetation as more red light is absorbed, whereas lower values correspond to non-vegetated areas where more red light is reflected.

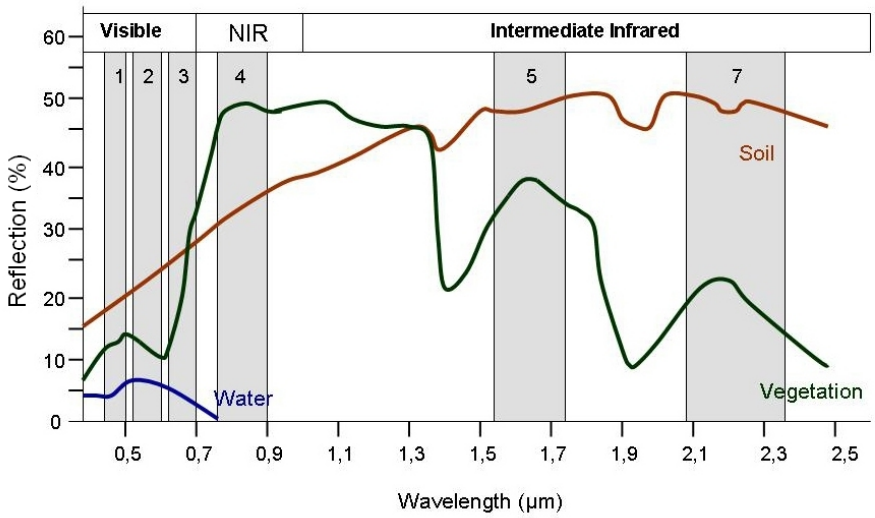


Figure 2.3: Spectral signatures of soil, green vegetation, and water across different wavelengths, representing the portion of incident radiation that is reflected by each material as a function of wavelength. The grey bands indicate the spectral ranges (channels) of Landsat TM satellite. Bands 1-3 capture visible light (Blue, Green, Red), Band 4 captures near-infrared (NIR), and Bands 5 and 7 cover parts of the intermediate infrared spectrum. These spectral bands allow for the differentiation of various surface materials based on their unique reflectance properties —source: Siegmund and Menz (2005) as cited and modified by SEOS (2014).

2.2 Machine Learning

Machine learning techniques such as neural networks, random forests, and support vector machines have long been applied for spatial data analysis and geographic modelling (Haddaway et al., 2022; Lavallin & Downs, 2021). Compared to using indices alone, machine learning techniques enhance the accuracy and efficiency of data analysis and interpretation processes, making it possible to analyze large volumes of data effectively. This is particularly useful for handling the high complexity and dimensionality of remote sensing data. In recent years, the application of machine learning techniques in remote sensing has surged, driven by the increasing availability of large datasets and advancements in computational power (UN-GGIM:Europe, 2019; Y. Zhang et al., 2022).

These machine learning models can be grouped into four main types according to the aims of analyses: classification, clustering, regression, and dimension reduction. Table 2.1 describes this grouping and gives examples. It is important to note that recent trends in machine learning and remote sensing analyses use hybrid or ensemble approaches using a combination of these groups, for a thorough review of these methods see UN-GGIM:Europe (2019).

Table 2.1: Categories of machine learning methods grouped according to the analytic aim.

Analysis aim	Explanation
Classification	Assigning objects to known classes based on input variables. For example, categorizing pixels in an image into crop types using a model trained on known data.
Regression	Predict a numeric (discrete or continuous) response variable based on input variables, similar to classification but with numeric outputs. An example is predicting crop yield from Earch Observation image data.
Clustering	Groups objects based on input variables without pre-defined classes, identifying similarities among the objects. This can help in grouping pixels in an image for further inspection.
Dimension reduction	Reduces a large set of variables to a smaller set that retains most of the original information. This can simplify analysis or generate new variables like indices (e.g., Vegetation Index) for interpretation.

Note:

Adapted from UN-GGIM:Europe (2019) and Haddaway et al.(2022).

Performance metrics are used to verify these analyses, which for classification tasks involve creating a confusion matrix — a cross-tabulation of class labels assigned to model predictions and reference data (ground truth). In a confusion matrix, the correctly classified instances are on the diagonal, and the off-diagonal cells indicate which classes are confused (i.e., incorrectly classified). In remote sensing applications, accuracy assessments are undertaken on a pixel, group of pixels (e.g. block), or an object level (Stehman & Foody, 2019).

Table 2.2: Confusion matrix of four classes

	Predictions					
Reference	**Class 1**	**Class 2**	**Class 3**	**Class 4**	**Total**	**Producer's accuracy**
Class 1	m_{11}	m_{12}	m_{13}	m_{14}	$m_{1.}$	$m_{11}/m_{1.}$
Class 2	m_{21}	m_{22}	m_{23}	m_{24}	$m_{2.}$	$m_{22}/m_{2.}$
Class 3	m_{31}	m_{32}	m_{33}	m_{33}	$m_{3.}$	$m_{33}/m_{3.}$
Class 4	m_{41}	m_{42}	m_{43}	m_{44}	$m_{4.}$	$m_{44}/m_{4.}$
Total	$m_{.1}$	$m_{.2}$	$m_{.3}$	$m_{.4}$	m	
User's accuracy	$m_{11}/m_{.1}$	$m_{22}/m_{.2}$	$m_{33}/m_{.3}$	$m_{44}/m_{.4}$		

Note:

The rows (r) represent the reference (observed) classification and the columns (c) represent the predicted classes. m_{rc} is the number of instances in reference (observered) class r and predicted class c, and m is the total number of instances (i.e., the number of pixels/objects classified).

From this matrix, performance measures such as overall accuracy are derived (FAO, 2016; Stehman

& Foody, 2019; UN-GGIM:Europe, 2019) where the overall accuracy is the total number of successful classifications s over the total number of instances, m (q is the number of classes).

$$\text{Overall Accuracy (OA)} = \frac{\sum_{r=1}^{q} m_{rr}}{m} = \frac{s}{m} \tag{2.1}$$

If the unit of accuracy assessment is a pixel, then overall accuracy is the proportion of pixels classified correctly. Other metrics include reliability (User's accuracy) and sensitivity (recall or Producer's accuracy). Reliability is the correct classification for a particular class divided by the column total ($m_{.c}$), and sensitivity is the correct classification over the row total ($m_{r.}$). It is important to consider the map's purpose when evaluating its accuracy, as overall accuracy may not reflect the accuracy of specific classes. Factors such as sample size, class stability, class proportions, and landscape variability influence the overall accuracy (FAO, 2016; see UN-GGIM:Europe, 2019).

2.3 Australia Land Cover Mapping

As an example of how remote sensing data and machine learning can be used to support ecologically sustainable development, Owers et al. (2022) developed an approach to monitor and map land cover across Australia using techniques. Their study used Landsat sensor data archived through Digital Earth Australia to generate annual land cover maps from 1988 to 2020 at a 25-meter resolution. The study used random forest and artificial neural networks to classify individual pixels according to the FAO's Land Cover Classification System (LCCS) framework.

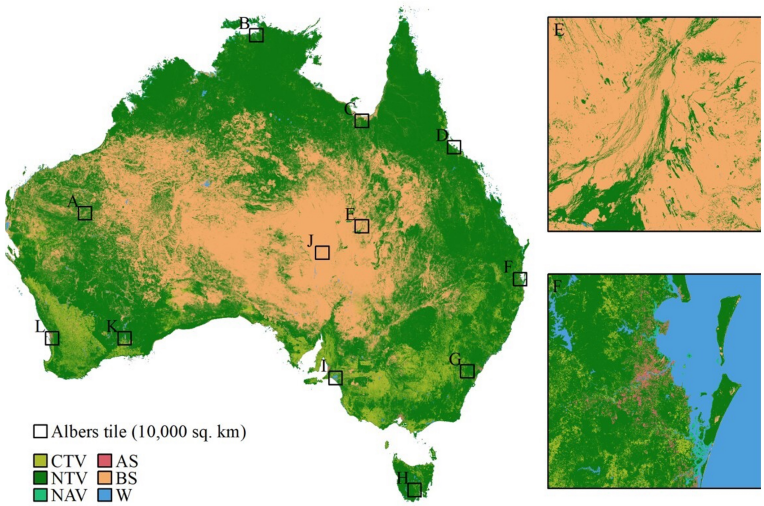


Figure 2.4: Land cover mapping created by Owers et al. (2022) using Landstat data to make continent-wide classifications using the LCCS frame work which differentiates six (classes) land cover types: cultivated terrestrial vegetation (CTV), natural terrestrial vegetation (NTV), natural aquatic vegetation (NAV), artificial surfaces (AS), bare surfaces (BS), and water bodies (W).

Producing such detailed maps through traditional topographical field surveys would be impractical, given Australia's size ($7,688,287$ km^2). While field surveys are considered the most accurate method for generating training sample data, they are labor-intensive, time-consuming, and costly (C. Zhang & Li, 2022). For example, surveying just 20 hectares (0.2 km^2) would take a team of four people approximately five days to complete, though the resulting topographical map would have a high resolution of 0.5 meters (L.A. Mbila, personal communication, January 26, 2024). In Owers et al. (2022), experts visually inspected the satellite imagery to validate the training and test data. While this method is less labor-intensive, costly, and time-consuming than field surveys, it still demands significant effort and expertise.

In contrast to the limitations of field surveys, remote sensing provides an efficient means for continuously monitoring large, often inaccessible areas (Owers et al., 2022; C. Zhang & Li, 2022). The potential applications of this technology are vast, including land use and degradation monitoring, forestry, biodiversity assessment, agriculture, disaster prediction, water resource management, public health, urban planning, poverty tracking, and the management and preservation of world heritage sites (Anshuka et al., 2019; Campbell & Wynne, 2011; Ekmen & Kocaman, 2024; O. Hall et al., 2023; Lavallin & Downs, 2021; Maso et al., 2023).

2.4 Previous Reviews

Numerous studies have previously examined the application of remote sensing for SDG monitoring. However, existing reviews are typically either limited to specific contexts, such as the use of satellite data for poverty estimation (O. Hall et al., 2023) or focus on descriptive results (see Yin et al., 2023). The existing reviews either apply methodology that aligns more closely with Synthesis Without Meta-Analysis (Campbell et al., 2020) —for example, Thapa et al. (2023) and Ekmen & Kocaman (2024) — or apply unweighted meta-analysis techniques, such as Khatami et al. (2016) and O. Hall et al. (2023). In an unweighted meta-analysis all studies are treated equally regardless of their sample size, quality, or variance (J. A. Hall & Rosenthal, 2018). However, it is more common in traditional applications of meta-analysis, to use the sample sizes when aggregating individual studies (J. A. Hall & Rosenthal, 2018). However, to my knowledge, no examples of a weighted meta-analysis applied to predictive performance in remote sensing data have been conduced, highlighting a gap that this study aims to address.

Methods

The methods adopted in this study are outlined in steps following the framework proposed by Debray et al. (2017). Additionally, efforts were made to follow the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines (Page et al., 2021), however due to the nature of this research, strict adherence was not always possible. For the statistical analyses `metafor` (Viechtbauer, 2010) and `dmetar` (Harrer et al., 2019) packages were used. The code was executed using R version 4.2.3 (2023-03-15).

3.1 Formulating the review question and protocol

The PICOTS (population, intervention, comparison, outcome, timing, and setting) system was used to frame the review aims for this analysis (Debray et al., 2017). As outlined by Table 3.1, using this framework, the research question was formulated to examine (1) the overall performance (2) and heterogeneity of machine learning models applied to remote sensing in the context of SDGs, and (3) to assess the influence of specific study features on model performance.

Table 3.1: PICOTS framework

Item	**Explanation**
Population	Studies monitoring SDGs.
Intervention	Application of machine learning models to remote sensing data.
Comparison	Comparison of different ML models and methodologies used in remote sensing applications.
Outcomes	Variability in the overall accuracy of machine learning models in monitoring SDGs.
Timing	Studies that focused on predicting current conditions rather than predicting future changes
Setting	Various geographic locations and environmental settings where remote sensing data is applied for SDG monitoring.

Note:

PICOTS framework items and the corresponding role in structuring this review.

To address this question, peer-reviewed articles published between January 2018 and December 2023 were gathered (on January 15 and 16, 2024) from several academic databases, including ScienceDirect and Taylor & Francis Online, as shown in Figure 3.3. The search terms were "*remote sensing* AND

machine learning AND *sustainable development goals*". The search results from these databases were downloaded in RIS format and imported into Zotero for further processing. Duplicate articles were handled using Zotero's "merge duplicates" function.

Several academic databases were used to reduce potential bias from database coverage (Hansen et al., 2022a; Tawfik et al., 2019). While Google Scholar can be useful for supplementary searches and grey literature, it is generally considered unsuitable as the primary source for systematic reviews (Gusenbauer & Haddaway, 2020). Furthermore, Google Scholar search results are not fully reproducible (Gusenbauer & Haddaway, 2020) and search result references that cannot be downloaded in batches, therefore the decision was made not to use Google Scholar to search for papers.

3.2 Specific inclusion and exclusion criteria

After removing review articles and non-research papers, a total of 811 relevant articles remained. Of these potentially relevant papers, 35% were published in 2023, highlighting the growth of research in this field. The trend, as illustrated in Figure 3.1, is consistent with other similar research, for example, Ekmen & Kocaman (2024), which reported a sharp increase in publications related to machine learning and remote sensing for SDG monitoring.

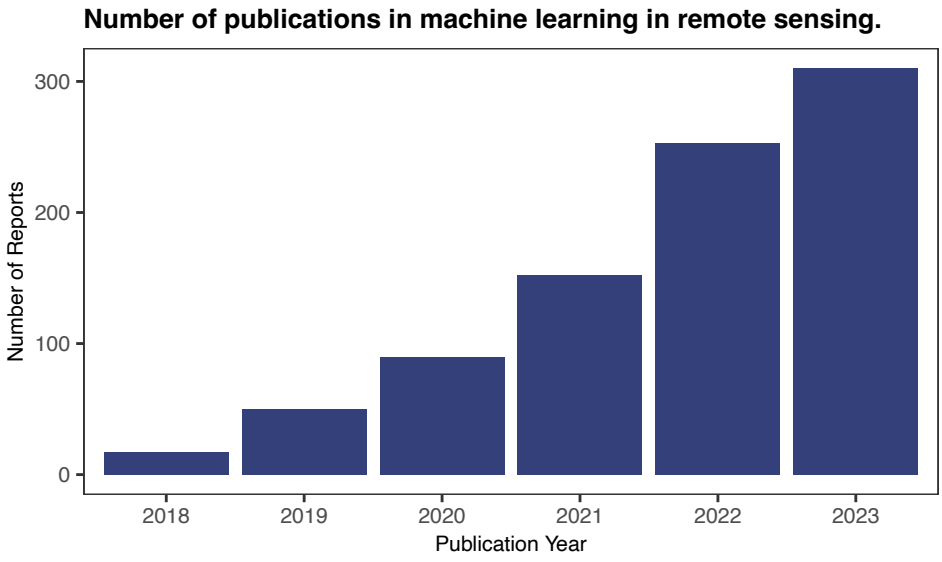


Figure 3.1: Publication increase between 2018 and 2022.

Due to the large number of papers remaining, a random sample of 200 articles was drawn for title and abstract screening. These potentially relevant articles were screened independently by three reviewers (the author and two internal supervisors) using the R package `metagear` (Lajeunesse, 2016). The papers were selected according to the following criteria: a) publications utilizing remote sensing and machine learning techniques, (b) indication of a quality assessment for example overall accuracy. Table 3.2 shows

the words highlighted in the abstract screening phase to aid the reviewers and Figure 3.2 shows the user interface highlighting these keywords.

Table 3.2: Keywords highlighted by the `metagear` user interface during abstract screening phase as a visual cue to speed up the screening process

Category	**Keywords**
General	empirical, result, predictive, analysis, sustainable development goal, sustainable development
Data related	remotely sensed, remote sensing, satellite, earth observation
Models	deep learning, machine learning, classification, classifier, regression, supervised, test set, training set, cart, svm, rf, ann, random forest, support vector machine, regression tree, decision tree, neural network, boosting, bagging, gradient, bayes
Quality metrics	overall accuracy, accuracy, coefficient of determination, rmse, mse, f1, precision, auc, roc, recall, sensitivity, specificity, mean absolute error, error, mae
To omit	systematic review, meta-analysis, review

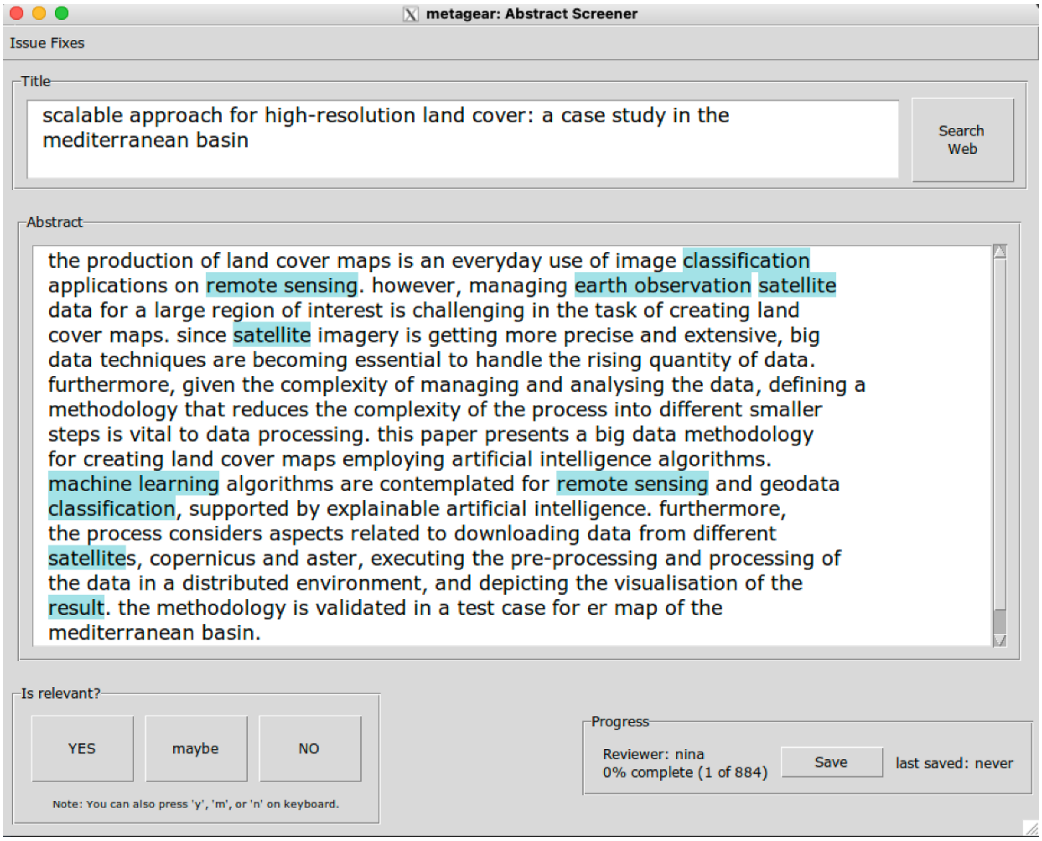


Figure 3.2: Metagear graphical user interface: Example of the metagear abstract screener interface highlighting keywords. On the bottom left the reviewer can select whether the paper is relevant.

As shown in Figure 3.3, of the 200 abstracts screened 57 were deemed potentially relevant by all three reviewers. To have comparable performance metrics it decided to focus on papers related to classification. The titles and abstracts of the 57 articles were screened using `metagear` dividing them to classification (40) and regression (17) papers. In the 40 papers, overall accuracy was the most commonly reported outcome metric and therefore it was decided to include all papers that report overall accuracy.

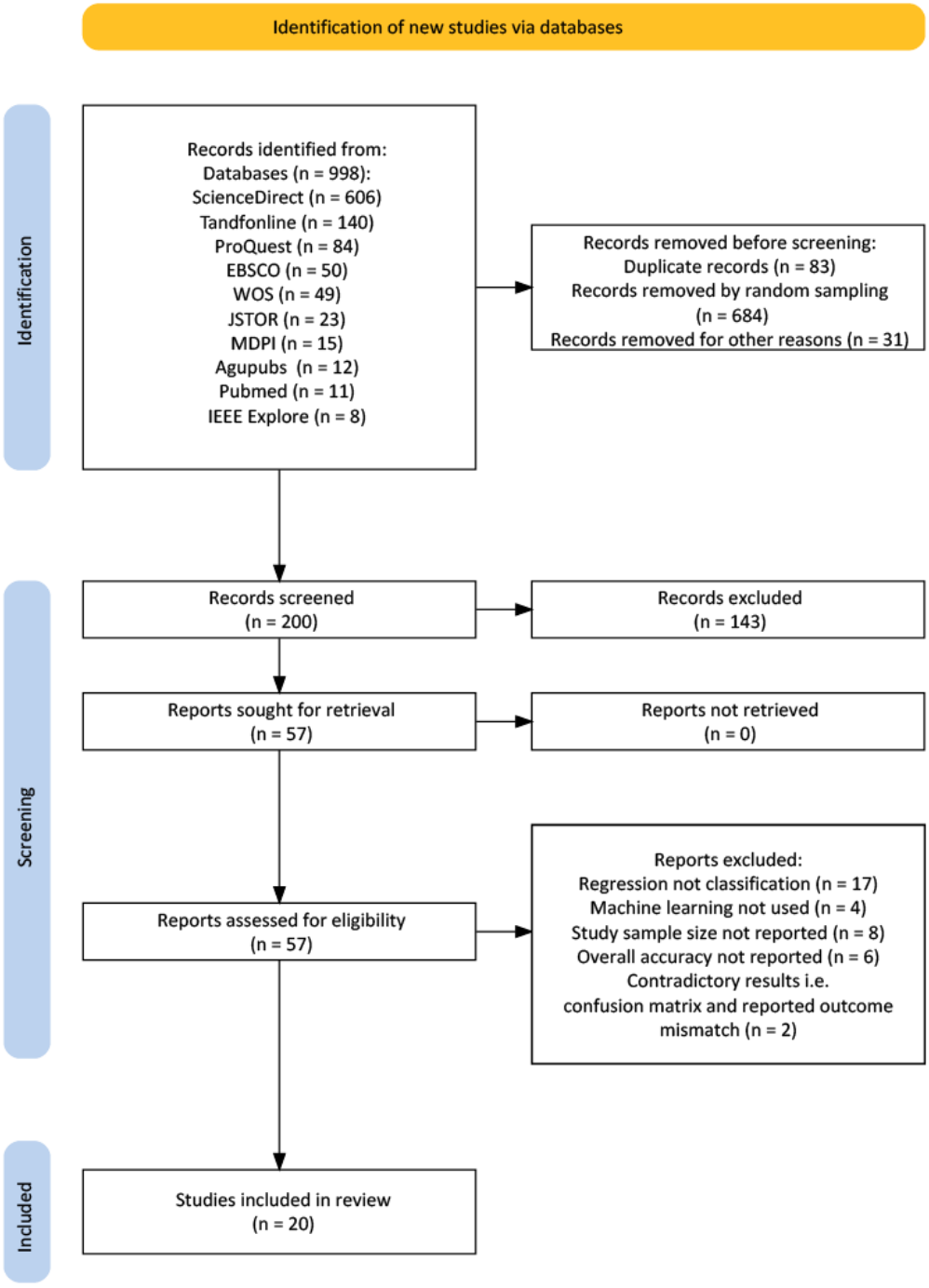


Figure 3.3: PRIMSA flow diagram of manuscript selection. The records were identified from databases including Web of Science (WOS), ScienceDirect, PubMed, Journal Storage (JSTOR), American Geophysical Union Publications (Agupubs), EBSCO, IEEE Xplore, Multidisciplinary Digital Publishing Institute (MDPI), ProQuest, and Taylor & Francis Online (Tandfonline), no papers were gathered from official registers.

3.3 Feature collection

Using the selected papers and previous systematic reviews a list of potential study features was created and structured in a table for data extraction. Table 3.3 outlines all the extracted features and study identification information. The features in the table are grouped according to their use in the analysis. The most frequently reported performance metric, overall accuracy is used as the effect size of interest. The sample size (m) is important for the weighted meta-analysis and was also used as a feature, as larger sample sizes should influence overall accuracy. The other features describe the methodology and data characteristics, which provide information about the complexity of the classification tasks (e.g., the number of output classes) and the proportion of the majority class, indicating potential class imbalance issues that can affect the performance of classification models. Remote sensing-specific information was also gathered, including the type of devices, spectral bands, and spatial resolution to assess how data collection impacts performance. Of the extracted features, the number of spectral bands and spatial resolution were categorized due to high levels of non-reporting. The type of remote sensing device was excluded because only one study did not use satellite data, and the specifics of the spectral bands used were too different to make meaningful groups. Several potentially useful features were not recorded, including temporal resolution (the frequency of data collection) and pre-processing steps, which also impact the performance of the model. These were excluded as the differences between papers were too large to make groups. The number of citations was gathered using the Local Citation Network web app, which collects article metadata from OpenAlex—a bibliographic catalogue of scientific papers (Priem et al., 2022)[1].

[1]The idea to add the number of citations was added after the analysis was mostly completed. This suggestion was made during a discussion of the project after the preliminary results were presented to the methodology team at the CBS.

Table 3.3: Extracted features

Feature	Definition	Ranges/Categories Adopted
Study Identification and Information		
DOI	Paper ID	-
Authors	Name(s) of authors	First author and publication year used as study label.
Title	Title of the article	-
Publication Name	Name of journal that published the paper	-
Location	Location of the data used (country level)	-
Used in Intercept-only Model		
Overall Accuracy	Effect size of interest	0.65 - 1.00
Sample Size	The sample size (i.e.: number of pixels, or objects)	259 - 75,782,016
Features Added to Mixed Effect Model		
Publication Year	Year of publication	2018 - 2023
SDG Theme	Area of research	SDG2: Zero Hunger, SDG11: Sustainable Cities, SDG15: Life on Land
Classification Type	Unit of analysis in the primary study	Object-level, Pixel-level, Unclear
Model Group	Exact algorithm recorded, grouped for analysis	Tree-Based Models, Neural Network, Other
Ancillary Data	Use of non-RS data in the model	Remote Sensing Only, Ancillary Data Included
Indices	Use of indices to enhance analysis	Used, Not Used
Remote Sensing Type	Category of remote sensing	Active, Passive, Combined, Not Reported
Device Group	Specific device extracted, then grouped	Landsat, Sentinel, Other, Not Reported
Number of Spectral Bands	Number of spectral bands used	Low, Mid, Not Reported
Spatial Resolution	Spatial resolution in meters	30, 15-25, 10, <1, Not Reported
Confusion Matrix	Whether a confusion matrix was present	Reported, Not Reported
Number of Classes	The number of classes predicted	2 - 13
Majority-class Proportion	The proportion of the largest class	0.142 - 0.995
Number of Citations	Number of times the study has been cited	0 - 68
Features Excluded		
Device	Type of remote sensing device	Satellite, Aerial Photographic Images
Spectral Bands	Special bands used	-

Note:

The intercept-only model and mixed effect model are described in the following section.

3.4 Statistical analysis

A meta-analysis is a statistical method that aggregates results from several primary studies to assess and interpret the collective evidence on a specific topic or research question. Specifically, the aim is to (a) determine the summary effect, (b) establish the degree of heterogeneity between effect sizes, and (c) access if study characteristics can explain any of the heterogeneity of the effect sizes (Cheung, 2014). In this case the effect size (dependent variable) of interest is the overall accuracy. Let $\hat{\theta}_{ij}$ be the $i-$th observed effect size in study j (where $i = 1, ..., k_j$, $j = 1, ..., n$). From Equation 2.1, the overall accuracy is the proportion of correctly classified instances, therefore, the effect size is:

$$\begin{aligned} \hat{\theta}_{ij} &= \frac{s_{ij}}{m_{ij}} \\ v_{ij} &= \frac{\hat{\theta}_{ij}(1-\hat{\theta}_{ij})}{m_{ij}} \end{aligned} \tag{3.1}$$

s_{ij} is the number of successful predictions and m_{ij} is total number of pixels or objects classified, and v_{ij} is the variance.

Weighted Approach

Before conducting the meta-analysis, first the structure of the collected data and assumption of independence of effect sizes need to be addressed. In the context of this research, dependencies are introduced since all reported effect sizes from each study are included. The degree of dependence between effect sizes can be categorized as either known or unknown (Cheung, 2014). Multivariate meta-analytic techniques use known dependencies reported in the primary studies, such as reported correlation coefficients (Cheung, 2014). However, dependency estimates between outcomes are rarely reported (Assink & Wibbelink, 2016). Therefore, to model these unknown dependencies a 3-level random-effects meta-analytic model is used (Cheung, 2014). The three-level meta-analysis approach models three different variance components distributed over three levels:

At level 1, the sampling variance of the effect sizes is modelled as:

$$\begin{aligned} &\text{Level 1: } \hat{\theta}_{ij} = \theta_{ij} + \epsilon_{ij}, \\ &\epsilon_{ij} \sim \mathcal{N}(0, v_{ij}). \end{aligned} \tag{3.2}$$

The observed overall accuracy $\hat{\theta}_{ij}$ is an estimate of overall accuracy from experiment i in study j and is modelled as the true overall accuracy, θ_{ij} and error component ϵ_{ij} which is normally distributed with mean 0 and known variance v_{ij}. A model that only takes into account sampling variance is referred to as a fixed-effects model, where it is assumed that all studies included in the meta-analysis share a single

true effect size, and therefore, the only source of variation between effect sizes is the sampling variance. The fixed-effects model assumes homogeneity across studies and allows for conditional inference about the specific set of studies included in the analysis, without accounting for variability that might arise from differences between studies. The inclusion of the random effects (at level 2 and 3) means that as well as sampling variance, the heterogeneity due to differing between and within study features are also taken into account (Harrer et al., 2022; Schwarzer et al., 2015, p. 34; Wang, 2023). Therefore, the addition random effect components allow one to make unconditional inferences about the population from which the included studies are a random sample.

At level 2, within-study heterogeneity (σ^2_{level2}) is modelled as:

$$\begin{aligned} &\text{Level 2: } \theta_{ij} = \kappa_j + \zeta_{ij}, \\ &\zeta_{ij} \sim \mathcal{N}(0, \sigma^2_{\text{level2}}). \end{aligned} \tag{3.3}$$

The true overall accuracy θ_{ij}, is modelled as the average overall accuracy κ_j of study j and study-specific heterogeneity ζ_{ij} which is normally distributed with mean 0 and variance σ^2_{level2}.

Lastly, level 3, the variance between heterogeneity (σ^2_{level3}) is modelled as:

$$\begin{aligned} &\text{Level 3: } \kappa_j = \mu + \xi_j, \\ &\xi_j \sim \mathcal{N}(0, \sigma^2_{\text{level3}}). \end{aligned} \tag{3.4}$$

The average overall accuracy κ_j of study j is modelled as the average population effect μ and between-study heterogeneity ξ_j, which is normally distributed with mean 0 and variance σ^2_{level3}. Combined, the three-level meta-analysis models the observed effect size modelled as the sum of the average population effect μ and these three error components:

$$\hat{\theta}_{ij} = \mu + \xi_j + \zeta_{ij} + \epsilon_{ij}. \tag{3.5}$$

For the expected value of the observed effect size to be the population average, $\mathbb{E}(\hat{\theta}_{ij}) = \mu$, the random effects at the different levels and the sampling variance are assumed independent: $\text{Cov}(\xi_j, \zeta_{ij}) = \text{Cov}(\xi_j, \epsilon_{ij}) = \text{Cov}(\zeta_{ij}, \epsilon_{ij}) = 0$. Therefore, unconditional sampling variance of the effect size is the sum of level 3 and level 2 heterogeneity, and the known sampling variance: $\text{Var}(\hat{\theta}_{ij}) = \sigma^2_{\text{level3}} + \sigma^2_{\text{level2}} + v_{ij}$, the effect sizes within the same study share the same covariance $\text{Cov}(\hat{\theta}_{ij}, \hat{\theta}_{lj}) = \sigma^2_{\text{level3}}$, and the effect sizes in different studies are independent $\text{Cov}(\hat{\theta}_{ij}, \hat{\theta}_{zu}) = 0$ (Cheung, 2014)[2].

The random-effects model can be extended to a mixed-effects model (also referred to as a meta-

[2]Like i, l refers to an effect size within the same study j. z and u refer to effect sizes in different clusters, where $u \neq j$ effect sizes are independent.

regression) by including study features as covariates (predictors). Let x denote the value covariate, where b' refers to the number of covariates included in the model. These covariates can be either x_{ij} for a level-2 covariate or x_j for a level-3 covariate. The mixed-effect model defined as:

$$\hat{\theta}_{ij} = \mu + \beta_1 x_{ij1} + ... + \beta_{b'} x_{jb'} + \xi_j + \zeta_{ij} + \epsilon_{ij} \tag{3.6}$$

The assumptions for Equation 3.6 remain the same as Equation 3.5, but the heterogeneity ($\sigma^2_{\text{level3}}, \sigma^2_{\text{level2}}$) is the variability among the true effects which is not explained by the included covariates (Cheung, 2014; Viechtbauer, 2010). The aim of the mixed-effects model is to examine the extent to which the included covariates in the model influence the overall summary effect (population average) μ and the heterogeneity σ^2_{level3} and σ^2_{level2} (Viechtbauer, 2010). Figure 3.4 illustrates this structure of the three-level random-effects meta-analysis model used to account for both within-study and between-study heterogeneity.

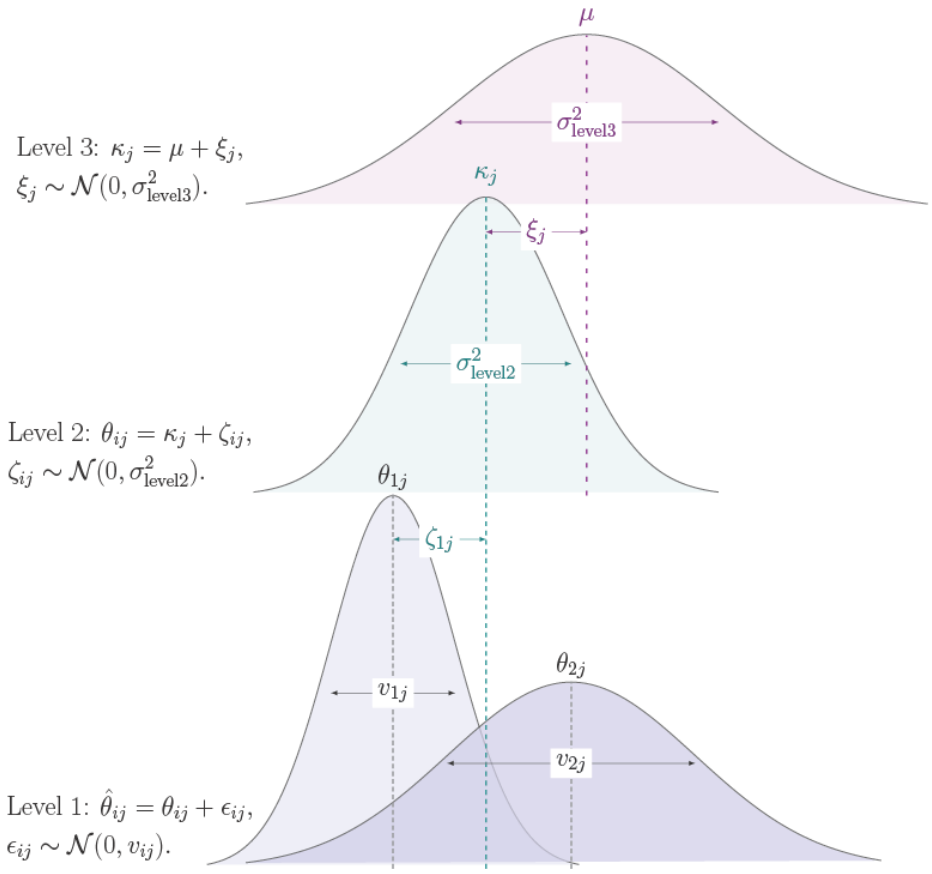


Figure 3.4: Three-level random-effects meta-analysis model. At level 1, observed effects $\hat{\theta}_{ij}$ are modeled with known sampling variance v_{ij}, where larger sample sizes m_{ij} have smaller sampling variances, represented by the narrower distribution around $\hat{\theta}_{1j}$ compared to $\hat{\theta}_{2j}$. At level 2, the true effects θ_{ij}, from each study are modeled as normally distributed with mean κ_j and within-study variance σ^2_{level2}. Lastly, at level 3, study average effects are modeled as normally distributed with mean μ and between-study variance σ^2_{level3}.

In this way, meta-analytic models are essentially, special cases of the general linear (mixed effects) model with heteroscedastic sampling variances which are assumed to be known (Viechtbauer, 2010). Therefore, the random- and mixed-effects models are fit by first by estimating the amount of (residual) heterogeneity (σ^2_{level2} and σ^2_{level3}), and then, the parameters defined above are estimated via weighted least squares with weights. There are several methods to estimate σ^2_{level2} and σ^2_{level3} heterogeneity — see Veroniki et al. (2015) for different methods and specifics. This study uses the (restricted) maximum likelihood method (ML and REML). The estimated heterogeneity terms are then used to aggregate the primary study results using inverse-variance weighting (Borenstein et al., 2009). In inverse-variance weighting, the effect size estimates with the lowest variance (higher sample sizes) are given more weight because they are more precise (Viechtbauer, 2010). If the model was only taking into account the sampling variance then the weights are equal to $w_{ij} = 1/v_{ij}$. In this case there are three sources of heterogeneity the sum of which the is the model implied variances of the estimates: $w_{ij} = 1/(\hat{\sigma}^2_{\text{level3}} + \hat{\sigma}^2_{\text{level2}} + v_{ij})$. However, covariance between the effects needs to be taken into account, therefore the marginal variance-covariance matrix of the estimates.

To calculate the weights, let $\mathbf{y}$ be a the vector of observed effects ($\hat{\theta}_{ij}$) of length k ($\mathbf{y} = \hat{\theta}_1, ..., \hat{\theta}_k$). The observations are organized as a series of independent groups, where the marginal variance-covariance matrix ($\mathbf{M}$) of the estimates account for the variance structure of the data. Since the effect sizes from different studies are assumed to be independent, the matrix takes a block-diagonal form. Where each block corresponds to a single study, with the diagonal elements representing the total variance for each outcome, and the off-diagonal elements within each block representing the shared between-study variance. The blocks themselves are independent, reflecting the assumption that there is no covariance between outcomes from different studies.

$$\mathbf{M} = \begin{pmatrix} \hat{\sigma}^2_{\text{level3}} + \hat{\sigma}^2_{\text{level2}} + v_1 & \hat{\sigma}^2_{\text{level3}} & ... & 0 & 0 \\ \hat{\sigma}^2_{\text{level3}} & \hat{\sigma}^2_{\text{level3}} + \hat{\sigma}^2_{\text{level2}} + v_2 & ... & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & ... & \hat{\sigma}^2_{\text{level3}} + \hat{\sigma}^2_{\text{level2}} + v_{\text{k-1}} & \hat{\sigma}^2_{\text{level3}} \\ 0 & 0 & ... & \hat{\sigma}^2_{\text{level3}} & \hat{\sigma}^2_{\text{level3}} + \hat{\sigma}^2_{\text{level2}} + v_{\text{k}} \end{pmatrix} \tag{3.7}$$

Let $\mathbf{W} = \mathbf{M^{-1}}$ be the weight matrix, where, w_{rc} correspond to the r-th row and the c-th column of $\mathbf{W}$ and let $\hat{\theta}_r$ denote the r-th estimate, with $r = 1, ..., k$[3]. Then the estimate of summary effect size $\hat{\mu}$ for the random-effects model, without covariances, i.e., intercept-only model, is given by (Pustejovsky, 2020; Viechtbauer, 2020)

[3]From this point the index r is used for conciseness rather than indexing for the effect size number within each study.

$$\hat{\mu} = \frac{\sum_{r=1}^{k}(\sum_{c=1}^{k} w_{rc})\hat{\theta}_r}{\sum_{r=1}^{k}\sum_{c=1}^{k} w_{rc}}$$
$$\text{with} \tag{3.8}$$
$$\bar{\sigma}^2 = \text{Var}(\hat{\mu}) = \frac{1}{\sum_{r=1}^{k}\sum_{c=1}^{k} w_{rc}}$$

This is equivalent to the generalized least squares estimate for the fixed effects (Viechtbauer, 2020);

$$\mathbf{b} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{y} \tag{3.9}$$

$\mathbf{X}$ is the design matrix corresponding to the fixed effects, in the random-effects model case this is a single column of 1's as there are no predictors, but in the mixed effects model, $\mathbf{X}$ has $b' + 1$ columns. In the mixed effects case the estimated parameters are μ and b''s in $\mathbf{b}$. Following the recommendation of Assink & Wibbelink (2016), t-distribution was applied to assess the significance of individual regression coefficients in meta-analytic models, as well as to construct confidence intervals.

Heterogeneity tests

To assess the significance of heterogeneity in the true effect sizes, the Cochran's Q statistic is used, with the null hypothesis assuming homogeneity of effect sizes. As defined by Cheung (2014):

$$H_0 : \theta_r = \theta$$
$$Q = \sum_{r=1}^{k} w_r(\hat{\theta}_r - \hat{\mu}_{\text{fixed}})^2$$
$$\text{where } w_r = \frac{1}{v_r}, \tag{3.10}$$
$$\hat{\mu}_{\text{fixed}} = \frac{\sum_{r=1}^{k} w_r\hat{\theta}_r}{\sum_{r=1}^{k} w_r}$$

Under the null hypothesis Cochran's Q has an approximate chi-squared distribution with $k-1$ degrees of freedom. Note, under the null hypothesis there are no cluster effects (no effect of the dependence) therefore the random effect terms are not considered for w_r (Cheung, 2014). The magnitude heterogeneity can be assessed using Higgins and Thompson (2002) I^2, which reflects the proportion of total variation that is not attributable to sampling error (i.e., due to within- and between- study heterogeneity). Therefore I^2_{level2} and Level 3 I^2_{level3} are defined as follows (Cheung, 2014):

$$I^2_{\text{level2}} = \frac{\hat{\sigma}^2_{\text{level2}}}{\hat{\sigma}^2_{\text{level2}} + \hat{\sigma}^2_{\text{level3}} + \tilde{v}}$$
$$I^2_{\text{level3}} = \frac{\hat{\sigma}^2_{\text{level3}}}{\hat{\sigma}^2_{\text{level2}} + \hat{\sigma}^2_{\text{level3}} + \tilde{v}} \tag{3.11}$$

where $\tilde{v}$ is the typical sampling variance. Since the sampling variance differ across studies the typical variance is needed to estimate the magnitude. There are different ways to define the total variation (Cheung, 2014). Here $\tilde{v}$ defined using Higgins and Thompson (2002):

$$\tilde{v} = \frac{(k-1)\sum_{r=1}^{k} \frac{1}{v_r}}{(\sum_{r=1}^{k} \frac{1}{v_r})^2 - \sum_{r=1}^{k} \frac{1}{v_r^2}} \tag{3.12}$$

Lastly, the percentage of variance explained by the mixed-effects can be quantified using R^2 (Cheung, 2014);

$$R^2_{\text{level2}} = 1 - \frac{\hat{\sigma}^2_{\text{level2(1)}}}{\hat{\sigma}^2_{\text{level2(0)}}}$$
$$R^2_{\text{level3}} = 1 - \frac{\hat{\sigma}^2_{\text{level3(1)}}}{\hat{\sigma}^2_{\text{level3(0)}}} \tag{3.13}$$

where, the variance is compared before$_{(0)}$ and after$_{(1)}$ including predictors.

Model Selection

The multi-model inference function from the R package `dmetar` was used to select the best combination of covariates (i.e., the best model). Instead of sequentially adding or removing covariates (step-wise regression methods) this technique models all possible covariate combinations. The number of models fit depends on the number and type of covariates; for b numeric or binary covariates, 2^b models are generated, however, if categorical covariates have multiple levels, the total models increase accordingly — e.g., for 3 categorical variables with 4 levels each and 2 numeric covariates, $4^3 \times 2^2$ (256) models would be fit.

The models are then compared using an information-theoretic approach such as Akaike's Information Criterion (AIC) (Harrer et al., 2022, Chapter 8). The `dmetar` package uses the AIC$_c$ (Corrected Akaike Information Criterion). The AIC$_c$ accounts for small sample sizes, a frequent scenario in meta-analyses or subgroup analyses; in large samples, AIC$_c$ converges to the AIC, so the difference between them diminishes as the sample size increases. AIC($_c$) provides a means for model comparison by balancing model fit with the complexity of the model, where lower AIC values indicate better-performing

models. In addition to the AIC_c, the importance of each covariate is assessed, by summing the Akaike weights (or probabilities) of the models in which the covariate appears (Viechtbauer, 2022). Covariates that frequently appear in high-weight models are assigned higher importance values, indicating their consistent inclusion in the best-performing models(Harrer et al., 2022, Chapter 8; Viechtbauer, 2022). It is important to note that the models will be refit from an REML to ML to make these comparisons (see Harrer et al., 2022, Chapter 8).

Unweighted Approach

The unweighted least squares gives an estimate of the simple (unweighted) average of the population effect, given by (Laird & Mosteller, 1990)

$$\hat{\mu}_{\text{unweighted}} = \frac{\sum \hat{\theta}_r}{k} \tag{3.14}$$

Unlike in the weighted approach methods, the observations from the primary studies, $\hat{\theta}_r$ are not assumed to originate from a distribution. The study results are the unit of analysis rather than the sample components, therefore the level 1 variance component is ignored. The unweighted effects model, focuses on between-study variance (J. A. Hall & Rosenthal, 2018). It achieves standard meta-analysis goals, such as describing central tendency, variance, and moderator effects, through an unconditional random effects approach(J. A. Hall & Rosenthal, 2018). A practical advantage of the unweighted model is that the effect sizes can be analyzed using standard descriptive and inferential statistics, t-tests, ANCOVA (see Khatami et al., 2016) and regression(see O. Hall et al., 2023).

Assumption of normality

It is important to note that the methods outlined above assume that the distribution the effect sizes is approximately normal. If the number of studies collected is sufficiently large and the observed proportions are centred around 0.5, proportions follow an approximately symmetrical binomial distribution, making the normal distribution a good approximation (Wang, 2023). However, in practice observed proportional data is rarely centred around 0.5 (Wang, 2023). In this context in particular, the distribution of overall accuracy is likely skewed to the left as models are designed to maximize predictive power. Although the performance is dependent on the complexity and the quality of the data and some models could perform worse than random, their accuracies will not be much lower than 0.5, while well-performing models can achieve significantly higher accuracies, causing the center of accuracies to be pulled toward 1. In Khatami et al. (2016), the range of collected overall accuracy was between 14.0 to 98.7%, with a median overall accuracy of 81.1% (IQR = $68.9, 89.7$).

To address skewed observed proportions, transformation methods are applied, most commonly the logit or log-odds transformation. However, this method may not be appropriate in cases where the observed proportions are extremely low (near 0) or extremely high (near 1), as the transformations and their sampling variances can become undefined. In such cases, the Freeman-Tukey (FT) transformation is more appropriate, providing a more robust approach to dealing with skewed distributions of overall accuracy, especially when dealing with extreme values (Borges Migliavaca et al., 2020; Wang, 2023). The FT transformation is calculated as (Freeman & Tukey, 1950; Viechtbauer, 2024a):

$$\hat{\theta}_r^{\mathrm{FT}} = g(\hat{\theta}_r) = \frac{1}{2} \cdot \left(\arcsin\sqrt{\frac{s_r}{m_r + 1}} + \arcsin\sqrt{\frac{s_r + 1}{m_r + 1}} \right) \tag{3.15}$$

where $\hat{\theta}_r^{\mathrm{FT}}$ denotes the transformed $\hat{\theta}_r$, with variance:

$$\mathrm{Var}(\hat{\theta}_r^{\mathrm{FT}}) = v_r = \frac{1}{4m_r + 2} \tag{3.16}$$

These transformed parameters replace Equation 3.1, while the rest of the analysis remains the same. To return to the pooled effect sizes natural scale, the Barendregt et al. (2013) back transformation is used, as instructed by Wang (2023):

$$\hat{\mu} = \frac{1}{2} \left(1 - \mathrm{sign}(\cos(2\hat{\mu}^{\mathrm{FT}})) \cdot \sqrt{1 - \left(\sin(2\hat{\mu}^{\mathrm{FT}}) + \frac{\sin(2\hat{\mu}^{\mathrm{FT}}) - 1/\sin(2\hat{\mu}^{\mathrm{FT}})}{1/\overline{\sigma}^2_{\mathrm{FT}}} \right)^2} \right) \tag{3.17}$$

where $\hat{\mu}^{\mathrm{FT}}$ is the summary statistic — pooled overall population average— and $\bar{\sigma}^2_{\mathrm{FT}}$ is the pooled variance, from Equation 3.8 but in the transformed scale (Wang, 2023).

Results

4.1 Descriptive Statistics

A total of $n = 20$ studies with $k = 86$ effect sizes were included in this analysis, with each primary study reporting between one and 27 results ($1 \leq k_j \leq 27$). The research area of these studies span 18 countries; Figure 4.1 shows a map indicating the location of each effect size. These primary studies were grouped into three different SDG goals: SDG 2 Zero Hunger, SDG 11 Sustainable Cities, and SDG 15 Life on Land.

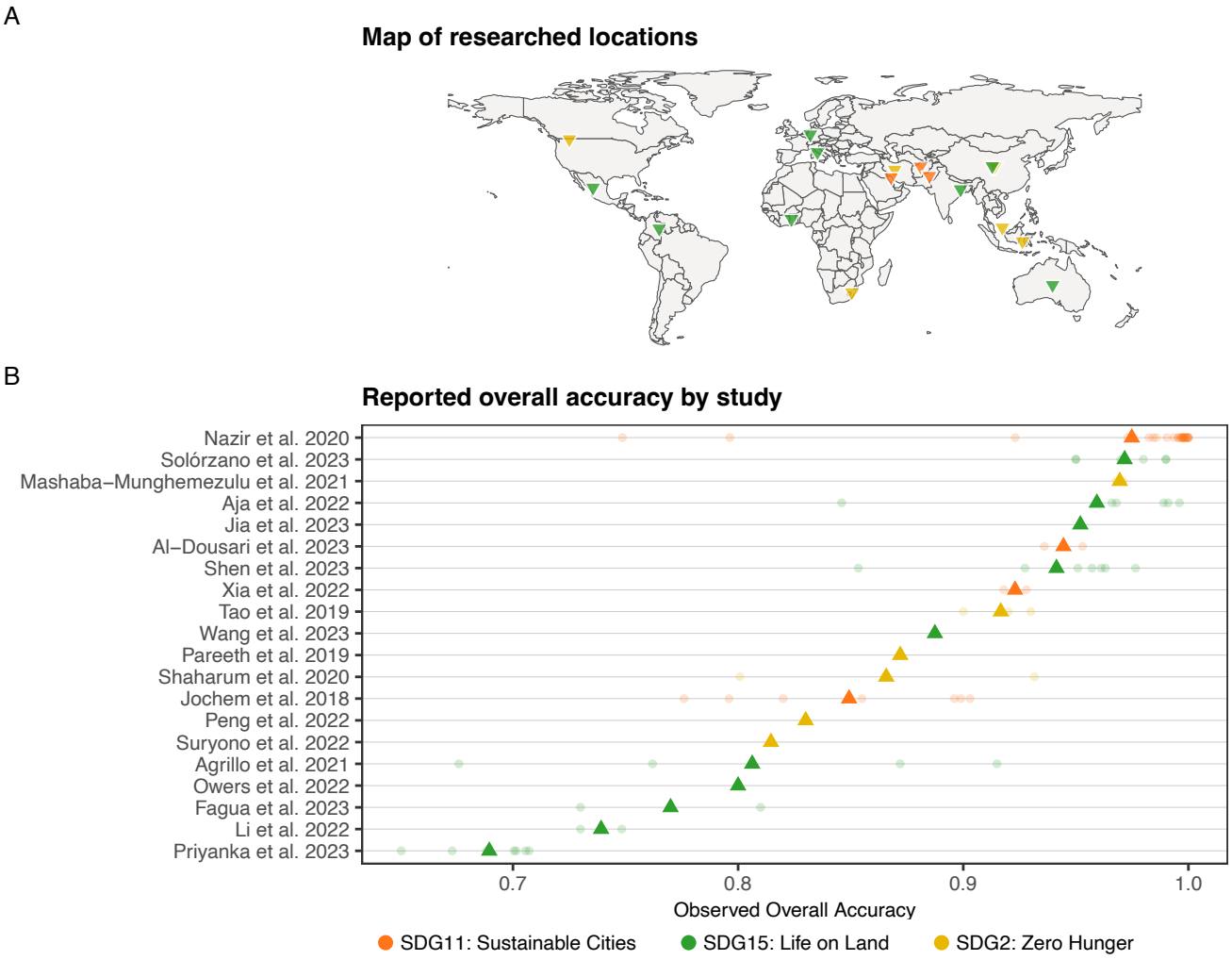


Figure 4.1: A. Country of research interest of the primary study. B. Range of reported overall accuracy, individual outcomes shown as points and mean overall accuracy represented by triangles. In both plots points are colour-coded by SDG goal.

Figure 4.1 and Table 4.1 (bellow) show the reported overall accuracies are not centered around 0.5.

Therefore, a transformation is required. Figure 4.2 shows the distribution of observed overall accuracy as well as the logit and FT transformation values. FT visually performs better than the Logit transformation. However, the Shapiro-Wilk Normality test shows that the distribution of the FT transformed overall accuracy still departed significantly from normality ($W = 0.93$, p-value < 0.01). Nevertheless, conducting a meta-analysis remains justified, as these statistical models are generally robust against violations of normality (McCulloch & Neuhaus, 2011).

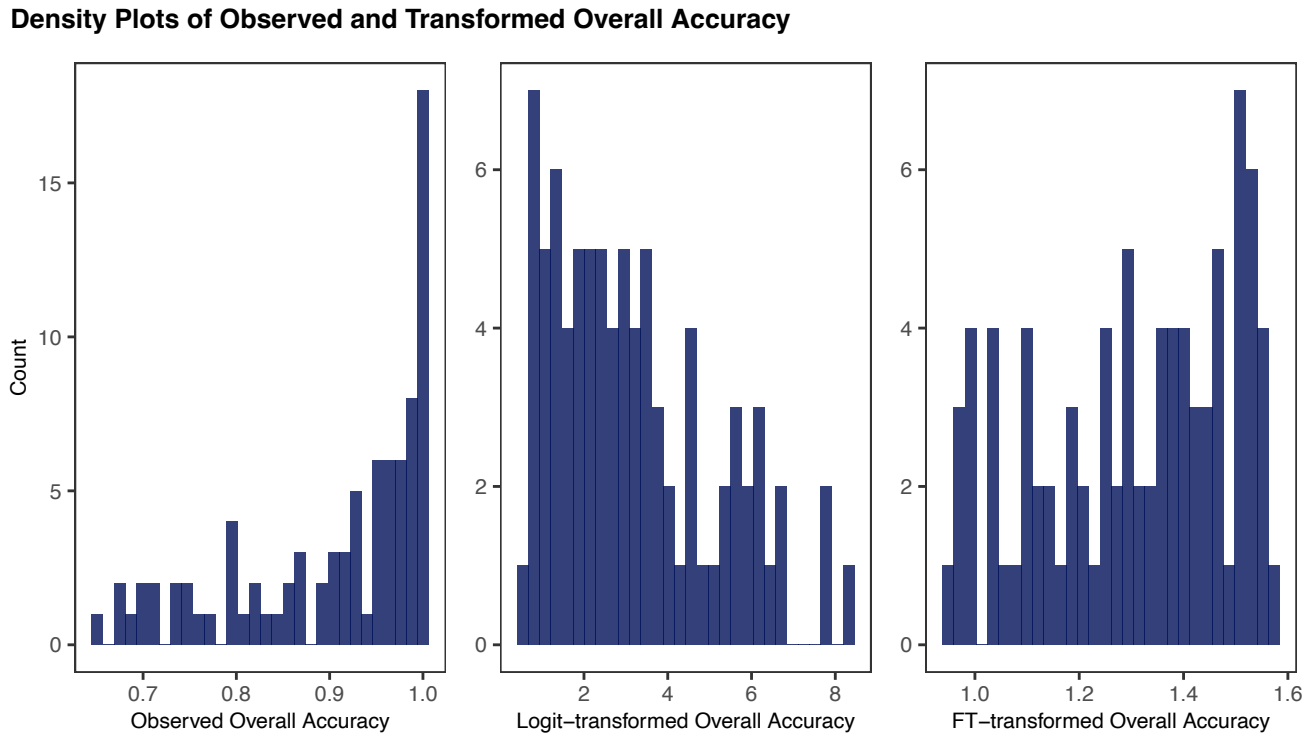


Figure 4.2: Distribution of the observed overall accuracy and transformed by logit and FT transformation.

Table 4.1 summarises the overall accuracy (effect size of interest), study sample size and the collected study features, including the study features such as sample size, overall accuracy, types of machine learning models used and SDG goal targeted. For the meta-analysis the range of the sample size (259 - 75782016) and overall accuracy (0.6504 - 1) are of importance. Most studies used Neural Networks (48%), followed by Tree-Based Models (45%), and a small portion used other types of models (7%). Regarding SDGs, 44% of the studies aimed at SDG 11 (Sustainable Cities), 43% targeted SDG 15 (Life on Land), and 13% focused on SDG 2 (Zero Hunger). Figure 4.3 and Figure 4.3 visualise the distribution of these features.

Table 4.1: Summary table

Feature	Statistic
Overall Accuracy	0.90 (0.65 - 1.00)
Study Features	
Numeric	
Sample Size	6,401,352 (259 - 75,782,016)
Number of Citations	15 (2 - 68)
Number of Classes	4 (2 - 13)
Majority-class Proportion	0.72 (0.14 - 1.00)
Publication Year	
2018	7 (8.1%)
2019	4 (4.7%)
2020	30 (35%)
2021	6 (7.0%)
2022	13 (15%)
2023	26 (30%)
Categorical	
SDG Theme	
SDG11: Sustainable Cities	38 (44%)
SDG15: Life on Land	37 (43%)
SDG2: Zero Hunger	11 (13%)
Classification Type	
Object-level	46 (53%)
Pixel-level	36 (42%)
Unclear	4 (4.7%)
Model Group	
Neural Networks	41 (48%)
Other	6 (7.0%)
Tree-Based Models	39 (45%)
Ancillary Data	
Remote Sensing Only	71 (83%)
Ancillary Data Included	15 (17%)
Indices	
Not Used	23 (27%)
Used	63 (73%)
Remote Sensing Type	
Active	11 (13%)
Combined	7 (8.1%)
Not Reported	7 (8.1%)
Passive	61 (71%)
Device Group	
Landsat	15 (17%)
Not Reported	7 (8.1%)
Other	44 (51%)
Sentinel	20 (23%)
Number of Spectral Bands	
Low	18 (21%)
Mid	26 (30%)
Not Reported	42 (49%)
Spatial Resolution	
<1 metre	7 (8.1%)
10-30 metres	39 (45%)
Not Reported	40 (47%)
Confusion Matrix	
Not Reported	23 (27%)
Reported	63 (73%)

Note:

[a] Effect size of interest. The statistic reported here are mean (range) for numeric[b] predictors and for categorical[c] variables number of effect sizes (percentage)

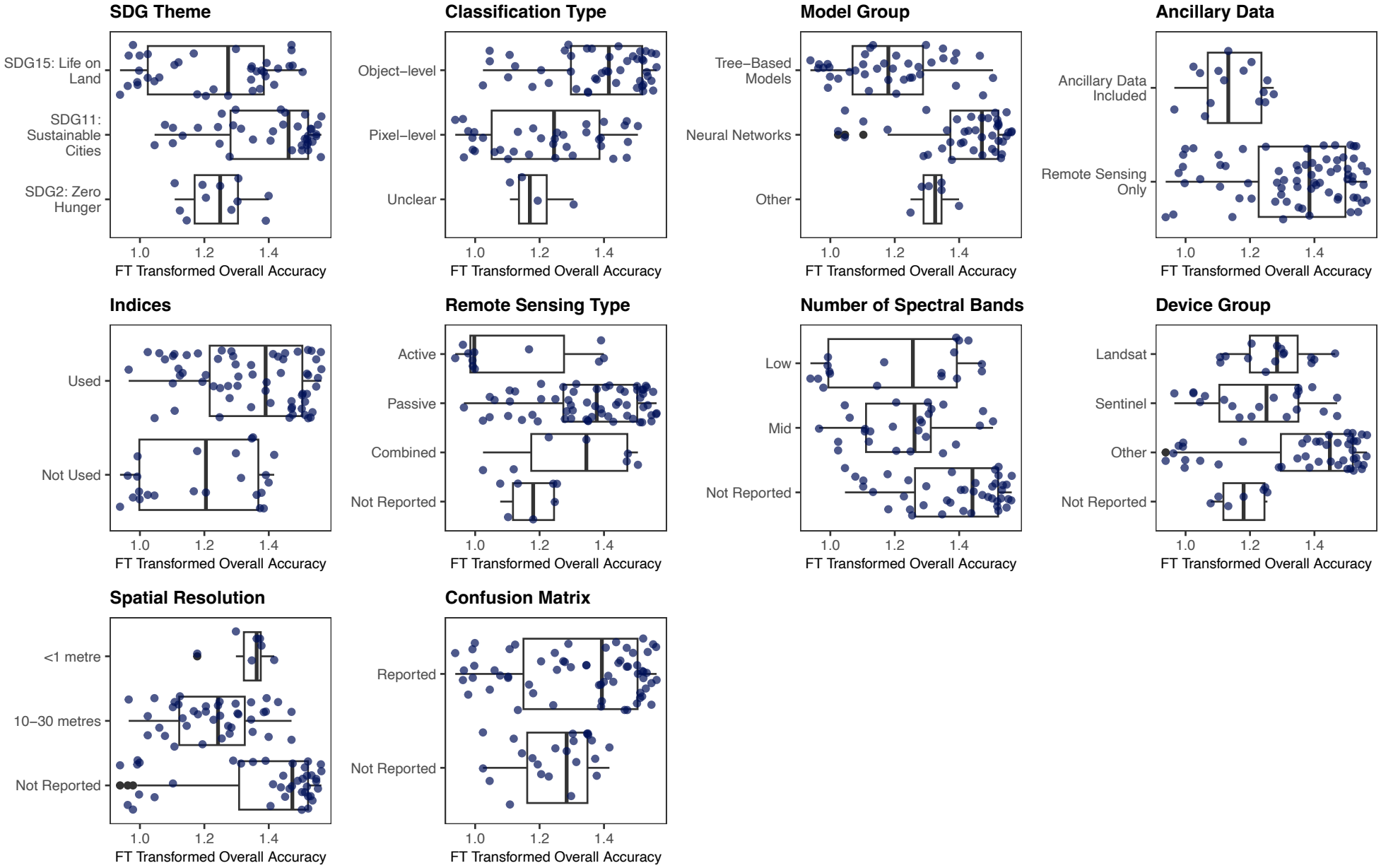


Figure 4.3: Categorical study features

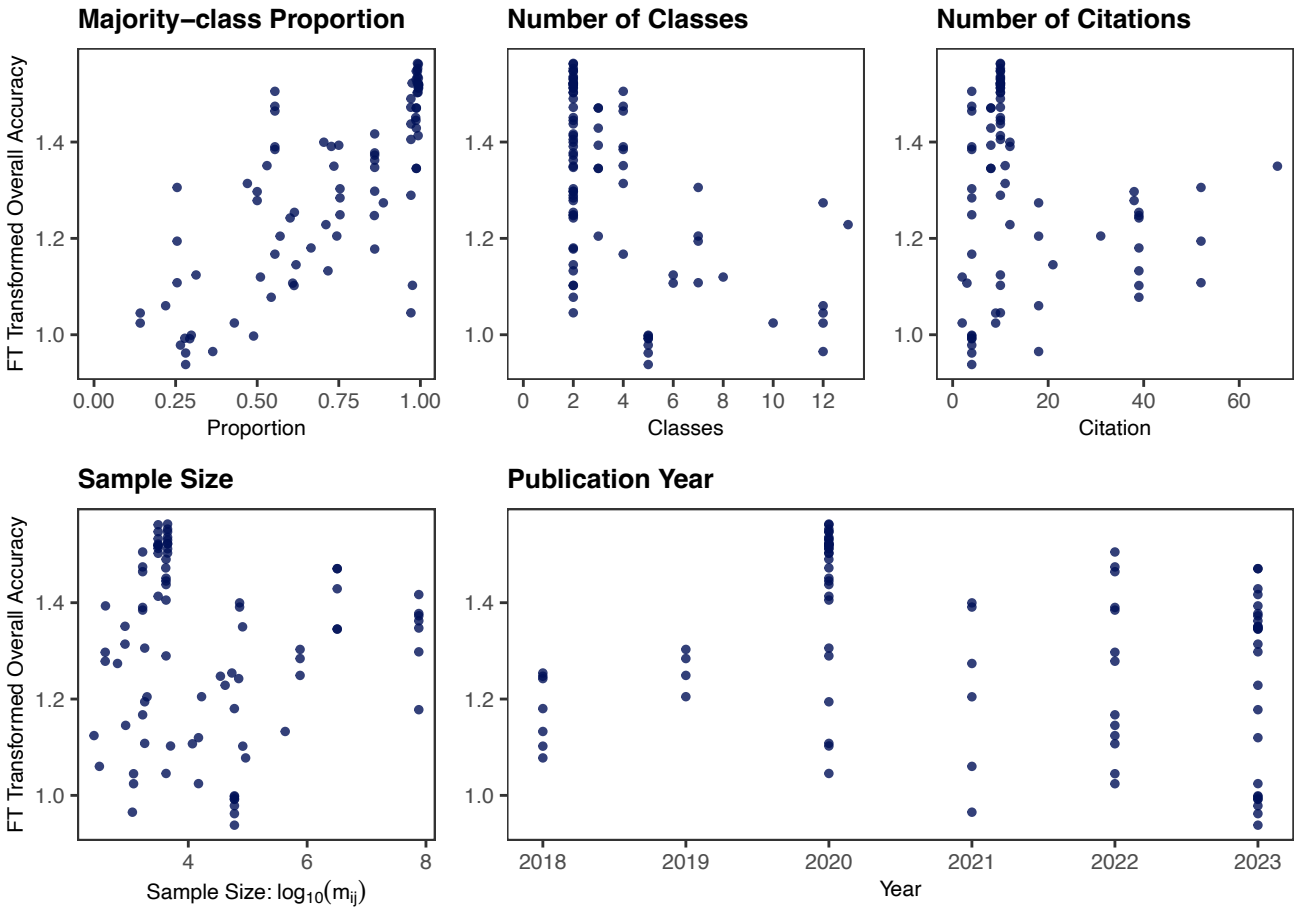


Figure 4.4: Numeric study features

4.2 Meta-analysis

The forest plot below (Figure 4.5) compares the overall accuracy effect size across studies using both weighted and unweighted models, with error bars which correspond to the weighted model — at this scale there is no discernible difference between the error bars of the two models. Each study is given with the number of estimates per study k_j, and study average effect size (κ_j), with 95% confidence intervals (CI), both for the weighted and unweighted model. Of the 20 primary studies included, six reported only one effect. Based on the unweighted model, the average accuracy of machine learning methods applied to remote sensing data is 0.90 (95% CI[0.85; 0.94]). While the three-level meta-analytic model produced an average accuracy of 0.89 (95% CI[0.85; 0.93]). This implies, that on average, the machine learning methods correctly classify around 90% of the time when applied to remote sensing data.

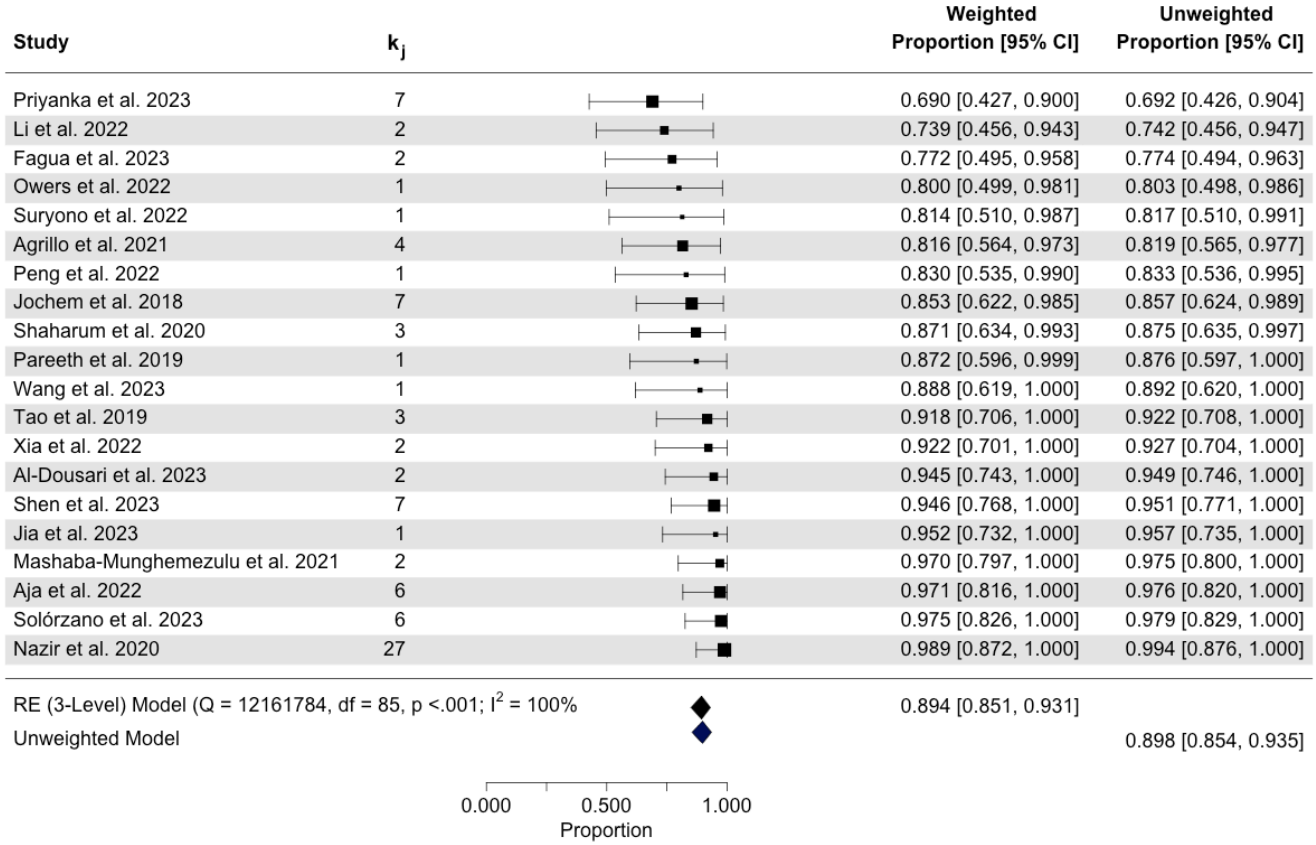


Figure 4.5: Forest plot for both the weighted and unweighted model. k_j is number of reported overall accuracy estimates per study, the corresponding average effect size(κ_j) and confidence interval per study for both models is given on the right. The pooled summary effect size based on the three-level RE meta-analytic and unweighted model are given on the bottom.

The heterogeneity metrics Cochran's Q indicate significant heterogeneity of the reported overall acccuracies. The percentage of the variance attribution is $I^2_{\text{level3}} = 63.62\%$ which is the fraction of the variation that can be attributed to between-study, and $I^2_{\text{level2}} = 36.38\%$ which is within-study heterogeneity, with

negligible fixed effect variance (variance due to sampling error). The I^2 value of 100% indicates that all the observed variability in effect sizes across studies is due to heterogeneity rather than sampling error, suggesting substantial differences between the studies and a high degree of variation in their results.

Model Selection

Using the multi-model inference function, a total of 31,298 models converged. Figure 4.6, illustrates the predictor importance after evaluating all possible combinations of predictors to identify which combination provides the best fit and which predictors are most influential. Higher importance values indicate more consistent inclusion in high-weight models. The majority class proportion is the most important predictor, followed by the inclusion of ancillary data. Less influential predictors include the use of indices, sample size, publication year, and the number of classes in the study. Meanwhile, factors such as classification type, SDG goal, machine learning group, spatial resolution, and citation count have minimal importance in the overall model performance (i.e., were not included in the models top performing models according to AIC_c).

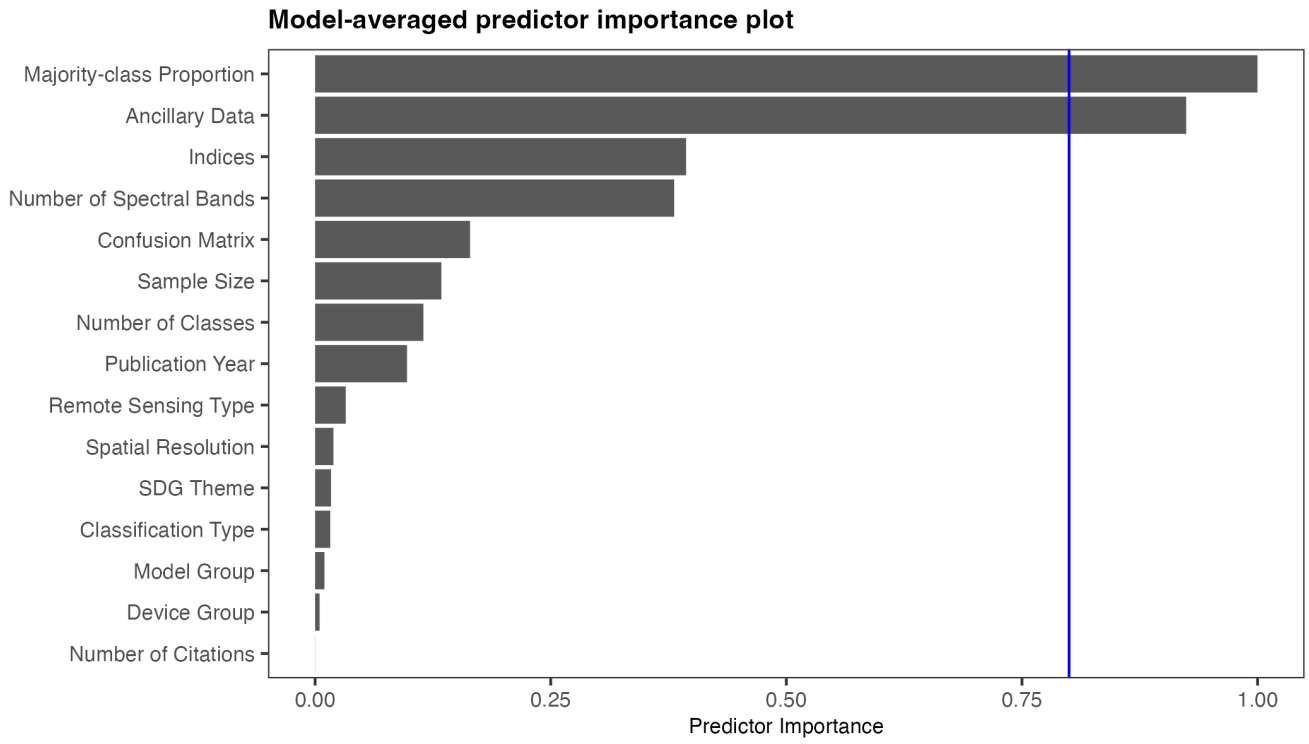


Figure 4.6: Model-averaged predictor importance plot with a reference line at 0.8 a commonly used as a threshold to indicate important predictors.

In the multimodel inference analysis, the five best-performing models were identified based on their AIC_c scores. The selected top models consistently included key predictors such as the majority-class proportion and the use of ancillary data. Table 4.2 shows the results of the multi-model inference. The significant study features are the majority-class proportion and the inclusion of ancillary data.

Interestingly, the use of ancillary data has a negative effect on overall accuracy in the FT transformed scale. Table 4.3 shows the five best performing models and the intercept-only model (before adding the predictors), note that the AIC_c is very similar among the top five. The Akaike weights shown are derived from the total pool of models.

Table 4.2: Multi-model inference coefficients and feature importance.

Importance	**Feature (Reference Category)**	**Comparison Category**	**b**	**SE**	**z**	**p**
1.00	Intercept		1.29	7.85	0.16	0.869
	Majority-class Proportion		0.47	0.08	6.15	<0.001
0.92	Ancillary Data (Remote Sensing Only)	Ancillary Data Included	-0.12	0.05	2.33	0.020
0.39	Indices (Not Used)	Used	0.03	0.04	0.67	0.500
0.38	Number of Spectral Bands (Low)	Mid	0.05	0.06	0.72	0.471
		Not Reported	0.02	0.04	0.55	0.581
0.16	Confusion Matrix (Not reported)	Reported	0.01	0.02	0.29	0.776
0.13	Sample Size		0.00	0.00	0.10	0.922
0.11	Number of Classes		0.00	0.00	0.19	0.846
0.10	Publication Year		0.00	0.00	0.06	0.952
0.03	Remote Sensing Type (Active)	Passive	0.00	0.02	0.16	0.870
		Combined	0.01	0.03	0.17	0.869
		Not Reported	0.00	0.02	0.04	0.971
0.02	Spatial Resolution (<1 metre)	10-30 metres	0.00	0.07	0.01	0.990
		Not Reported	0.00	0.07	0.00	0.996
0.02	SDG Theme (SDG11:Sustainable Cities)	SDG2: Zero Hunger	0.00	0.01	0.08	0.939
		SDG15: Life on Land	0.00	0.01	0.10	0.924
0.02	Classification Type (Object-level)	Pixel-level	0.00	0.01	0.06	0.955
		Unclear	0.00	0.01	0.05	0.958
0.01	Model Group (Neural Networks)	Tree-Based Models	0.00	0.01	0.05	0.961
		Other	0.00	0.01	0.04	0.965
0.00	Device Group (Landsat)	Sentinel	0.00	0.01	0.05	0.959
		Not Reported	0.00	0.01	0.05	0.956
		Other	0.00	0.00	0.05	0.963
0.00	Number of Citations		0.00	0.00	0.01	0.995

Note:

Importance of each feature, the reference and comparison categories given with their estimated coefficients(b), standard errors (SE) on the FT transformed scale with corresponding z- and p-values.

Table 4.3: Set of 5 best-ranked models and intercept only model ordered by AIC_c.

Candidate models	**df**	**AIC_c**	**Akaike weights**
Ancillary Data + Majority-class Proportion + Indices	5	-115.46	0.39
Ancillary Data + Majority-class Proportion + Number of Spectral Bands	6	-114.42	0.23
Ancillary Data + Majority-class Proportion	4	-114.13	0.20
Ancillary Data + Confusion Matrix + Majority-class Proportion + Number of Spectral Bands	7	-113.08	0.12
Ancillary Data + Majority-class Proportion + Number of Spectral Bands + Sample Size	7	-111.65	0.06
Intercept-Only	2	-41.93	0.00

Table 4.4 shows the estimated coefficients for the best-fit model — i.e., the model with the lowest AIC_c value among the candidate models. The coefficients are presented both in the FT-transformed scale($\mathbf{b}$) and on the natural (back-transformed) scale. The results highlight that the proportion of the majority class has the largest positive effect ($b = 0.39$, $b^{BT} = 0.15$, $p < .001$). Suggesting that increasing the majority-class proportion significantly improves overall accuracy. While, the inclusion of ancillary data has a small negative effect on the FT-transformed scale ($b = -0.11, p = 0.029$) but shows a slight positive effect once back-transformed ($b^{BF} = 0.01$). The use of indices has a minimal and non-significant effect ($p = 0.131$).

Table 4.4: Table of estimated coefficients for the best-fit model.

					back-transformed scale	
Predictor	**b**	**SE**	**t**	**p**	**b^{B-FT}**	**CI**
Intercept	0.99	0.06	17.22	<.0001	0.70	[0.58, 0.80]
Majority-class Proportion	0.39	0.08	4.93	<.0001	0.15	[0.05, 0.27]
Ancillary Data: Included	-0.11	0.05	-2.22	0.029	0.01	[0.04, 0.00]
Indices: Used	0.06	0.04	1.53	0.131	0.00	[0.00, 0.02]

Note:

The estimated coefficients (b), standard errors (SE) on the FT transformed scale, with corresponding t-statistics and p-values. Additionally, the coefficients (b^{B-FT}) and corresponding confidence intervals (CI) are shown on the back-transformed scale.

To assess the impact of the study features on the estimated heterogeneity the features included in the best-fit model are fitted as sole covariates. Table 4.5 shows the parameter estimates from the meta-analysis, comparing the intercept-only model with four mixed-effects models, one for each of the features in the best-fit model and the best-fit model itself.

Table 4.5: Results for heterogeneity and covariates tests for the intercept-only model, individual covariates, as well as the best-fit model.

Model	$\sigma^2_{\textbf{level2}}$	$\sigma^2_{\textbf{level3}}$	$Q_E \times 10^7$	**df**	p_Q	F	**df**	p_F	$I^2_{\textbf{level2}}$	$I^2_{\textbf{level3}}$	$R^2_{\textbf{level2}}$	$R^2_{\textbf{level3}}$
Intercept-only	0.010	0.017	12.16	85	<.0001				36.38	63.62		
Majority-class Proportion	0.009	0.007	11.46	84	<.0001	27	1	<.0001	57.29	42.71	7.85	60.71
Ancillary Data	0.010	0.015	12.04	84	<.0001	3	1	0.117	40.47	59.53	-1.44	14.66
Indices	0.010	0.018	11.99	84	<.0001	3	1	0.100	34.26	65.74	3.60	-5.75
Combined model[a]	0.009	0.005	11.44	82	<.0001	13	3	<.0001	63.46	36.54	8.64	69.92

Note:

Test statistics, degrees of freedom, and respective p-values are provided for the intercept-only model, single-predictor models for each of the predictors in the best model, as well as the combined model (Q_E values are divided by 10^7 for compactness).

[a] Combined model: Ancillary Data + Majority-class Proportion + Indices

As shown in Table 4.5, the majority-class proportion explains a greater proportion of the between-study heterogeneity, as indicated by the reduction in σ^2_{level2} between the intercept-only model and the model with the Majority-class Proportion. In contrast, the use of Ancillary Data explains relatively little between-study heterogeneity and negligible within-study heterogeneity.

The combined model (best-fit model) explains the most heterogeneity overall, as reflected in the shift in I^2 values. The total I^2, consistently at 100% across all models, suggests that nearly all the variation in effect sizes is due to differences between the studies, rather than sampling error. This observation raises the possibility of an "apples and oranges" problem (see the discussion section), where the included studies may be too heterogeneous to be meaningfully compared.

All models show significant heterogeneity, with Cochran's Q test results being significant ($p < 0.001$). The R^2 values indicate that the covariates in the combined mixed-effects model account for 69.9% of the variance at level 3 (between-study level) and 8.6% of the variance at level 2 (within-study level).

Figure 4.7 illustrates the relationship between the proportion of the majority class and the overall accuracy of the individual studies included in the meta-analysis. The plot is based on the combined mixed-effects model, where the solid black line represents the fitted regression line, and the shaded area indicates the 95% confidence interval. Each point (or bubble) represents an individual study, with the size of each bubble proportional to the weight it received in the analysis (i.e., larger points represent studies that had more influence on the overall results). The plot demonstrates a clear trend: as the proportion of the majority class increases, overall accuracy tends to improve, indicating a positive correlation between these two variables.

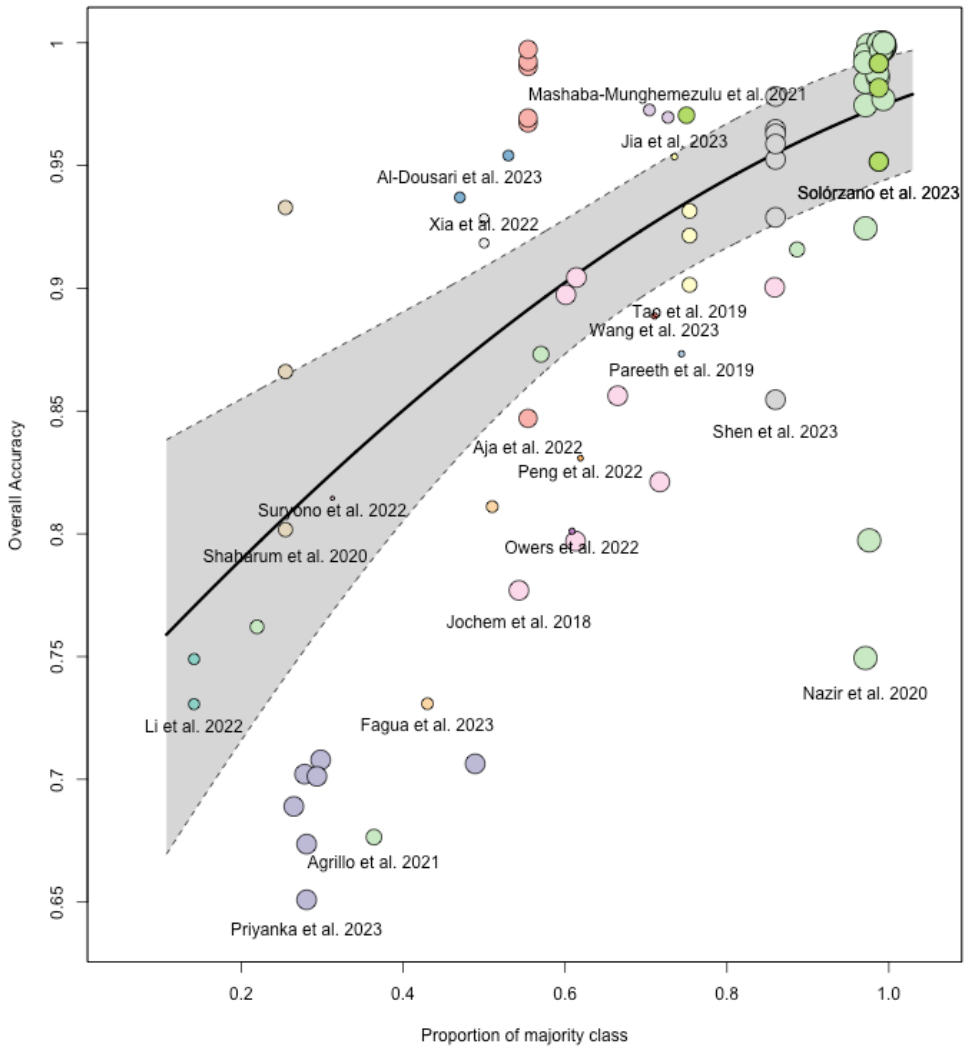


Figure 4.7: Bubble plot showing the observed effect size (overall accuracy) of the individual studies plotted against the proportion of the majority class. Based on the mixed-effects model, the plot displays overall accuracy as a function of the majority class proportion, with corresponding 95% confidence interval bounds. The size of the points is proportional to the weight that each observation received in the analysis, while the color of the points is unique to each study. The lowest overall accuracy from each study is labeled with the first author and publication year.

The size of the points in the bubble plot illustrates the benefit of incorporating the structure of the data into meta-analytic weighting. Specifically, the difference in the size of the bubbles is not excessive. Figure 4.8 highlights this by plotting the weights for each study from a fixed-effects model, a random-effects model with two levels, and the structure used here, the random-effects model with three levels. As shown, the fixed-effects model is problematic, particularly as one study is heavily weighted, which can distort the overall results. In contrast, the two-level and three-level models distribute the weights more evenly across studies, reflecting the importance of accounting for between-study heterogeneity and within-study variation.

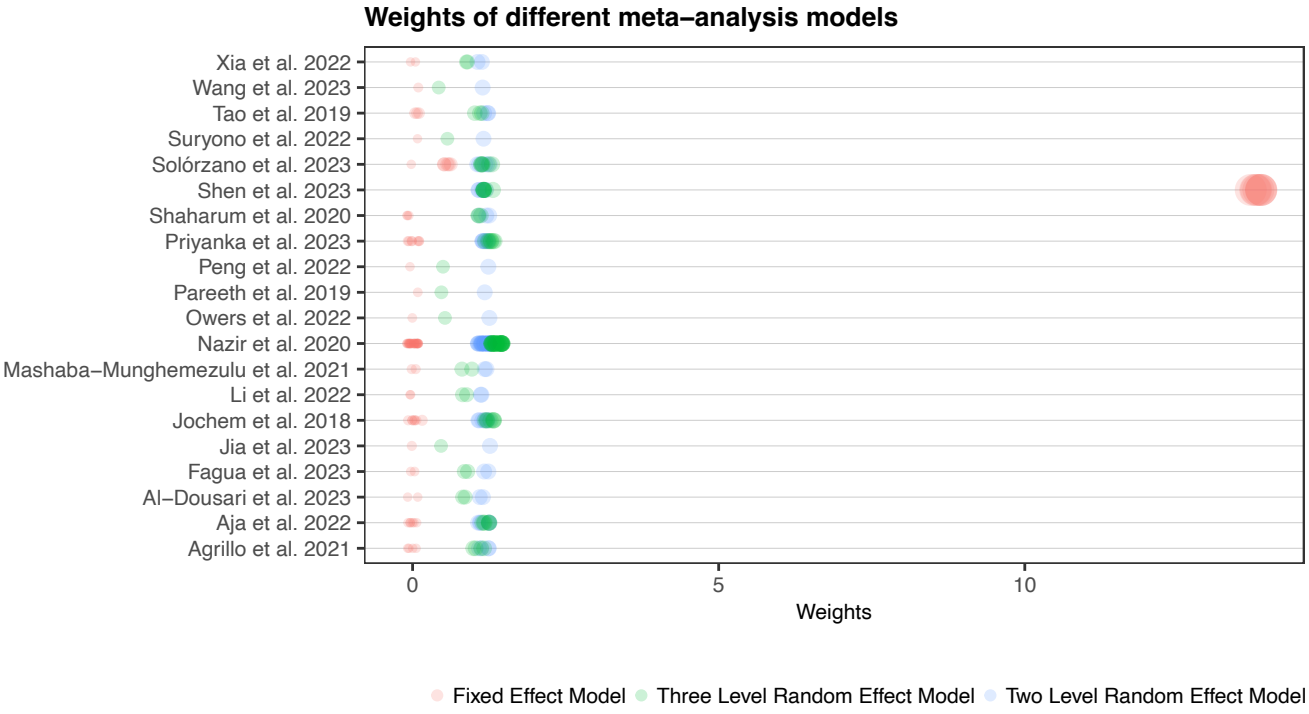


Figure 4.8: Intercept only models at three different levels plot to compare weighting. The size of the points corresponds to the weights of each of the effect sizes.

Lastly, Figure 4.9 is a plot of the observed overall accuracy against the predicted overall accuracy from the combined meta-regression model. The points are coloured based on whether ancillary information was included in the primary study. As Figure 4.9 illustrates, the meta-regression model tends to overestimate overall accuracy — the fitted regression line (in grey) lies above the line of perfect agreement (y = x, in black), indicating that the model's predictions are generally higher than the observed accuracy values.

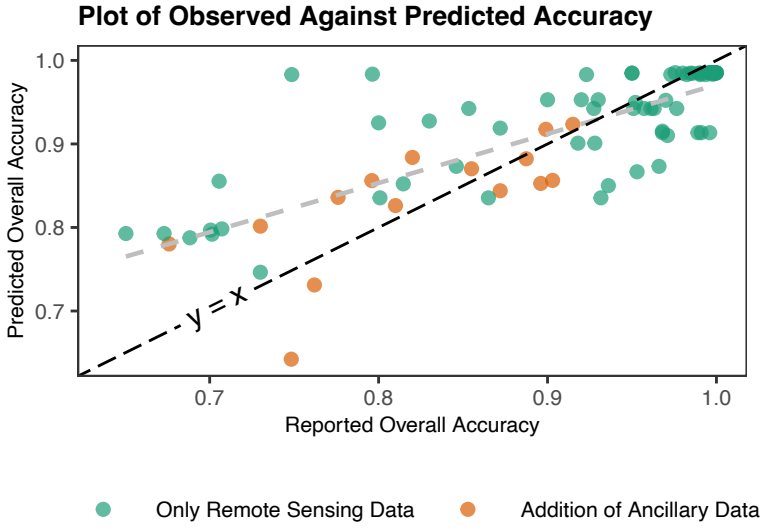


Figure 4.9: Observed and predicted overall accuracy. The colour indicates the addition of ancillary data in the primary study's model. The black dashed line represents perfect agreement between observed and predicted values $y = x$. The grey dashed line is a simple linear regression of observed versus predicted values, showing the overestimation of overall accuracy by the meta-regression model.

Discussion

This meta-analysis aimed to evaluate the performance of machine learning models applied to remote sensing for SDG monitoring. Specifically, the study aimed to estimate the average performance, determine the level of heterogeneity within and across studies, assess whether specific study features influence model performance, and lastly compare the sample-weighted and unweighted estimate summary effect. While previous meta-analyses on machine learning models for remote sensing have predominantly relied on unweighted approaches (O. Hall et al., 2023; e.g., Khatami et al., 2016), this study found that incorporating a weighted approach did not significantly alter the results. Both the weighted and unweighted estimates showed similar average performance metrics, suggesting that weighting by sample size may not dramatically influence the outcomes in this context.

The results from this meta-analysis show that the overall accuracy of machine learning models applied to remote sensing is consistently high. The estimated average overall accuracy of $\hat{\mu}_{\text{unweighted}} = 0.90$ and $\hat{\mu}_{\text{weighted}} = 0.89$. The results also demonstrate a considerable variability in the predictive performance of machine learning models applied to remote sensing data for SDGs. Some of this variability could be attributed to the proportion of the majority class as well as the inclusion of ancillary data. The type of model, whether neural networks and tree-based models or the SDG studied, showed no differences in overall accuracy. Unsurprisingly, the proportion of the majority class significantly affected the overall accuracy of machine learning models. While the use of ancillary data in primary studies has a small but significant positive effect on overall accuracy performance. No other significant effects were found in the study features examined in this study.

The findings of this study regarding the use of ancillary data aling with Khatami et al. (2016) and Hanadé Houmma et al. (2022) who found the use of ancillary data did improve model performance. Some effect of the choice of machine learning model was found by previous research. For example, Khatami et al. (2016) noted that while support vector machines and neural networks performed well, differences between other model types were not significant. Notably, no study was found that explicitly corrected for class imbalance (proportion of the majority class) when assessing the difference in performance between groups. While Khatami et al. (2016) employed pairwise comparisons, which does ensure that models are compared within the same data context, this study goes further in directly highlighting the influence of class proportion on overall accuracy.

Limitations

1. **Number of reviewers:** From the 200 studies randomly sampled, three reviewers assessed whether full-text screening should be conducted. Only 57 papers were agreed upon by all three reviewers, while each reviewer thought between 77 and 81 studies could have been included. This highlights the subjectivity of the selection process and the importance of having multiple reviewers. The full-text screening was only conducted by one person which means that this subjectivity or potential mistakes were missed in the final dataset. This issue is exasperated by the inconsistent reporting on methods in this field. For example, one feature that could not be included in the analysis was whether the results reported were derived from the training or test set because it was very unclear in some of the selected studies.

2. **Sample size:** This analysis included a total of 20 studies. While several simulation studies suggest that a three-level meta-analysis can yield accurate results with as few as 20 to 40 studies (Hedges et al., 2010), this analysis is at the lower bound, and the included studies exhibit considerable variability, making the statistical power a concern. Polanin (2014) suggests a minimum of 40 studies is generally recommended to ensure robust results. Furthermore, a relatively high proportion of the studies (6 out of 20) reported only one result ($k_j = 1$), limiting the ability to assess within-study variability. The small sample size inherently increases the potential for bias and may affect the reliability of the findings (Polanin, 2014).

3. **Choice of effect size:** While overall accuracy is widely used, it does not capture the complexity of model performance, especially in studies with imbalanced classes. To illustrate the problem, if 99% of the data belongs to class A, a model that always predicts class A—without any regard to the predictors—will achieve an overall accuracy of 99%, despite essentially doing nothing and failing to capture meaningful patterns. For more specific details on the issues related to the use of overall accuracy, see Foody (2020) and Stehman & Foody (2019). Alternative metrics include Matthews' correlation coefficient, F1 score, Somers' D, and average precision. Unfortunately, these metrics are rarely reported in the studies analyzed here. Moreover, some of these alternatives are also sensitive to class imbalance and must be corrected to ensure comparability across studies (Burger & Meertens, 2020).

4. **Publication bias:** This study only examined published results, which introduces publication bias—a well-documented effect where studies with positive results are more likely to be published, while negative or neutral findings remain unpublished (Borenstein et al., 2009; Bozada et al., 2021; Hansen et al., 2022b; Harrer et al., 2022). This bias can lead to an overestimation of effects, as demonstrated in this study, where the average overall accuracy is around 90%.

5. **Study features included:** The analysis would have benefited from the inclusion of more study

features. It is also important to note that most of the study features included in this research were between-study covariates and did not differ within studies, which explains why only the between-study heterogeneity was reduced. Furthermore, due to the small sample size, it was necessary to aggregate the study features into broad categories, which limited the granularity of the analysis.

6. **Apples and oranges problem:** The I^2 result of effectively 100% may indicate that the included studies are too different to statistically compare. This is often referred to as the "apples and oranges problem" (Harrer et al., 2022, Chapter 1). The extent to which primary studies can differ while still being meaningfully combined in a meta-analysis is debated. However, when Robert Rosenthal, a pioneer in meta-analysis, was asked whether combining studies with significant differences is valid his response was *"combining apples and oranges makes sense if your goal is to produce a fruit salad"* (Borenstein et al., 2009, Chapter 40, pp. 357). In this case, despite the diverse research aims of the included studies, the objective is to draw general conclusions about machine learning applications in remote sensing for SDG monitoring. This approach can be viewed as a "fruit salad" with potential for broad applicability across different SDG contexts. However, this again raises the issue of sample size, as a large sample is required to ensure sufficient statistical power to draw confident conclusions.

7. **Cochran's Q and large sample sizes:** Another limitation is the reliance on Cochran's Q for testing heterogeneity. While widely used, the power of the Q-statistic is dependent on the number of included effect sizes (k) and the precision of the studies i.e., the sample size of that study (m_{ij}). In cases with large sample-sizes, the Q-statistic becomes highly sensitive to even minor differences between studies. The Q-statistic is "overpowered", which results in the detection of statistically significant heterogeneity even when the actual differences between studies are small. Little research has been done on the effect of very large primary-sample-sizes since meta-analyses typically compile studies who's unit of analysis are human patients. Primary sample sizes in the millions is not a common issue.

8. **Transformation of the effect size:** In general model selection at the transformed level presents limitations, as the relevance of features is assessed on the transformed scale, which may not directly translate to the original effect size after back-transformation. This complicates the interpretation of results, since conclusions drawn on the transformed scale may not have the same meaning when applied to the original scale. This effect is seen with the covariate: use of ancillary data. Additionally, the use of FT transformation is contested in the literature because of several important limitations (Doi & Xu, 2021; Lin & Xu, 2020; Röver & Friede, 2022; Schwarzer et al., 2019). First, the FT is notably unintuitive, specifically the calculation of variance which relies on the structure of an arcsine function's derivative. Second, back-transforming the pooled effect size using certain methods—such as the harmonic mean of primary sample sizes—can lead to

misleading results (Doi & Xu, 2021; Lin & Xu, 2020; Röver & Friede, 2022; see Schwarzer et al., 2019; Wang, 2023). In this analysis, the pooled variance, rather than the harmonic mean, was used for back-transformation, which seems to address the main concern debated in the literature. Nevertheless, the choice of back-transformation method significantly influences the outcome, and justifying a specific method is especially challenging in a multilevel data structure (Röver & Friede, 2022). Lastly, in a random-effects model the true (transformed) proportion is assumed to follow a normal distribution between studies, the FT transformation potentially violates this assumption as the arcsine function has a bounded domain (Röver & Friede, 2022).

Implications for Future Research

The limitations identified in this meta-analysis suggest several directions for future research that can enhance the robustness and generalisability of findings related to machine learning applications in remote sensing for SDG monitoring.

1. **Sample size and model complexity**: One of the primary limitations of this meta-analysis was the small sample size. Future research should aim to expand the pool of included studies. This would mean that interaction effects between the collected study features could also be included in the analysis. The structure of the random effects can also be explored with the application of more sophisticated variance-covariance structures for random effects. This approach, sometimes referred to as dose-response meta-analysis (Viechtbauer, 2024b, p. 269), would provide insights into how specific study characteristics influence effect sizes over time or across varying conditions.

2. **Broader inclusion of performance metrics**: This meta-analysis primarily focused on overall accuracy, a commonly used but potentially misleading performance metric, particularly in imbalanced datasets. Future studies should expand the range of performance metrics, incorporating class-specific precision, recall, F1-score, Matthews' correlation coefficient, F1 score, or Somers' D to provide a more comprehensive evaluation of model performance (Burger & Meertens, 2020). More than one effect size can be modeled using network meta- analysis models (Harrer et al., 2022, Chapter 12). The inclusion of more performance metrics would offer a more nuanced understanding of how models perform under different conditions.

3. **Exploring additional study features and moderators**: The present study focused on a limited set of study features. Future research should investigate a broader range of potential moderators, such as model complexity, data preprocessing techniques, and environmental or socio-economic factors specific to SDG challenges. By including a more extensive set of features, researchers can better understand the drivers of performance variability and refine model selection for specific applications.

4. **Effect of large sample size in primary studies**: Simulation studies could provide insights into the sensitivity of Cochran's Q in the context of large sample sizes. Developing less sensitive methods for assessing heterogeneity would improve the reliability of meta-analytic findings, especially when studies involve substantial sample sizes, which can exaggerate minor differences between studies.

5. **Data extraction:** In the time frame of this research, the ChatGPT virtual assistant showed significant improvements in data extraction capabilities. Initially, in January 2024, ChatGPT struggled to extract meaningful features. By May 2024, it was capable of accurately filling in all study features directly from the provided papers (in PDF format). Although the improvement was not formally assessed in this study, the difference was striking. Some research has already examined the potential accuracy of large language models (LLMs) in data extraction for meta-analyses, with promising results (Mahuli et al., 2023). However, for this thesis, ChatGPT was not used for formal data extraction. Instead, traditional manual extraction methods were employed to ensure accuracy. Further investigation into the accuracy of LLMs for meta-analysis is required. LLMs can expedite the data extraction process, potentially addressing challenges related to the limited number of included studies. Another unrelated recommendation to improve data extraction would be for journals to require results and specific features to be submitted separately in addition to the manuscript so that the journals themselves can report trends in outcomes.

Conclusion

This meta-analysis provides insights into the variability of machine learning models used for remote sensing in SDG monitoring. First, (1) the average performance of machine learning models was found to be high, but strongly influenced by class imbalance. This finding reinforces the limitations of overall accuracy as a metric for assessing model performance. It highlights the need for a shift towards more balanced and nuanced performance metrics in future SDG monitoring studies. Second, (2) the three-level random-effects model showed a substantial degree of heterogeneity across outcomes. Third (3), the role of specific study features was notable: although no significant differences were observed between model types (e.g., neural networks or tree-based models), the proportion of the majority class and the inclusion of ancillary data were important factors. Finally, the comparison of sample-weighted and unweighted models (4) revealed no substantial difference in summary effect size, though the weighted model uncovered significant heterogeneity. Lastly, more research is needed to assess the robustness and applicability of meta-analyses methods to this field. In particular, the use of Cochran's Q-statistic is questionable in the context of this analysis, as the very large sample sizes might make the Q-statistic overly sensitive. This can result in the detection of statistically significant heterogeneity, even when the heterogeneity may not be practically meaningful.

References

Anshuka, A., Ogtrop, F. F. van, & Willem Vervoort, R. (2019). Drought forecasting through statistical models using standardised precipitation index: A systematic review and meta-regression analysis. *Natural Hazards*, *97*(2), 955–977. https://doi.org/10.1007/s11069-019-03665-6

Assink, M., & Wibbelink, C. J. M. (2016). Fitting three-level meta-analytic models in R: A step-by-step tutorial. *The Quantitative Methods for Psychology*, *12*(3), 154–174. https://doi.org/10.20982/tqmp.12.3.p154

Barendregt, J. J., Doi, S. A., Lee, Y. Y., Norman, R. E., & Vos, T. (2013). Meta-analysis of prevalence. *Journal of Epidemiology and Community Health (1979-)*, *67*(11), 974–978. https://doi.org/10.1136/jech-2013-203104

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Wiley.

Borges Migliavaca, C., Stein, C., Colpani, V., Barker, T. H., Munn, Z., Falavigna, M., & on behalf of the Prevalence Estimates Reviews Systematic Review Methodology Group (PERSyst). (2020). How are systematic reviews of prevalence conducted? A methodological study. *BMC Medical Research Methodology*, *20*(1), 96. https://doi.org/10.1186/s12874-020-00975-3

Bozada, T., Borden, J., Workman, J., Del Cid, M., Malinowski, J., & Luechtefeld, T. (2021). Sysrev: A FAIR platform for data curation and systematic evidence review. *Frontiers in Artificial Intelligence*, *4*, 685298. https://doi.org/10.3389/frai.2021.685298

Burger, J., & Meertens, Q. (2020). The algorithm versus the chimps:on the minima of classifier performance metrics. In L. Cao, W. Kosters, & J. Lijffijt (Eds.), *BNAIC/BeneLearn 2020 proceedings* (pp. 38–55). BNAIC/BeneLearn. https://bnaic.liacs.leidenuniv.nl/bnaic2020proceedings.pdf

Burke, M., Driscoll, A., Lobell, D. B., & Ermon, S. (2021). Using satellite imagery to understand and promote sustainable development. *Science*, *371*(6535), eabe8628. https://doi.org/10.1126/science.abe8628

Campbell, McKenzie, J. E., Sowden, A., Katikireddi, S. V., Brennan, S. E., Ellis, S., Hartmann-Boyce, J., Ryan, R., Shepperd, S., Thomas, J., Welch, V., & Thomson, H. (2020). Synthesis without meta-analysis (SWiM) in systematic reviews: reporting guideline. *BMJ*, *368*, l6890. https://doi.org/10.1136/bmj.l6890

Campbell, & Wynne, R. H. (2011). *Introduction to remote sensing* (5th ed). Guilford Press.

Cheung, M. W. L. (2014). Modeling dependent effect sizes with three-level meta-analyses: A structural equation modeling approach. *Psychological Methods*, *19*(2), 211–229. https://doi.org/10.1037/a0032968

Debray, T. P. A., Damen, J. A. A. G., Snell, K. I. E., Ensor, J., Hooft, L., Reitsma, J. B., Riley, R. D.,

& Moons, K. G. M. (2017). A guide to systematic review and meta-analysis of prediction model performance. *BMJ*, i6460. https://doi.org/10.1136/bmj.i6460

Doi, S. A., & Xu, C. (2021). The Freeman–Tukey double arcsine transformation for the meta-analysis of proportions: Recent criticisms were seriously misleading. *Journal of Evidence-Based Medicine*, *14*(4), 259–261. https://doi.org/10.1111/jebm.12445

Ekmen, O., & Kocaman, S. (2024). Remote sensing for UN SDGs: A global analysis of research and collaborations. *The Egyptian Journal of Remote Sensing and Space Sciences*, *27*(2), 329–341. https://doi.org/10.1016/j.ejrs.2024.04.002

FAO, F. and A. O. (2016). *Map accuracy assessment and area estimation practical guide.*, http://www.fao.org/3/a-i5601e.pdf

Foody, G. M. (2020). Explaining the unsuitability of the kappa coefficient in the assessment and comparison of the accuracy of thematic maps obtained by image classification. *Remote Sensing of Environment*, *239*, 111630. https://doi.org/10.1016/j.rse.2019.111630

Freeman, M. F., & Tukey, J. W. (1950). Transformations Related to the Angular and the Square Root. *The Annals of Mathematical Statistics*, *21*(4), 607–611. https://doi.org/10.1214/aoms/1177729756

Gusenbauer, M., & Haddaway, N. R. (2020). Which academic search systems are suitable for systematic reviews or meta-analyses? Evaluating retrieval qualities of Google Scholar, PubMed, and 26 other resources. *Research Synthesis Methods*, *11*(2), 181–217. https://doi.org/10.1002/jrsm.1378

Haddaway, N. R., Bannach-Brown, A., Grainger, M. J., Hamilton, W. K., Hennessy, E. A., Keenan, C., Pritchard, C. C., & Stojanova, J. (2022). The evidence synthesis and meta-analysis in R conference (ESMARConf): Levelling the playing field of conference accessibility and equitability. *Systematic Reviews*, *11*(1), 113. https://doi.org/10.1186/s13643-022-01985-6

Hall, J. A., & Rosenthal, R. (2018). Choosing between random effects models in meta-analysis: Units of analysis and the generalizability of obtained results. *Social and Personality Psychology Compass*, *12*(10), e12414. https://doi.org/10.1111/spc3.12414

Hall, O., Dompae, F., Wahab, I., & Dzanku, F. M. (2023). A review of machine learning and satellite imagery for poverty prediction: Implications for development research and applications. *Journal of International Development*, *35*(7), 1753–1768. https://doi.org/10.1002/jid.3751

Hanadé Houmma, I., El Mansouri, L., Gadal, S., Garba, M., & Hadria, R. (2022). Modelling agricultural drought: A review of latest advances in big data technologies. *Geomatics, Natural Hazards and Risk*, *13*(1), 2737–2776. https://doi.org/10.1080/19475705.2022.2131471

Hansen, C., Steinmetz, H., & Block, J. (2022a). How to conduct a meta-analysis in eight steps: A practical guide. *Management Review Quarterly*, *72*(1), 1–19. https://doi.org/10.1007/s11301-021-00247-4

Hansen, C., Steinmetz, H., & Block, J. (2022b). How to conduct a meta-analysis in eight steps: a practical guide. *Management Review Quarterly*, *72*(1), 1–19. https://doi.org/10.1007/s11301-021-

00247-4

Harrer, M., Cuijpers, P., Furukawa, T. A., & Ebert, D. D. (2022). *Doing meta-analysis with r: A hands-on guide*. CRC Press/Taylor & Francis Group. https://bookdown.org/MathiasHarrer/Doing_Meta_Analysis_in_R/

Harrer, M., Cuijpers, P., Furukawa, T., & Ebert, D. D. (2019). *Dmetar: Companion r package for the guide 'doing meta-analysis in r'*. http://dmetar.protectlab.org/

Hedges, L. V., Tipton, E., & Johnson, M. C. (2010). Robust variance estimation in meta-regression with dependent effect size estimates. *Research Synthesis Methods*, *1*(1), 39–65. https://doi.org/10.1002/jrsm.5

Higgins, J. P. T., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine*, *10*, 1539–1558. https://doi:10.1002/sim.1186

Holloway, J., & Mengersen, K. (2018). Statistical Machine Learning Methods and Remote Sensing for Sustainable Development Goals: A Review. *Remote Sensing*, *10*(9), 1365. https://doi.org/10.3390/rs10091365

Iliescu, D., Rusu, A., Greiff, S., Fokkema, M., & Scherer, R. (2022). Why We Need Systematic Reviews and Meta-Analyses in the Testing and Assessment Literature. *European Journal of Psychological Assessment*, *38*(2), 73–77. https://doi.org/10.1027/1015-5759/a000705

Khatami, R., Mountrakis, G., & Stehman, S. V. (2016). A meta-analysis of remote sensing research on supervised pixel-based land-cover image classification processes: General guidelines for practitioners and future research. *Remote Sensing of Environment*, *177*, 89–100. https://doi.org/10.1016/j.rse.2016.02.028

Laird, N. M., & Mosteller, F. (1990). Some Statistical Methods for Combining Experimental Results. *International Journal of Technology Assessment in Health Care*, *6*(1), 5–30. https://doi.org/10.1017/s0266462300008916

Lajeunesse, M. J. (2016). *Facilitating systematic reviews, data extraction, and meta-analysis with the metagear package for r*. *7*, 323–330.

Lavallin, A., & Downs, J. A. (2021). Machine learning in geography–Past, present, and future. *Geography Compass*, *15*(5), e12563. https://doi.org/10.1111/gec3.12563

Lin, L., & Xu, C. (2020). Arcsine-based transformations for meta-analysis of proportions: Pros, cons, and alternatives. *Health Science Reports*, *3*(3), e178. https://doi.org/10.1002/hsr2.178

Mahuli, S. A., Rai, A., Mahuli, A. V., & Kumar, A. (2023). Application ChatGPT in conducting systematic reviews and meta-analyses. *British Dental Journal*, *235*(2), 90–92. https://doi.org/10.1038/s41415-023-6132-y

Maso, J., Zabala, A., & Serral, I. (2023). Earth Observations for Sustainable Development Goals. *Remote Sensing*, *15*(10), 2570. https://doi.org/10.3390/rs15102570

McCulloch, C. E., & Neuhaus, J. M. (2011). Misspecifying the shape of a random effects distribution:

Why getting it wrong may not matter. *Statistical Science*, *26*(3), 388–402. https://doi.org/10.1214/11-STS361

NASA. (2019). *What is Remote Sensing?* https://www.earthdata.nasa.gov/learn/backgrounders/remote-sensing

Owers, C. J., Lucas, R. M., Clewley, D., Tissott, B., Chua, S. M. T., Hunt, G., Mueller, N., Planque, C., Punalekar, S. M., Bunting, P., Tan, P., & Metternicht, G. (2022). Operational continental-scale land cover mapping of Australia using the Open Data Cube. *International Journal of Digital Earth*, *15*(1), 1715–1737. https://doi.org/10.1080/17538947.2022.2130461

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., … Moher, D. (2021). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*, n71. https://doi.org/10.1136/bmj.n71

Polanin, J. R. (2014). *An introduction to multilevel meta-analysis,*. https://www.youtube.com/watch?v=rJjeRRf23L8&t=1358s; Campbell Colloquium.

Priem, J., Piwowar, H., & Orr, R. (2022). *OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts*. https://arxiv.org/abs/2205.01833

Pustejovsky, J. E. (2020). *Weighting in multivariate meta-analysis*. https://jepusto.com/posts/weighting-in-multivariate-meta-analysis/.

Röver, C., & Friede, T. (2022). Double arcsine transform not appropriate for meta-analysis. *Research Synthesis Methods*, *13*(5), 645–648. https://doi.org/10.1002/jrsm.1591

Schwarzer, G., Carpenter, J. R., & Rücker, G. (2015). *Meta-Analysis with R*. Springer International Publishing. https://doi.org/10.1007/978-3-319-21416-0

Schwarzer, G., Chemaitelly, H., Abu-Raddad, L. J., & Rücker, G. (2019). Seriously misleading results using inverse of freeman-tukey double arcsine transformation in meta-analysis of single proportions. *Research Synthesis Methods*, *10*, 476–483. https://doi.org/10.1002/jrsm.1348

SEOS. (2014). *Introduction to remote sensing*. https://seos-project.eu/remotesensing/remotesensing-c01-p06.html

Stehman, S. V., & Foody, G. M. (2019). Key issues in rigorous accuracy assessment of land cover products. *Remote Sensing of Environment*, *231*, 111199. https://doi.org/10.1016/j.rse.2019.05.018

Tawfik, G. M., Dila, K. A. S., Mohamed, M. Y. F., Tam, D. N. H., Kien, N. D., Ahmed, A. M., & Huy, N. T. (2019). A step by step guide for conducting a systematic review and meta-analysis with simulation data. *Tropical Medicine and Health*, *47*(1), 46. https://doi.org/10.1186/s41182-019-0165-6

Thapa, A., Horanont, T., Neupane, B., & Aryal, J. (2023). Deep Learning for Remote Sensing Image Scene Classification: A Review and Meta-Analysis. *Remote Sensing*, *15*(19), 4804. https://doi.org/10.3390/rs15194804

UCS. (2021). *Union of Concerned Scientists (UCS) Satellite Database*. https://www.ucsusa.org/resources/satellite-database

UN DESA. (2023). *The Sustainable Development Goals Report 2023: Special Edition*. United Nations. https://doi.org/10.18356/9789210024914

UN-GGIM:Europe. (2019). *The territorial dimension in SDG indicators: Geospatial data analysis and its integration with statistical data*. Instituto Nacional de Estatística. https://un-ggim-europe.org/wp-content/uploads/2019/05/UN_GGIM_08_05_2019-The-territorial-dimension-in-SDG-indicators-Final.pdf

United Nations. (2017). *Earth observations for official statistics: Satellite imagery and geospatial data task team report*. https://unstats.un.org/bigdata/task-teams/earth-observation/UNGWG_Satellite_Task_Team_Report_WhiteCover.pdf

United Nations. (2024). *The sustainable development goals report 2024*. https://unstats.un.org/sdgs/report/2024/The-Sustainable-Development-Goals-Report-2024.pdf

Veroniki, A. A., Jackson, D., Viechtbauer, W., Bender, R., Bowden, J., Knapp, G., Kuss, O., Higgins, J. P., Langan, D., & Salanti, G. (2015). Methods to estimate the between-study variance and its uncertainty in meta-analysis. *Research Synthesis Methods*, *7*(1), 55–79. https://doi.org/10.1002/jrsm.1164

Viechtbauer, W. (2010). Conducting Meta-Analyses in R with the metafor Package. *Journal of Statistical Software*, *36*, 1–48. https://doi.org/10.18637/jss.v036.i03

Viechtbauer, W. (2020). *Weights in models fitted with the rma.mv() function*. https://www.metafor-project.org/doku.php/tips:weights_in_rma.mv_models.

Viechtbauer, W. (2022). *Metafor: Model selection using the glmulti and MuMIn packages*. https://www.metafor-project.org/doku.php/tips:model_selection_with_glmulti_and_mumin#variable_importance.

Viechtbauer, W. (2024a). *Frequently asked questions [the metafor package]: Freeman-tukey transformation of proportions*. https://www.metafor-project.org/doku.php/faq#how_is_the_freeman-tukey_trans.

Viechtbauer, W. (2024b). *metafor: Meta-Analysis Package for R*. https://doi.org/10.32614/CRAN.package.metafor

Wang, N. (2023). Conducting Meta-analyses of Proportions in R. *Journal of Behavioral Data Science*, *3*(2), 64–126. https://doi.org/10.35566/jbds/v3n2/wang

Yin, C., Peng, N., Li, Y., Shi, Y., Yang, S., & Jia, P. (2023). A review on street view observations in support of the sustainable development goals. *International Journal of Applied Earth Observation and Geoinformation*, *117*, 103205. https://doi.org/10.1016/j.jag.2023.103205

Zhang, C., & Li, X. (2022). Land Use and Land Cover Mapping in the Era of Big Data. *Land*, *11*(10), 1692. https://doi.org/10.3390/land11101692

Zhang, Y., Liu, J., & Shen, W. (2022). A Review of Ensemble Learning Algorithms Used in Remote Sensing Applications. *Applied Sciences*, *12*(17), 8654. https://doi.org/10.3390/app12178654

Zhao, Q., Yu, L., Du, Z., Peng, D., Hao, P., Zhang, Y., & Gong, P. (2022). An Overview of the Applications of Earth Observation Satellite Data: Impacts and Future Trends. *Remote Sensing*, *14*(8), 1863. https://doi.org/10.3390/rs14081863