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 2024-01-01

MASTER THESIS

STATISTICS AND DATA SCIENCE

UNIVERSITEIT LEIDEN

[1]Department of Research and Development, Statistics Netherlands (CBS), Heerlen, The Netherlands
[2]Department of Methodology and Statistics, Leiden University, Leiden, The Netherlands

Table of contents

Forword

Abstract

Objective: This meta-analysis aims to evaluate machine learning methods in remote sensing applications for monitoring Sustainable Development Goals (SDGs). Specifically the aims are (a) estimate the average performance (population effect size); (b) determine whether performance varies within and across studies (establish the degree of heterogeneity); (c) assess whether study features are related to the performance of machine learning models; and (d) 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 which 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 20 studies with 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 is 0.90 (95% CI [0.85; 0.94]), which is not substantially different from the weighted model at 0.89 (CI 95% [0.85, 0.94]. The weighted models found substantial heterogeneity between results. The proportion of the majority class was identified as the most important factor affecting the overall accuary, followed by the inclusion of ancillary data. However, type of machine learning model (e.g., neural networks, tree-based models) or SDG goal did not have a significant effect. When the important study features were included (in the mixed effects model) the average overall accuracy dropped 0.70 (CI 95% CI [0.58, 0.80]).

Conclusion: This study demonstrates the high variability model performance in remote sensing applications. As well as the effect that the proportion of the majority class 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

The following table....

Notation	Section	Definition
m_{rc}	Chapter 2	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.
$m_{r.}, m_{.c}$	Chapter 2	The total number of instances of class r (sum of row r). The total number of instances predicted as class c (sum of column c).
n	Chapter 2	The total number of instances in the dataset (sum of all cells).
s	Chapter 2	Correctly classified instances
i, j	Chapter 3	
n		

Table of Abbreviation

Abbreviation	Section	Definition?
SDGs	Chapter 1	Sustainable Development Goals
NSI		National Statistical Institute

Chapter 1

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., 2016a; 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 of 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., 2016a). Earth Observation satellites alone can measure 42% of the SDGs targets (Y. Zhang et al., 2022).

Despite the increased research, machine learning and remote sensing for SDG monitoring, there is still a lack of comprehensive understanding regarding the factors that determine the performance of these models across different contexts. The success of machine learning models in remote sensing depends on various factors, including the quality and resolution of input data, the choice of algorithm, sample's representativeness, and the complexity of the landscape (Heydari & Mountrakis, 2018; Lu & Weng, 2007). Additionally, model performance is often evaluated using localized datasets which can limit the generalisability of findings and the ability to apply these models in broader contexts (Burke et al., 2021; Khatami et al., 2016a; United Nations, 2017).

Although the uptake of remote sensing data by NSIs has been slow, 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, RS, 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). Despite these advances, this paradigm shift from traditional statistical methods—such as counting and measuring by humans—towards estimation from sensors, simulation, and modeling, presents challenges (United Nations, 2017). It 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).

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). The literature examined in this study found that previous meta-analyses investigating the performance of machine learning models on remote sensing data have predominantly relied on unweighted approaches. This meta-analysis is performed on peer-reviewed research articles that use machine learning methods and remote sensing data to monitor SDGs. The aims are (a) estimate the average performance (population effect size); (b) determine whether performance varies within and across studies (establish the degree of heterogeneity); (c) assess whether study features are related to the performance of machine learning models; and (d) compare the sample-weighted and unweighted estimate summary effect.

Chapter 2

Background

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).

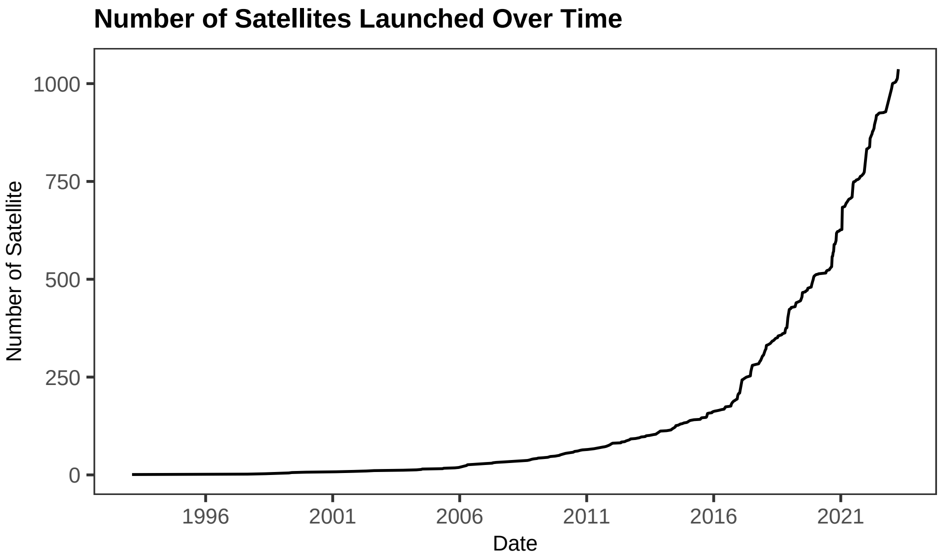

Figure 2.1: Number of active satellites by date of launch. Data acquired from UCS (2021).

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 with the Earth's surface (NASA, 2019).

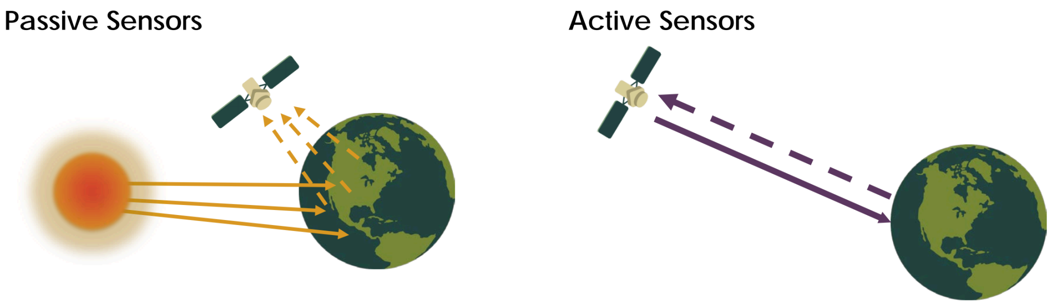

Figure 2.2: Illustration of a passive sensor and an active sensor. Source: NASA (2019) Applied Passive Sciences Remote Sensing Training Program.

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) —given by $\frac{NIR-Red}{NIR+Red}$ — 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.

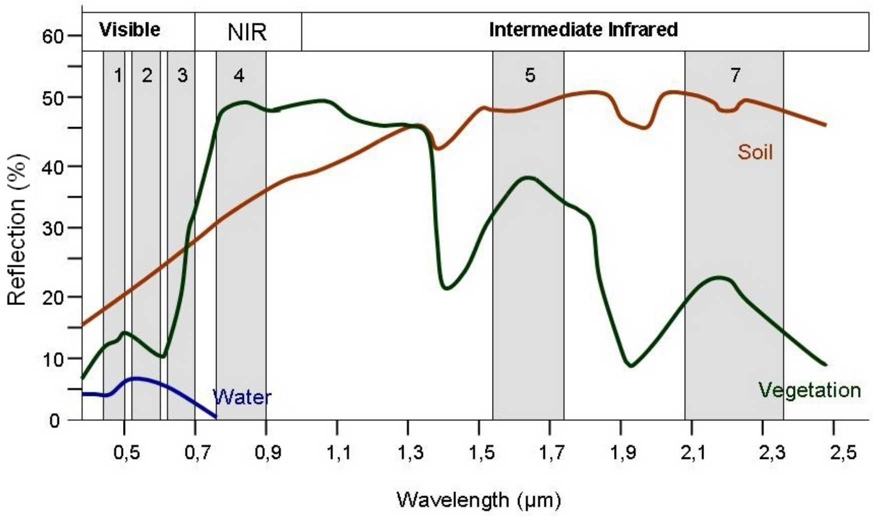

Figure 2.3: Spectral signatures of soil, green vegetation, and water across different wavelengths, representing the portion of incident radiation that is reflected by each material as a function of wavelength. The grey bands indicate the spectral ranges (channels) of Landsat TM satellite. Bands 1-3 capture visible light (Blue, Green, Red), while 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 modeling (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. Which is particularly useful for handling the high complexity and dimensionality of remote sencing 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 as well as giving 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 (UN-GGIM:Europe, 2019). 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).

To verify these analyses performance metrics are used. For classification tasks, this involves 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. are 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	$m_{.1}$	$m_{.2}$	$m_{.3}$	$m_{.4}$	n	
User's accuracy	$m_{11}/m_{.1}$	$m_{22}/m_{.2}$	$m_{33}/m_{.3}$	$m_{44}/m_{.4}$		

Note: Confusion matrix for a classification with four classes, where the rows (r) represent the reference (observed) classification and the columns (c) represent the predicted classes. m_{rc} is the number of instances predicted in reference class r and predicted class c, and n 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 total number of instances, n.

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

If the unit of accuracy assessment is a pixel, then overall accuracy is the proportion of pixels classified correctly. Other metrics include the reliability (User's accuracy) and sensitivity (recall or Producer's accuracy). Reliability is the correct classifications for a particular class divided by the column total ($m_{.c}$) and sensitivity is correct classifications over the row total ($m_{r.}$). It is important to consider the purpose of the map 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

To illustrate how remote sensing data and machine leaning can be used to support ecological sustainable development, Owers et al. (2022) developed an approach to monitor and map land cover across Australia using techniques.

The study utilized Landsat sensor data archive 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.

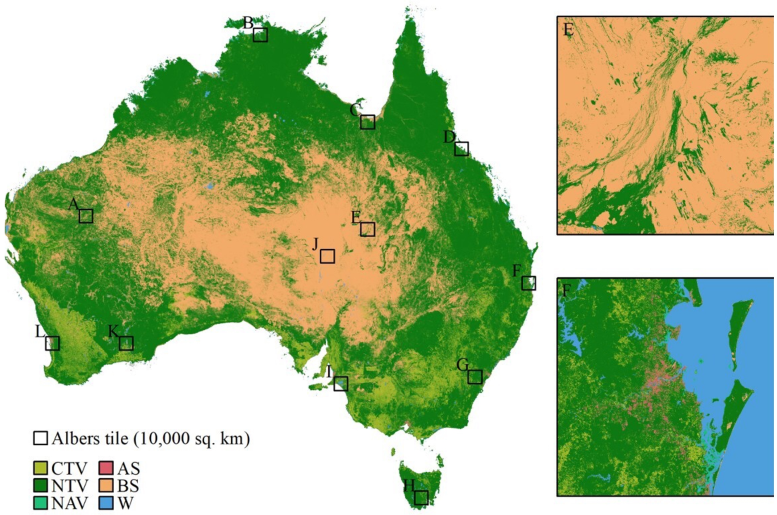

Figure 2.4: Land cover mapping created by Owers et al. (2022) using Landstat data to make continent-wide classifications using the LCCS frame work which differentiates six (classes) land cover types: cultivated terrestrial vegetation (CTV), natural terrestrial vegetation (NTV), natural aquatic vegetation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

To produce such maps using a topographical field survey is impractical, given Australia's size ($7,688,287$ km^2). While field surveys are the most accurate method of generating training sample data, they are labor-intensive, time-consuming, and expensive (C. Zhang & Li, 2022). A topographical survey of just 20 hectares (0.2 km^2) takes a team of four people approximately five days to complete, even 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 is a less labor-intensive, costly and time-consuming than field surveys it still requires significant effort and expertise.

In contrast to the challenges associated with field surveys, remote sensing provides an efficient method for the continuous monitoring of large areas that would otherwise be inaccessible (Owers et al., 2022; C. Zhang & Li, 2022). Thefore, the potential applications are numerous. Examples include monitoring of land use and degradation, forestry, biodiversity, agriculture, disaster prediction, water resources, public health, urban planning, poverty, 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. (2016a) and O. Hall et al. (2023)). In 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.

Chapter 3

Methods

The methods adopted in this study are delineated in sequential steps, following the framework proposed by Debray et al. (2017). Additionally, all procedures and reporting were conducted in compliance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines (Page et al., 2021). For the statistical analyses `metafor` (Viechtbauer, 2010) and `dmetar` (Harrer et al., 2019) packages the were used.

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). Based on this framework, the question was formulated as follows: In studies focused on SDGs, how heterogeneous is the performance of ML applied to various remote sensing applications, and what study features account for any observed differences in model performance?

Table 3.1: PICOTS framework

Item	Explanation
Population	Studies monitoring SDGs.
Intervention	Application of ML models to remote sensing data.
Comparison	Comparison of different ML models and methodologies used in remote sensing applications.
Outcomes	Variability in the OA of ML 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: Key items of the PICOTS framework used to determine the framing of the review.

The data collected for this report was extracted from peer-reviewed articles published between January 2018 and December 2023. These articles were gathered (on January 15 and 16, 2024) from several academic databases, including ScienceDirect and Taylor & Francis Online, as shown in Figure 3.3. To reduce potential bias from database coverage (Hansen et al., 2022a; Tawfik et al., 2019), several academic databases were used. 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.

While Google Scholar can be useful for supplementary searches and grey literature, it is generally considered unsuitable as the primary sourse for systematic reviews (Gusenbauer & Haddaway, 2020). Furthermore, Google Scholar searches results are not fully reproducible (Gusenbauer & Haddaway, 2020) and search result references that cannot be downloaded in batches.

Duplicate articles were handled using Zotero's "merge duplicates" function. 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 ML and RS for SDG monitoring.

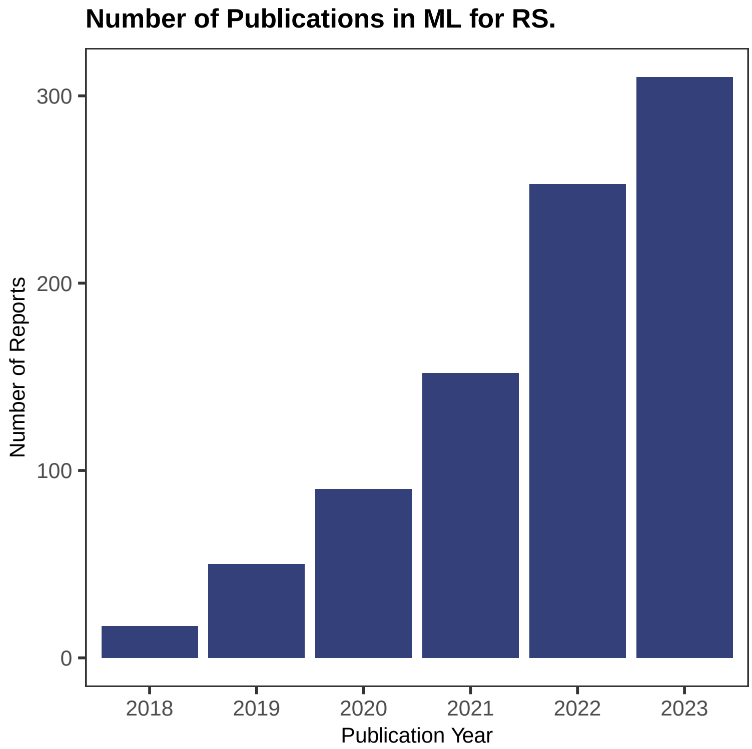

Figure 3.1: Publication increase between 2018 and 2022.

3.2 Specific inclusion and exclusion criteria

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 ML 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 used

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

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

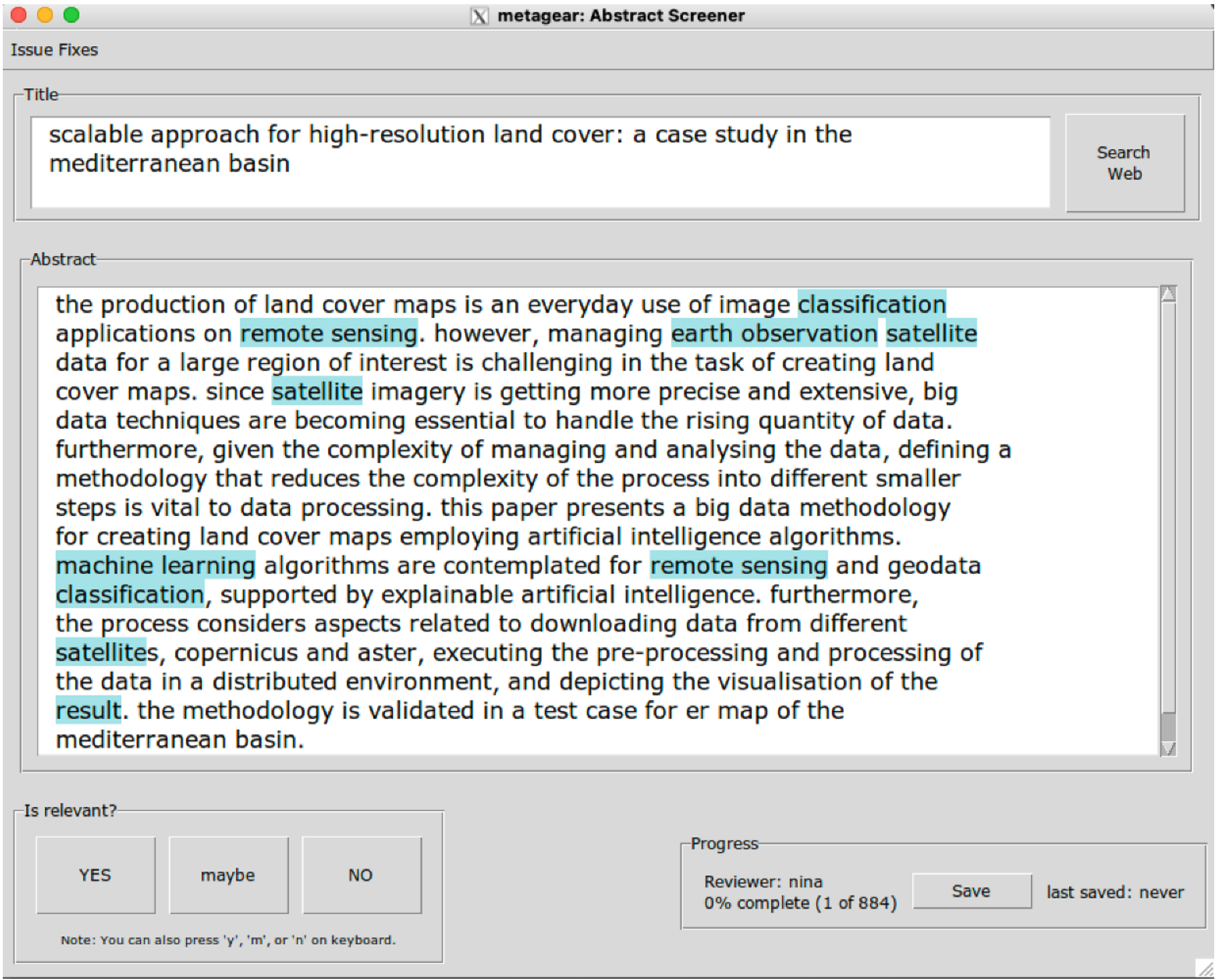

Figure 3.2: Metagear graphical user interface: Example of the metagear abstract screener interface, with key words highlighted. On the bottom left the reviewer can select whether the paper is relevant.

As shown in Figure 3.3, of the 200 abstracts screened only 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.

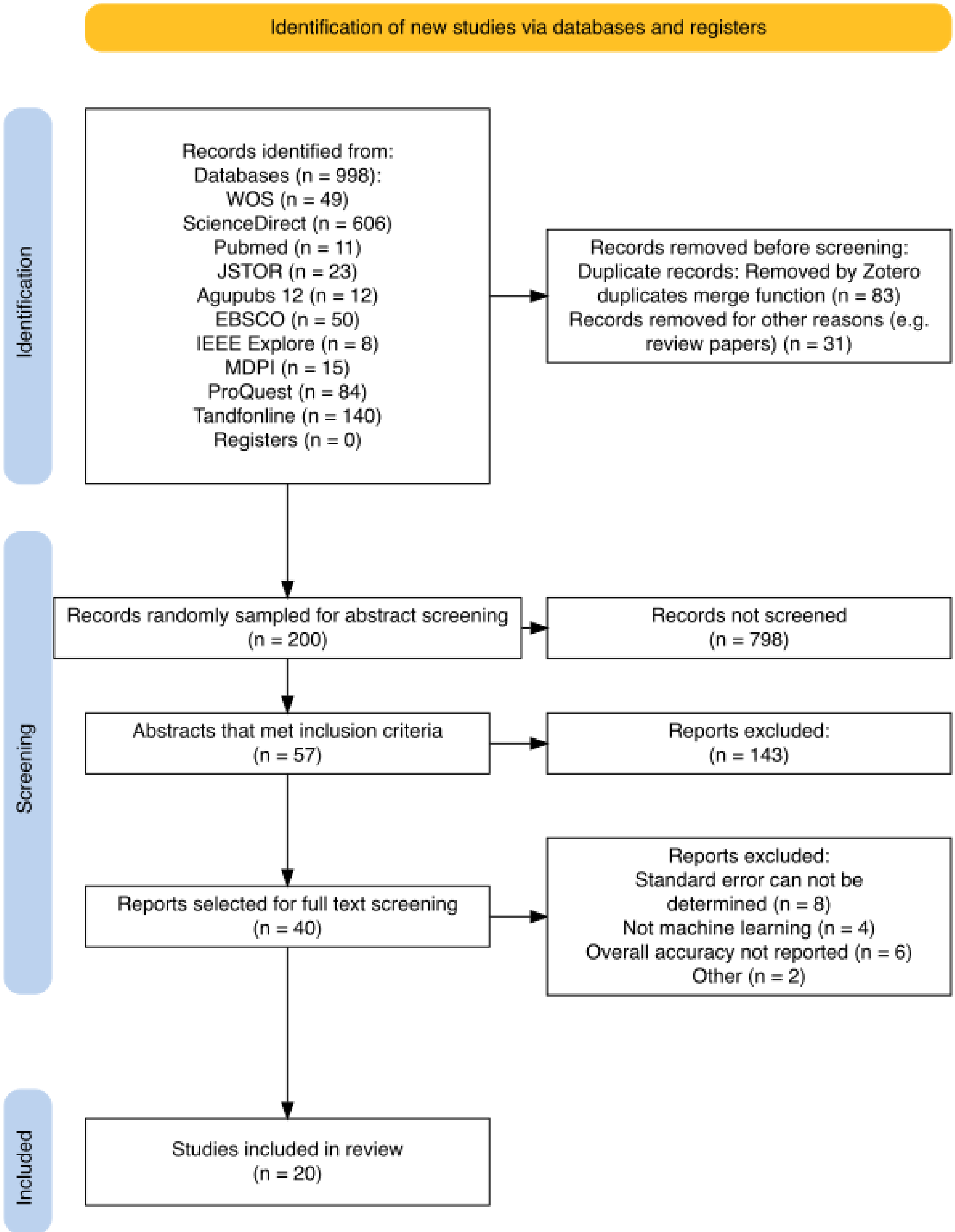

Figure 3.3: PRIMSA flow diagram of manuscript selection. The records were identified from databases including Web of Science (WOS), ScienceDirect, PubMed, Journal Storage (JSTOR), American Geophysical Union Publications (Agupubs), EBSCO, IEEE Xplore, Multidisciplinary Digital Publishing Institute (MDPI), ProQuest, and Taylor & Francis Online (Tandfonline), no papers were gathered from official registers. Note: number of records removed four where not journal articles and 27 were omitted for being reviews. A random sample of 200 of the total 884 was drawn and reviewed by three independent reviewers. A total of 57 records were left, 40 of which were deemed to be classification papers and the full text screened.

3.3 Feature collection

Using the first 10 papers and previous systematic reviews, a list of potential study features was made and structured in a table for data collection. Table 3.3 outlines all the extracted features, including including study identification information, as well as the number of citations, which was gathered using the Local Citation Network web app. This tool collects the articles' metadata from OpenAlex—a bibliographic catalog of scientific papers (Priem et al., 2022)[1]. The methodology and data characteristics includes information regarding 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, which can affect the performance of classification models. Remote sensing specific information focuses on the type of devices, spectral bands, and spatial resolution, to assess how the data collection impacts the performance. Several potential useful features were excluded including temporal resolution (the frequency of the data collection) and pre-processing steps which also impact the performance of the model. These where excluded as the differences between papers were to large to make meaningful groups.

[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	
Citation	Name(s) of authors and publication year	
Title	Title of the article	
Publication name	Name of journal that published paper	
Citation number	Number of citations of the paper	0-68
Research Context		
Location	Location of the data used (country level)	
SDG	Area of research	SDG 2: Zero Hunger, SDG 11: Sustainable Cities and Communities, SDG 15: Life on Land
Methodology and data characteristics		
Model group	Algorithm group used	Tree-Based Models, Neural Network, Other
Number classes	The number of classes to predict	
Proportion of the majority class	The fraction of the largest class	
Total	Sample size, number of pixels, or objects	
Training/Test set	Results based on training or test set	test, train, Not Reported
Ancillary data	Use of non-RS data in the model	0 - only remote sensed data, 1 - additional data used
RS device	Type of RS device	Satellite , Drone, Plane, Unmanned Aerial Vehicles (UAVs)
Remote Sensing (RS) Specific Information		
RS device group	Specific device name	Landsat, Sentinel, Other, and Not Reported
RS bands	Specital bands used	
RS spectral bands number	Number of bands used	
RS spatital resolution	Spatital resolution in meters	30, 15-25, 10, <1, Not Reported
Performance Metrics		
Accuracy metric	Measure used to assess the predictive performance of the ML method applied	Overall Accuracy (OA)

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 average (summary) effect, (b) establish the degree of heterogeneity, and (c) access if study characteristics can explain any variability 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. From Equation 2.1, the overall accuracy is the proportion of correctly classified instances, therefore, the effect size is:

$$\begin{aligned} \hat{\theta}_{ij} &= \frac{s_{ij}}{n_{ij}} \\ \text{with } \mathrm{Var}(\hat{\theta}_{ij}) &= \frac{\hat{\theta}_{ij}(1-\hat{\theta}_{ij})}{n_{ij}} \end{aligned} \tag{3.1}$$

where s_{ij} is the number of successful predictions and n_{ij} is total number of instances (i.e. sample size).

Weighted Approach

Before conducting the meta-analysis, first the structure of the collected data and assumption of independence of effect sizes needs 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. 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:

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

The observed overall accuracy $\hat{\theta}_{ij}$ is an estimate of overall accuracy from experiment i in study j 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} .[2]

[2]A model accounting only for sampling variance is referred to as a fixed-effects model, where it is assumed that all

At level 2, within-study variance (σ_ζ^2):

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

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_ζ^2.

Lastly, level 3, the variance between studies (σ_ξ^2):

$$\begin{aligned} &\text{Level 3: } \kappa_j = \mu + \xi_j \\ &\xi_{ij} \sim \mathcal{N}(0, \sigma_\xi^2) \end{aligned} \tag{3.4}$$

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_ξ^2. Combined, the three-level meta-analyses 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: $\mathrm{Cov}(\xi_j, \zeta_{ij}) = \mathrm{Cov}(\xi_j, \epsilon_{ij}) = \mathrm{Cov}(\zeta_{ij}, \epsilon_{ij}) = 0$. Therefore, (1) unconditional sampling variance of the effect size is the sum of Level 3 and Level 2 heterogeneity, and the known sampling variance: $\mathrm{Var}(\hat{\theta}_{ij}) = \sigma_\xi^2 + \sigma_\zeta^2 + v_{ij}$, (2) the effect sizes within the same study share the same covariance $\mathrm{Cov}(\hat{\theta}_{ij}, \hat{\theta}_{kj}) = \sigma_\xi^2$, and (3) the effect sizes in different studies are independent $\mathrm{Cov}(\hat{\theta}_{ij}, \hat{\theta}_{mn}) = 0$ (Cheung, 2014).[3] In other words, observations are organized as a series of independent groups, where the marginal variance-covariance matrix (**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

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). The addition random effect components allows one to make unconditional inferences about the population from which the included studies are a random sample.

[3] where k indexes the effect size study j and m effect size in study n where $n \neq j$

variance. The blocks themselves are independent, reflecting the assumption that there is no covariance between outcomes from different studies.

Let each study j have two reported outcomes ($N_j = 2$) and J represent the total number of studies, then $\mathbf{M}$ is structured as:

$$\mathbf{M} = \begin{pmatrix} \hat{\sigma}_\xi^2 + \hat{\sigma}_\zeta^2 + v_{11} & \hat{\sigma}_\xi^2 & 0 & 0 & \dots & 0 \\ \hat{\sigma}_\xi^2 & \hat{\sigma}_\xi^2 + \hat{\sigma}_\zeta^2 + v_{21} & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \hat{\sigma}_\xi^2 + \hat{\sigma}_\zeta^2 + v_{1\mathrm{J}} & \hat{\sigma}_\xi^2 \\ 0 & 0 & 0 & 0 & \hat{\sigma}_\xi^2 & \hat{\sigma}_\xi^2 + \hat{\sigma}_\zeta^2 + v_{2\mathrm{J}} \end{pmatrix} \tag{3.6}$$

The random-effects model can be extended to a mixed-effects model (also referred to as a meta-regression) by including study characteristics as covariates (predictors). Let x denote the value covariate, where p' refers to the number of covariates included in the model. These covariates can be either be x_{ij} for a Level 2 covariate, or Level 3 covariate x_j if the same for all effect sizes in study j . The mixed-effect model is

$$\hat{\theta}_{ij} = \mu + \beta_1 x_{ij1} + ... + \beta_{p'} x_{jp'} + \xi_j + \zeta_{ij} + \epsilon_{ij} \tag{3.7}$$

The assumptions here remain the same as Equation 3.5, but the heterogeneity $(\sigma_\xi^2, \sigma_\zeta^2)$ 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 population average μ and the heterogeneity σ_ξ^2 and σ_ζ^2 (Viechtbauer, 2010).

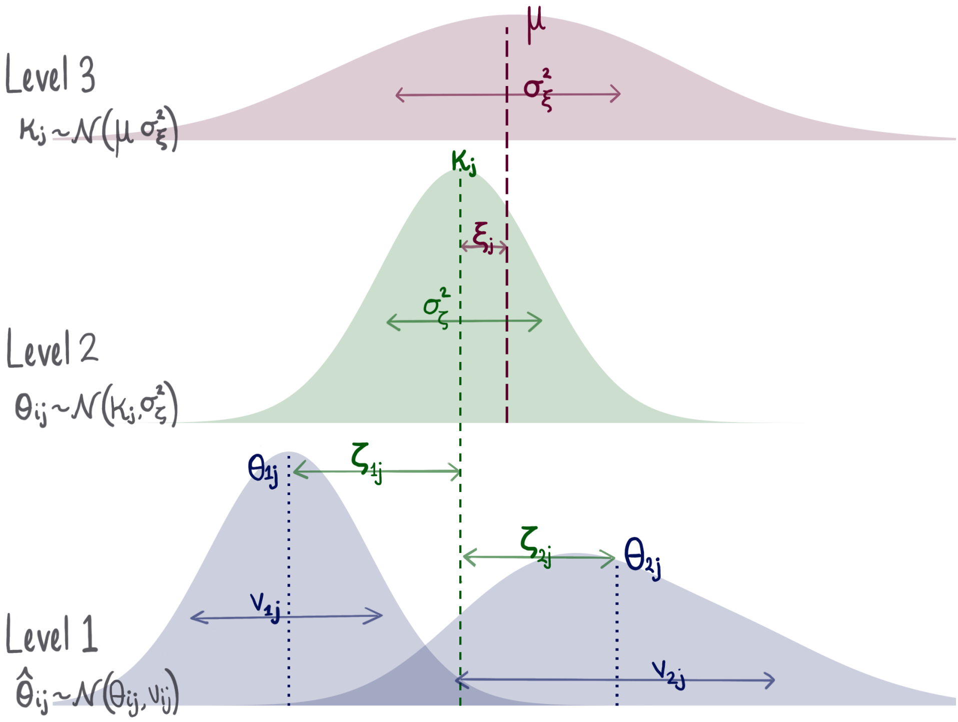

Figure 3.4: (Place holder) Illustration of the 3 level random effects meta analysis model. At Level 1: The observed effects $\hat{\theta}_{ij}$ are modelled as random draws from a normal distribution centred around the true effect size θ_{ij}, with known sampling variance v_{ij}. Observations from larger sample sizes n_{ij} have smaller sampling variances, which are 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 modelled as normally distributed with mean κ_j and within-study variance σ_ζ^2. Large deviations of θ_{ij} from κ_j indicate substantial within-study differences. In the mixed-effects model, the inclusion of Level 2 covariates x_{ij} aiims to reduce within-study heterogeneity σ_ζ^2 by explaining part of this variability. Lastly, at Level 3, study average effects are modelled as normally distributed with mean μ and between-study variance σ_ξ^2. A large σ_ξ^2 suggests substantial differences across studies, and the inclusion of Level 3 covariates x_j aims to explain this heterogeneity.

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_ζ^2 and σ_ξ^2), and then, the parameters defined above are estimated via weighted least squares with weights. There are several methods[4] to estimate σ_ζ^2 and σ_ξ^2 heterogeneity. This study uses the restricted maximum likelihood (REML) method. 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

[4]For the different methods and specifics see (Veroniki et al., 2015)

three sources of heterogeneity the sum of which the is the model implied variances of the estimates, therefore $w_{ij} = 1/(\hat{\sigma}^2_{\xi} + \hat{\sigma}^2_{\zeta} + v_{ij})$. However covariance between the effects needs to be taken into account, therefore the marginal variance-covariance matrix of the estimates is used (see Equation 3.6). Let $\mathbf{W} = \mathbf{M^{-1}}$ be the weight matrix, and, w_{ij} correspond to the i-th row and the j-th column of $\mathbf{W}$, and let $\hat{\theta}_i$ be the $i-$th estimate (with $i = 1, ..., N$[5]). Then the estimate of overall population average (summary effect) $\hat{\mu}$ for the random-effects model (without covariances, i.e. intercept-only model), is given by (Pustejovsky, 2020; Viechtbauer, 2020)

$$\begin{aligned} &\hat{\mu} = \frac{\sum_{i=1}^{N}(\sum_{j=1}^{N} w_{ij})\hat{\theta}_{ij}}{\sum_{j=1}^{N}\sum_{i=1}^{N} w_{ij}} \\ &\text{with} \\ &\overline{\sigma}^2 = \mathrm{Var}(\hat{\mu}) = \frac{1}{\sum_{i=1}^{N}(\sum_{j=1}^{N})w_{ij}} \end{aligned} \tag{3.8}$$

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

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

Where $\mathbf{y}$ is the vector of observed effects ($\hat{\theta}_{ij}$), $\mathbf{X}$ 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 $p' + 1$ columns. In the mixed effects case the estimated parameters are μ and $\beta_{p'}$'s ($\mathbf{b}$). 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.

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):

[5]Note $N = \sum_{j=1}^{J} N_j$ i.e. the total number of included effect sizes.

$$
\begin{aligned}
&H_0 : \theta_{ij} = \theta \\
&Q = \sum_{ij=1}^{n} w_{ij}(\hat{\theta}_{ij} - \hat{\mu}_{\text{fixed}})^2 \\
&\text{where} \\
&w_{ij} = \frac{1}{v_{ij}}, \\
&\hat{\mu}_{\text{fixed}} = \frac{\sum w_{ij}\hat{\theta}_{ij}}{\sum w_{ij}}
\end{aligned}
\tag{3.10}
$$

Here N is the total number of effect sizes. Under the null hypothesis Cochran's Q has an approximate chi-squared distribution with $N-1$ 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_{ij} (Cheung, 2014). The magnitude heterogeneity can be assessed using Higgins and Thompson's 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 $I^2_{(\text{Level 2})}$ and Level 3 $I^2_{(\text{Level 3})}$ are defined as follows (Cheung, 2014):

$$
\begin{aligned}
I^2_{(2)} &= \frac{\hat{\sigma}^2_{\zeta}}{\hat{\sigma}^2_{\zeta} + \hat{\sigma}^2_{\xi} + \tilde{v}} \\
I^2_{(3)} &= \frac{\hat{\sigma}^2_{\xi}}{\hat{\sigma}^2_{\zeta} + \hat{\sigma}^2_{\xi} + \tilde{v}}
\end{aligned}
\tag{3.11}
$$

Since the sampling variance varies across studies there are different ways to define the total variation (Cheung, 2014). Here $\tilde{v}$ is the typical sampling variance, defined as:

$$
\tilde{v} = \frac{(N-1)\sum w_{ij}}{(\sum w_{ij})^2 - \sum w_{ij}^2}, \tag{3.12}
$$

Lastly, the percentage of variance explained by the mixed-effects can be quantified using R^2, where $R^2_{(2)}$ and $R^2_{(3)}$ are the variance explained at Level 2 and Level 3 (Cheung, 2014);

$$
\begin{aligned}
R^2_{(2)} &= 1 - \frac{\hat{\sigma}^2_{\zeta(1)}}{\hat{\sigma}^2_{\zeta(0)}} \\
R^2_{(3)} &= 1 - \frac{\hat{\sigma}^2_{\xi(1)}}{\hat{\sigma}^2_{\xi(0)}}
\end{aligned}
\tag{3.13}
$$

where, $\hat{\sigma}^2_{\zeta(0)}$ and $\hat{\sigma}^2_{\zeta(1)}$ represent Level 2 variances, and $\hat{\sigma}^2_{\xi(0)}$ and $\hat{\sigma}^2_{\xi(1)}$ represent Level 3 variances, before and after including predictors.

The multi-model inference function from the `dmetar` package was used to select the best combination of covariates (i.e., the best model). This technique models all possible covariate combinations and compares them using an information-theoretic approach such as Akaike's Information Criterion (AIC) or Bayesian Information Criterion (BIC) (Harrer et al., 2022, Chapter 8). Additionally, it assesses the importance of each covariate. Covariate importance is calculated 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)[6].

Unweighted Approach

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

$$\hat{\mu}_{\text{UW}} = \frac{\sum \hat{\theta}_{ij}}{N} \tag{3.14}$$

Unlike in the weighted approach methods, the observations from the primary studies, $\hat{\theta}_{ij}$ 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., 2016b) and regression(see O. Hall et al., 2023).

Assumption of normality

The methods outlined assume that the distribution the effect size; 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,

[6]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)

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. (2016b), 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 Transformation (FTT) 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 FTT is calculated as follows (Freeman & Tukey, 1950; Viechtbauer, 2024):

$$\hat{\theta}_{ij}^{\mathrm{FT}} = g(\hat{\theta}_{ij}) = \frac{1}{2} \cdot \left(\arcsin\sqrt{\frac{s_{ij}}{n_{ij}+1}} + \arcsin\sqrt{\frac{s_{ij}+1}{n_{ij}+1}} \right) \tag{3.15}$$

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

$$\mathrm{Var}(\hat{\theta}_{ij}^{\mathrm{FT}}) = v_{ij} = \frac{1}{4n_{ij}+2} \tag{3.16}$$

To return to the pooled effect sizes natural scale, the Barendregt et al. (2013) back transformation is used, as instructed by Wang (2023):

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

where $\hat{\mu}^{\mathrm{FT}}$ is the (pooled) overall population average and $\overline{\sigma}_{\mathrm{FT}}^2$ is the pooled variance in the transformed scale (Wang, 2023).

Chapter 4

Results

4.1 Descriptive Statistics

A total of 20 studies with 86 effect sizes were included in this analysis, with each primary study reported between one and 27 results. These studies span 18 countries, Figure 4.1 shows a map of distribution and number of results per study.

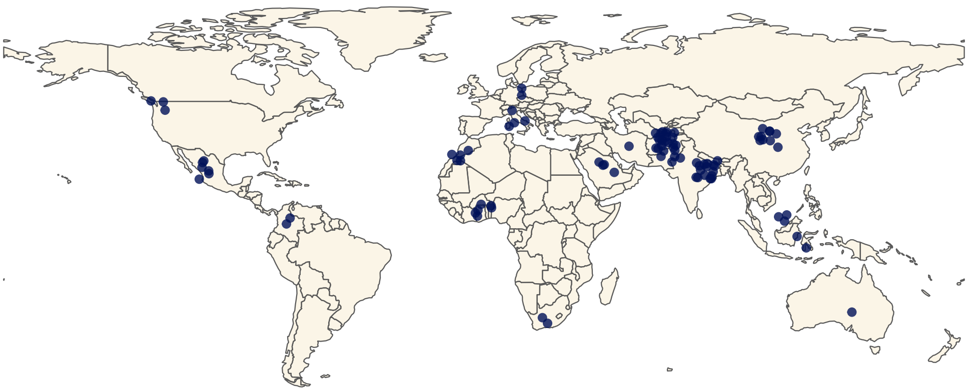

Figure 4.1: map of study location

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 and Communities), 43% targeted SDG 15 (Life on Land), and 13% focused on SDG 2 (Zero Hunger).

Table 4.1: Summary Statistics of Study Features and Performance Metrics

Feature	Overall
Overall Accuracy	0.90 (0.65 - 1.00)
Sample Size	6,401,352.08 (259.00 - 75,782,016.00)
Majority Class Propotion	0.72 (0.14 - 1.00)
Machine Learning Model Group	
Neural Networks	41 (48%)
Tree-Based Models	39 (45%)
Other	6 (7.0%)
SDG Goal	
SDG 11: Sustainable Cities and Communities	38 (44%)
SDG 15: Life on Land	37 (43%)
SDG 2: Zero Hunger	11 (13%)
Classification Type	
Object-level	46 (53%)
Pixel-level	36 (42%)
Unclear	4 (4.7%)
Indices Used	
0	23 (27%)
1	63 (73%)
Ancillary Data	
0	71 (83%)
1	15 (17%)
RS Device Group	
Sentinel	20 (23%)
Landsat	15 (17%)
Other	44 (51%)
Not Reported	7 (8.1%)
RS Spatial resolution (m)	
>1	7 (8.1%)
10	15 (17%)
15-25	16 (19%)
30	8 (9.3%)
Not Reported	40 (47%)
Citation Number	14.84 (2.00 - 68.00)

Note*:* RS: remote sensing, SDG: Sustainable development goals

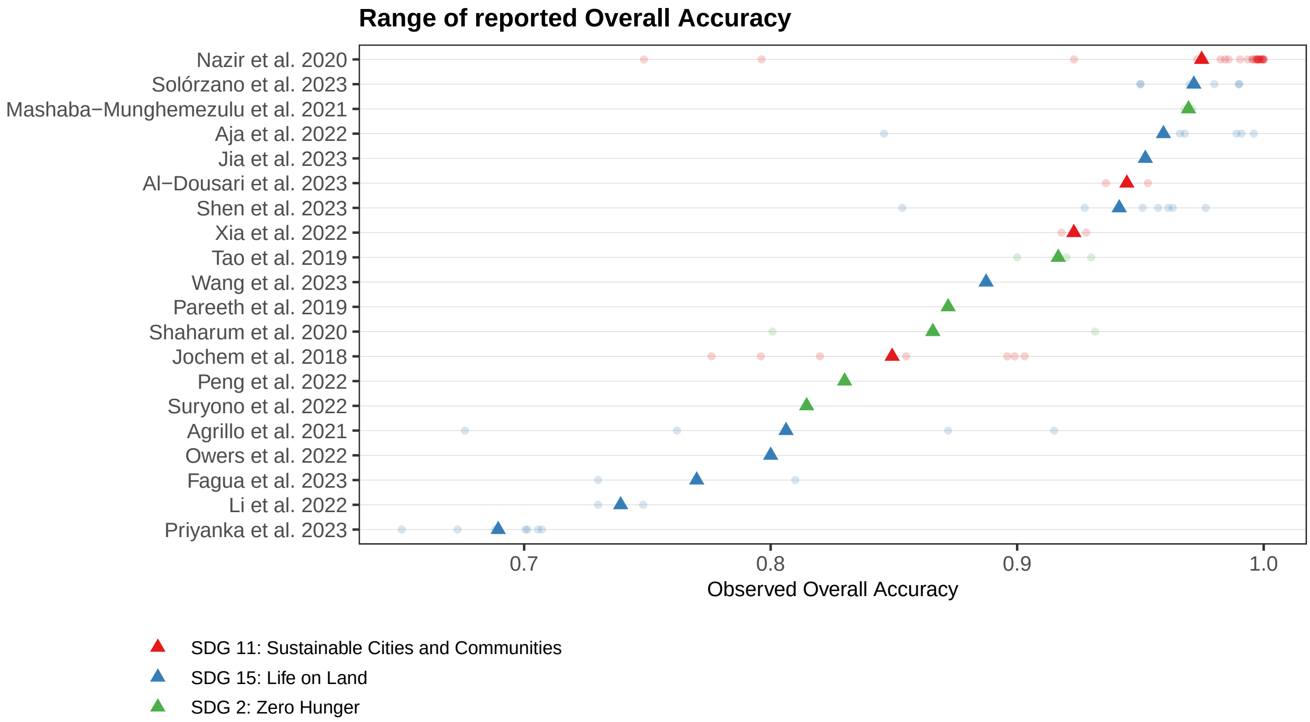

Figure 4.2: Reported overall accuracy from each of the included studies. Individual results are depicted with a circle point and the mean overall accuracy with triangles. The studies are colour-coded according to the SDG goal that the study aims fell under.

As Figure 4.2 and Table 4.1 shows that overall accuracies are not centered around 0.5. Therefore, a transformation is required, Figure 4.3 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 OA still departed significantly from normality ($W = 0.93$, p-value < 0.01)[1].

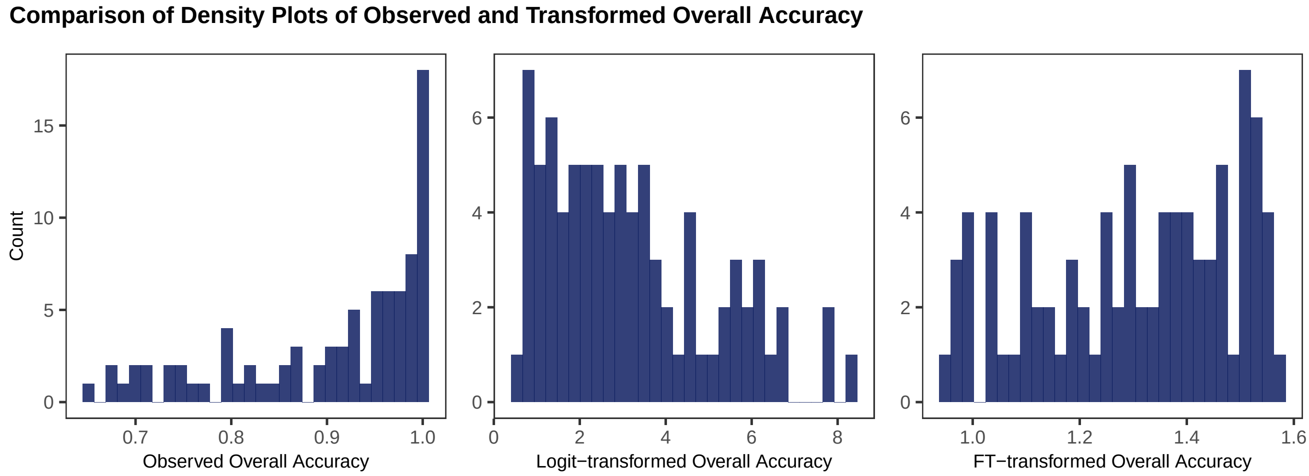

Figure 4.3: Distribution of the observed overall accuracy and transformed by logit and FT transformation.

[1]The limitations of the FT as well as the departure from the normal distribution are discussed in the next chapter Q: However continued anyway? how should I justify that?

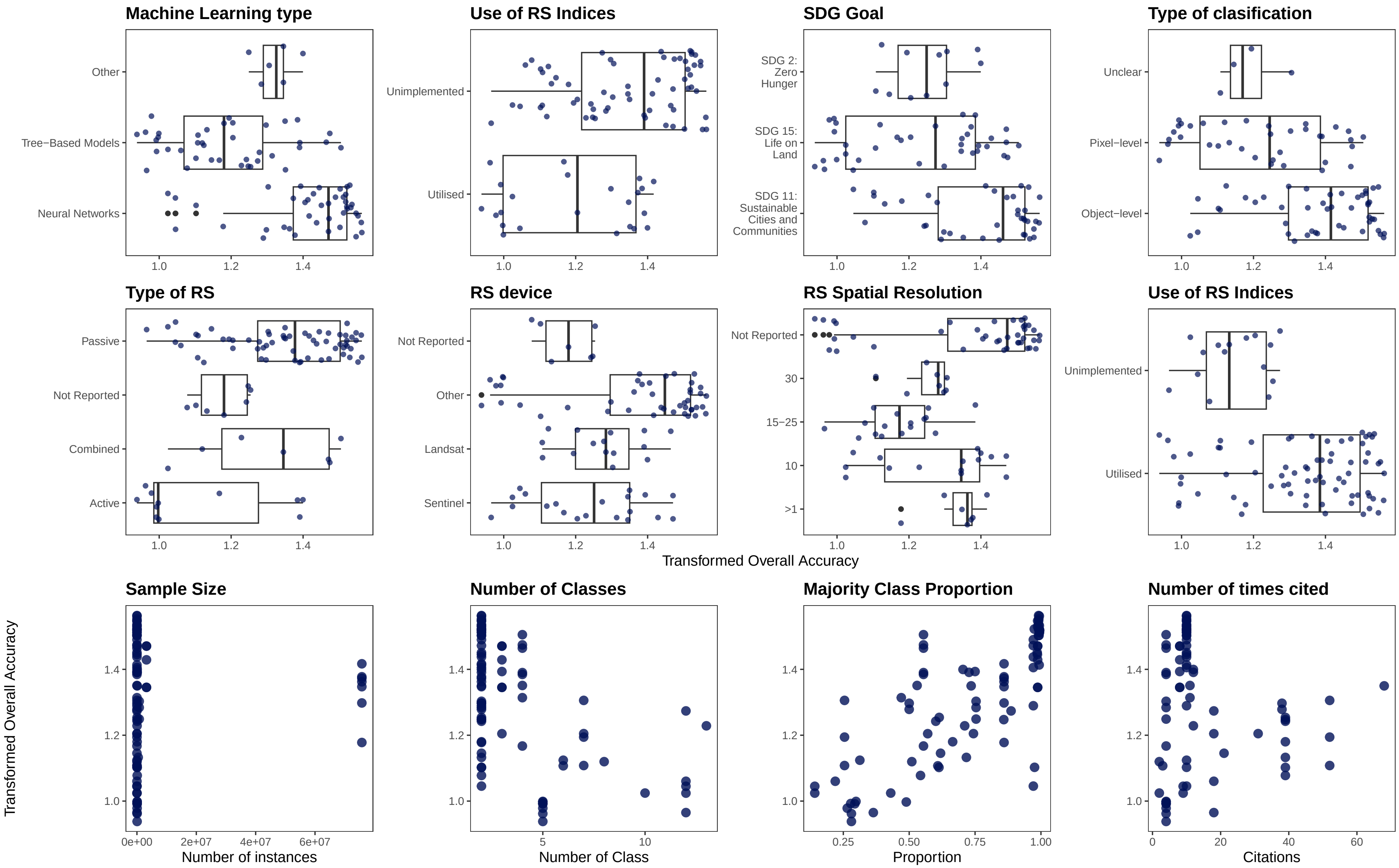

Figure 4.4: Distribution of study features used to try and explain the variation

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. Each study is given with the number of estimates per study, 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.

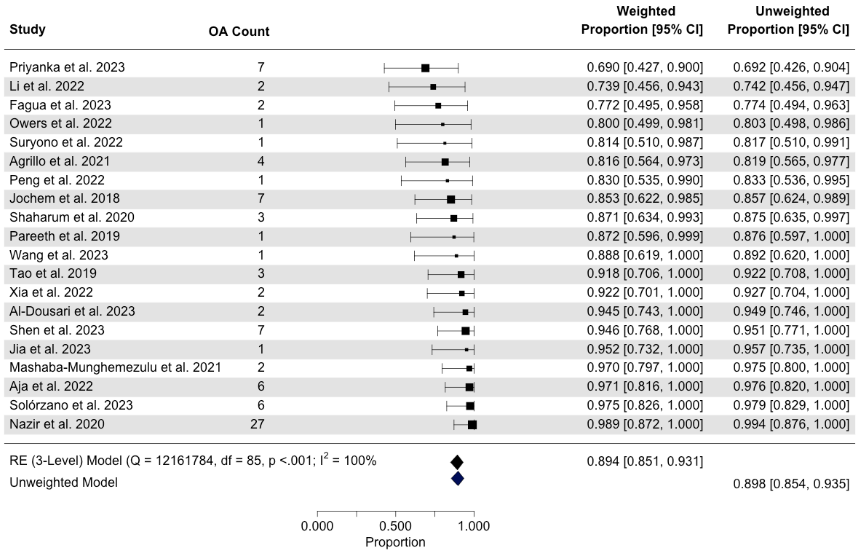

Figure 4.5: Forest plot for both the weighted and unweighted model. OA Count is number of overall accuracy estimates per study, the corresponding average effect size(κ_j) and confidence interval per study for both models is given on the right. The pooled summary effect size based on the three-level RE meta-analytic and unweighted model are given on the bottom. Note, the error bars here are only the ones corresponding to the weighted model but at this scale there is no discernible difference.

On the bottom left of Figure 4.5, the heterogeneity metrics Cochran's Q indicate significant heterogeneity. The percentage of the variance attribution are I^2 Level 3 = 63.62% which is the total variation can be attributed to between-study, and I^2 at Level 2 = 63.62% which is within-study heterogeneity, with negligible fixed effect variance (variance do 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 inconsistency in their results. Table 4.4 shows these results in more detail.

Some of the variance in the model accuracy can be explained by the collected study features. 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 and the use of indices. Less influential predictors include 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. where not included in the models top performing models according to AIC).

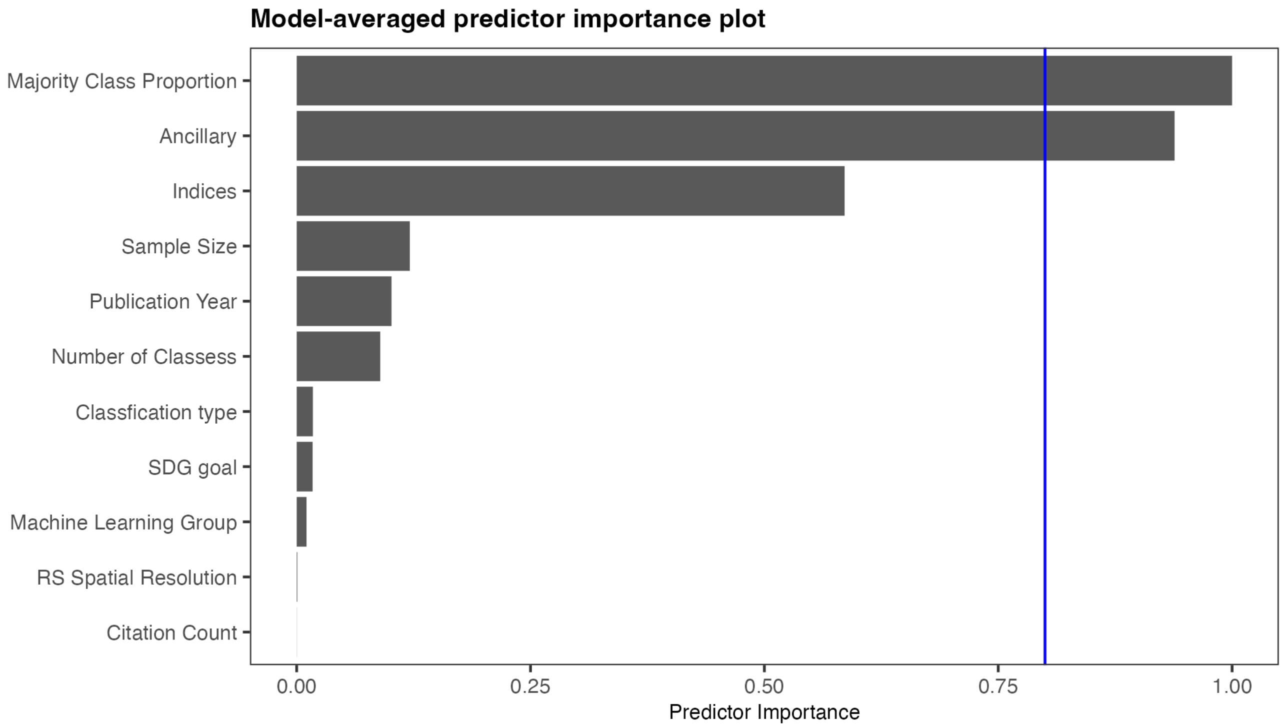

Figure 4.6: Model-averaged predictor importance plot. The averaged importance of each predictor across all models is displayed. The majority class proportion is the most important predictor, followed by the inclusion of ancillary data and the use of indices.

Table 4.2 compares the intercept-only model (a model without covariates) and the full model (which includes all available features), along with the two best-performing models based on multi-model inference. The best-performing model was *Intercept + Majority class proportion + Ancillary+ Indices* (Model 22); followed closely by *Intercept + Majority class proportion* (Model 6). The table presents the degrees of freedom (df), AIC, BIC, and p-values from ANOVA tests comparing the models. All models that include covariates significantly improve upon the intercept-only model ($p < 0.001$). Model 6 has the lowest AIC and BIC, indicating it is the best model overall. However, when comparing Model 6 to Model 22, the difference is not significant at the 0.05 level, meaning the addition of "ancillary data" does not offer substantial improvement ($p = 0.045$). Similarly, the full model does not provide further improvements over the simpler models (p = 0.540).

Table 4.2: Model Comparison

				Comparisions	
Model	df	AIC	BIC	Intercept only	Model 6
Intercept only	3	-111.04	-103.71	NA	NA
Model 6	4	-127.28	-117.56	0.000	NA
Model 22	6	-125.18	-110.74	0.000	0.045
Full model	17	-88.85	-50.39	0.004	0.540

Note*:* Comparison of model performance based on degrees of freedom (df), Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and p-values from likelihood ratio tests (ANOVA). Model 6: *Intercept + Majority class proportion*; Model 22: *Intercept + Majority class proportion + Ancillary + Indices*; The Full Model: *Intercept + Majority class proportion + Ancillary + Sample Size + Publication Year + Number of Classes + Classification type + SDG goal + Machine Learning Group + RS Spatial Resolution + Citation Count.* Lower AIC and BIC values indicate better model fit, and p-values less than 0.05 suggest significant improvements over simpler models.

Table 4.3 shows the estimated coefficients of Model 22, both in the FT transformed scale (b) and on the natural scale (b back-transformed). This shows that the proportion of majority class has the largest positive impact on the model's outcome ($b = 0.39$, $p < .001$), while the inclusion of ancillary data has a small but significant negative effect ($b = -0.11$, $p = 0.029$). The use of indices has a minimal and non-significant effect ($b = 0.06$, $p = 0.131$). The intercept is also significant, indicating a strong baseline effect ($b = 0.99$, $p < .001$).

Table 4.3: Coef

Predictor	b	SE	t	p	b back transformed
intercept	0.986	0.06	17.22	0.000	0.696
ancillary1	-0.113	0.05	-2.22	0.029	0.011
fraction_majority_class	0.394	0.08	4.93	0.000	0.145
indices1	0.065	0.04	1.53	0.131	0.003

Figure 4.7 illustrates the relationship between the proportion of the majority class and overall accuracy of the individual studies included in the meta-analysis. The plot is based on Model 22, with the solid black line representing the fitted regression line and the shaded area indicating the 95% confidence interval. Each point (bubble) represents a study, with its size proportional to the weight it received in

the analysis (larger points indicate studies with more influence). The plot shows that as the proportion of the majority class increases, overall accuracy tends to improve.

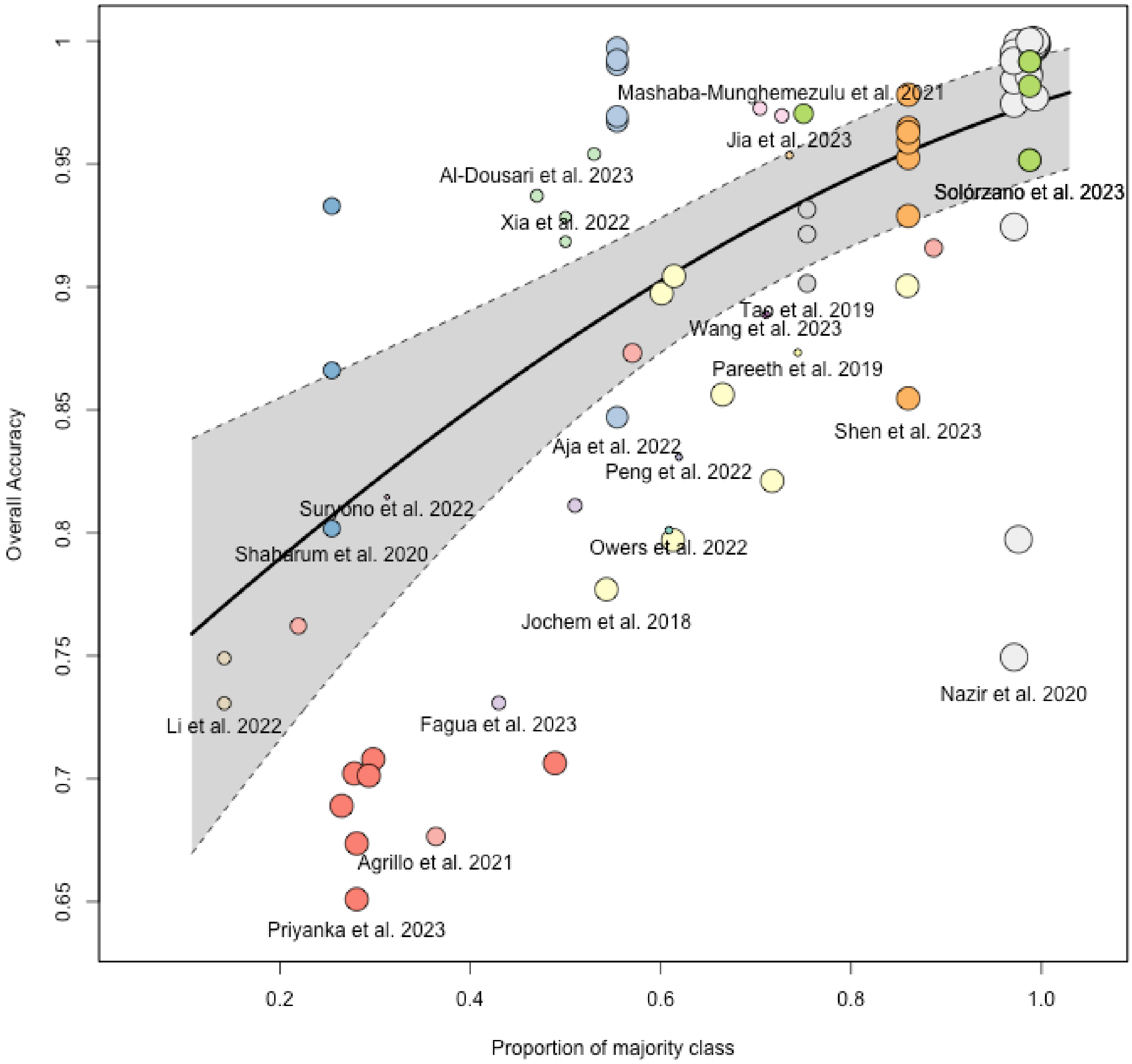

Figure 4.7: Bubble plot showing the observed effect size, overall accuracy of the individual studies plotted against a the proportion of the majority class. Based on the mixed-effects meta-regression model, the overall accuracy as a function of proportion of the majority with corresponding 95% confidence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

Figure 4.8 shows the observed overall accuracy against the predicted overall accuracy's made by Model 22. The points are coloured by the addition of ancillary information in the primary study. It appears that the addition of ancillary information leads to a lower overall accuracy, however, this could be due to a number of unmeasured factors, such as study's with more complicated classifications (more similar classes) adding accuracy data. As Figure 4.8 shows Model 2 over estimates the overall accuracy — the fit regression line (in grey) is above the line of perfect agreement ($y = x$, in black).

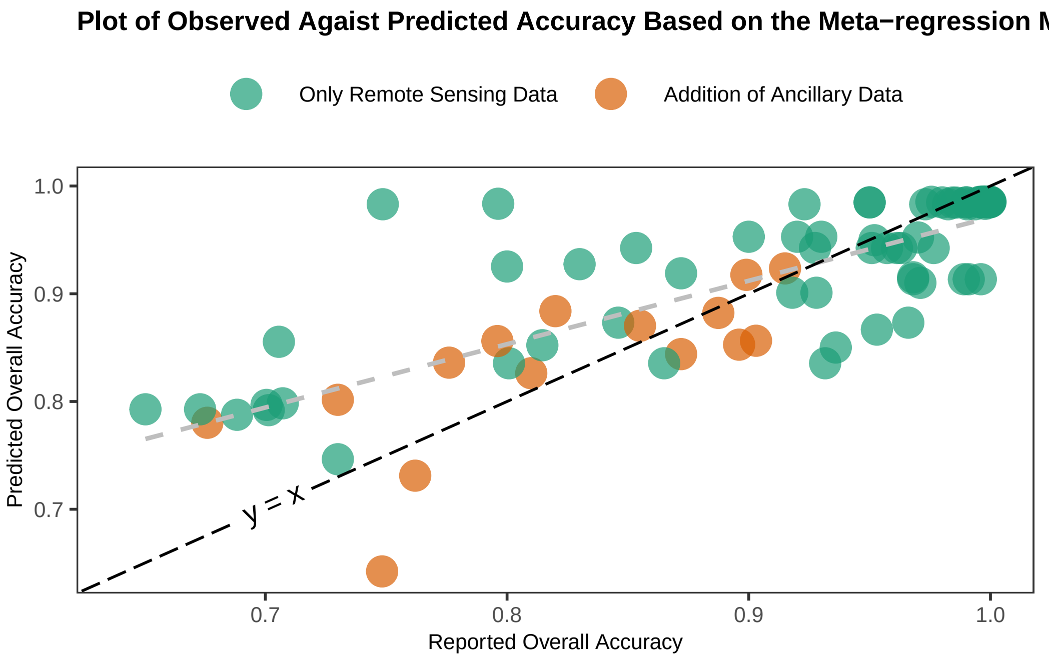

Figure 4.8: Observed and predicted overall accuracy. The colour indicates whether the addition of ancillary data in the primary study's model. The line of perfect agreement $y = x$ is in black and fit regression line in grey.

Lastly, Table 4.4 shows the parameter estimates of the meta-analysis comparing the Intercept Only Model and the Mixed Effects Model(Model 22). The Intercept Only Model estimates a population effect size $\hat{\mu}$ of 0.894, with estimated variance components $\sigma^2_{\zeta} = 0.017$ and $\sigma^2_{\xi} = 0.01$. Model 22, reduces the population effect size to 0.696 (95% CI: [0.579, 0.801]). The within-study variance (0.009) remains similar, while the between-study variance is reduced to 0.005, indicating less variability between studies when accounting for covariates. This shifts is also reflect in the I^2 Level 2 goes from 36.38% to 63.46% and at Level 2 from 63,62% to 36.54%, indicating less variability between studies when accounting for covariates. The total I^2 consistently being 100% in both models indicates that almost none of the variation between effect sizes can be attributed to sampling error, this might suggest that the included studies are too different from each to compare (see discussion for apples and oranges problem). Both models show significant heterogeneity (Cochran's Q, $p < 0.001$) results. The R^2values show that the covariates in the Mixed Effects Model explain 69.9% of the variance at Level 3 and 8.6% at Level 2.

Table 4.4: Heterogeneity Metrics

Parameter	Intercept Only Model	Mixed Effects Model
mu	0.894 [0.848, 0.933]	0.696 [0.579, 0.801]
within-study sig	0.01 [0.007, 0.014]	0.009 [0.007, 0.013]
between-study sig	0.017 [0.009, 0.038]	0.005 [0.001, 0.016]
Cochran's Q	Q(85) = 12161783.98, p <.001	Q(82) = 11440330.6, p <.001
I2 Level 2	36.38	63.46
I2 Level 3	63.62	36.54
R2		Level 2 : 0.086 Level 3 : 0.699

Chapter 5

Discussion

Central Findings

The results of this meta-analysis demonstrate the 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. However, 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 negative effect on overall accuracy performance, which is counterintuitive. Perhaps this is explained by other variables not captured in this study, for example researchers addressing more complex classification problems use models with ancillary data.

To the best of available knowledge, this is the only meta-analysis of remote sensing methods that utilized weighted estimates. The use of a three-level random effects model enabled the decomposition of variance into within-study and between-study components, offering insights into the observed heterogeneity. No meaningful difference was found between the weighted and unweighted approaches.

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 subjectively 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 study included a total of 20 studies. While several simulations 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 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, 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).

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 around 90%. Accuracy, is easy to understand and compute, as addressed it does not take into account class imbalance. When models are developed and tuned to maximize accuracy on training data, they often perform poorly on unseen data, inflating the performance metrics, but the reporting of training and test results was inconsistent.

5. **Study features included:** The analysis would have benefited from the inclusion of more study features. For example, to better understand the effect of ancillary data, a feature representing the complexity of the problem addressed by the primary study could explain the negative effect of additional information on a prediction model. 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 organs 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"* (**?**, 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 circles back to the issue of the sample size. A large sample size is needed to have a high enough power to make confident conclusions.

7. **Cochran's Q and large sample sizes:** Another limitation of this analysis, is the use of Cochran's Q for testing heterogeneity. The power of the Q-statistic is dependent on the number of included effect sizes (n) and the precision of the studies (i.e., the sample size with in that study). In cases with large primary-sample-sizes, the Q statistic becomes highly sensitive to even minor differences between studies. The Q-statistic is "overpowered", which result in the detection of statistically significant heterogeneity even when the actual differences between studies are small. This sensitivity may exaggerate the extent of heterogeneity, potentially lead to misleading conclusions about the variability among the included studies. Little research has been done on the effect of very large primary-sample-sizes since meta-analysis typically compile studies who's unit of analysis in on a patient level. Primary-sample-sizes in the millions is not a common issue.

8. **Transformation of the effect size:** Although the distribution of the transformed overall accuracy was closer to a normal distribution, it remained significantly non-normal. Furthermore, the use of FT transformation is highly contested in the literature (Doi & Xu, 2021, 2021; Lin & Xu, 2020; Röver & Friede, 2022, 2022; **?**) because of several important limitations. First, the FT is notably unintuitive, notably the calculation of variance relies on the structure of an arcsine function's derivative. As seen in Equation 3.17, the FT transformation has a complex back-transformation. In this analysis, the pooled variance was used for the back-transformation, addressing main issue debated in the literature—namely, that using harmonic mean of primary-sample-sizes to back-transform the pooled effect size can lead to misleading results (Doi & Xu, 2021; Lin & Xu, 2020; Röver & Friede, 2022; Wang, 2023; for details see **?**). The choice of

back-transformation method significantly influences the outcome, and justifying specific method becomes especially challenging in a multilevel data structure. 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 generalizability 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, with only 20 studies included. Future research should aim to expand the pool of included studies. This would mean that interaction effects of 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, n.d.), 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 used, incorporating precision, recall, F1-score, and AUC to provide a more comprehensive evaluation of model performance. More than one effect size can be modeled using network meta- analysis models. The inclusion of more permortance 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, including the proportion of the majority class and the inclusion of ancillary data. 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 timeframe 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 formate). Although the improvement was not formally assessed in this study, the difference was striking. Some research has allready examined the potential accuracy of large language models (LLMs) in data extraction for meta-analyses, with promising results (Mahuli et al., 2023).However, 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 tends in outcomes.

Chapter 6

Conclusion

This meta-analysis highlights the considerable variability in the predictive performance of machine learning models applied to remote sensing data for SDGs. While the proportion of the majority class and the inclusion of ancillary data were significant factors influencing overall accuracy, no significant differences were observed between the types of models used (e.g., neural networks or tree-based models) or the specific SDGs studied. Interestingly, the negative impact of ancillary data on model performance, though counterintuitive, suggests the need for further investigation. The use of a three-level random effects model here provided a deminstraction of the potential insights into the sources of heterogeneity, however because of the limited data few conclusions can currently be made. The strongest finding of this paper is the limitation of overall accuracy as a performance metric, as seen from the reduction in the average effect size once the proportion of the majority class was controled for the overall accuaary dropped from 0.90 to 0.70. Future studies and guildlines should expanding the pool of studies, and incorporate a broader range of performance metrics, and exploring additional study features. Furthermore, more research is needed to improve the robustness and applicability of meta-analyses methods to this field. In conclusion, despite the limitations, this study provides a valuable starting point for understanding the variability in machine learning model performance for remote sensing applications in SDG monitoring.

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

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

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

Heydari, S. S., & Mountrakis, G. (2018). Effect of classifier selection, reference sample size, reference class distribution and scene heterogeneity in per-pixel classification accuracy using 26 Landsat sites. *Remote Sensing of Environment*, *204*, 648–658. https://doi.org/10.1016/j.rse.2017.09.035

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. (2016b). 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

Khatami, R., Mountrakis, G., & Stehman, S. V. (2016a). 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

Lu, D., & Weng, Q. (2007). A survey of image classification methods and techniques for improving classification performance. *International Journal of Remote Sensing*, *28*(5), 823–870. https://doi.org/10.1080/01431160600746456

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

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

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. (n.d.). *metafor: Meta-Analysis Package for R.* https://doi.org/10.32614/CRAN.package.metafor

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. (2024). *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.

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