nature portfolio

Corresponding author(s): Keiichiro Kanemoto

Last updated by author(s): Apr 20, 2026

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Data collection and preprocessing were conducted using Google Earth Engine (GEE), QGIS, and R. These tools were used for satellite image processing, feature extraction, spatial data integration, and geospatial analysis. Custom code developed in this study is available at https://github.com/yutongcheng/mining-rf-classification-example and archived at https://doi.org/10.5281/zenodo.19648354.
Data analysis	Data analysis was conducted using Google Earth Engine (GEE), QGIS, and R for machine learning classification and spatial analysis. The hierarchical random forest (HRF) model was implemented in R using the randomForest package and trained on 26 ASTER band ratios for both the ASTER and HISUI datasets. Geospatial analyses, including deforestation and biodiversity risk calculations, were performed using QGIS and R. Statistical validation of the classification model was carried out using 10-fold cross-validation, implemented through the caret and DMwR2 packages. Synthetic Minority Over-sampling Technique (SMOTE) was applied for class balancing using smotefamily. Additional statistical and geospatial processing was performed using gee, stats, and dplyr, with ggplot2 used for data visualisation.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Author-generated example datasets, together with grid-level commodity classification results in GPKG format, are publicly available through the Zenodo archive (https://doi.org/10.5281/zenodo.19648354). Additional publicly shareable materials are provided in the Supplementary Information and Supplementary Data 1–3, as cited in the main text. The example datasets and grid-level results are provided under CC BY-SA 4.0. The full training dataset cannot be publicly shared because part of it contains data derived from the licensed commercial S&P Global Market Intelligence SNL Metals & Mining Database, whose redistribution is restricted by the data provider's terms of use. Previously published third-party datasets analysed in this study are available from their original sources, as cited in the manuscript.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	This study does not involve human participants, their data, or biological material. Therefore, sex and gender were not considered in the study design, and no disaggregated sex or gender data were collected.
Reporting on race, ethnicity, or other socially relevant groupings	This study does not involve human participants, and no race, ethnicity, or other socially relevant categorization variables were used. As such, there was no classification based on social groupings, and no confounding variables related to these aspects were analyzed.
Population characteristics	As the study does not involve human participants, there are no covariate-relevant population characteristics such as age, genetic information, or health-related variables. This criterion does not apply to the present research.
Recruitment	No human participants were recruited for this study. The dataset used consists of spatial and remote sensing data obtained from publicly available sources and proprietary datasets, without any direct involvement of individuals or their personal information.
Ethics oversight	Since this study does not involve human participants, their data, or biological materials, no ethics approval was required. The research focuses on the classification of mining areas using remote sensing and machine learning techniques, ensuring that all data used comply with ethical and legal standards.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☐ Behavioural & social sciences ☒ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	This study integrates remote sensing, machine learning, and cloud computing to classify 71,041 mining sites globally by commodity type. It quantifies mining-induced deforestation and biodiversity risks, producing spatially explicit commodity-specific maps. The study assesses 20 extracted commodities and their associated environmental impacts from 2001 to 2022.
Research sample	The study examines 71,041 mining polygons covering a total of 113,068 km². These polygons were obtained from Maus et al. (2022) and Tang and Werner (2023) and supplemented with external datasets, including the open-pit locations dataset (Cheng et al., 2025), the SNL Metals & Mining dataset, USGS data, and datasets from Jasansky et al. (2023) and Franks et al. (2021).
Sampling strategy	Mining areas were initially extracted from Maus et al. (2022) and Tang and Werner (2023), and commodity labels were assigned to polygons through spatial joins with datasets from SNL Metals & Mining, USGS, Jasansky et al. (2023), and Franks et al. (2021). Open-pit locations within these polygons were identified using the open-pit locations dataset (Cheng et al., 2025). Commodity classification was then performed using a hierarchical random forest (HRF) model, which utilised ASTER-derived spectral features (band ratios) for both the ASTER and HISUI datasets. The HRF structure categorised commodities into hierarchical levels to improve classification accuracy and robustness.
Data collection	The study utilises ASTER and HISUI satellite imagery (14 spectral bands and 185 hyperspectral bands), mining polygon datasets from

Data collection	Maus et al. (2022) and Tang and Werner (2023), the open-pit locations dataset from Cheng et al. (2025), and mining site databases from SNL Metals & Mining, USGS, Jasansky et al. (2023), and Franks et al. (2021). Commodity classification was performed using machine learning-based hierarchical random forest (HRF) models trained on 26 ASTER band ratios, applied to both the ASTER and HISUI satellite imagery. Environmental impact data included deforestation records from 2001 to 2022, sourced from the Hansen Global Forest Change dataset, and biodiversity risk indices derived from IUCN Red List species threat data (RLI).
Timing and spatial scale	The study assesses mining activities and their environmental impact on deforestation from 2001 to 2022. The dataset includes 71,041 classified mining sites distributed globally across tropical, subtropical, temperate, and other regions. Key mining locations include the Amazon Basin in South America, Indonesia and the Philippines in Southeast Asia, West and Central Africa, and Australia.
Data exclusions	Mining polygons that neither contained open pits nor included dams, waste rock dumps, or facilities, and were not located within 2 km of a known mining polygon with a pit area, were excluded from commodity classification. Underground mines were also excluded, as the Maus et al. (2022) and Tang and Werner (2023) polygon datasets capture only mining-related activities visible in satellite imagery, and ASTER and HISUI imagery are primarily suited to detecting surface mining operations. Deforestation prior to 2001 was not considered, as the Hansen Global Forest Change dataset provides annual deforestation data only from 2001 onward.
Reproducibility	The study utilises publicly available datasets, including ASTER and HISUI imagery, mining polygon datasets from Maus et al. (2022) and Tang and Werner (2023), the open-pit locations dataset from Cheng et al. (2025), and mining site datasets from USGS, Jasansky et al. (2023), and Franks et al. (2021), as well as Global Forest Change data. Commercially licensed data from SNL Metals & Mining were also incorporated. Machine learning models were trained and validated using an 80:20 data split and 10-fold cross-validation. Model accuracy was assessed using external mining datasets, including the Global Iron Ore Mines Tracker, the Global Coal Mine Tracker and Northey et al. (2017). A stratified random-sampling validation was additionally conducted, with accuracy estimates and confidence intervals summarised in Section S8 of SI Part 1. The methodology is fully replicable using the same data sources and processing workflow.
Randomization	The study does not involve experimental groups or treatment conditions requiring randomisation. During the training sample preparation phase, a spatial join was performed to integrate the commodity point datasets with the mining polygons, assigning commodity labels to polygons where spatial overlap was detected. To reduce potential misassignments, only polygons that fully contained exactly one verified point were retained, ensuring a one-to-one correspondence between polygons and commodity types. As a result, 2,212 polygons were labelled with commodities across 20 classes and used as the training dataset for classification (Table S5 of SI Part 1). Training samples for both the open-pit identification model and the commodity classification model were then randomly split into training and validation/test subsets using an 80:20 cross-validation approach.
Blinding	Not applicable – the study does not involve human participants, subjective assessments, or experimental groups requiring blinding.

Did the study involve field work? ☐ Yes ☒ No

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Plants

Seed stocks	*Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.*
Novel plant genotypes	*Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.*
Authentication	*Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosiacism, off-target gene editing) were examined.*