EARTH OBSERVATION LAND ANIMAL SURVEY

TAIM-2
AI Accuracy Framework Design Document
V1.0

Copyright

CONTENTS

1. OVERVIEW

AI driven computer vision technologies are transforming the way maps are created. It is now possible to process vast areas of imagery which would previously be prohibitive due to the effort required in creation. Using techniques such as machine learning it is possible to classify features or predict values over an extent near instantaneously. This has potential to help emerging nature markets operate with high levels integrity, by providing accurate maps reflecting the current state of terrestrial ecosystem assets. However, AI approaches have a distinct disadvantage when compared to manual creation methodologies as the technologies are new, do not have years of heritage, the scientific and methods used to construct algorithms for analysis and presentation of data are not transparent, therefore AI approaches are not trusted.

The Trustable AI Mapping – Phase 2 (TAIM-2) project aims to increase trust in AI derived mapping by establishing a framework for the direct measurement of accuracy versus ground conditions. By doing this in line with existing geospatial industry practices, the project will ensure that these new technologies fit within existing guidance, whilst also establishing core parameters for accuracy measurement for comparison of results. This will create transparency in the analysis, learning models and their processes, and subsequently increase trust in this new and transformative field.

The framework needs to consider all the errors inherent in a geospatial AI classification system, and how these relate to end-to-end accuracy of the algorithm outputs vs ground condition (obtained from ground survey).

Categories of error under consideration include:

- AI-induced:
    - Errors resulting from the model and data used to train it.
    - Errors resulting from a poorly chosen model.
    - Errors from unrepresentative training data. For example, it is common for training datasets to be imbalanced, resulting in poor performance for under-represented classes.
- Geospatial:
    - Positional error from poor image rectification, errors resulting from image resolution.
    - Errors from variation in illumination and shading. For example, using a 10m pixel to identify peat gullies that are 1-2m wide will induce error from mixed spectral responses.

- Ecological:
    - Errors that result from uncertainties in estimation of parameters such as structural diversity, heterogeneity and species richness from remote sensing data.

Turning to the error evaluation framework, this document will look across the different error categories and identify common metrics can be applied across all use cases as well as more specific error metrics that may only be suitable for a subset of use cases.

Evaluation of AI induced error is typically limited to evaluation of map outputs against ground data with specific metrics applied depending on the AI algorithm and use case. For example, the intersect over union metric is typically only applied to object detection classifications and would not make sense for evaluating a continuous map. There can also be statistical errors or overestimates of model performance, due to improper selection or splitting of the training and validation data. Inclusion of sample data from points that are at the same location or near one another can result in spatial autocorrelation effects, making the model look better than it is. However, these sources of error are rarely evaluated or reported on.

Geospatial positional errors can be evaluated by comparison against known ground control points, while resolution errors can be estimated by directly comparing image pixels against the detailed ground data. Illumination errors are harder to evaluate as conditions vary day to day, but areas of poor signal to noise can be identified in raw image data. Ecological errors can be evaluated by direct comparison between fieldwork data and AI retrievals, for example by comparing AI-based tree height retrievals against ground data. Given their complexity, a research paper focused on ecological error will be delivered during the project and will be used to enhance the evaluation of these errors within the framework.

2. AI MAPPING USE CASES IN REMOTE SENSING

Remote sensing, or Earth Observation (EO), is the use of sensors that use measurements of the electro-magnetic spectrum to infer properties of the surface of the Earth. These sensors can be passive (e.g. cameras and multispectral imagers) or active (e.g. lidar and radar) and can be carried by aircraft, drones or satellites. Over the last decade there has been a significant increase in the number and diversity of EO satellites (see Figure 1), with a corresponding increase in the volume of EO data. With the advent of constellations such as Planet daily high-resolution (i.e. sub 10m) observations are now viable.

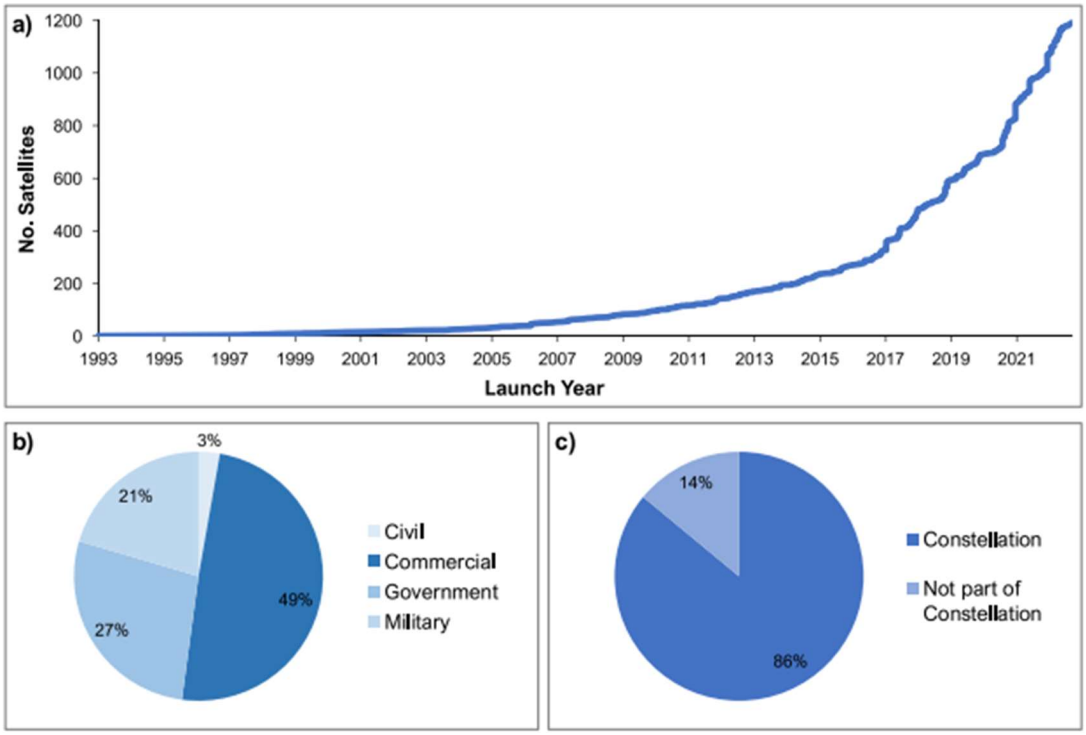

Figure 1: Active EO satellites: a) total number launched, b) percentage of satellites in different sectors, c) percentage of EO satellites in constellations. Adapted from (Wilkinson, et al., 2024)

AI, in the form of machine learning, has a well-established lineage in remote sensing, being used for use cases such as land cover mapping, infrastructure, forest biomass estimation and change detection. This is largely driven by the high spatial, spectral and temporal dimensionality of remote sensing data, meaning that there are many more variables to evaluate than are practical for a human analyst. The complexity of AI models applied to remote sensing use cases has increased substantially over recent years, tracking developments in the wider field. A review by Janga (2023) identified the following key techniques used in remote sensing:

- Conventional Machine Learning:
  - Ensemble decision-tree-derived classifiers
  - Random forest
  - Support vector machines
  - Extreme Gradient Boosting (XGBoost)
- Deep learning:
  - Deep Convolutional Neural Networks (DCNNs)
  - Deep Residual Networks (ResNets)
  - You Only Look Once (YOLO)
  - Faster Region-Based CNN (R-CNN)
  - Self-Attention Methods
  - Long Short-Term Memory, LSTM
- Other AI methods:
  - Generative adversarial networks (GAN's)
  - Super-resolution generative adversarial network
  - Deep Reinforcement Learning (DRL)

Considering the types of outputs, remote sensing derived maps can be separated into two groups: categorical and continuous. Categorical maps are split into pre-defined categories on the understanding that discrete objects or features can be identified in the landscape. Examples include landcover maps, such as the UKCEH Landcover Map (Figure 2) or Habitat Map of Scotland (Figure 3), and object detection, such as solar panel identification (Figure 4). Continuous maps instead show how a continuous variable, such as canopy height (
Figure 5), changes spatially.

Regardless of the AI technique applied, these will typically require large, training datasets to teach the model how best to interpret the remote sensing data for a given use case. Training data may consist of a combination of pre-existing datasets, field survey data and manually labelled imagery. It is also common to hold back randomly selected sets of training data for evaluation of the model and output maps.

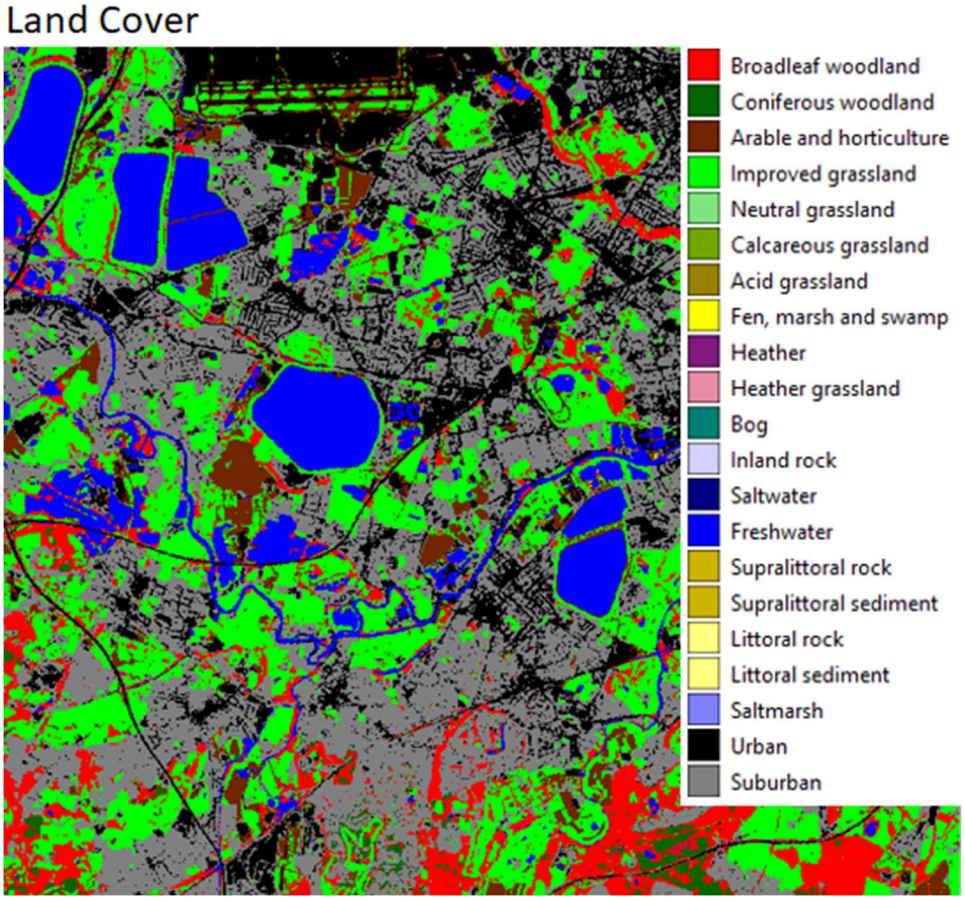

Figure 2: excerpt of the UKCEH Landcover map[1]

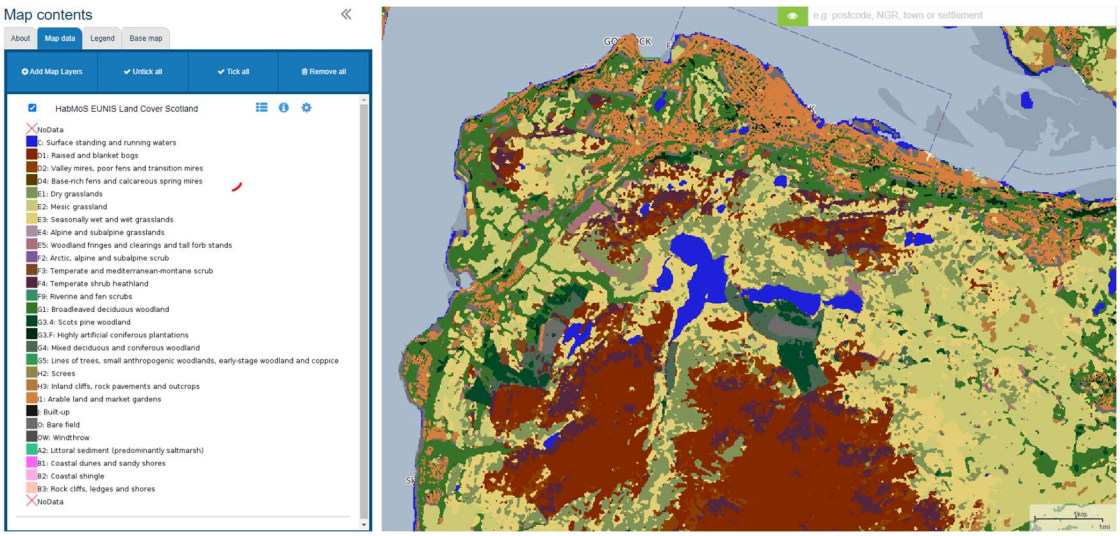

Figure 3: Excerpt of the Habitat Map of Scotland[2]

[1] https://www.ceh.ac.uk/latest-land-cover-map-provides-greater-detail-about-british-landscape
[2] https://map.environment.gov.scot/

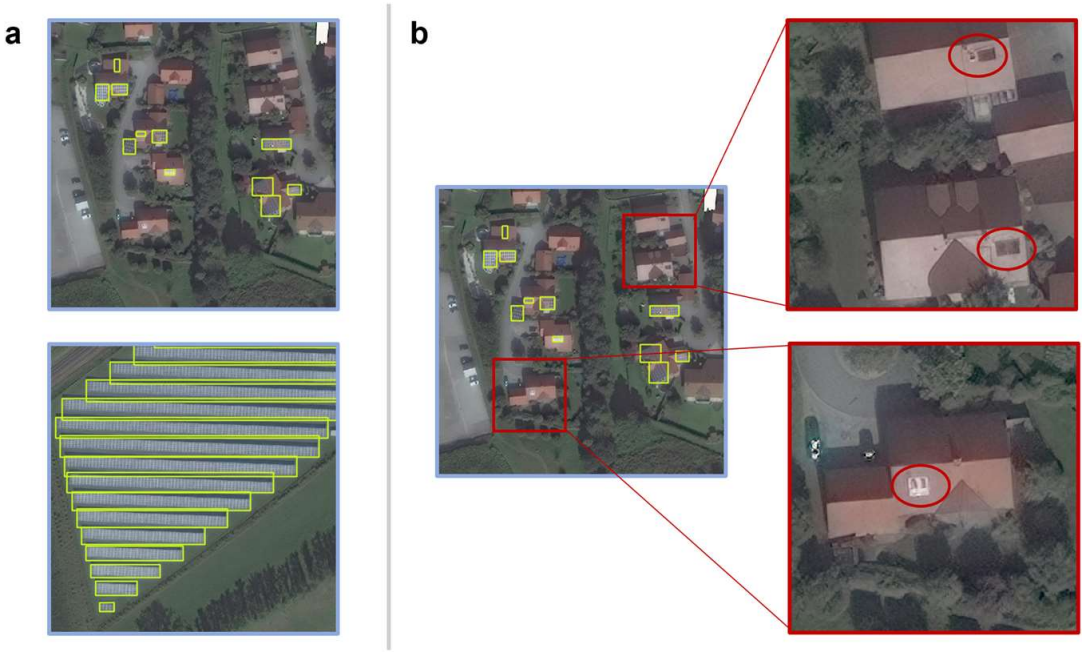

Figure 4: Examples of solar panel objects and non-solar panel objects. (a) Single solar panels in residential areas were labelled with a unique bounding box, labelled in yellow, where individual panels were determined by differences in size, shape, and spacing (top). For large ground-mounted panels, one label constitutes one row (bottom). The HD labelling windows illustrating confirmed panels are outlined in blue. (b) Objects on residential roofs with less than three distinct panel shapes did not meet the necessary criteria to be confidently identified as solar panels and were treated as non-solar panel objects. The windows illustrating non-solar panel objects are outlined in red, with specific non-solar panel objects circled in red. Adapted from Clark & Pacifici (2023)

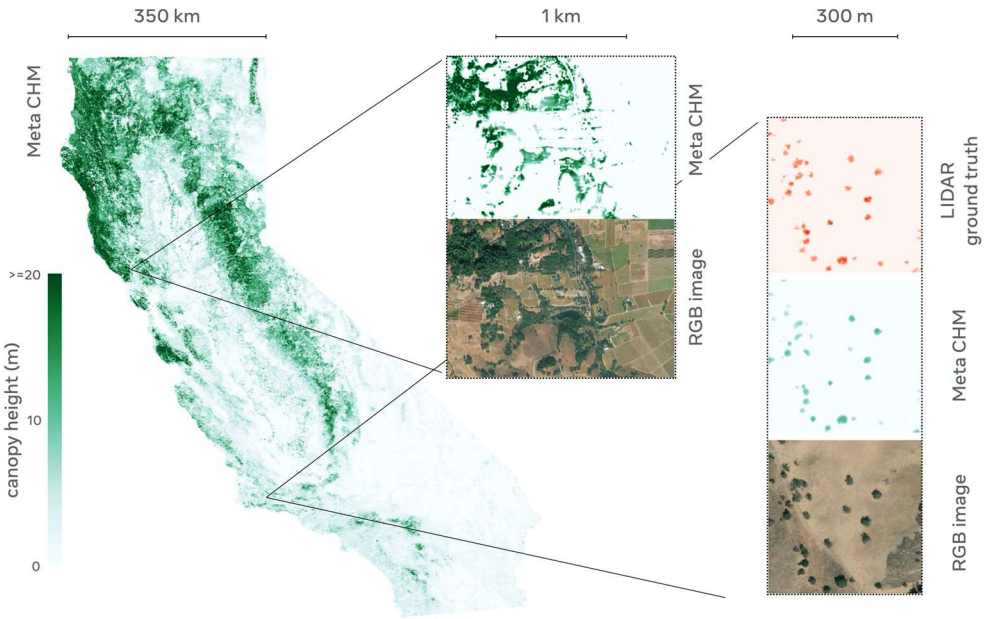

Figure 5: Canopy Height Map (CHM) for California, with inset showing zoomed-in region with input RGB imagery and LIDAR ground truth[3]

[3] Meta, https://research.facebook.com/blog/2023/4/every-tree-counts-large-scale-mapping-of-canopy-height-at-the-resolution-of-individual-trees/, last accessed 19/06/2024

The accuracy of the AI mapping that is achieved depends on the AI method chosen and if the maps are categorical or continuous. For categorical maps the foundation of accuracy assessment is the confusion matrix in which predicted classes are compared against known reference data. Figure 6 shows a hypothetical confusion matrix with correct classifications falling along the diagonal (blue cells), and errors for each class can be analysed in terms of errors of commission (orange cells) and omission (yellow cells). Standard error metrics that can be computed from the confusion matrix such as overall accuracy, users accuracy (complement or commission error), user accuracy (complement of omission error), Kappa index of agreement and the F1 score, (see Foody (2002) for a more complete discission of accuracy metrics). For continuous maps, the scatter plot is at the core of accuracy assessment with predicted values compared against observed values from a reference dataset. Standard metrics of error include R^2, root mean standard error (RMSE) and mean absolute deviation (MAD), all of which quantify how well the predictions match reference data.

		Reference image			
		A	B	C	Total
Classified image	a	37	3	7	∑a = 47
	b	9	25	5	∑b = 39
	c	11	2	43	∑c = 56
	Total	∑A = 57	∑B = 30	∑C = 55	N = 142

Figure 6: Example of confusion matrix showing the numbers of samples of each predicted class and the correct label for that observation[4].

Providing an assessment of final map accuracy using some or all of the accuracy metrics outlined above provides some insight into how well the EO and AI derived map functions but does not provide insight into where in the processing chain the error has originated. Elmes *et al* (2020) provided an excellent breakdown of how errors can propagate from training data, highlighting the need to directly assess errors in training data to communicate their impact on AI generated maps. In many ways this project can be seen as an extension of these ideas to include errors from throughout processing chain.

[4] http://www.50northspatial.org/classification-accuracy-assessment-confusion-matrix-method/

3. SOURCES OF ERROR

The workflows of AI-derived mapping projects vary significantly based upon the EO data, models and use case. However, across all these projects there are some common steps that we expect to be present, shown in Figure 7 .

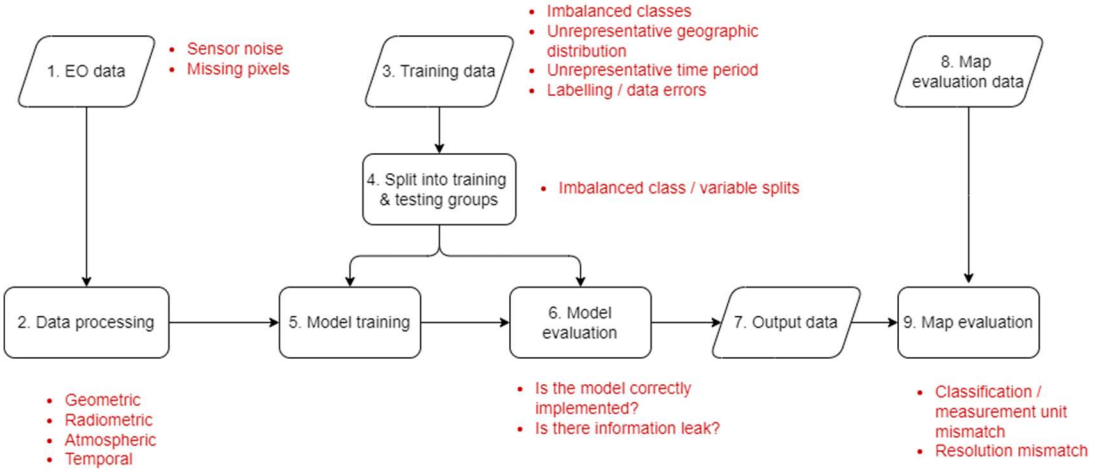

Figure 7: Simplified AI mapping project flow with example errors for the different stages shown in red.

Each of the workflow stages, and the associated sources of error, are discussed in detail below.

1. EO data. Can consist of satellite imagery, aerial photography, lidar point clouds, or any combination of the above.
2. Data Processing. The correction and harmonization of EO data. The exact corrections undertaken will depend on the data source, but common ones, with associated errors, are:
    - Geometric. Locates data in a real-world coordinate system. Errors will result in misplaced EO data, and can be non-uniform in scale across the scene.
    - Radiometric / atmospheric. Converts raw at-sensor data to meaningful measurements. Cloud masking can be seen as part of this step. Errors, especially missed clouds, can lead to incorrect measurements and impact map accuracy.
    - Temporal compositing / harmonization. Creating temporal composites, e.g. monthly, to improve the coverage and quality of EO measurements. Can lead to short duration events (e.g. crop harvest) being missed if compositing period is too long.
3. Training data. Data that characterizes the target mapping features. Can be collected in the field or remotely by image interpretation. Will differ in nature depending on whether map is categorical or continuous.

Errors in training data collection, such as incorrect labels or imbalanced classes, will impact the quality of the trained model.

4. Split data into training and testing groups.
5. Model training. A subset of the training data is used to teach the model relationships between EO data and target mapping features.
6. Model evaluation. A different subset of the training data is used to test the performance of the model. Typically, this is held-back training data and provides the primary source of map accuracy assessment. However, any errors in the training data, such as class imbalances or poor geographic coverage, will continue to the model evaluation.
7. Output data. The output map data.
8. Map evaluation data. This data should be fully separate from the training data, and ideally sourced through a different methodology to avoid training data capture errors.
9. Map evaluation. An independent assessment of the quality of the final map from the perspective of the user. As discussed in point 6, this step is rarely performed as the held back training data will usually be used as the map accuracy assessment and may lead to an incorrect assessment of map accuracy. If fully independent map assessment is used, there may be sources of error that will arise from translating classification schema and measurement units to the EO derived map. Different spatial resolutions can also introduce errors.

Sources of error, and ways of evaluating them, are discussed in more detail in the following section (Section 3.1).

3.1. GEOMETRIC ACCURACY

For all EO datasets the sensor measurements need to be transformed into real-world coordinates, and the exact translation performed will be dependent on measurement type. Examples of errors that could be introduced at this stage include:

- Platform position errors. If the platform location or pointing direction is incorrect, this can lead to substantial shifts in EO data location. This is particularly important for drone data as most drones do not carry high accuracy GPS units.
- Orthorectification. Often an accurate model of the Earth's surface is required to correctly position sensor data in areas with significant topography. Low resolution or incorrect elevation data can lead to errors that may be non-uniform across the image.
- Projection errors. Using the wrong datum during processing can lead to data shifts of hundreds of metres.

For satellite data it has become common practice for images to be provided orthorectified, with high quality ground control calibration already undertaken. For example, Sentinel-2 data is calibrated against a global reference image that produces positioning accuracy of sub 6m RMSE (Copernicus, 2023). For drone and aerial photography surveys geometric accuracy will be dependent on the quality of GPS data used during the survey to capture platform location, and whether additional ground control points (GCP) have been also collected.

Regardless of the data source, geometric accuracy can be independently evaluated by comparing the location of identifiable features in the imagery to known GCP's collected through ground survey or from high-accuracy reference datasets.

3.2. TRAINING DATA QUALITY

Training data are typically assumed to be a source of truth for AI mapping projects, but in practice training data will have multiple sources of uncertainty that result from sample design and collection errors.

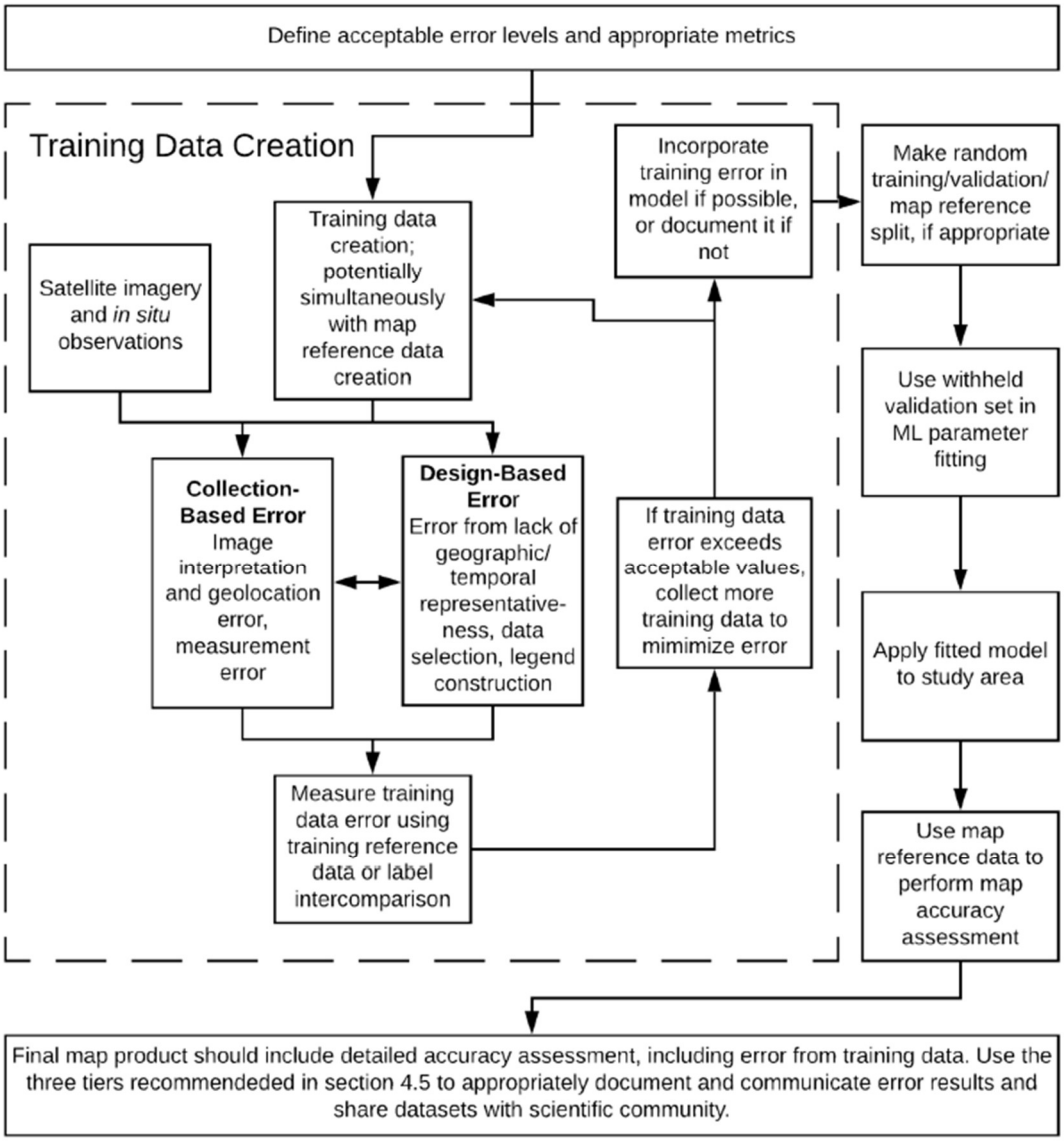

Figure 8: Flow chart of typical workflow for machine-learning applications in Earth observation data. From Elmes et al (2020)

Elmes *et al* (2020) provide a comprehensive review of how training data quality affects categorical and continuous mapping, showing how errors can propagate through projects and be difficult to unpick. In their paper they provide a workflow that outlines how training errors can be evaluated across a mapping project (see Figure 8). They also provided s set of suggested steps for minimising and accounting for training data error in AI mapping projects.

These steps are:

1. Define acceptable level of accuracy and choose appropriate metric:
    - As a starting point the minimum level of accuracy required should be defined, and the accuracy metric that best fits the research question should be selected.
    - For example, for a continuous variable where the absolute accuracy of the map is of most importance then mean square error should be preferred over R^2.
2. Minimise design related errors:
    - Sample design should be based on the needs of the ML algorithm and account for geographic distribution and class balance of samples. If classes are not well distributed spatially or if certain classes are over or under-represented, then this can lead to biases in the AI algorithm.
    - Training data sources should be temporally consistent with EO datasets if image interpretation is used to collect training data.
    - Training data features should be at least twice as large as the pixel resolution to ensure that pure pixel examples are available for the ML algorithm. This should influence minimum mapping unit definition.
3. Minimise collection-related errors:
    - As there is a great variety of training data collection methods, each with their own sources of error, there are many appropriate methods for minimising data collection errors.
    - However, several general approaches can be followed, with Elmes *et al* (2020) providing advice that focussed on image-interpreted training data.
4. Assess error in training data
    - Training data error should be measured directly and evaluated separately to map error.
    - For categorical mapping, label error can be evaluated against internal reference data such as independent survey data.
    - Note this implies that there is a separate source of evaluation data held back from model training and map evaluation that covers the same spatial area as the training data was sourced from. Indeed, Elmes *et al* (2020) recognise that production of additional reference data is a challenge for most projects.
5. Evaluate and communicate the impact of training data error:
    - Given the variety in EO data and AI mapping projects, no single protocol for treating training data error can be defined. However, the authors identify three tiers that can be considered:

i. Tier 1: Optimal training data accuracy assessment where error is evaluated against gold standard reference data.
ii. Tier 2: If Tier 1 evaluation is not possible, then introduce a plausible range of simulated error to the training data and evaluate impact on map accuracy.
iii. Tier 3: if tiers 1 & 2 are not possible, then the authors recommend that training and map reference data are published with metadata that describes sources of error and uncertainty.

- Uncertainty should be faithfully reported with maps and accompanying documents.

While the framework outlined by Elmes *et al* (2020) is comprehensive, much of it is best seen as a practical guide for project delivery, rather than a set of metrics that can be used to independently assess data accuracy.

A study by Rosli *et al* (2018) into training data quality for machine learning proposed as simpler approach that could be applied to a training data set directly without additional reference data. These tests are:

a) Duplicate data Definition: Two or more records that have the same measurement values associated with the same metric for the same entity.
b) Inconsistent data Definition: Two or more records that have different measurement values associated with the same metric for the same entity.
c) Missing data Definition: A record that does not have a measurement value for a given metric.
d) Incorrect data Definition: A record that has an implausible measurement value for a given metric.

These points, together with the training data design assessment points raised by Elmes *et al* (i.e. class balance, geographical spread and temporal consistency, could form a strong basis for an independent assessment of training data accuracy.

3.3. MISMATCH BETWEEN REFERENCE AND MAPPING DATA

Comparing independent reference data against map outputs can introduce uncertainty that arises purely from differences in mapping definitions. Key sources of error that will be address here are resolution, classification system and measurement unit mismatches.

Resolution mismatch

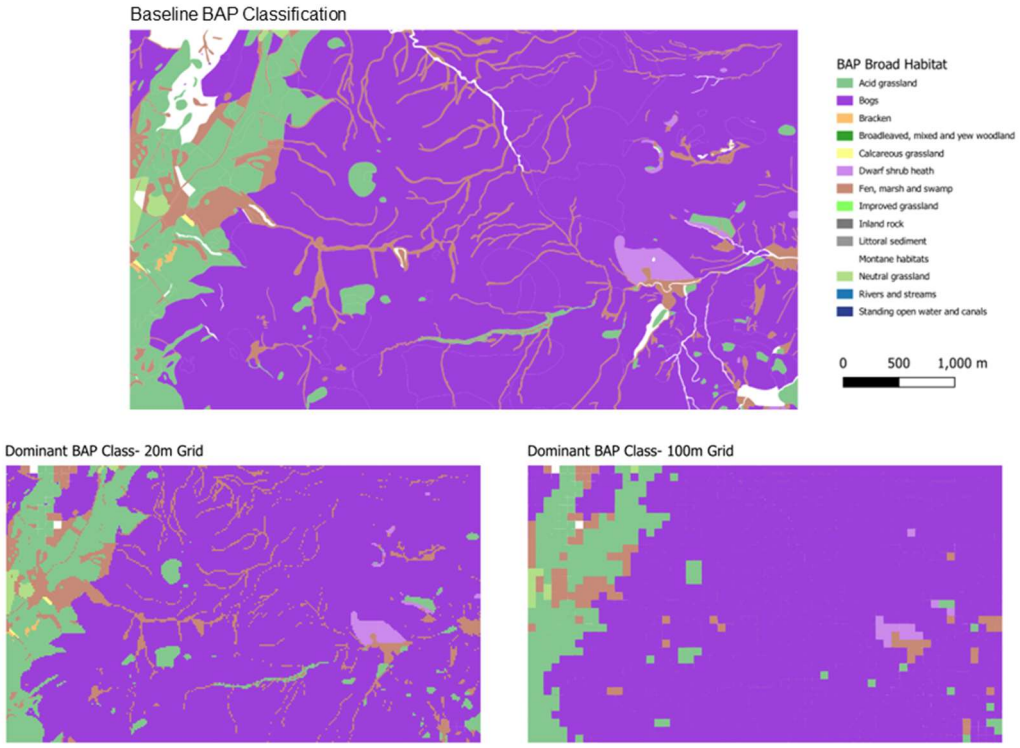

Figure 9: The effect of map resolution upon class representation. The baseline Biodiversity Action Plan (BAP) classification map (top) shows detailed classes that conform to natural features. When degraded to a 20m grid (bottom left) many of these features can still be identified, while at 100m (bottom right) most small-scale features are lost.

EO derived mapping will typically have a finest spatial resolution that is defined by the resolution of the EO data. Often a minimum mapping unit will be defined that is a multiple of the pixel size and reflects the fact that discrete features often need to be covered by multiple pixels in order to be classified.

In the case of satellite imagery, where pixels can be 10m across or larger, directly comparing coarse satellite data against fine scale ground mapping data will lead to a poor accuracy assessment for features that are at the scale of, or smaller than, the minimum mapping unit. Figure 9 illustrates this by demonstrating how fine scale features, such as the patches of 'Fen, marsh and swamp' that follow water courses, become somewhat degraded in appearance with a 20m pixel, and largely vanish at 100m. If the map accuracy

was assessed by direct comparison against the base map, this would lead to an unduly negative appraisal as it is not realistic for such fine scale features to be detected from coarse EO data. Indeed, Foody (2002) reported that such practices lead to a widespread underestimation in the capabilities of EO data.

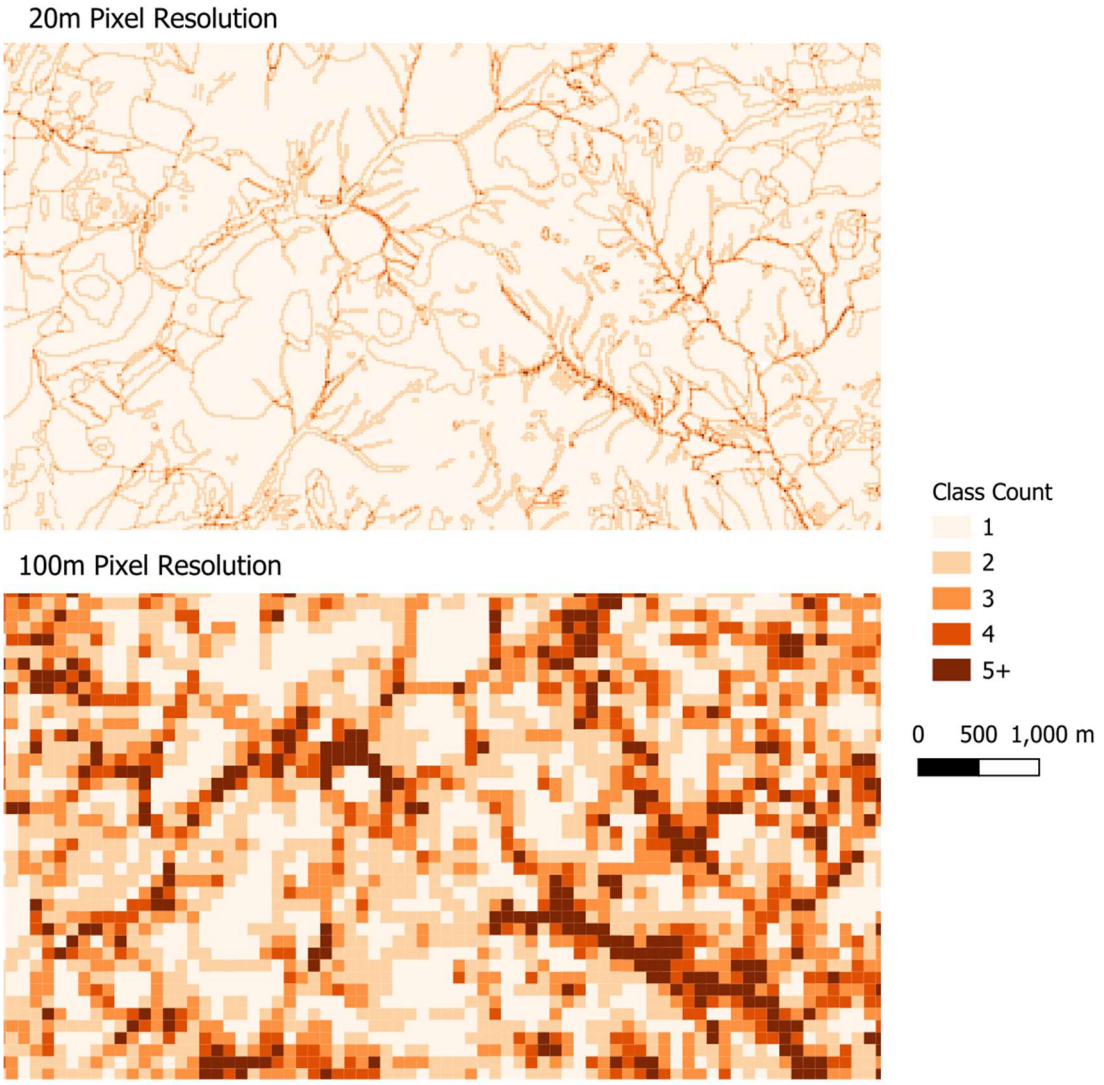

Figure 10: Number of classes found within each pixel when baseline BAP map is resampled to 20m (top) and 100m (bottom) pixels

As the spatial resolution, pixel gridding and minimum mapping unit of an EO derived map are known ahead of time, it is possible to resample baseline reference data to the resolution of the EO data and evaluate the expected uncertainty that will arise from this process. Figure 10 shows an example of this, with the number of classes from the baseline map found within each pixel when resampled to 20m and 100m resolution. A coarser pixel resolution results in more baseline classes per pixel, and a higher resolution induced uncertainty.

Classification schema mismatch

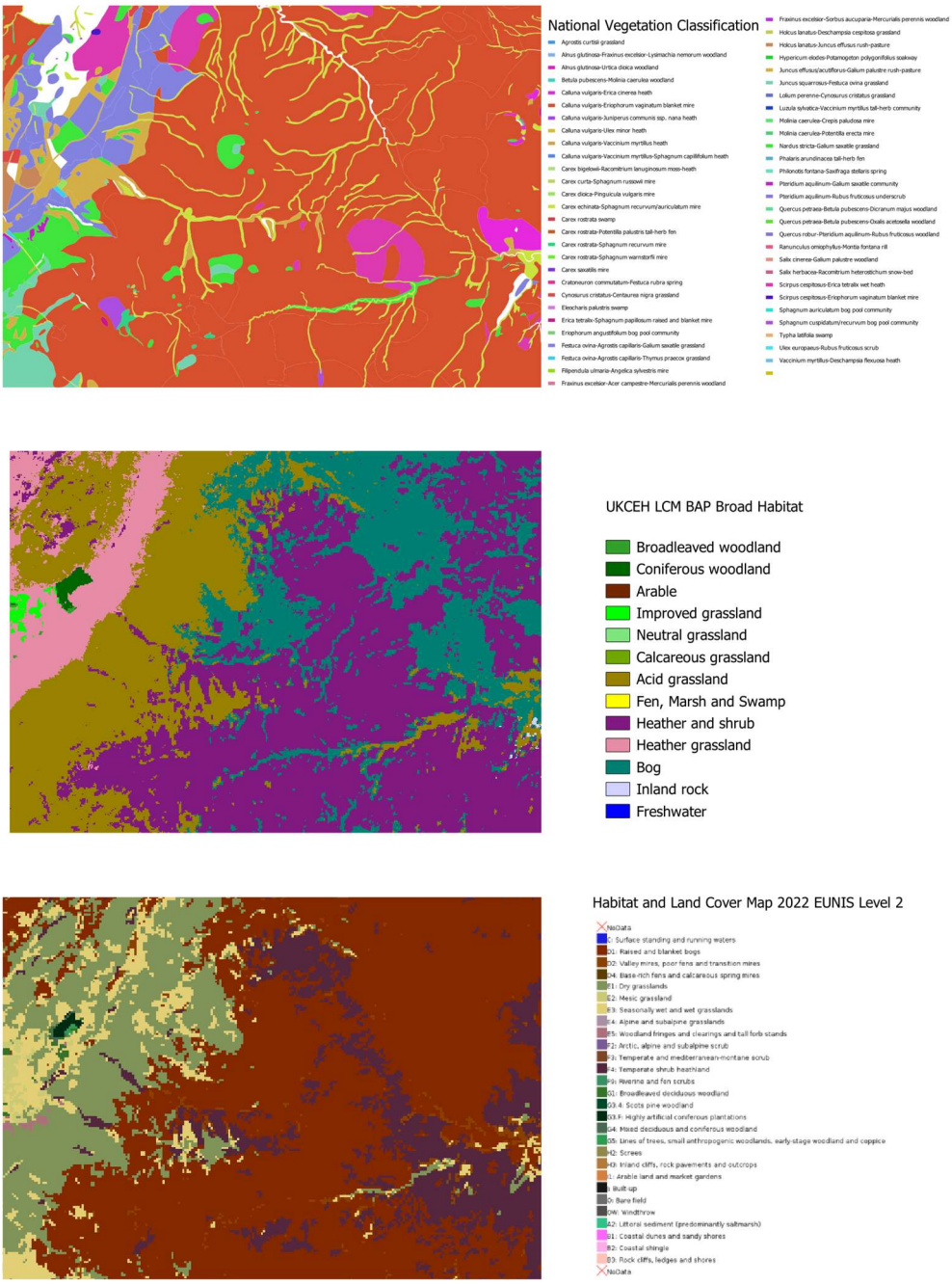

Figure 11: Comparison of National Vegetation Classification from NatureScot (top), Biodiversity Action Plan from UKCEH Landcover Map (middle) and EUNIS Level 2 from the NatureScot Habitat Map of Scotland (bottom).

The classification schema chosen for categorical classification is determined by the use case, where different schema may be equally valid. The National Vegetation Classification (NVC) is a detailed ecological survey that focuses on specific floristic assemblages. It is one of the key common standards for UK

nature conservation agencies and is adapted to British plant communities[5]. The high level of detail in NVC is typically not appropriate for landcover maps derived from EO data as it requires identification of specific plant species that cannot be detected in most EO data. Coarser landcover classification systems that are more appropriate for EO data also exist, such as the UK Biodiversity Action Plan (BAP) Broad / Priority Habitats[6] used in the CEH Landcover Map and EUNIS Habitat Classification Level 2[7] used by the NatureScot Habitat Map of Scotland. Figure 11 shows an example of these three classification schemas for the same area of land and illustrates that there is little agreement between the systems in terms of class description or the spatial location of transitions between classes.

The difference between classification systems presents challenges for the AI accuracy framework as it is not possible for ground survey to record landcover according to all the viable classification schema. JNCC produced a spreadsheet in 2008 that maps the correspondence between the following classifications:

- National Vegetation Classification (NVC)
- Phase 1 Habitat Classification
- UK BAP broad and priority habitat types (based on the list of habitats produced prior to the Species and Habitats Review in 2008)
- Vegetation communities of British lakes
- EUNIS habitat classification
- EU Habitats Directive Annex I habitat types
- Marine Habitat Classification
- OSPAR threatened and/or declining habitats

This spreadsheet provides the basis for mapping between different classifications, which is ambiguous for some specific classes, with one-to-many or many-to-many relationships between classes (illustrated in Figure 12).

UKHab[8] is a habitat classification system that attempts to provide a unified classification for the UK and allow the different nature agencies to report consistently on habitats of European and national significance. However, as UKHab is a more recent system it is not included in the JNCC habitat correspondences spreadsheet and there is no official habitat correspondence data available on the UKHab website.

[5] NVC | JNCC - Adviser to Government on Nature Conservation
[6] https://jncc.gov.uk/our-work/uk-bap-priority-habitats/
[7] https://www.eea.europa.eu/data-and-maps/data/eunis-habitat-classification-1/folder_contents
[8] https://ukhab.org/

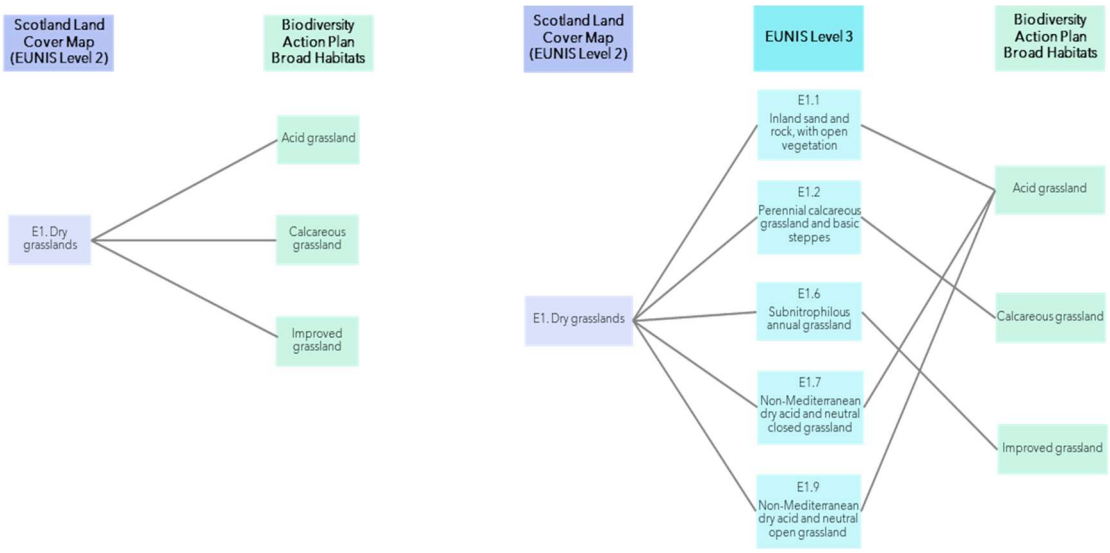

Figure 12: Mapping between EUNIS Level 2 and BAP Broad habitats for dry grassland. There is an ambiguous relationship when comparting the classifications directly (left). However, mapping landcover at the EUNIS Level 3 resolution (right) results in a non-ambiguous mapping between EUNIS Level 2 and BAP habitats.

The key observations for the AI accuracy framework are:

1. The EO classifications that are commonly used in the UK (e.g. BAP the for CEH Landcover Map and EUNIS Level 2 for the Habitat Map of Scotland) are not directly comparable.
2. Field data needs to be collected at a more granular level than is practical EO classification, such as EUNIS level-3 or NVC.
3. Additional work is needed to extend the JNCC habitat correspondence spreadsheet to newer classification systems such as UKHab.

Measurement unit mismatches

Differences in measurement units may be expected between reference data and map outputs. This may consist of map outputs being provided at a coarser measurement resolution, i.e. tree heights to the nearest metre, or measurements may record different biophysical variables such as tree carbon. If reference data is recorded to a high degree of precision (e.g. vegetation height to sub centimetre) then it can be degraded to match map outputs. Translating reference data to different biophysical variables will need to be considered on a case-by-case basis depending on the variable in question.

4. CONCEPTUAL FRAMEWORK

The AI accuracy framework must provide a system that is capable of differentiating between the different error sources discussed above. To separate EO and training data errors from AI algorithm errors will require access to these datasets in addition to the map outputs. Also, error metrics need to be adapted to the specific use case under consideration. Taking all these points together, we have laid out a conceptual AI Accuracy Assessment Framework in the form of a flow chart (Figure 13). Each of the steps in the flowchart, including user provided data, analysis steps and accuracy metrics, are discussed below.

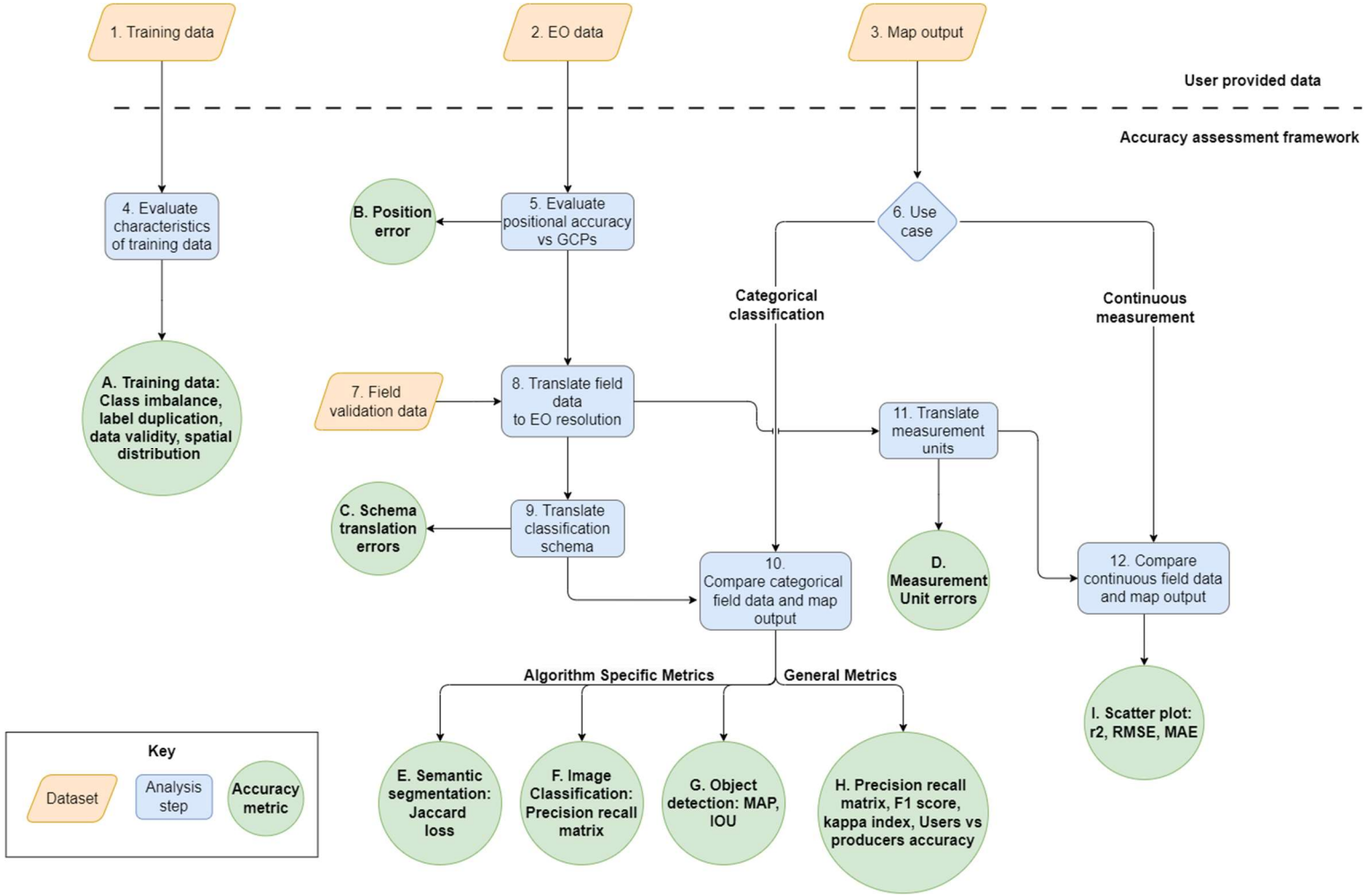

Figure 13: AI Accuracy Assessment conceptual framework

Data and analysis steps

1. Training data. This could consist of the training data directly uploaded to a website for assessment, or self-assessment conducted by the user if the training data is too large to upload.
2. EO data. Assuming that multiple sources / dates of EO data have been used, this should be a representative example.
3. Map outputs. The complete, final map output.
4. Evaluate characteristics of training data. Analysis of the training data to evaluate its overall quality.
5. Evaluate positional accuracy vs GCP's. Comparison of EO data against known GCP's to evaluate if there are positional errors in the data.
6. Evaluate use case. Identify whether map outputs are categorical or continuous and treat accordingly.
7. Field data. Independent reference data collected from field sites.
8. Translate field data to EO resolution. Modify field reference data to match spatial resolution of EO data.
9. Translate classification schema. Adjust reference data to match the spatial resolution and classification schema of output map data. Evaluate expected inconsistencies in terms of mixed classes and ambiguous relationships between reference and map classification. Produce a "best-case" classification from the reference data that best matches the classification schema and spatial resolution of the output map data.
10. Compare categorical field data and map output. Conduct a direct comparison between map outputs and reclassified reference data, with appropriate error metrics computed.
11. Translate measurement units. Translate continuous reference data to match the measurement units used in the output map. May involve quantising reference data, or converting to a different biophysical variable if required and feasible.
12. Compare continuous field data and map output. Conduct a direct comparison between map outputs and reclassified reference data, with appropriate error metrics computed.

Error metrics

A. Training data. Evaluate training data quality by scoring class imbalance, spatial distribution of training data, data duplication and data validity. These are mostly relative measures, and will likely result in indicative scores rather than absolute quality statements.
B. Position error. Reference GCP's will be used to evaluate spatial accuracy of EO data, with errors reported as RMSE in m.
C. Schema translation errors. Error in categorical map that is expected from translating the reference schema and spatial resolution to map output schema and resolution. Likely metrics will be:
    a. Class purity, i.e. how many pixels contain multiple classes. This could be used to filter map accuracy assessment (step 10), with map accuracy computed for high purity vs low purity areas.
    b. Schema compatibility, i.e. how many classes have one to one relationships vs one to many.
D. Measurement unit errors. Assessment of any loss of precision expected when translating continuous reference data to map resolution and biophysical variables.
E. Semantic segmentation. Computation of Jaccard loss.
F. Image classification. Computation of precision recall matrix.
G. Object detection. Computation of mean average precision and intersect over union.
H. General metrics. Computation of metrics that can apply across all AI algorithm types, such as F1 score, kappa index, users vs producers accuracy.
I. Scatter plot. Computation of accuracy metrics relevant to continuous variables such as R^2 and RMSE.

5. REFERENCES

Clark, C., & Pacifici, F. (2023). A solar panel dataset of very high resolution satellite imagery to support the Sustainable Development Goals. *Scientific Data*, 10 (636).

Copernicus. (2023). *Copernicus Sentinel-2 GRI as Database of GCPs in L1B & L1C -Product Handbook.*

Elmes, A., Alemohammad, H., Avery, R., Caylor, K., Eastman, J., Fishgold, L., . . . Laso Bayas, J. (2020). Accounting for Training Data Error in Machine Learning Applied to Earth Observations. *Remote Sensing*, 1034.

Foody, G. (2002). Status of land cover classification accuracy assessment. *Remote Sensing of Environment*, 185-201.

Janga, B. A. (2023). A review of practical ai for remote sensing in earth sciences. *Remote Sensing*, 4112.

Rosli, M. M., Tempero, E., & Luxton-Reilly, A. (2018). Evaluating the quality of datasets in software engineering. *Advanced Science Letters 24*, 7232-7239.

Wilkinson, R., Mleczko, M., Brewin, R., Gaston, K., Mueller, M., Shutler, J.D., . . . Anderson, K. (2024). Environmental impacts of earth observation data in the constellation and. *Science of the Total Environment*.