Basic Assignment Metrics

Table of contents

0.1 Overview

The following provides an assessment of the completed assignments undertaken by the full team of labellers, who were tasked with digitizing Class 2 and 4 sites, and whose quality was assessed against Class 1 sites. The labelling teams additionally remapped nearly 1000 sites from the Class 1 sample that were not including for quality control in the platform.

0.2 Initial transfer and processing

Data were initially extracted from `labeller` using code provided in the `labelreview` repository, and were transferred into this repository for processing.

Data on assignments related to assessed quality scores (from Q type assignments, i.e. where a labeller's digtizations were assessed Class 1 expert labels), completion time, assignment status, and type of assignment were joined to counts of the number of fields collected for each assignment (site).

0.3 Basic stats

0.3.1 Label quality

0.3.1.1 Q scores

Quality was measured using two basic approaches. The first was to assess each labeller using randomly assigned Q sites, which were labelling assignments at locations were the expert team had labelled the fields as part of the Class 1 labelling effort. The platform's built in scoring algorithm then compared the labeller's maps against the Class 1 labels, and calculating four metrics that contributed to an overall Score (the Q score):

- N = agreement between number of digitized between labeller and expert labeller;
- Edge = nearness of delineated labeller's digitized field edges to expert's field edges;
- Area = Overall agreement of area of fields digitized by labeller and by expert;
- Categorical = Agreement of categorical label assignment to fields.
- Score = The weighted mean of the previous 4 scores, here specified as:

$$Score = 0.225N + 0.1Edge + 0.55Area + 0.125Categorical$$

Edge and Categorical received relatively low weights because the former is a very difficult measure to get right, given the inherent difficulty of distinguishing precise boundaries within the resolution of Planet imagery, while Categorical accuracy is relatively unimportant because the team only labelled field/no-field, therefore mislabels only occurred in cases of complete false positives or false negatives. A more detailed explanation of these measures are provided in Estes et al. (2022).

The mean overall score dimension is shown in @tbl-qcomponents, and for each labeller in Figure 1.

Table 1: The average overall score in 4 label quality dimensions: Score = Overall accuracy; N = agreement between number of digitized between labeller and expert labeller; Edge = nearness of delineated labeller's digitized field edges to expert's field edges; Area = Overall agreement of area of fields digitized by labeller and by expert; Categorical = Accuracy of assigned labels

Score	N	Edge	Area	Categorical
0.609	0.33	0.049	0.753	0.925

The weekly average scores for each metric Figure 2 can also provide useful insight into increases or decreases in label quality, owing to increasing experience, pressure to meet labeling deadlines,

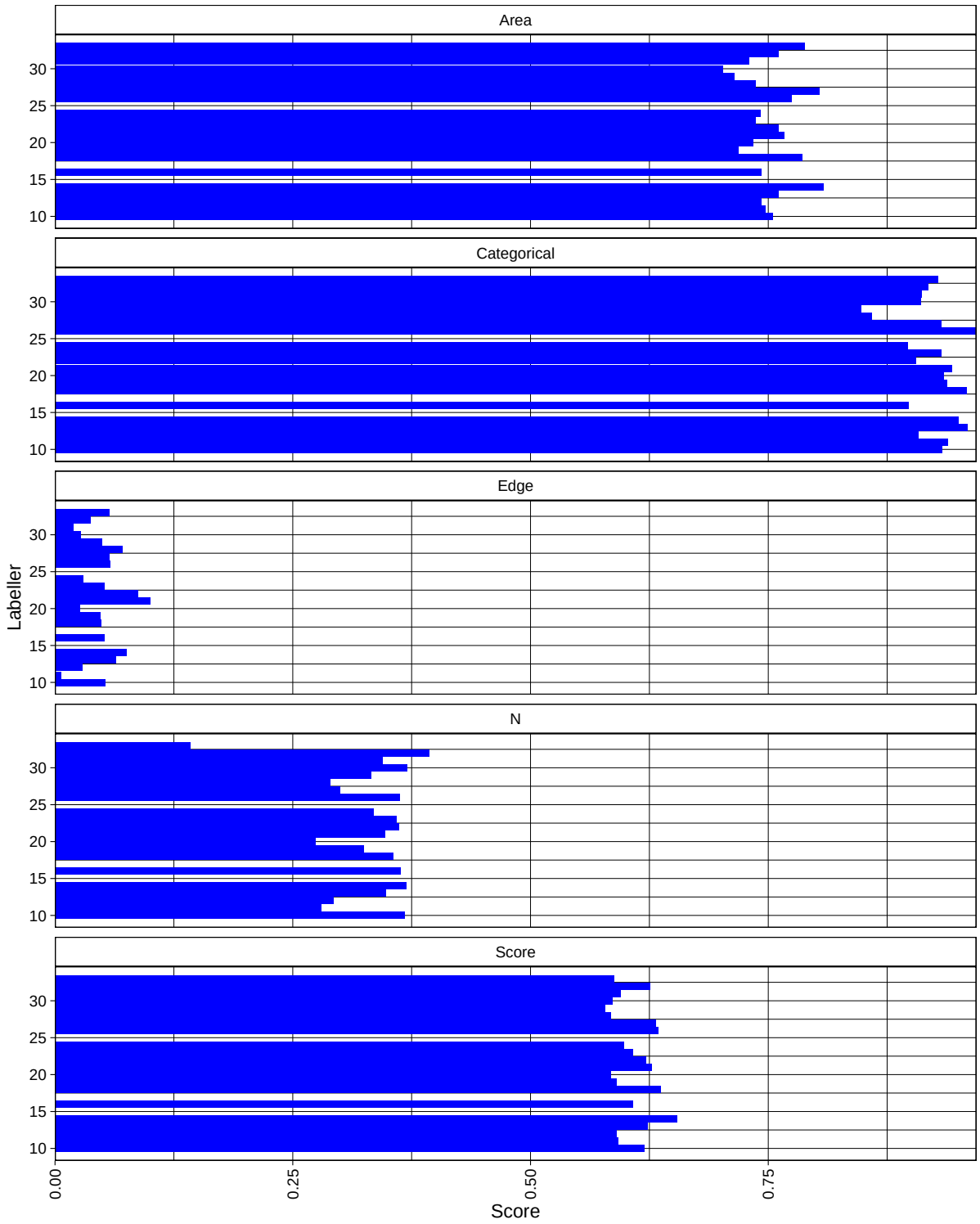

Figure 1: The average score per labeller in each of 4 label quality dimensions: N = agreement between number of digitized between labeller and expert labeller; Edge = nearness of delineated labeller's digitized field edges to expert's field edges; Area = Overall agreement of area of fields digitized by labeller and by expert; Categorical = Agreement of categorical label assignment to fields.

and other factors. The weekly component scores for each labeller are shown in Figure 3.

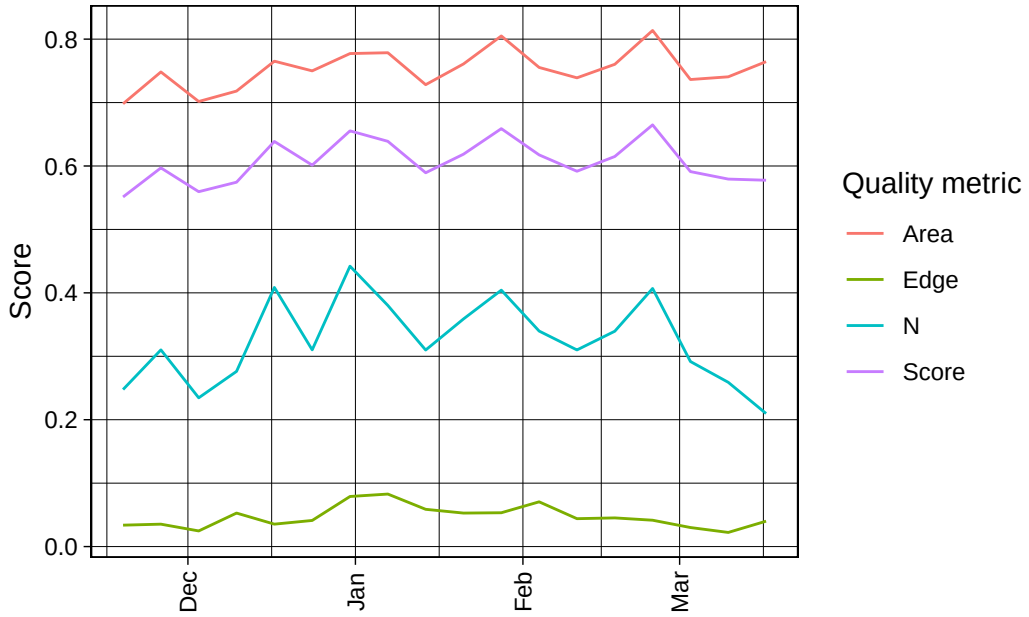

Figure 2: Weekly averages in each of the quality components (excluding categorical)

0.3.1.2 Assignment status

The labeling platform assigns a status to each label that includes the folllowing:

- Abandoned: Assignments begun by a labeller but not completed within 24 hours. These are returned to the system for remapping.
- Returned: Assignments that were returned to the system unmapped by the labeller, perhaps because of missing imagery, poor image quality, etc.
- Rejected: Q sites where the labeller's work was scored against the underlying Class 1 labels as being below the 0.4 threshold.
- Untrusted: F or N assignments completed at a time when the labeller's average score against the last 5 Q sites completed is below a pre-determined trust threshold.
- Approved: F or N sites completed by a labeller whose last 5 scores had an average score above the trust threshold.

The overall distribution of assignment status is shown in Table 2, and the weekly means of non-approved in Figure 4.

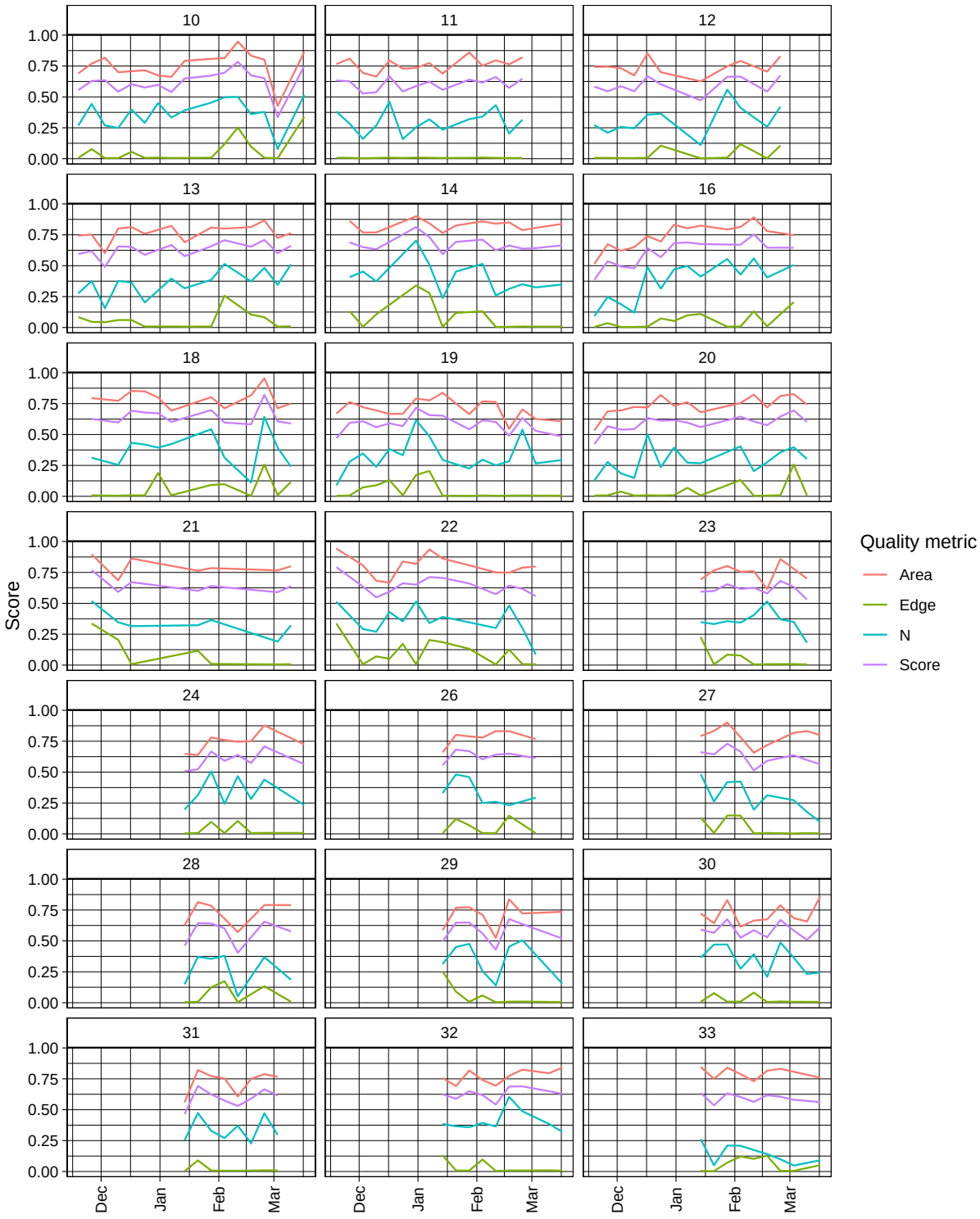

Figure 3: Weekly scores for each worker in each of the quality components (excluding categorical, and filtering out weeks where only or fewer Q sites were completed)

Table 2: Number of assignments in each status class.

Status	N
Abandoned	425
Approved	38801
Rejected	283
Returned	124
Untrusted	1547

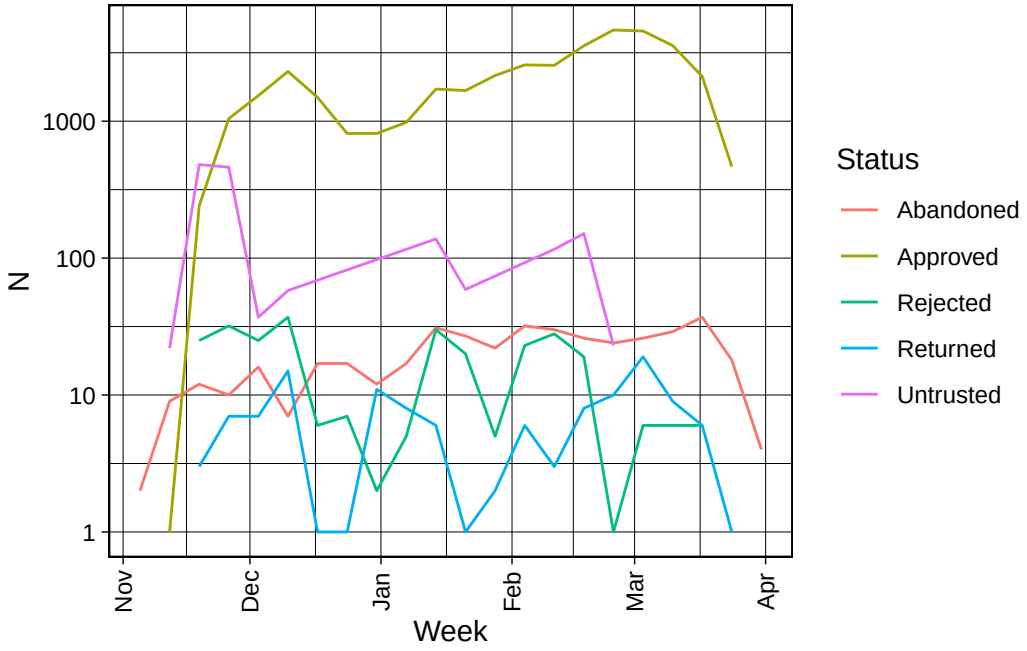

Figure 4: Assignment status summed by week, shown here logarthmically (base 10) scaled.

Returned and Abandoned assignments were hereafter excluded because they provide no valid label data. The results from an additional 445 assignments were dropped, as these corresponded to 253 sites where labellers reported missing or cloudy imagery.

0.3.1.3 Label reviews

Reviews of randomly selected sites were also conducted by two of the supervisory team, using the following rubric (also described [here](https://github.com/agroimpacts/labelreview#review-labels)):

Expert review rubric

The following definitions were used to visually review label quality against the Planet imagery:

- True positive (TP): A field that is correctly labelled as such;
- True negative (TN): A non-field area that is correctly left unlabelled;
- False negative (FN): An actual field that should have been mapped, but wasn't;
- False positive (FP): A non-field area that was incorrectly mapped as a field;
- Over-segmented (OS): A larger field that was incorrectly divided into many small fields (in these cases, the labeller is making up internal boundaries in the larger field that are not visible in the imagery);
- Under-segmented (US): Two or more smaller fields that were incorrectly grouped into one larger field, even though boundaries are visible that would enable the smaller fields to be correctly digitized.

Using those definitions, for sites where the imagery shows that there are fields in the imagery, assign one of the following categories to each reviewed site:

- 0: For cases where the labeller maps less than half the site correctly, either by:
    a. leaving 50% or more of the area covered by actual fields unlabelled (FN);
    b. incorrectly mapping more than twice the area of fields that are actually there (FP);
    c. correctly mapping the total area covered by fields, but grouping them into a larger field or fields that sum to less than half the total number of fields in the imagery (US);
    d. correctly mapping the total areas covered by fields, but falsely dividing them into more than twice the number of individual fields that are actually there (OS);
- 1: The labeller maps 50-70% of the site correctly, either by:
    a. leaving 30-50% of the area covered by actual fields unlabelled (FN);
    b. incorrectly labelling an areas that is 50 to 100% larger than the area of actual fields (FP);
    c. correctly mapping the total area covered by fields, but grouping them such that there are only 50-70% of the total number of fields in the imagery (US);
    d. correctly mapping the total areas covered by fields, but falsely dividing them into 50 to 100% more fields than are actually there (OS);

- 2: The labeller maps 70-90% of the site correctly, either by:
  a. leaving 10-30% of the area covered by actual fields unlabelled (FN);
  b. incorrectly labelling an areas that is 10 to 50% larger than the area of actual fields (FP);
  c. correctly mapping the total area covered by fields, but grouping them such that there are only 70-90% of the total number of fields in the imagery (US);
  d. correctly mapping the total areas covered by fields, but falsely dividing them into 10 to 50% more fields than are actually there (OS);
- 3: The labeller maps 90+% of the site correctly, such that:
  a. <10% of the area covered by actual fields is left unlabelled (FN);
  b. The labeled field areas is <10% larger than the actual field area (FP);
  c. the total number of correctly labelled fields is <10% smaller than the total number of actual fields (US);
  d. the total number of correctly labelled fields is <10% larger than the total number of actual fields (OS);

For sites where then are no fields visible in the imagery, and the labeller correctly classifies them as having no fields, assign a value of 4.

An evaluation of these score is provided in a separate notebook on Expert Reviews, including a quantitative comparison of the two experts' reviews for a subset of assignments that were reviewed by both, and the overall mean review scores for each labeller. Here the weekly mean review scores Figure 5 and weekly mean review scores per labeller Figure 6 are presented.

0.3.2 Number and area of fields

The distributions of the number of fields digitized per site, along with average size of digitized fields, is shown in Figure 7, indicating strongly right-skewed distributions, with the most common result being 0-5 fields and 0.5-1 ha per site. The average number of fields digitized per site was 19.4, with 3385 having no fields digitized (note: untrusted, returned, and rejected sites were excluded from the counts). Across all digitized polygons (7.79101×10^5 total), the average and median field sizes were 0.96 and 0.49 ha.

The weekly average number of fields digitized per site is shown in Figure 8, showing a peak in late December/early January, along with the average area, which shows a peak in mid-January. These latter two trends may indicate a tendency to under-segment during the last two months, or over-segment in the first two months.

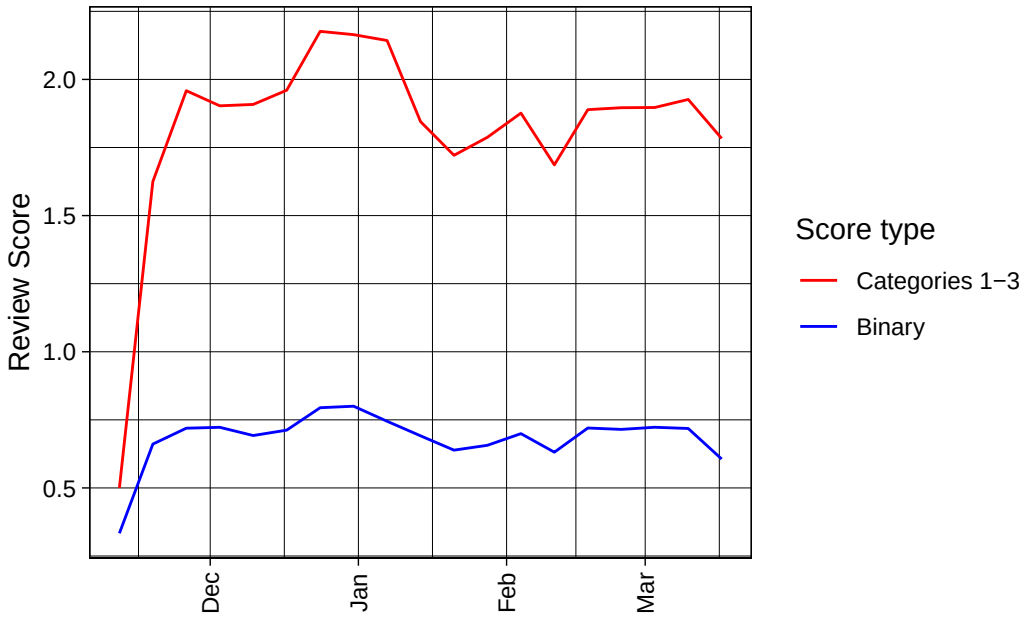

Figure 5: Averages weekly expert review scores for F type (Class 4) sites. The average score across categories 0-3 was calculated (category 4 was excluded), as well as the average of review score recoded to 0 (a 0 or 1 review score) or 1 (review score of 2-4).

0.4 Spatial distributions

The spatial distribution of the different sample classes and the number of times each location was mapped is shown in Figure 9.

Estes, Lyndon D., Su Ye, Lei Song, Boka Luo, J. Ronald Eastman, Zhenhua Meng, Qi Zhang, et al. 2022. "High Resolution, Annual Maps of Field Boundaries for Smallholder-Dominated Croplands at National Scales." *Frontiers in Artificial Intelligence* 4: 744863. https://www.frontiersin.org/article/10.3389/frai.2021.744863.

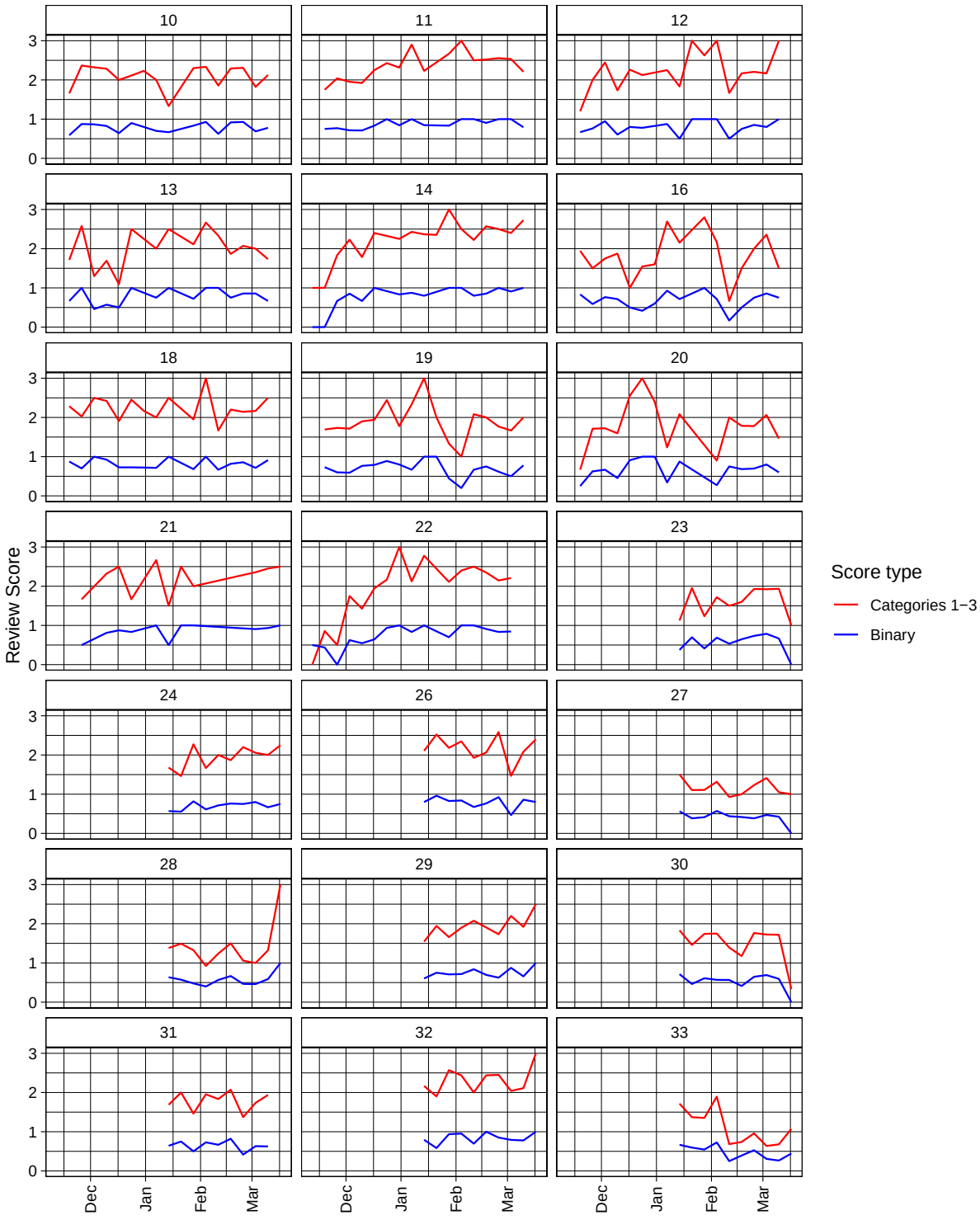

Figure 6: Averages weekly expert review scores for each labeller for F type (Class 4) sites. The average score across categories 0-3 was calculated (category 4 was excluded), as well as the average of review score recoded to 0 (a 0 or 1 review score) or 1 (review score of 2-4).

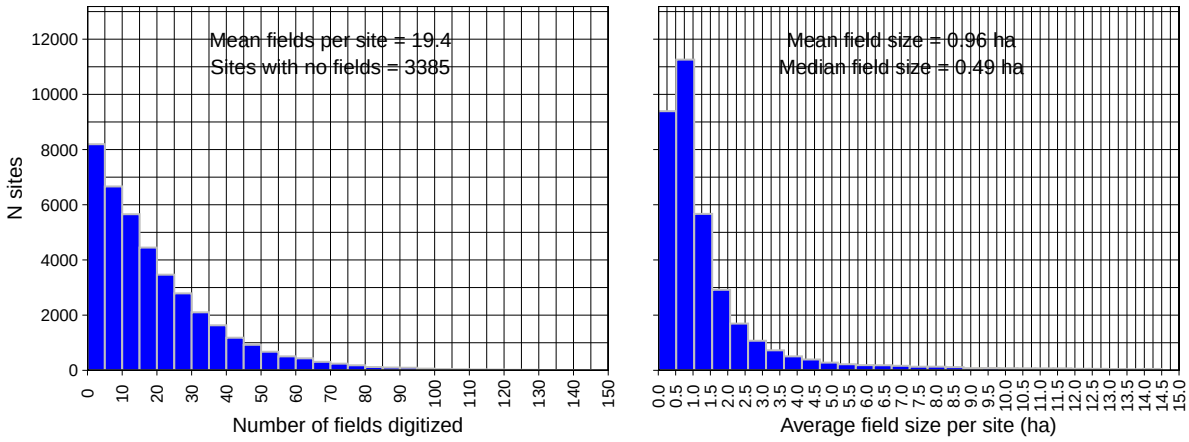

Figure 7: The distribution of the numbers of fields digitized per assignment (left), including the average number and the total number of sites that were assessed as having no fields, and the average size of fields (in hectares) per site (note: 2% of sites have mean sizes >15 ha, and are not shown here), along with the mean and median of digitized field sizes.

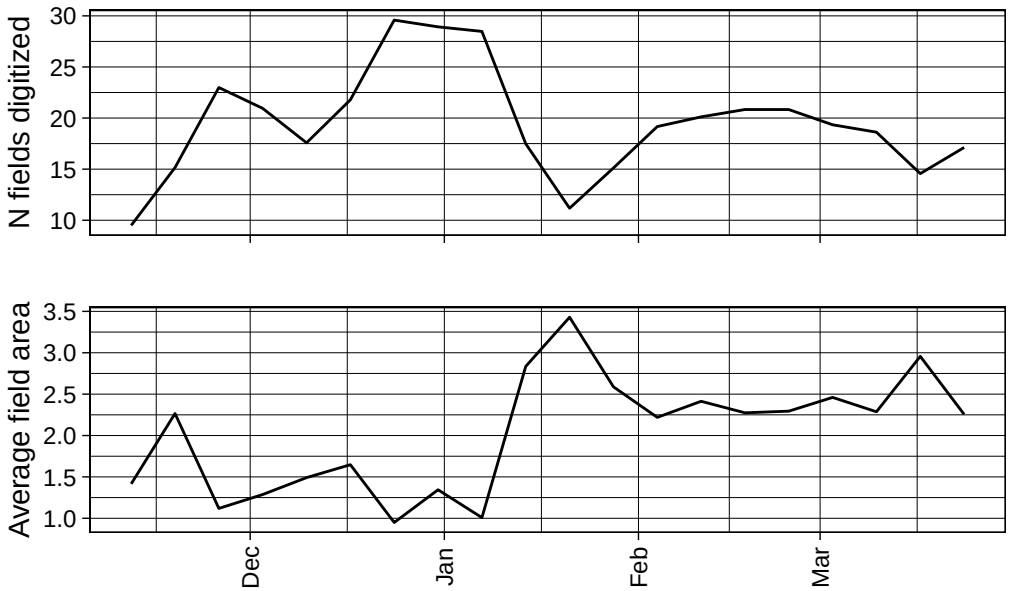

Figure 8: The average number (top) and area (bottom) of fields digitized per assignment by week.

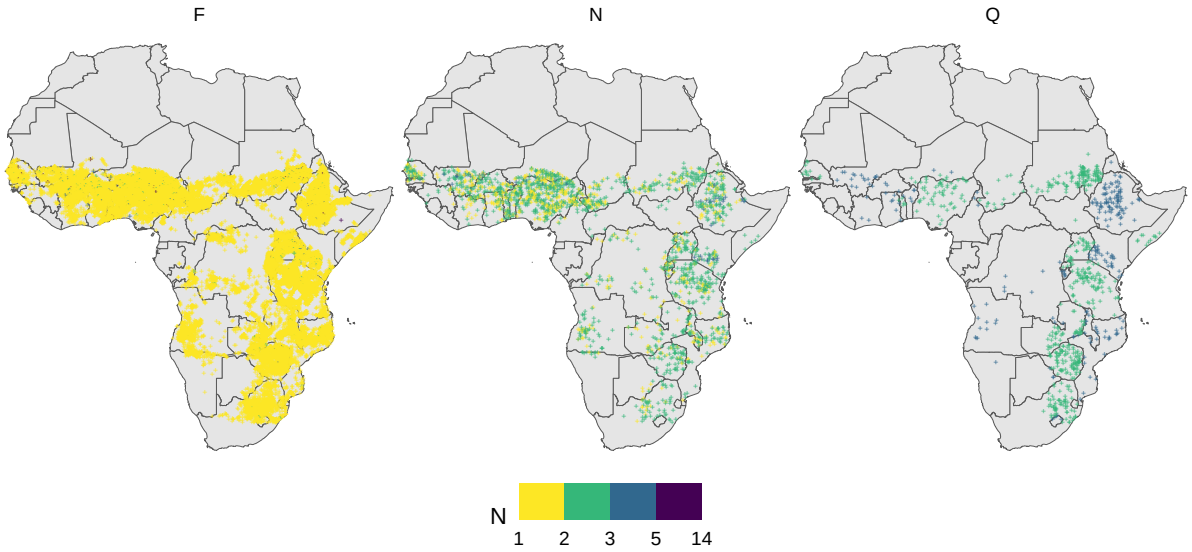

Figure 9: The distribution of sites mapped by assignment type (F = Class 2; N = Class 4; Q = quality control sites) and the number of times each were mapped.

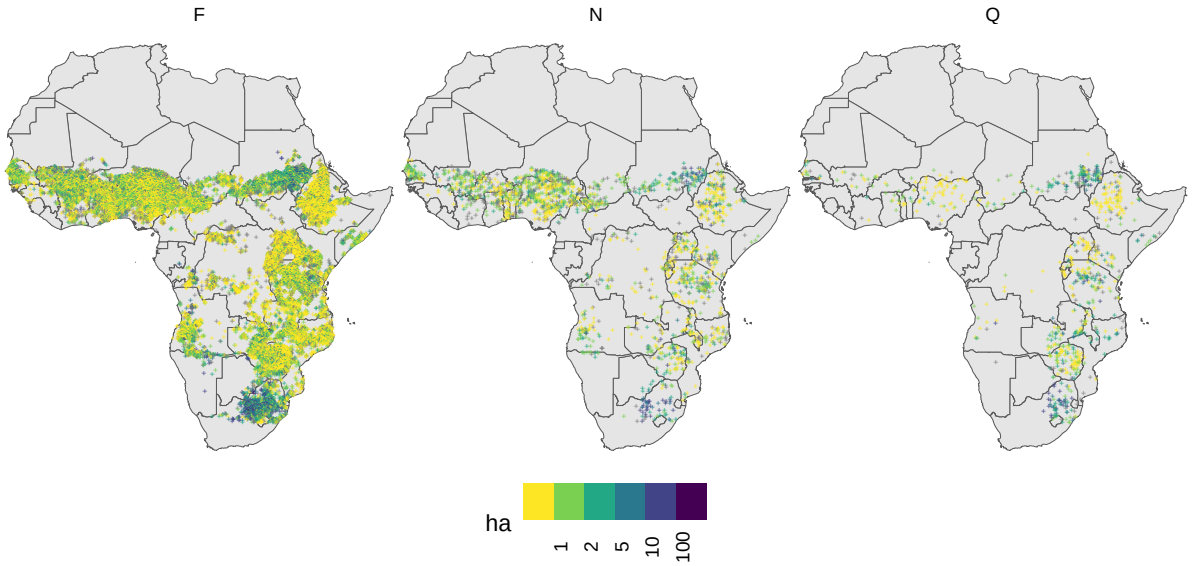

Figure 10: The distribution of sites mapped by assignment type (F = Class 2; N = Class 4; Q = quality control sites) and the average sizes of fields at each site.