Improved Vote Aggregation Techniques for the Geo-Wiki Cropland Capture Crowdsourcing Game

Baklanov Artem [1,2], Fritz Steffen [1], Khachay Mikhail [2], Nurmukhametov Oleg [2] Salk Carl [1],See Linda [1], and Shchepashchenko Dmitry [1].

1 — International Institute for Applied Systems Analysis (IIASA); Schlossplatz 1, Laxenburg, Austria, A-2361;

2 — N.N. Krasovskii Institute of Mathematics and Mechanics (Russian Academy of Sciences).

Challenge

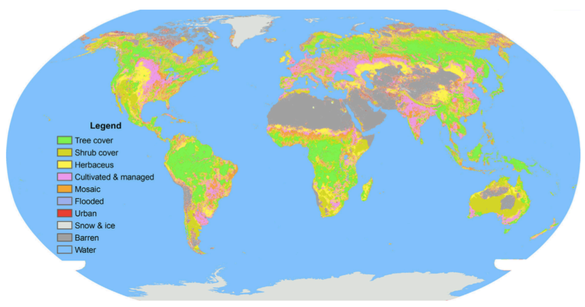
How to aggregate votes from non-experts?

The Cropland Capture Game (iOS, Android)

Over **5 million** opinions from non-experts

HOW?

Land cover map



Expert-quality decisions about 190 000 images

Results

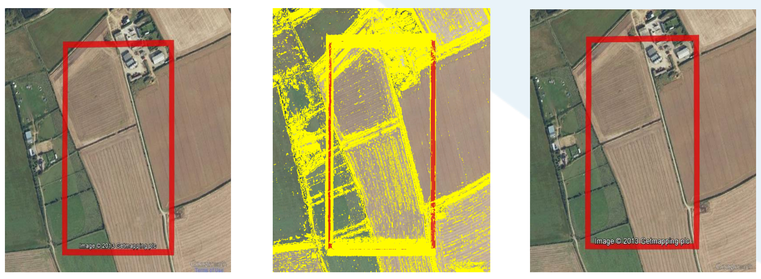
We increased the accuracy of "Cropland Capture" data from **76% to 91%**

- ✓ Improved quality of image dataset;
- ✓ Improved majority voting estimates;
- ✓ Benchmarked state-of-the-art algorithms;
- ✓ Demonstrated that these algorithms perform on a par with majority voting. **Explanation:** all volunteers are reliable, the task assignment is highly irregular.
- ✓ Accuracy is **96%** for images with more than 9 votes.

Approach

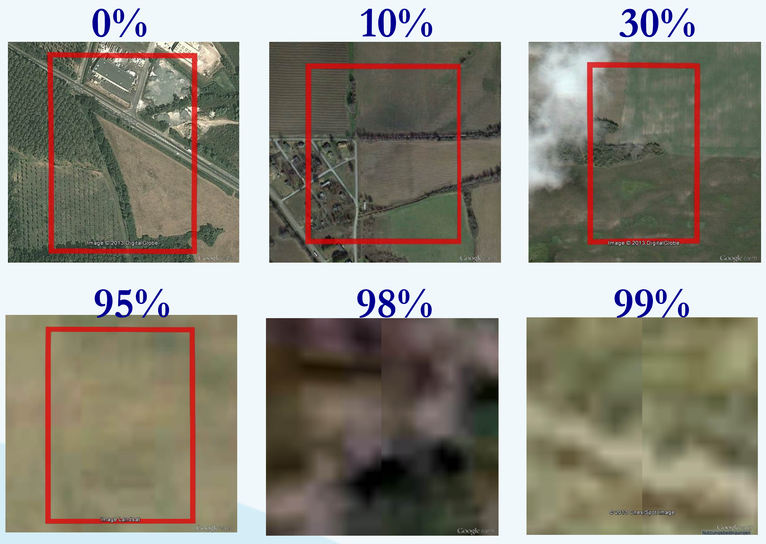
Data preprocessing

1) Detection of similar images using pHash (perceptual hash) [Zauner, 2010].



➔5% of images are not unique

2) Detection of low quality images using Blur detection algorithm [H Tang, 2012].



➔2% of images are discarded

Volunteers' ROCs

Individual performance of volunteers is studied with respect to the number of votes [Rayker, 2012].

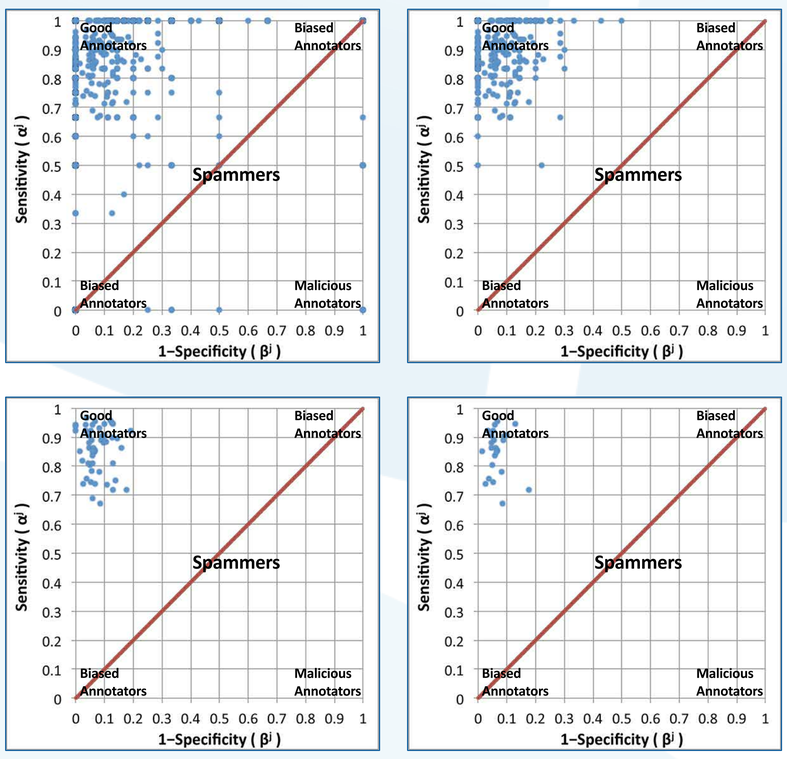


Fig. 1: In the figure we use notation introduced in [11]. Threshold = 0, 12, 44 , and 100 votes. These thresholds leave 1813, 262, 52, and 24 volunteers, respectively. ROCs of spammers lie on the red line.

- ✓ There are no spammers among volunteers with more than 12 votes;
- ✓ Good volunteers prevail;
- ✓ Volunteers with >100 votes show higher accuracy than any tested algorithm.

Benchmark

We compare machine learning algorithms and state-of-the-art vote aggregation algorithms:

EM [Dawid, 1979];

KOS, KOS+ [Karger, 2011];

Hard Penalty [Jagabathula, 2014].

Table 1: Baseline algorithms

Number of features	*Random Forest*	*LDA*	*AdaBoost*
5	**89.92**	87.60	89.15
14	89.14	**90.70**	89.92
35	88.37	89.53	**91.08**

Table 2: Accuracy for 'crowdsourcing' algorithms without image-vote thresholding

iteration	*MV*	*EM*	*KOS*	*KOS+*	*weighted MV*
Base	89.81	89.81	88.99	89.81	90.63
1	90.05	90.16	88.88	90.16	**91.45**
2	90.05	90.05	88.64	90.16	**91.45**
3	89.67	89.58	88.17	89.70	91.22

Table 3: Accuracy for 'crowdsourcing' algorithms with image-vote thresholding. Only images with at least 10 votes are left in the expert dataset. In this case we have 404 images annotated by 1777 volunteers.

iteration	*MV*	*EM*	*KOS*	*KOS+*	*weighted MV*
Base	94.55	94.55	94.06	94.55	95.05
1	94.55	94.55	93.81	94.55	95.05
2	94.55	94.55	93.81	94.55	95.05
3	94.55	94.55	94.06	94.55	95.05

*We use publicly available code (https://github.com/ashwin90/Penalty-based-clustering)