School mapping in ESRI imagery

Andrija Gorup, Dominik Jambrović, Marin Kačan, Siniša Šegvić

October 24th, 2024

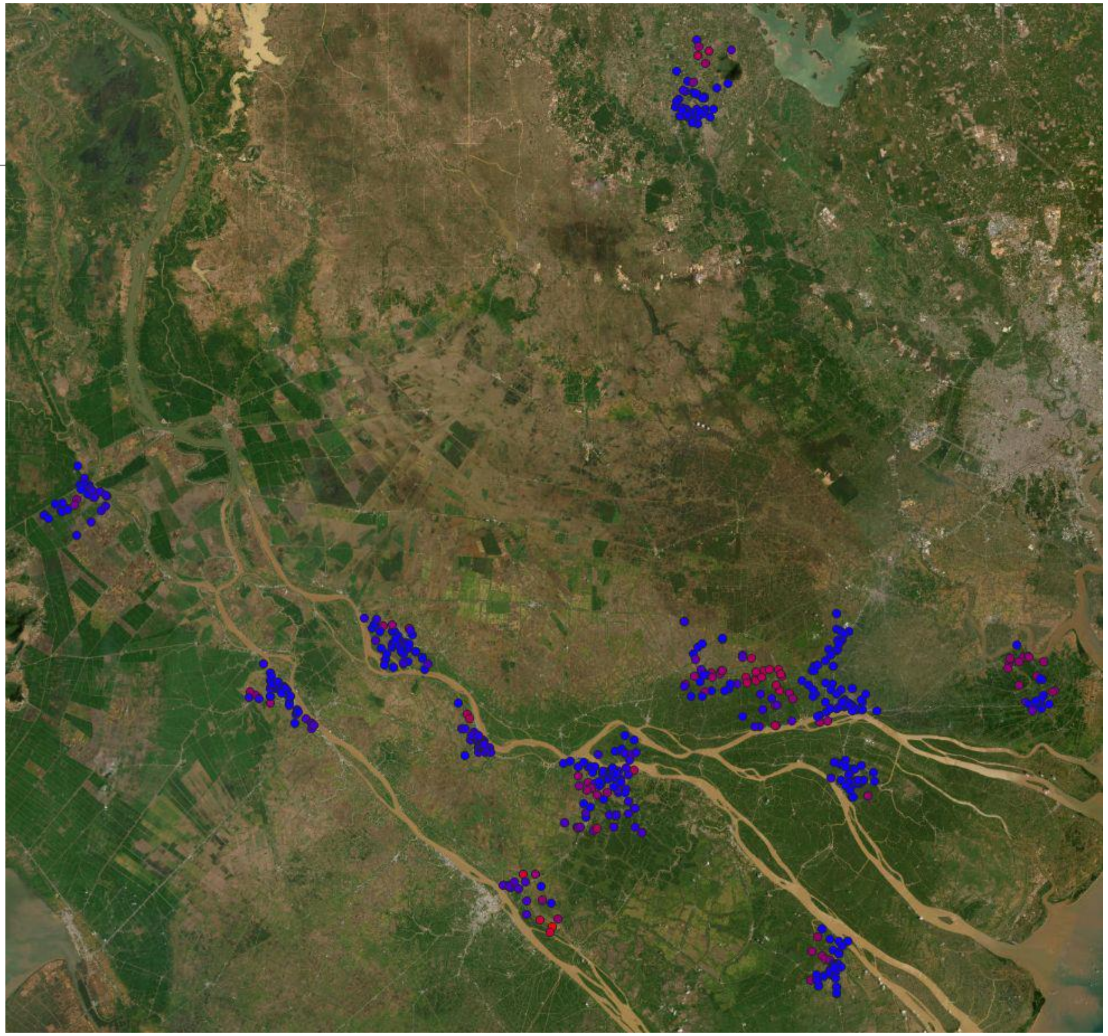
Recap from last meeting

2-stage training, fine-tune and eval on Anditi schools

- 2-fold cross-validation
  - 50:50 train/val split of school locations
  - Add equal number of non-school locations
    - Sampled randomly throughout Vietnam
  - Unusually high results – **F1: 93.91%**
    - In spite of potentially problematic outdated imagery

Potential problems

- **Outdated imagery**
  - Schools visible on Google Maps, but not in ESRI images
- Possible solutions:
  - Urban growth layer
  - ESRI metadata layer
  - Combination of the 2 methods
  - Compressed file size



Potential problems

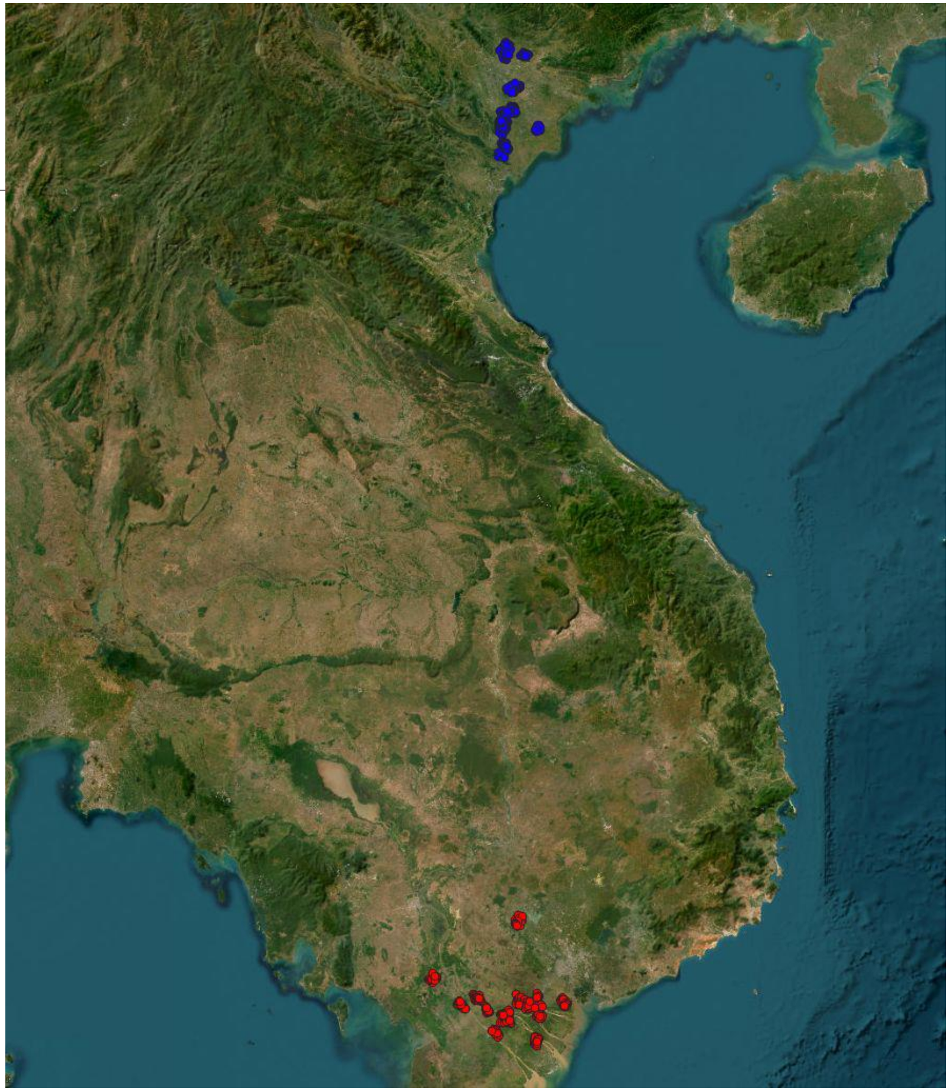
- **Many schools next to each other**
  - For cross-validation experiments, we do a random 50:50 train/val split
  - Data leakage – overlapping images get into both splits

Potential problems

- **Distinct appearance**
  - Images from different regions have large variations in appearance

Solution – split by clusters

- More realistic results
  - **F1: 81.55%**
  - P: 95.94%
  - R: 73.97%



New contributions

Non-school sampling

- Previously:
  - Non-schools sampled throughout Vietnam
  - Non-schools have no connection to the splits' schools
- **Distance-based non-school sampling**
  - Sample non-schools based on distance from schools
  - **Goal:** train model only on non-school images geographically close to corresponding school images

Distance-based non-school sampling

- Schools are split into two clusters
- For each cluster, calculate the corresponding **centroid**

- For each non-school (OSM Vietnam data), calculate the **Euclidean distance from both centroids**

$$d((x, y), (a, b)) = \sqrt{(x-a)^2 + (y-b)^2}$$

Distance-based non-school sampling

- Convert distances to **probabilities of sampling**

$$p_c\,(x) = \frac{1}{d_{c,x}+\varepsilon} \quad , \quad \varepsilon = 1 \, * 10^{-10}$$

- Sample non-schools for each cluster individually (**no repetition**)

Distance-based non-school sampling

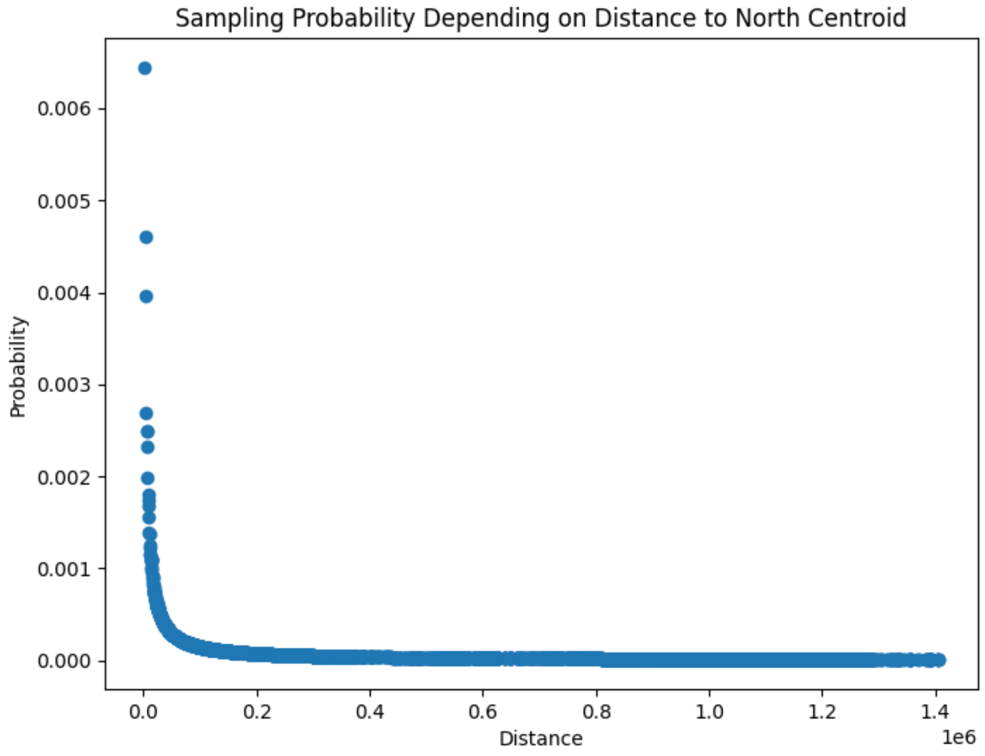
Algorithm 1 Distance-based non-school sampling

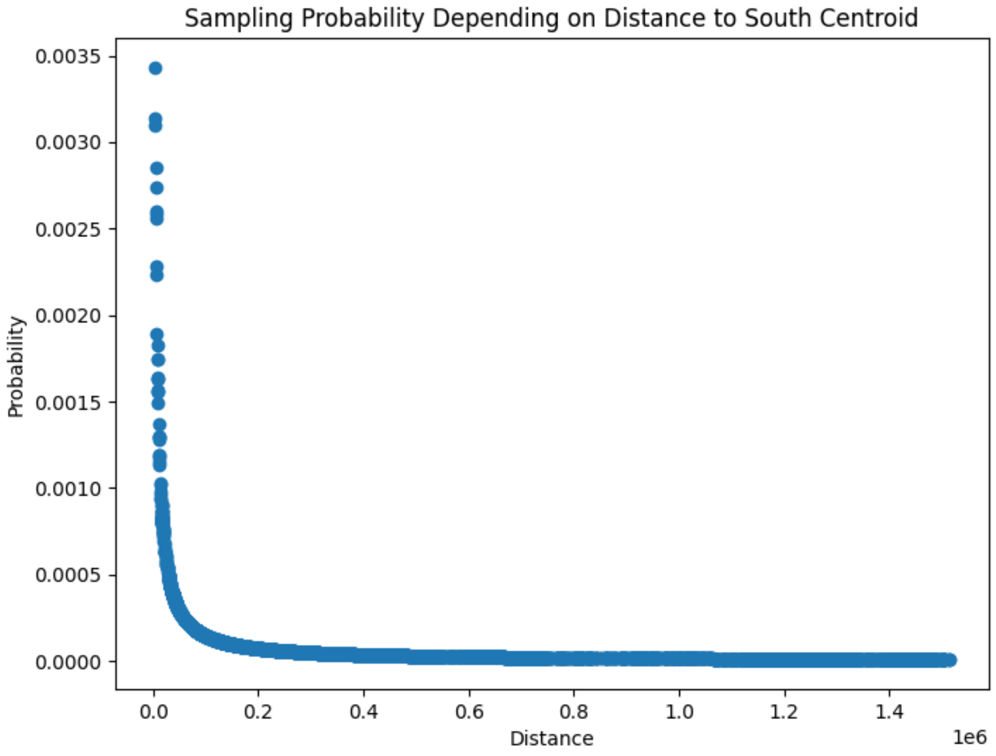
X_N = {North Cluster School Data}
X_S = {South Cluster School Data}
X_{NS} = {Non-school Data}
$\epsilon = 1 \cdot 10^{-10}$

C_N = (mean($X_N.long$), mean($X_N.lat$))
C_S = (mean($X_S.long$), mean($X_S.lat$))

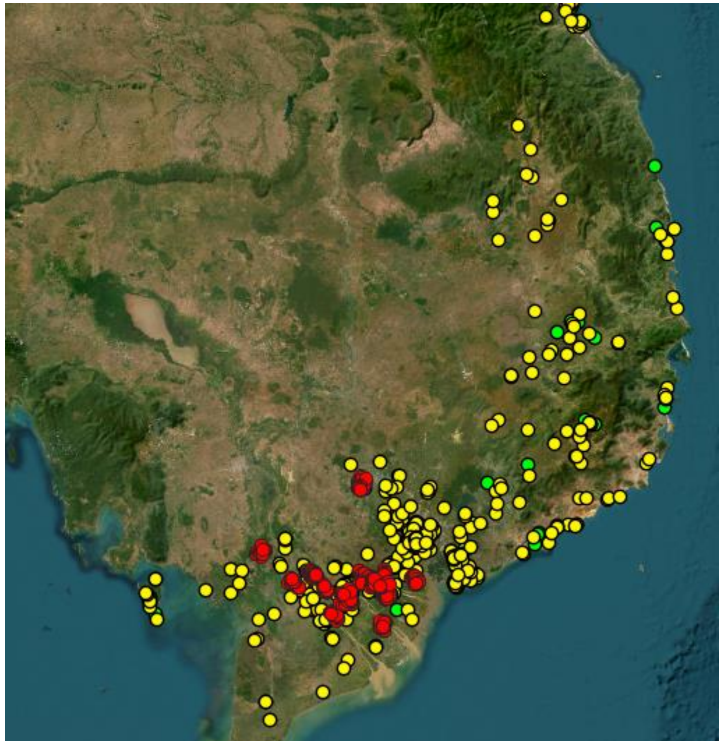
for (X_i, C_i) in ((X_N, C_N), (X_S, C_S)) **do**
  for (X_c in X_{NS}) **do**
    $X_c.d = \sqrt{(X_c.long - C_i.long)^2 + (X_c.lat - C_i.lat)^2}$
    $X_c.p = \frac{1}{X_c.d+\epsilon}$
  end for
  p_{sum} = sum($X_{NS}.p$)
  for (X_c in X_{NS}) **do**
    $X_c.p$ /= p_{sum}
  end for
  X_{i_NS} = sample_with_probs(data=X_{NS}, size=length(X_i), replace=False)
  $X_{NS} = X_{NS} \setminus X_{i\_NS}$
end for

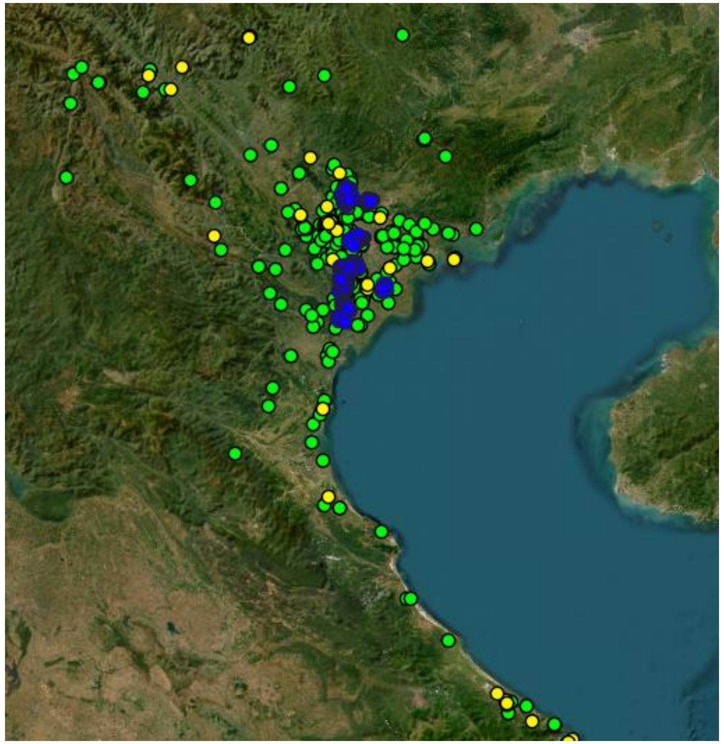
Distance-based non-school sampling





Distance-based non-school sampling





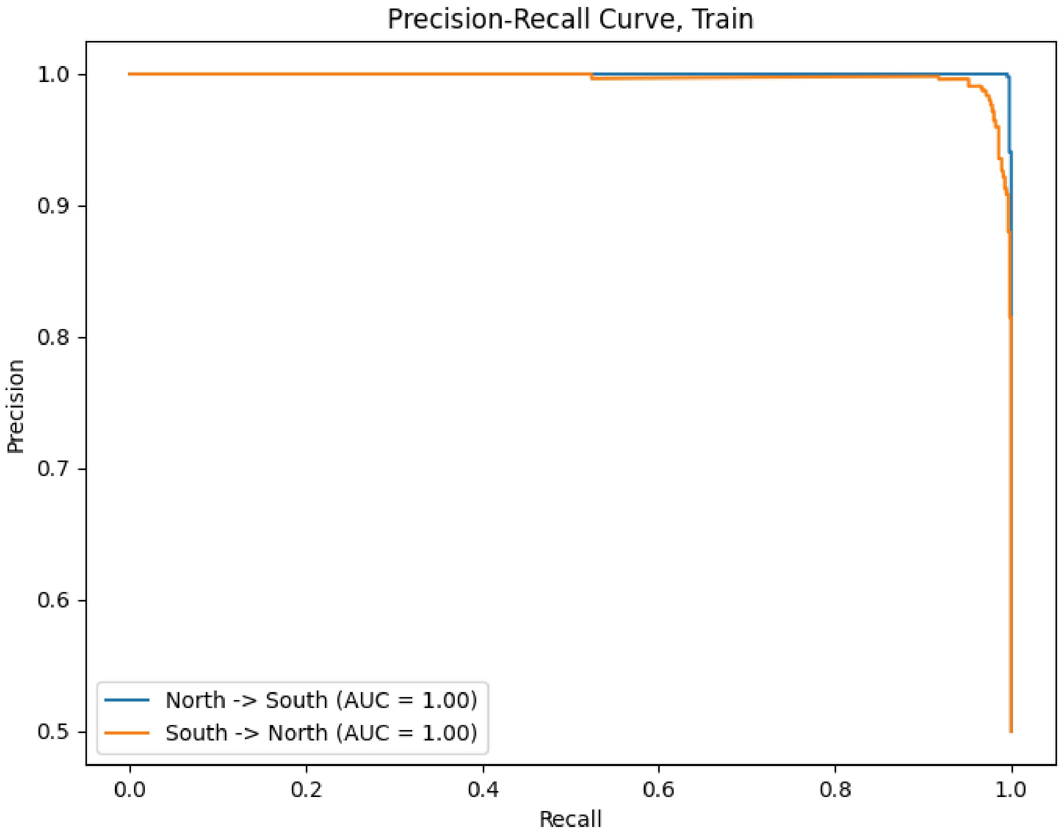
- South School Cluster
- South Sampled Non-Schools

- North School Cluster
- North Sampled Non-Schools

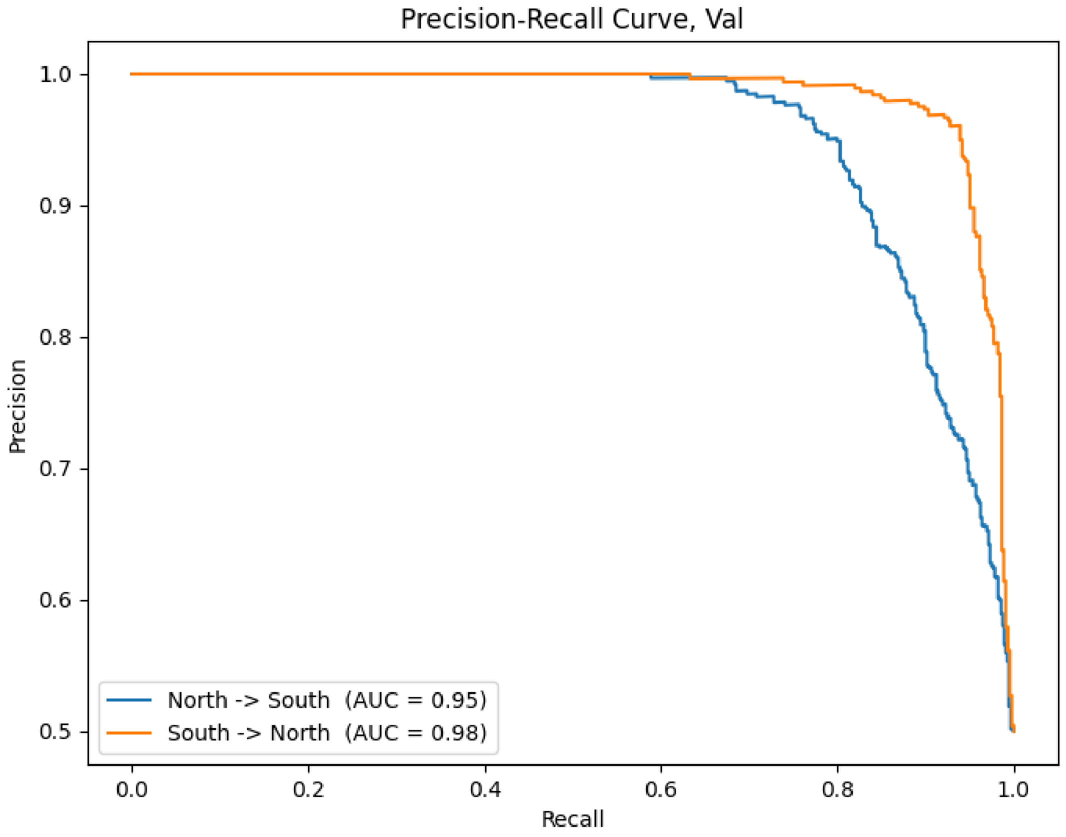
Results comparison

- Random non-school sampling
  - **F1: 81.55%**
  - P: 95.94%
  - R: 73.97%
- Distance-based non-school sampling
  - **F1: 89.03%**
  - P: 97.38%
  - R: 82.53%

Precision-recall curve



Precision-recall curve



Future work

- Dense inference (run newly fine-tuned model on tiles covering the chosen 26 districts)