School mapping in ESRI imagery

Andrija Gorup, Marin Kačan, Siniša Šegvić

October 3rd, 2024

Recap from last meeting

Outdated imagery

ESRI imagery outdated compared to Google Maps

- New schools (visible on Google Maps), might not be visible in ESRI images

ESRI

GOOGLE

Outdated imagery

ESRI imagery outdated compared to Google Maps

- New schools (visible on Google Maps), might not be visible in ESRI images

ESRI

GOOGLE

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 throughout Vietnam
- unusually high results - **F1: 93.91 pp**
  - in spite of potentially problematic outdated imagery

Data analysis

Outdated imagery

Newer schools

- visible on Google Maps but not visible in ESRI images

Outdated imagery

Newer schools

- visible on Google Maps but not visible in ESRI images

Find them efficiently through:

Outdated imagery

Newer schools

- visible on Google Maps but not visible in ESRI images

Find them efficiently through:

- urban growth layer (2014 - 2023) [Anditi]
  - new schools more likely in areas where urban growth was recorded
  - get an “urban growth score” for each school tile

Outdated imagery

Newer schools

- visible on Google Maps but not visible in ESRI images

Find them efficiently through:

- urban growth layer (2014 - 2023) [Anditi]
  - new schools more likely in areas where urban growth was recorded
  - get an “urban growth score” for each school tile
- ESRI metadata layer
  - get the date of capture of ESRI satellite imagery for a given school tile
  - get an “oldness score” for each school tile

Outdated imagery

Newer schools

- visible on Google Maps but not visible in ESRI images

Find them efficiently through:

- urban growth layer (2014 - 2023) [Anditi]
  - new schools more likely in areas where urban growth was recorded
  - get an “urban growth score” for each school tile
- ESRI metadata layer
  - get the date of capture of ESRI satellite imagery for a given school tile
  - get an “oldness score” for each school tile
- combine the two scores and sort schools from highest to lowest

Urban growth layer (2014 - 2023)

A binary mask; 500m resolution

- urban growth: 0/1

Urban growth layer (2014 - 2023)

A binary mask; 500m resolution

- urban growth: 0/1

Example (74.5% tile coverage):

Urban growth layer (2014 - 2023)

A binary mask; 500m resolution

- urban growth: 0/1

Example (74.5% tile coverage):

Urban growth layer (2014 - 2023)

A binary mask; 500m resolution

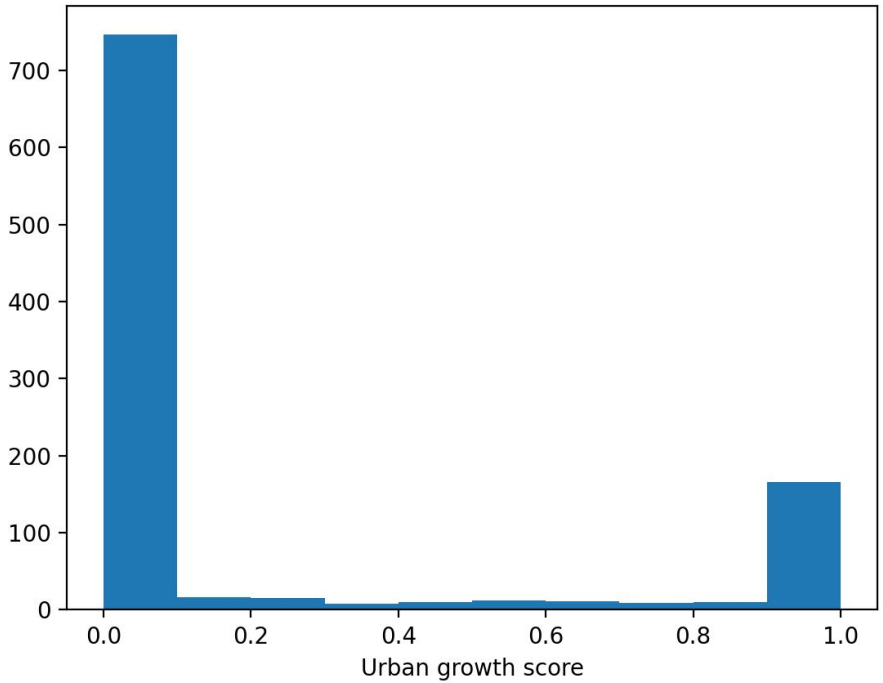
- urban growth: 0/1

For each school tile

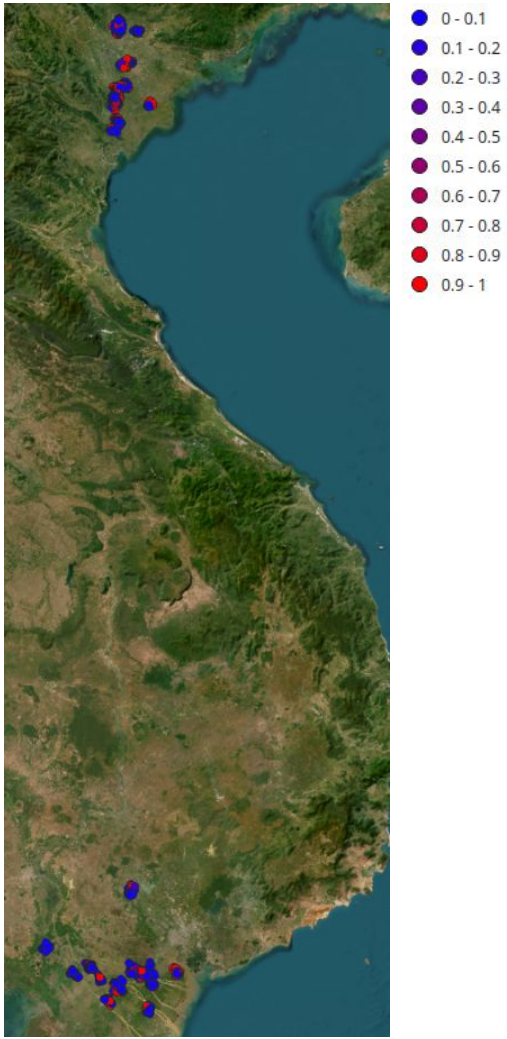
- **urban growth score = area covered by u.g. mask / total area of tile**
- scores in the interval [0,1]

Urban growth layer (2014 - 2023)

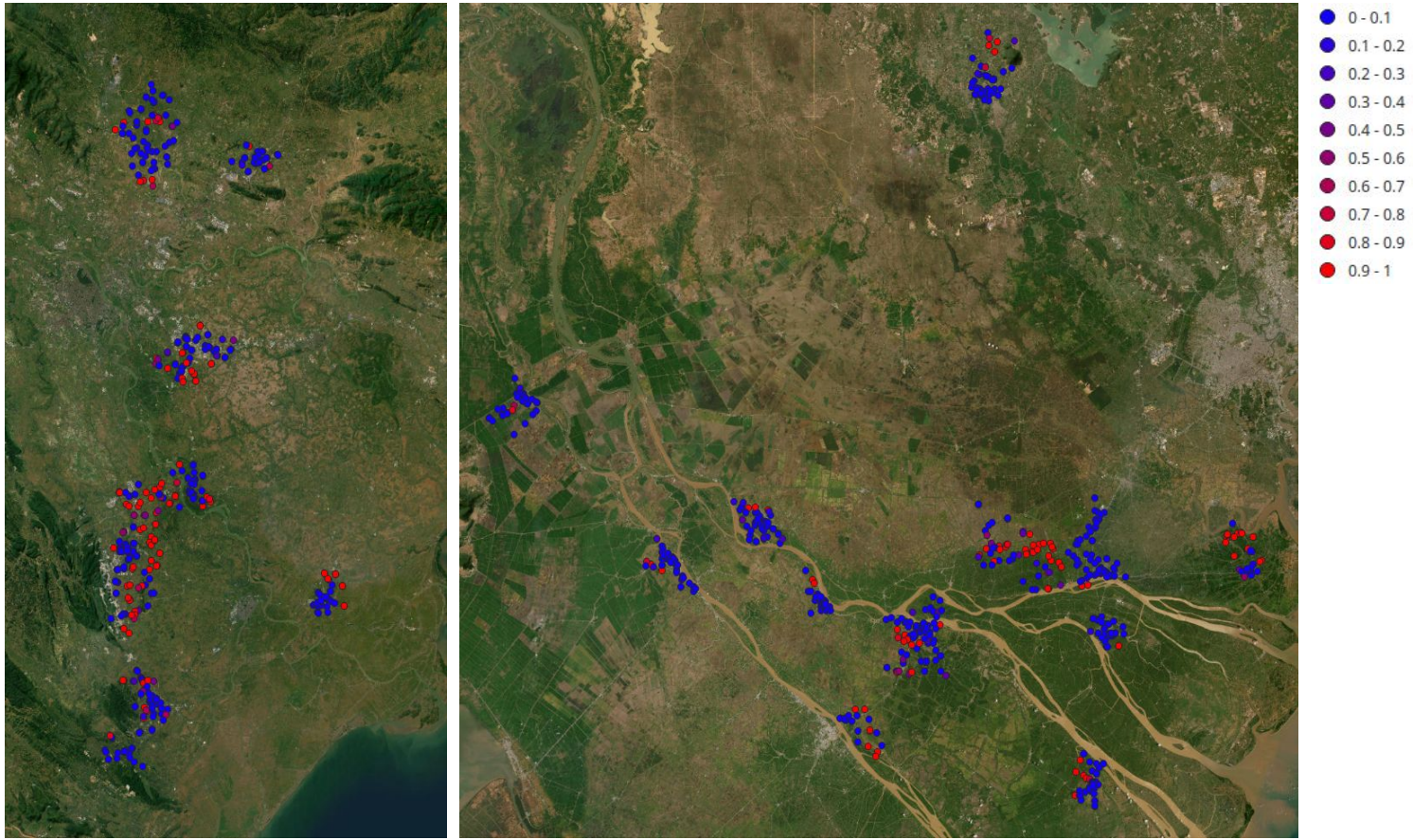
Histogram of urban growth scores



Urban growth layer (2014 - 2023)



Urban growth layer (2014 - 2023)

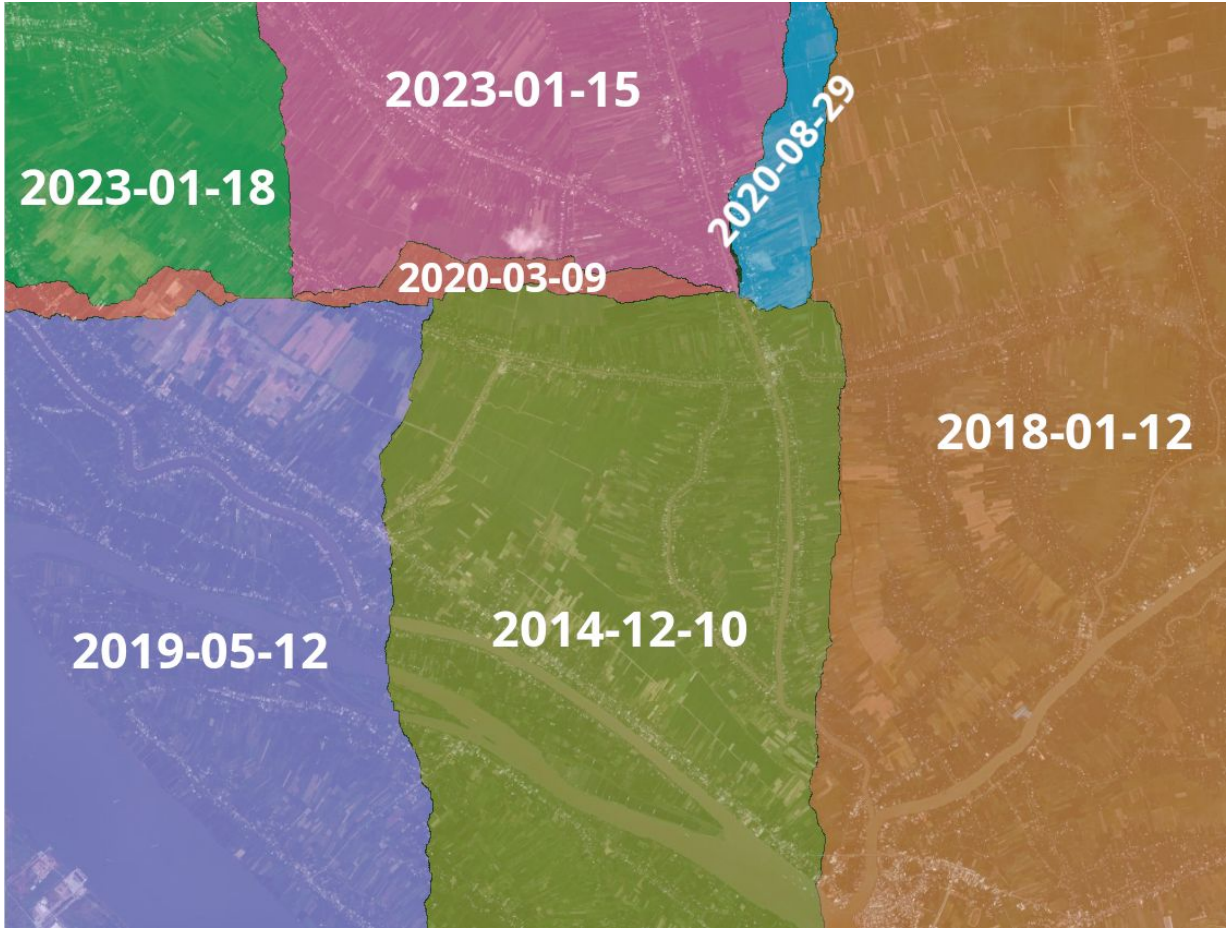


ESRI Metadata Layer

Imagery captured at varying dates (oldest 2014, newest 2023)

ESRI Metadata Layer

ESRI Metadata Layer



ESRI Metadata Layer

Oldness score

- how to score older tiles? linearly, logarithmically, ...?

ESRI Metadata Layer

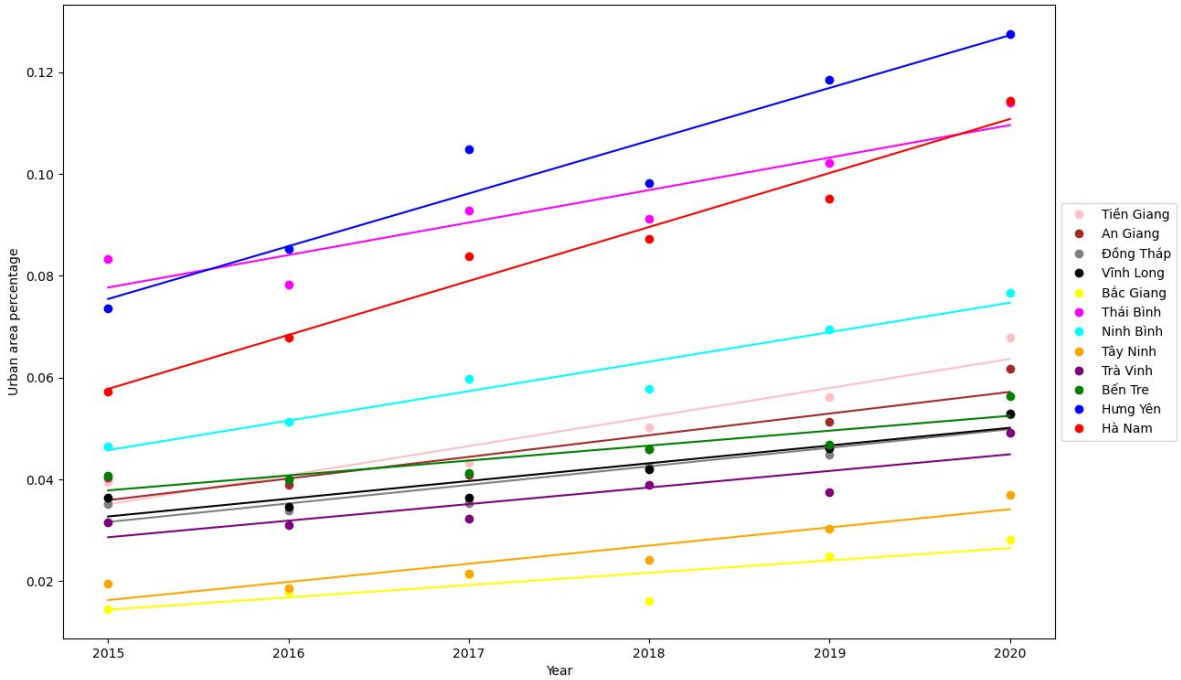
Oldness score

- how to score older tiles? linearly, logarithmically, ...?
- check HR LULC per-year urban growth (2015 - 2020) for relevant provinces

ESRI Metadata Layer

Oldness score

- how to score older tiles? linearly, logarithmically, ...?
- linear growth



ESRI Metadata Layer

Score linearly with the “age” of the imagery (in days)

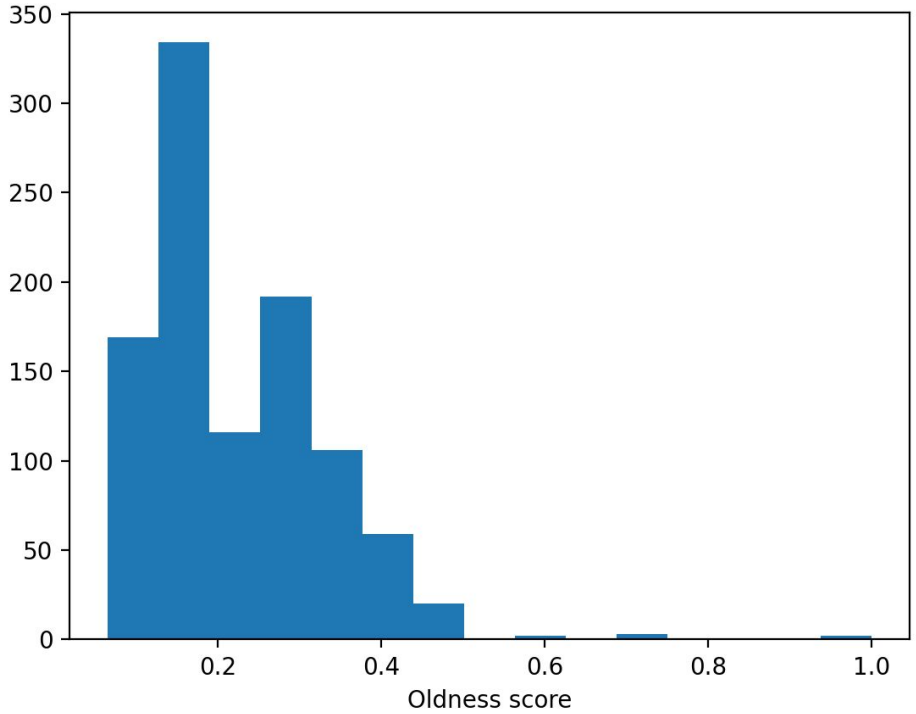
- normalize to [0,1]

ESRI Metadata Layer

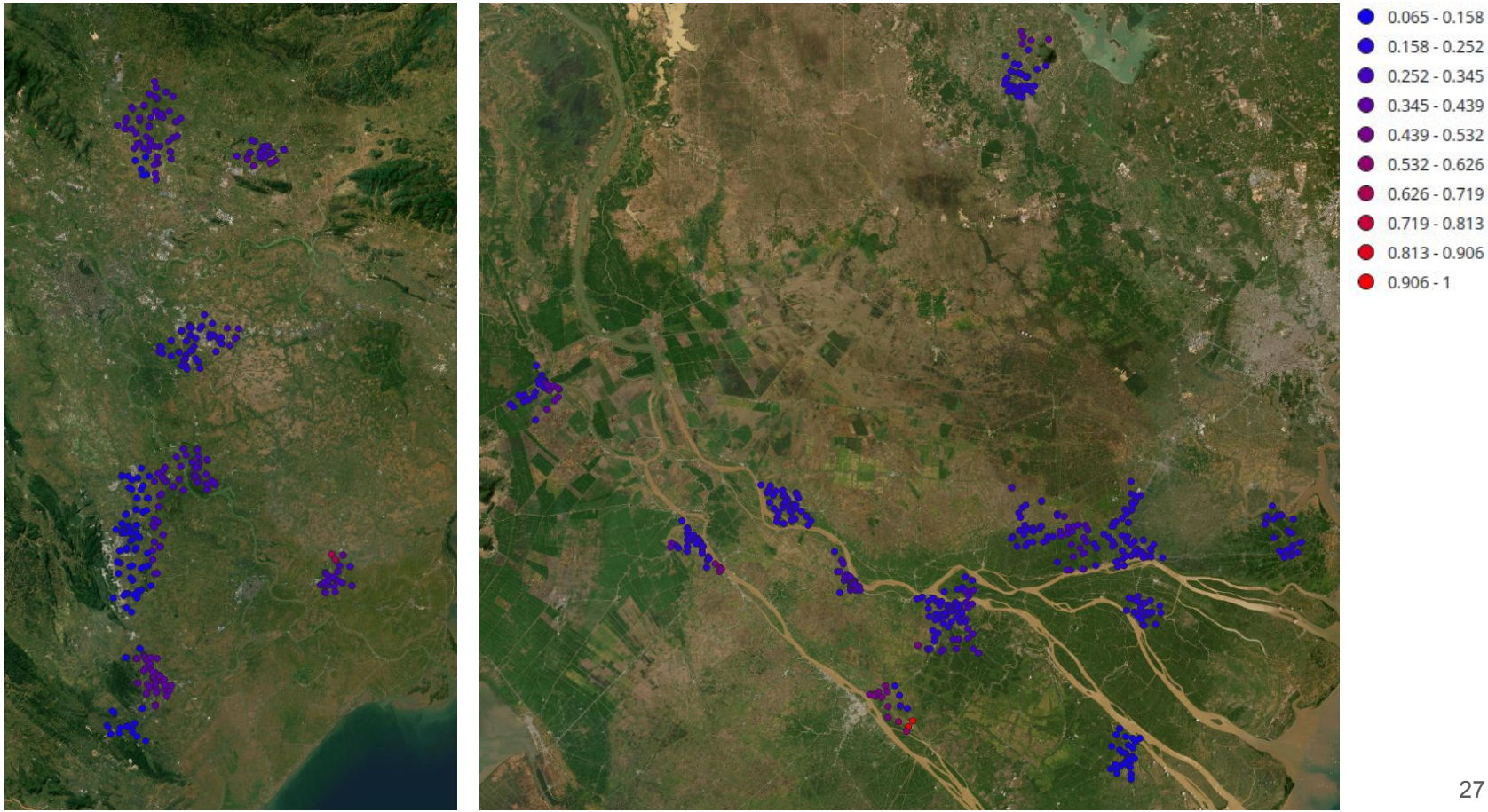
Score linearly with the “age” of the imagery (in days)

- normalize to [0,1]

Histogram of “oldness” scores



ESRI Metadata Layer



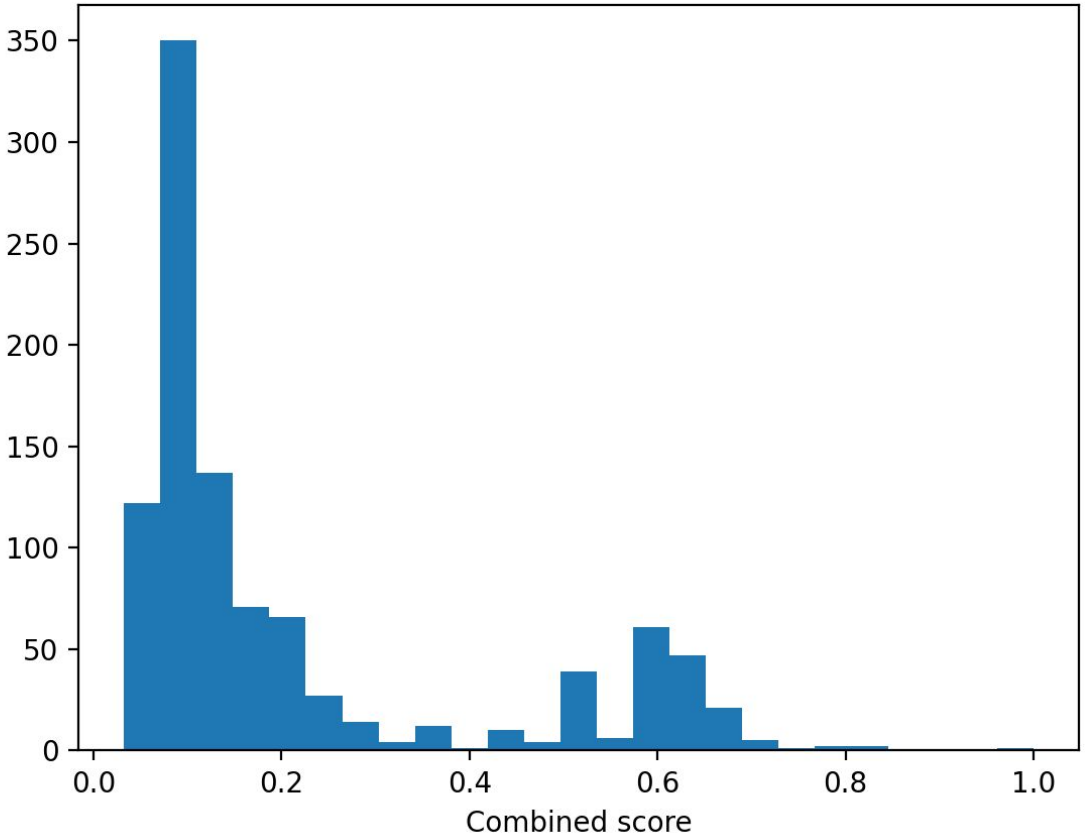
Combined score

UG - urban growth score
O - oldness score

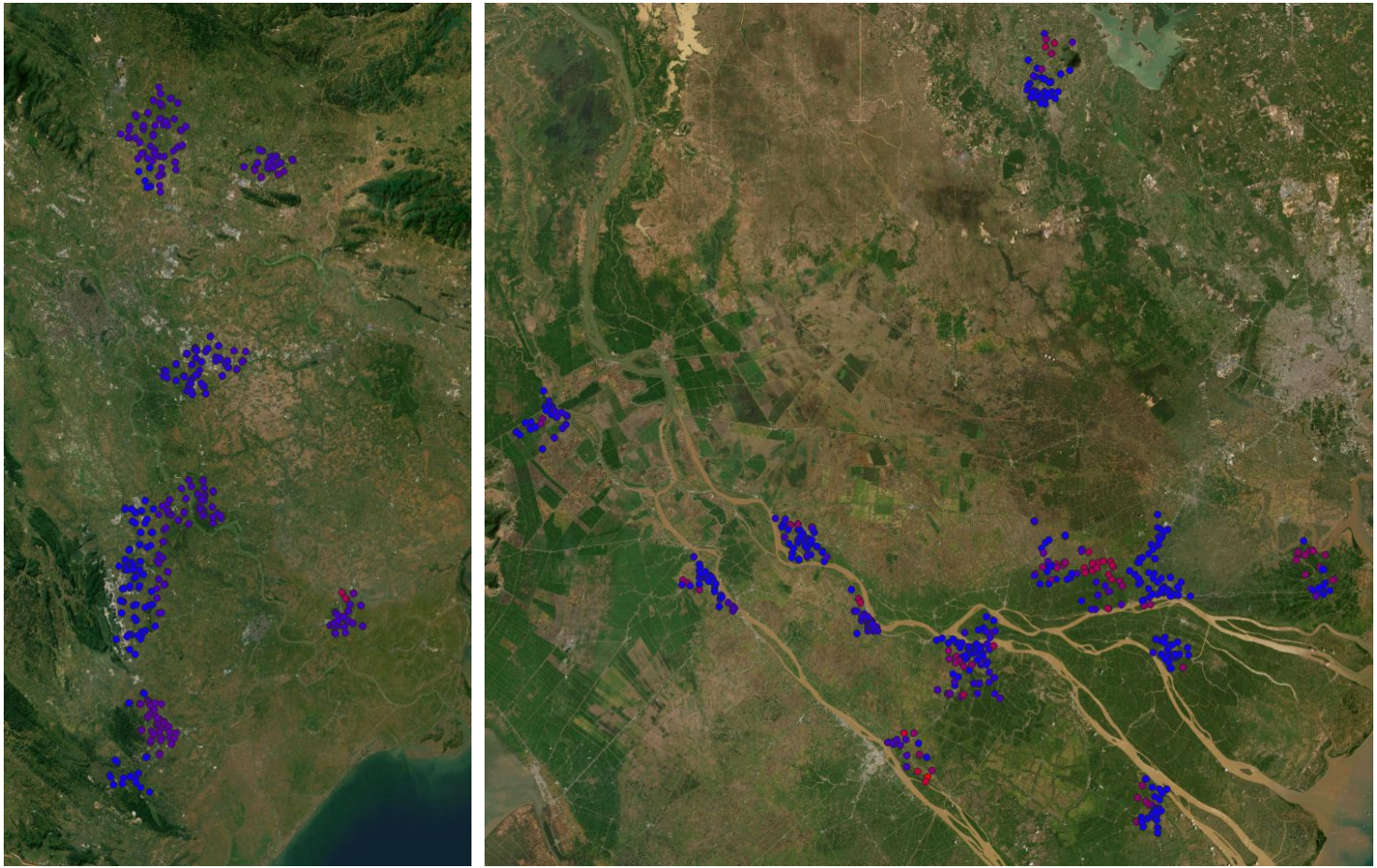
combined score (C)

- (UG + O) / 2

Combined score - histogram



Combined score



Find outdated tiles on areas with observed urban growth

Sort by combined score (from highest)

- checked top 40 schools

Find outdated tiles on areas with observed urban growth

Sort by combined score (from highest)
Examples:

ESRI (2021-05-30)

GOOGLE MAPS (2024)

Find outdated tiles on areas with observed urban growth

Sort by combined score (from highest)
Examples:

ESRI (2021-02-10)

GOOGLE MAPS (2024)

Only use oldness score (without urban growth)

Only use oldness score (without urban growth)

Estimated urban growth = 0

ESRI (2019-12-08)

GOOGLE MAPS (2024)

Only use oldness score (without urban growth)

Estimated urban growth = 0

ESRI (2020-03-09)

GOOGLE MAPS (2024)

Only use urban growth score (without oldness score)

Examples:

Only use urban growth score (without oldness score)

Examples:

ESRI (2022-10-29)

GOOGLE MAPS (2024)

Another approach - sort by compressed file size

Sort images by file size, look for smallest image

Another approach - sort by compressed file size

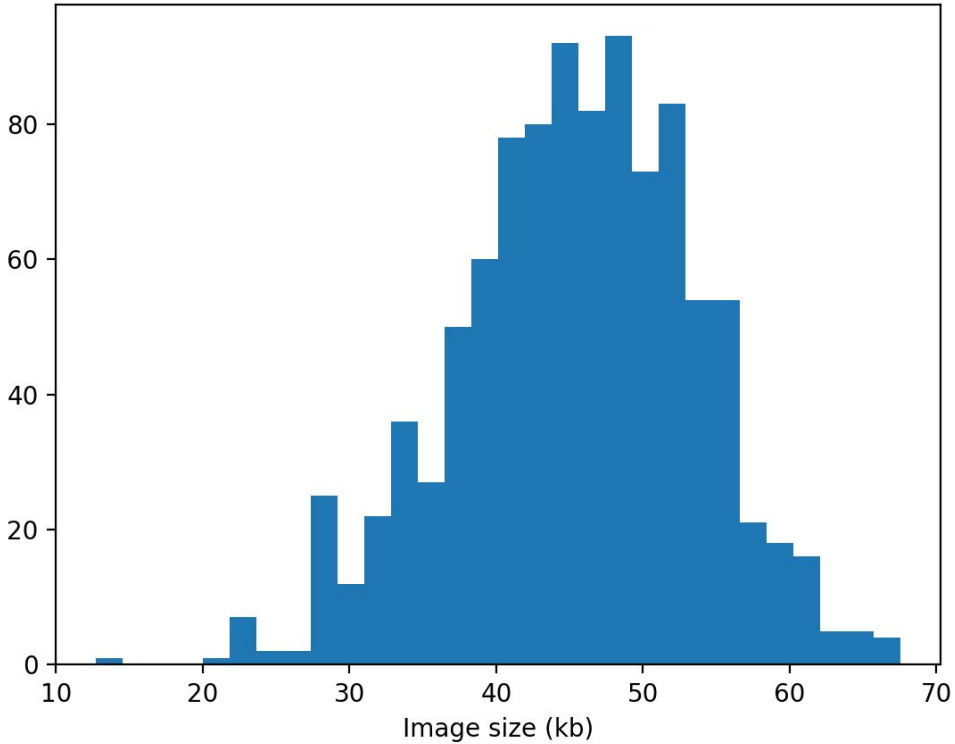
Sort images by file size, look for smallest image

- information theory - images with no urban area will often have lower entropy
    - consequently, their compressed file size will be smaller

Another approach - sort by compressed file size

Sort images by file size, look for smallest image

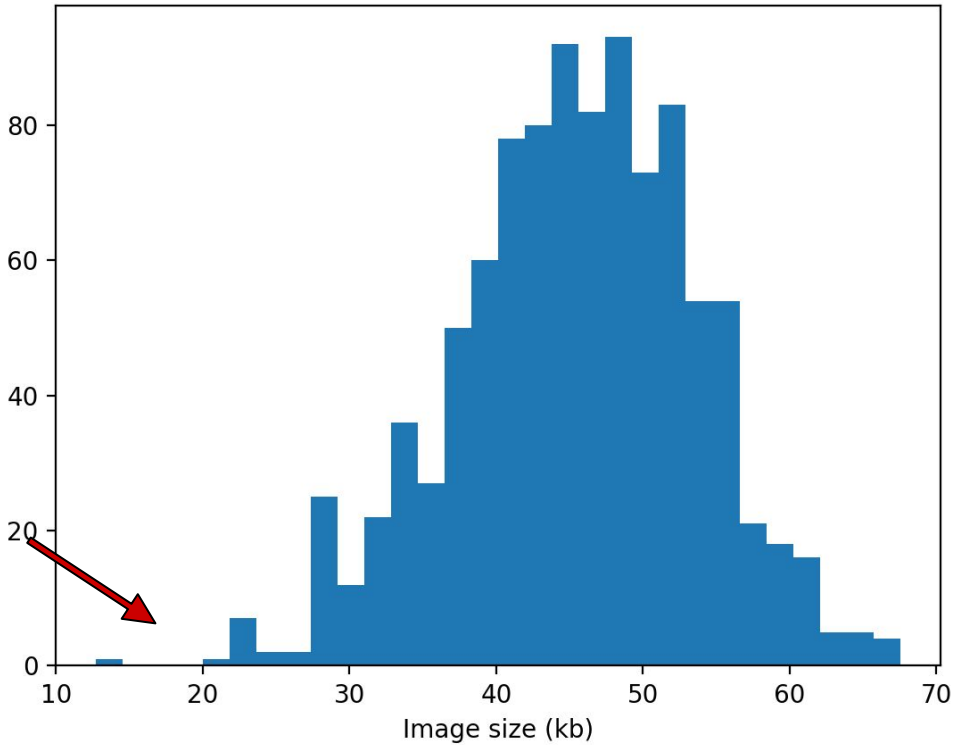
- information theory - images with no urban area will often have lower entropy



Another approach - sort by compressed file size

Sort images by file size, look for smallest image

- information theory - images with no urban area will often have lower entropy



Another approach - sort by compressed file size

Sort images by file size, look for smallest image

- information theory -- images with no urban area will often have lower entropy

Examples:

Another approach - sort by compressed file size

Sort images by file size, look for smallest image

- information theory -- images with no urban area will often have lower entropy

Smallest image

Another approach - sort by compressed file size

Sort images by file size, look for smallest image

- information theory -- images with no urban area will often have lower entropy

Smallest image - cloud-covered:

Another approach - sort by compressed file size

Sort images by file size, look for smallest image

- information theory -- images with no urban area will often have lower entropy

Smallest image - cloud-covered:

Another issue with ESRI - different appearance

Not just out of date, but also more pronounced variations in general appearance

Another problem with ESRI - different appearance

Not just out of date, but also more pronounced variations in general appearance

ESRI




GOOGLE MAPS




Issues in cross-validation experiments

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

2-fold cross-validation

- 50:50 train/val split
- add equal number of non-school locations
  - sampled throughout Vietnam
- **unusually high results - F1: 93.91 pp**
  - in spite of potentially problematic outdated imagery

Issues

Many schools next to each other

Issues - neighbouring schools

Distance: ~1.5 m

Issues - neighbouring schools

Distance: ~51 m

Issues - neighbouring schools

Distance: ~103 m

Issues - neighbouring schools

Distance: ~199 m

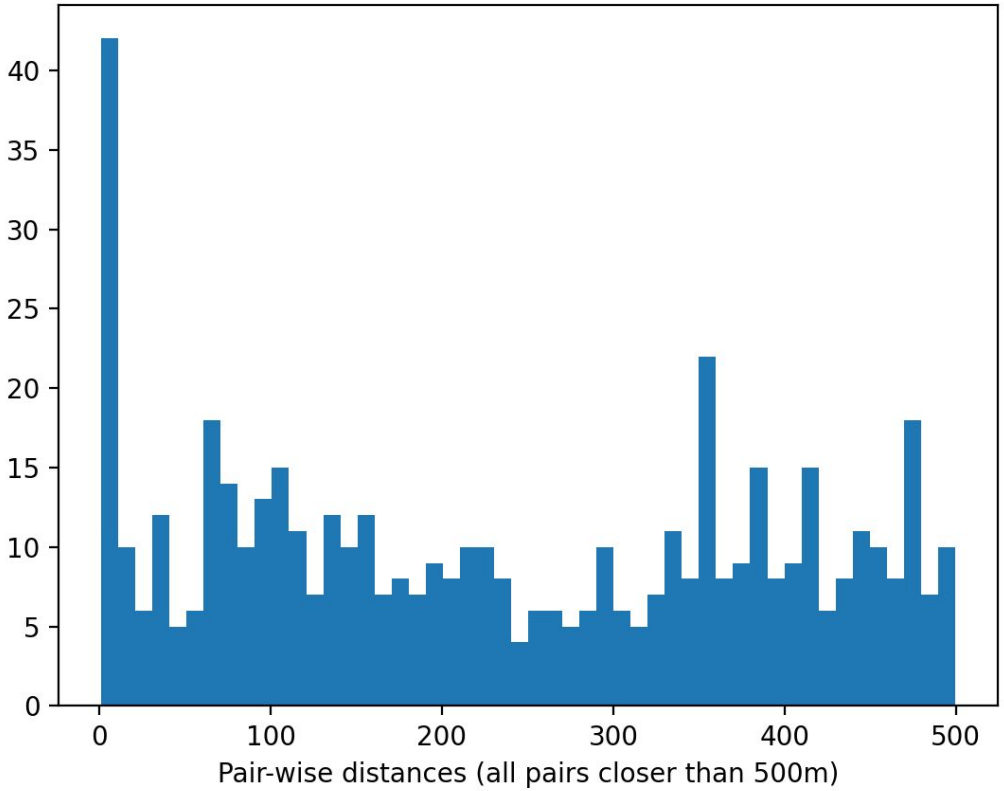
Issues - neighbouring schools

Outdated neighbouring schools (3 separate “schools”)



Issues - neighbouring schools

Histogram of pairwise distances (only distances <500m)



Issues - neighbouring schools

1) Many schools next to each other
- for cross-validation experiments, we do a random **50:50** train/val split
- **leakage** -- very overlapping images get into both splits
    - model can overfit to those examples

Issues - neighbouring schools

1) Many schools next to each other
- for cross-validation experiments, we do a random **50:50** train/val split
- **leakage** -- very overlapping images get into both splits
    - model can overfit to those examples

2) Schools in areas with distinct “appearance”
- not due to similar landscape/architecture, but source of imagery

Issues - neighbouring schools

1) Many schools next to each other
- for cross-validation experiments, we do a random **50:50** train/val split
- **leakage** -- very overlapping images get into both splits
    - model can overfit to those examples

2) Schools in areas with distinct “appearance”
- not due to similar landscape/architecture, but source of imagery

Issues - neighbouring schools

1) Many schools next to each other
- for cross-validation experiments, we do a random **50:50** train/val split
- **leakage** -- very overlapping images get into both splits
    - model can overfit to those examples

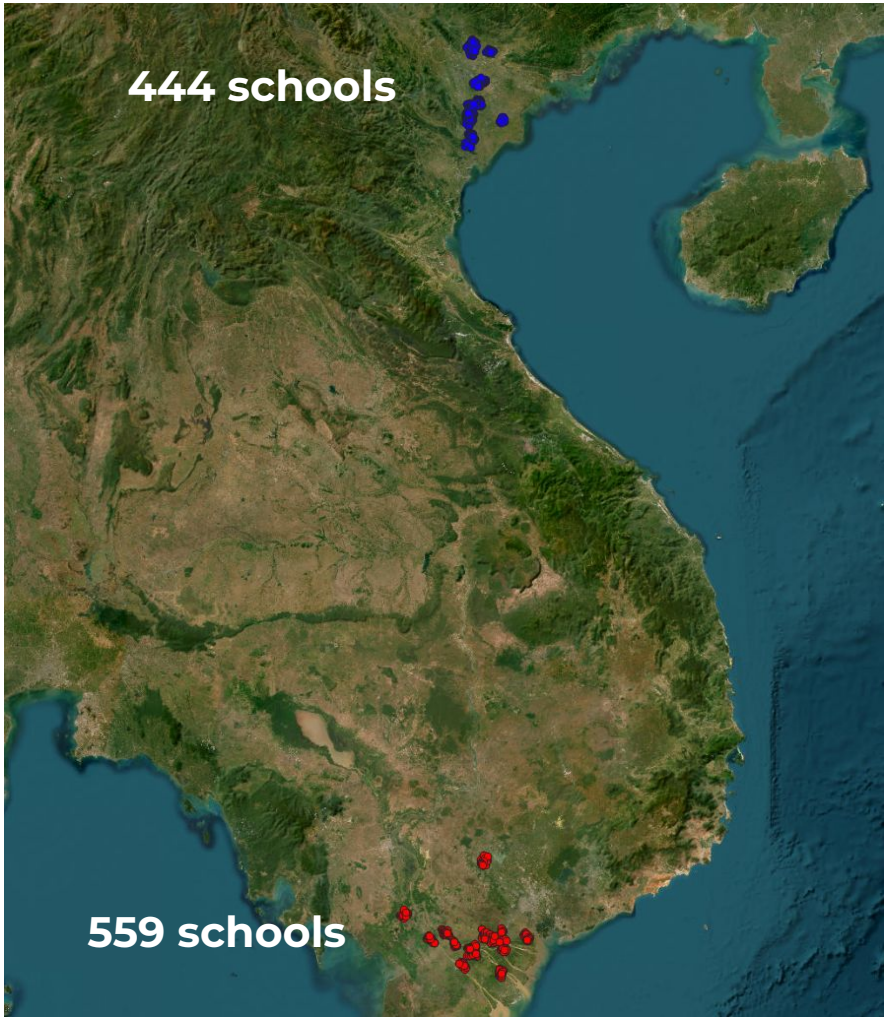
2) Schools in areas with distinct “appearance”
- not due to similar landscape/architecture, but source of imagery

- tiles with very similar appearance get into both splits
- non-schools are sampled from entire Vietnam
    - probably don’t have that area-specific appearance
    - model can overfit to that kind of area-specific appearance

Solution - split by clusters

Solution - split by clusters

Two very obvious clusters



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

More realistic results

- **F1: 81.55 pp**
- P: 95.94 pp
- R: 73.97 pp

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

More realistic results

- **F1: 81.55 pp**
- P: 95.94 pp
- R: 73.97 pp

Additional improvement (future work)

- sample non-school locations from areas where schools are located

Conclusion, future work

Google

- use this investigation to justify the request for Google’s satellite imagery

Conclusion, future work

Google
- use this investigation to justify the request for Google’s satellite imagery

Dense inference (run our model on tiles covering the chosen 26 districts)
- re-run the procedure (we now have a better setup for model selection)