Using Openstreetmap crowdsourced data and Landsat imagery for land cover mapping in the Laguna de Bay area of the Philippines

Brian A. Johnson*, Kotaro lizuka, lsao Endo, Damasa B. Magcale-Macandog, Milben Bragais

*Institute for Global Environmental Strategies (Hayama, Japan)

Institute for Sustainable Humanosphere, Kyoto University (Kyoto, Japan)

University of the Philippines Los Banos (Los Banos, Philippines)

IGES Institute for Global Environmental Strategies

What is crowdsourced geo-data?

- Geographic data provided by private citizens rather than government agencies.
- Examples
  - OpenStreetMap: free base map data on roads, land use, buildings, etc.
    - users digitize points/lines/polygons onto georeferenced satellite imagery (Bing Maps imagery) or upload GPS data taken in the field
    - Largest source of crowdsourced geo-data
  - GeoWiki: global land cover validation data
    - users label the land cover at random locations based on interpretation of high-res images.
  - Flickr (geotagged photos)
    - users upload georeferenced photos with tags

Potential uses of crowdsourced data for land cover mapping

- For accuracy assessment of land cover maps (e.g. GeoWiki)
- For extracting training data.
  - Benefit: land cover mapping can be done very quickly (no need to collect training data).
  - Challenge: the data contains various types of errors.
    - User errors: volunteer mislabels polygon or digitizes inaccurate boundary
    - Image errors: image not accurately georegistered or image outdated

Research questions

- What classification methods can handle the noisy training data extracted using OpenStreetMap (OSM)
  - “landuse” and “natural” polygon layers used in this study
- What level of classification accuracy can be achieved using this extracted training data?

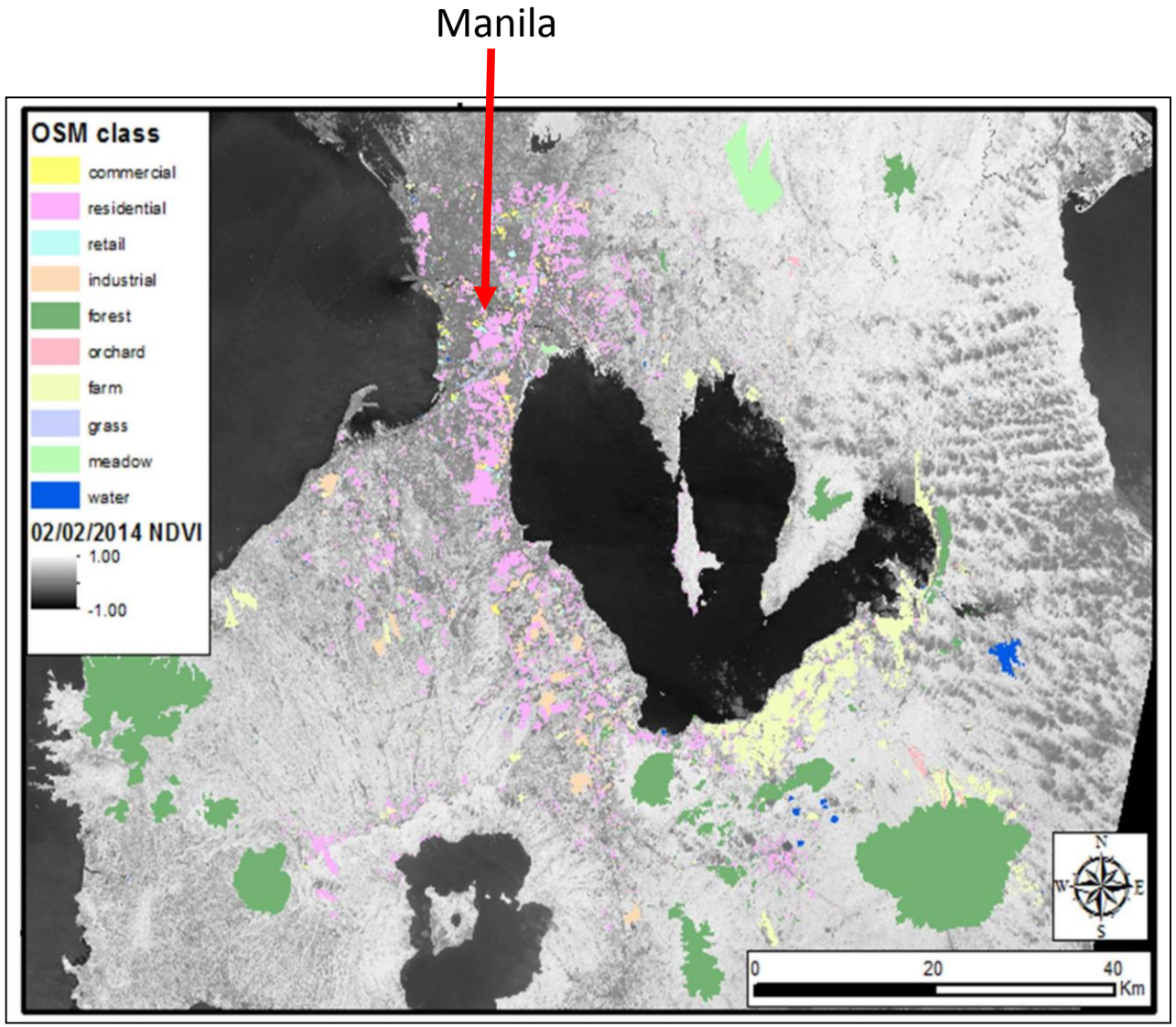
What is new?

- Other studies have used OSM for image classification, but they manually filtered the OSM data first to remove any errors (very time consuming). We try to use the noisy data without manual filtering.

Study area and data

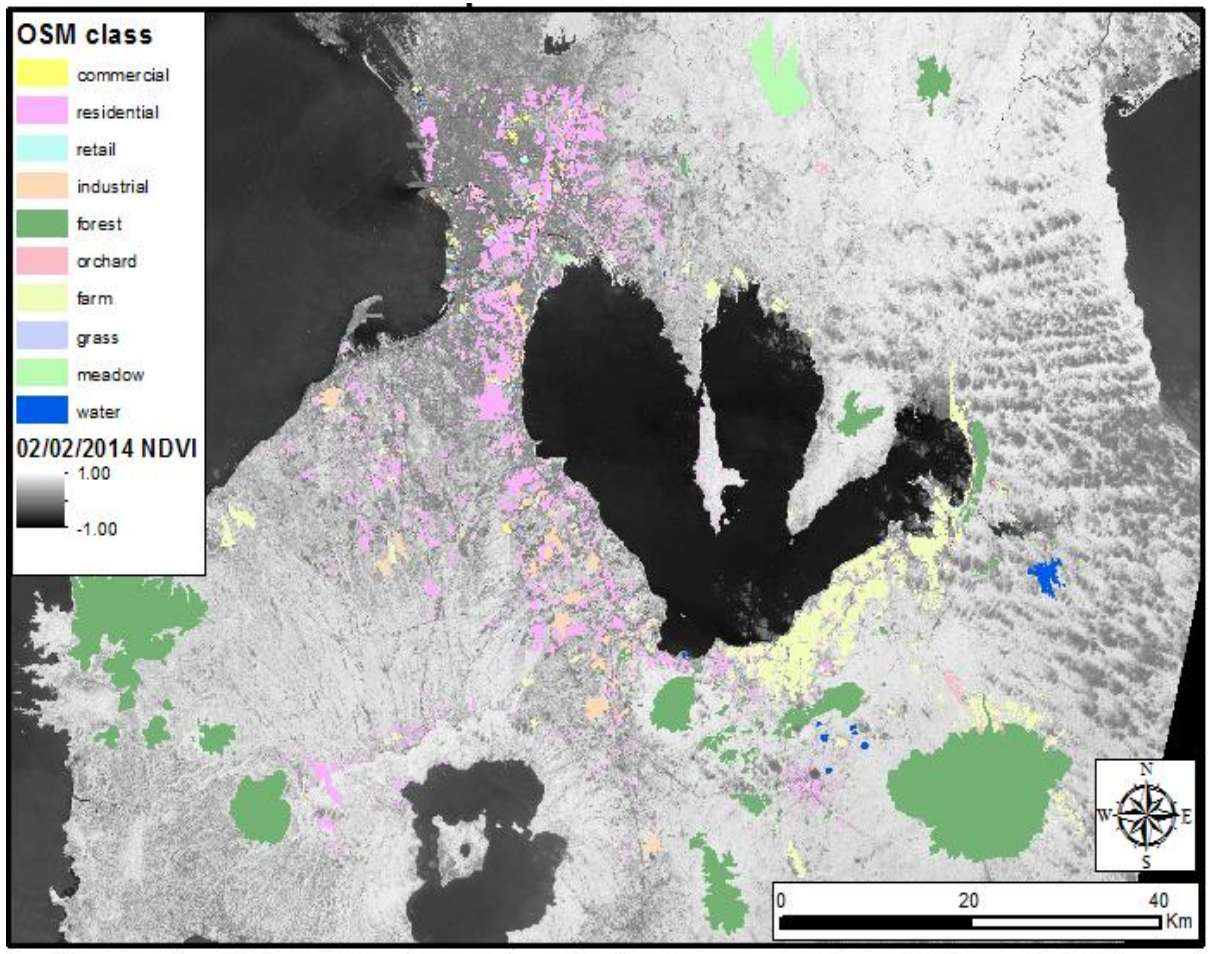
- Study area: Lake Laguna
  - Largest lake in the Philippines
  - Important water source for millions of people
- OSM data: “landuse” and “natural” polygon layers
- Image data: Landsat NDVI time-series data from 2014-2015

2014			2015	
6-Jan	28-Apr	5-Oct	9-Jan	15-Apr
22-Jan	14-May	21-Oct	25-Jan	1-May
7-Feb	30-May	6-Nov	10-Feb	17-May
23-Feb	15-Jun	22-Nov	26-Feb	2-Jun
27-Mar	1-Jul		14-Mar	20-Jul
12-Apr	18-Aug		30-Mar	



Extracting training data from Landsat images

- OSM classes converted to 6 land cover classes. Aggregated to 4 classes after classification.
- OSM polygons split 50/50 to generate training/validation data sets.
- Sample pixels (~10,000) extracted from within training polygons
- 300 points generated inside validation polygons, manually labelled using Google Earth images from 2014-2015.

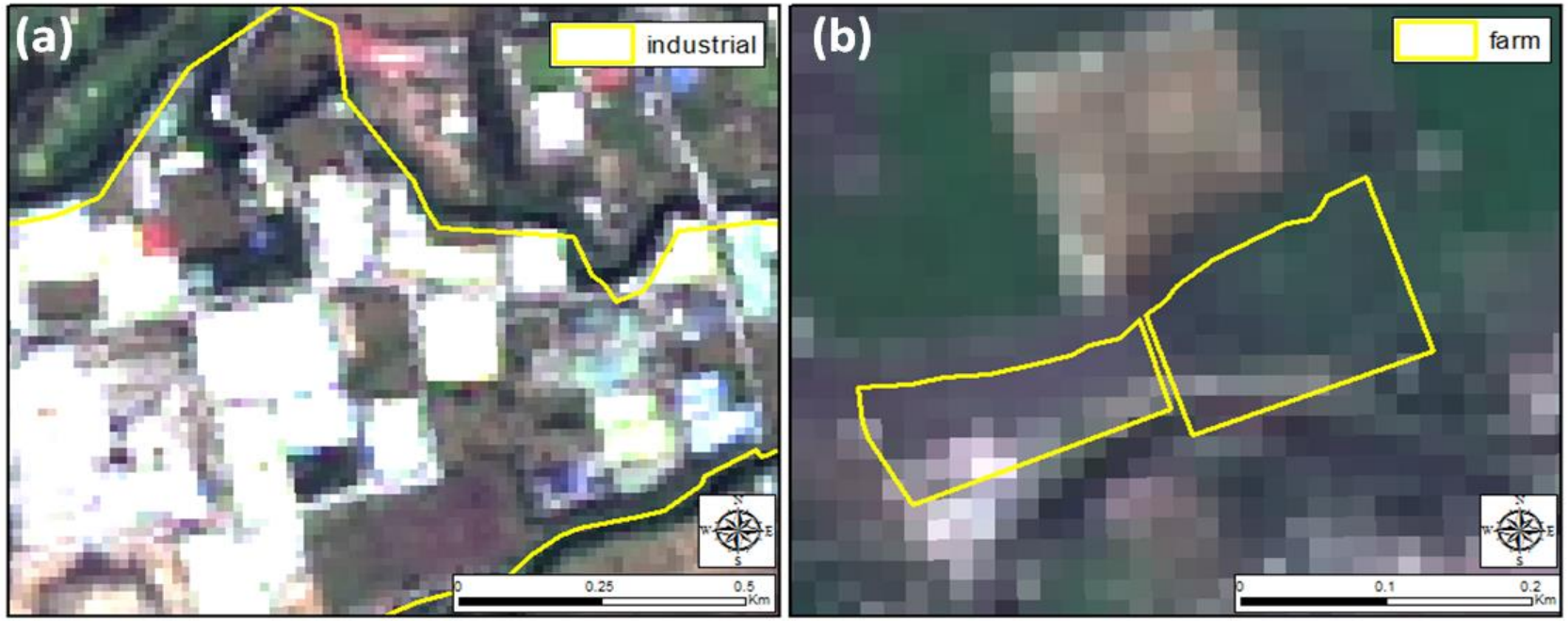


OSM class	LULC class
commercial →	impervious
residential →	impervious
retail →	impervious
industrial →	impervious
forest →	tree
orchard →	**orchard** → tree
farm →	**Farm** -> other vegetation
grass →	other vegetation
meadow →	other vegetation
water →	water

Final classes: impervious, tree, other vegetation, water

Common errors in extracted training data

- (a) pixels representing “impervious” land cover, extracted from “industrial” OSM class, contain vegetation. (class conversion error)
- (b) Inaccurate boundary of a farm in the OSM data (geolocation error in the Bing maps imagery)



Workflow

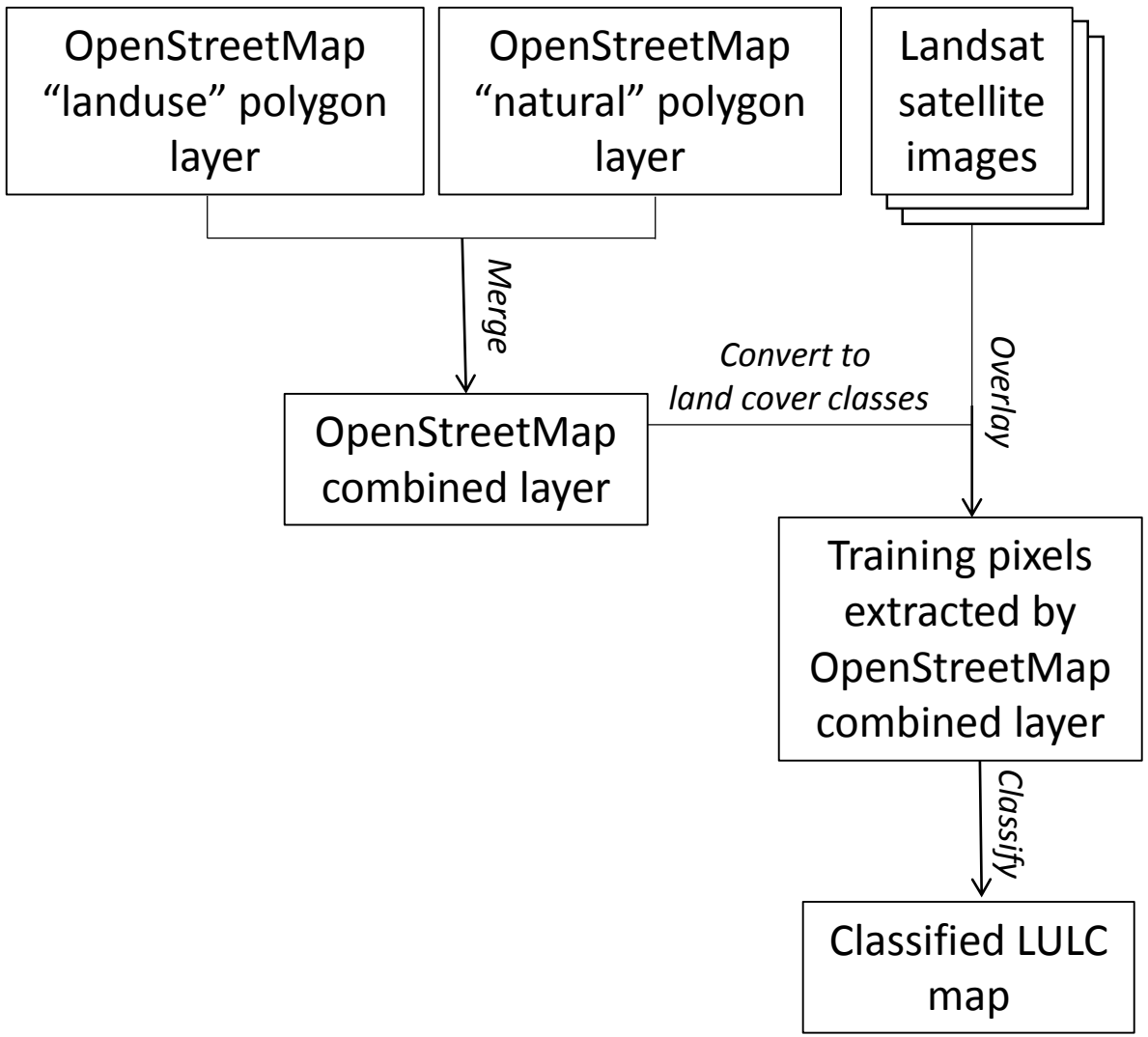

OpenStreetMap "landuse" polygon layer
OpenStreetMap "natural" polygon layer
Landsat satellite images
Merge
Convert to land cover classes
Overlay
OpenStreetMap combined layer
Training pixels extracted by OpenStreetMap combined layer
Classify
Classified LULC map

Image classification

- 3 noise-tolerant algorithms tested for classification
  - C4.5 (decision tree)
  - Naïve bayes (probabilistic)
  - Random forest (ensemble decision tree)

- Synthetic minority class over-sampling technique (SMOTE) used to balance training data.
  - High class imbalance in training data set due to different number/size of OSM polygons for each land cover class (classes with larger coverage have more training pixels).
  - Example: Forest =7431 training pixels, water = 205 training pixels
  - SMOTE generates artificial training samples in the feature space between training pixels to ensure classes have equal # of training samples.

Classification accuracies

- NB and SMOTE-RF had highest overall accuracies (OA).
- NB more accurate for “tree” class (class with most validation samples), but SMOTE-RF more accurate for all other classes.

Classification algorithm	OA (four-class system)
Naïve bayes (NB)	81.3%
C4.5	66.0%
Random forest (RF)	80.3%
SMOTE-NB	80.0%
SMOTE-C4.5	71.3%
SMOTE-RF	**84.0%**

NB

		True LULC					
		I	T	V	W	Sum	UA (%)
Classified	I	33	1	6	0	40	82.5
	T	0	112	13	0	125	89.6
	V	5	21	62	1	89	69.7
	W	3	2	4	37	46	80.4
	Sum	41	136	85	38	300	
	PA (%)	80.5	82.4	72.9	97.4		
	OA (%)						**81.3**

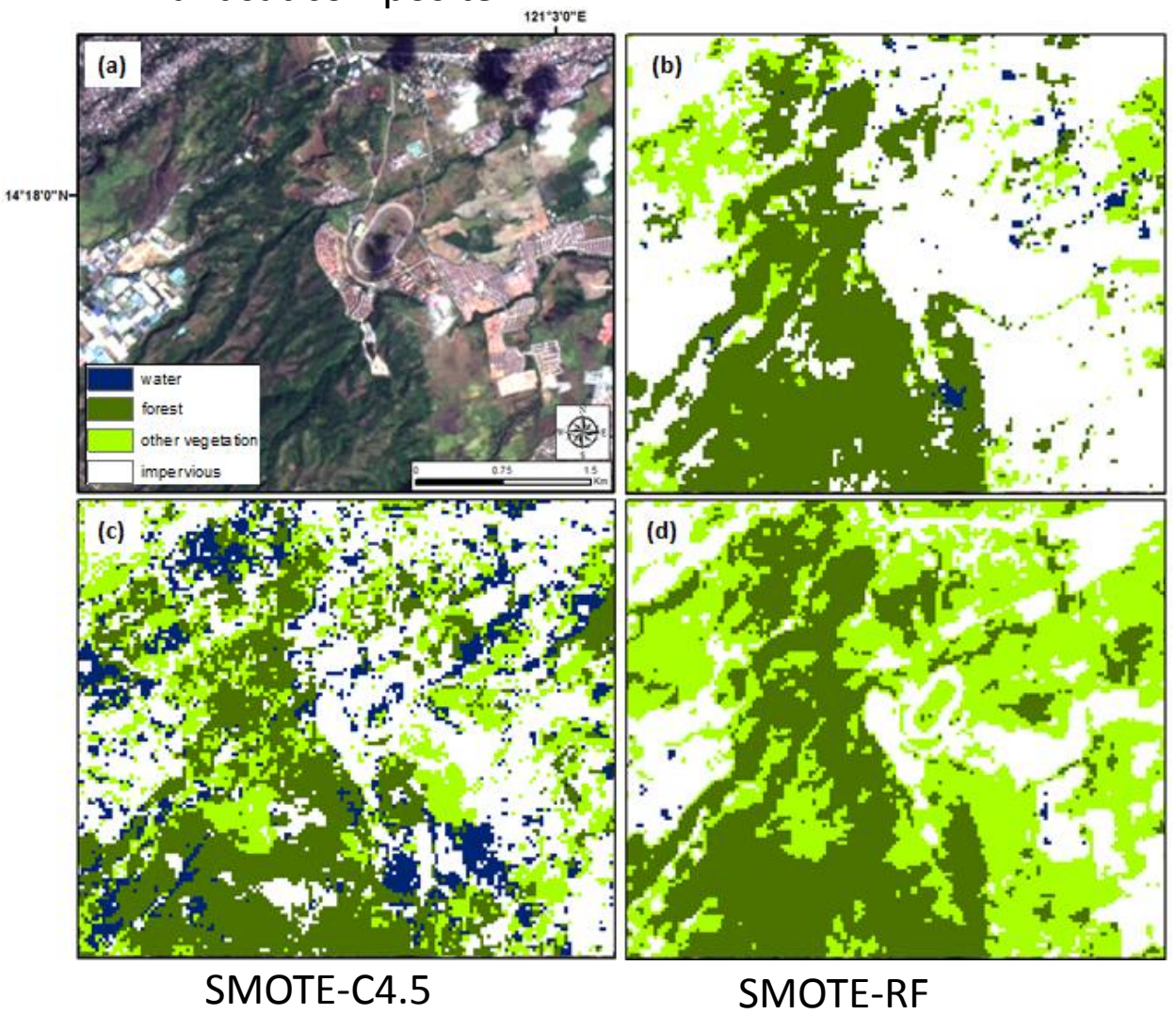
SMOTE-RF

		True LULC					
		I	T	V	W	Sum	UA (%)
Classified	I	38	1	1	0	40	95.0
	T	1	104	20	0	125	83.2
	V	5	12	72	0	89	80.9
	W	3	2	3	38	46	82.6
	Sum	47	119	96	38	300	
	PA (%)	80.9	87.4	75.0	100		
	OA (%)						**84.0**

I = “impervious”, T = “tree”, V = “other vegetation”, W = “water””

Visual comparison of classification results

- NB overestimated impervious area, but better at discriminating between trees and other vegetation.
- C4.5 performed worst and produced noisy result.
- Random forest performed best for impervious class, but some confusion between trees and other vegetation



Conclusions

- Naïve bayes and random forest classifiers could produce moderately accurate (>80% OA) land cover maps using training pixels extracted automatically from OpenStreetMap layers.
  - Possibly lower accuracy than if training data was gathered the traditional way (due to errors in the OSM-extracted training data), but faster and more automated.
  - May be useful if budget or time is limited
- SMOTE could overcome some of the impacts of class imbalance in the training data, particularly for C4.5 and RF algorithms.

Future work

- Test additional classification algorithms
- Evaluate different filtering methods to automatically identify and remove errors in the OSM-extracted training data.

Thank you for your attention!!!

*Funding provided by Climate Change Resilient Low Carbon Society Network (CCR-LCSNet)", Japanese Ministry of the Environment.

Sunset view from IGES.