Disaggregating Census Data in Somalia Using Random Forests

Zhen Liu (zhenliu3@clarku.edu)

Geographic Information Science for Development and Environment, Clark University

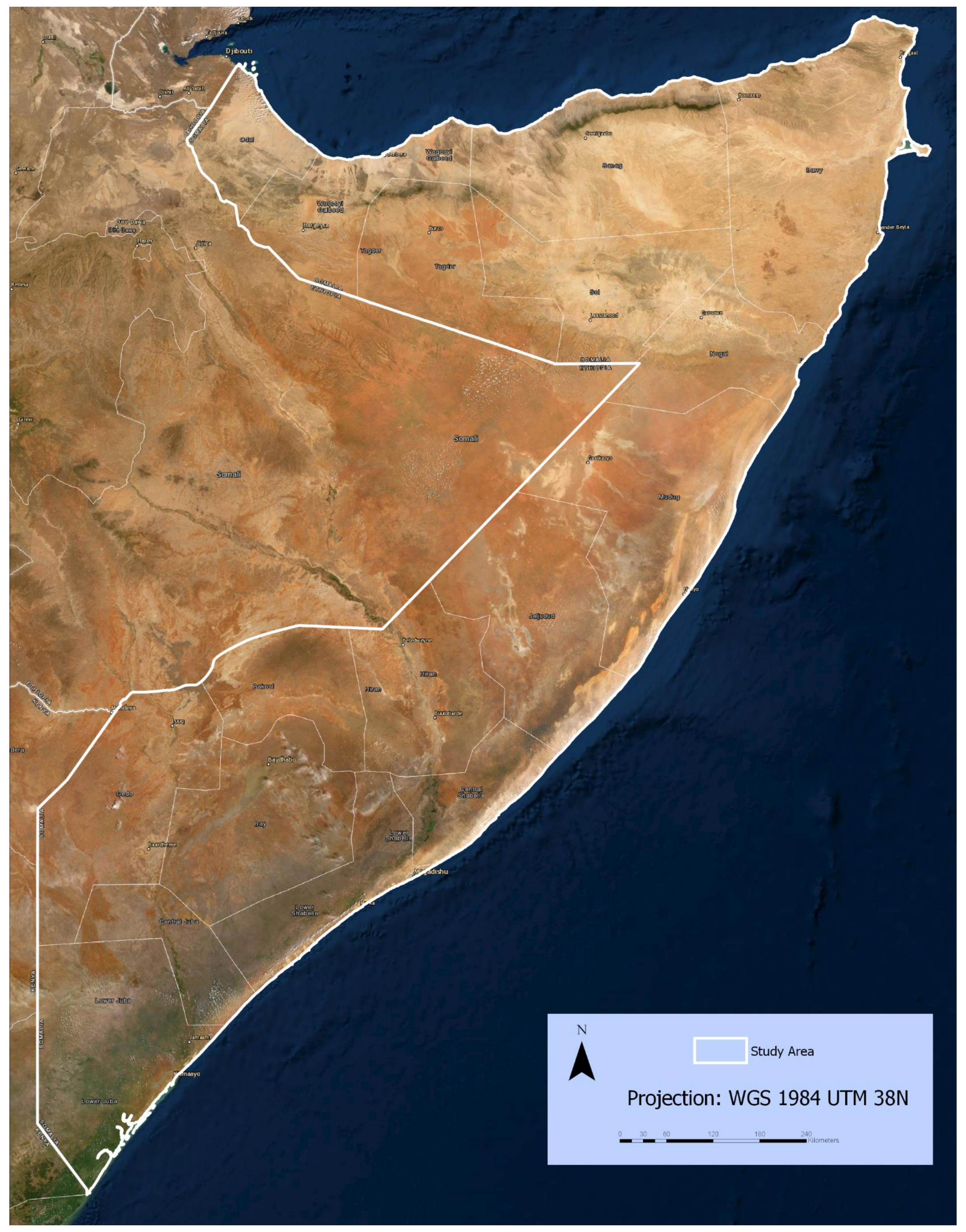
Disaggregating Census Data in Somalia

In 2014, the Population Estimation Survey (PESS) became the first extensive household sample survey to be carried out among the Somalia population in decades. In the meantime, millions of people in Somalia are suffering from the resource shortage, natural disasters, and disease.

For those countries like Somalia that do not have the reliable census data collection, the use of remotely sensed data is especially important for mapping population. This project is to generate the gridded population map in 2018 in Somalia by using Random Forest and an improved method. It will be critical in health, economic, and environmental fields across various temporal and spatial scales.

Figure 1. Displaced population in Somalia (credit IOM)

Study Area - Somalia (the Horn of Africa)



The study area is the whole country of Somalia. The training unit is the level 2 administrative unit in Somalia. The study year is 2018. (Because only 2014 census data is accessible, the census data in 2018 is estimated by the data in 2014 and the growth rate of population.)

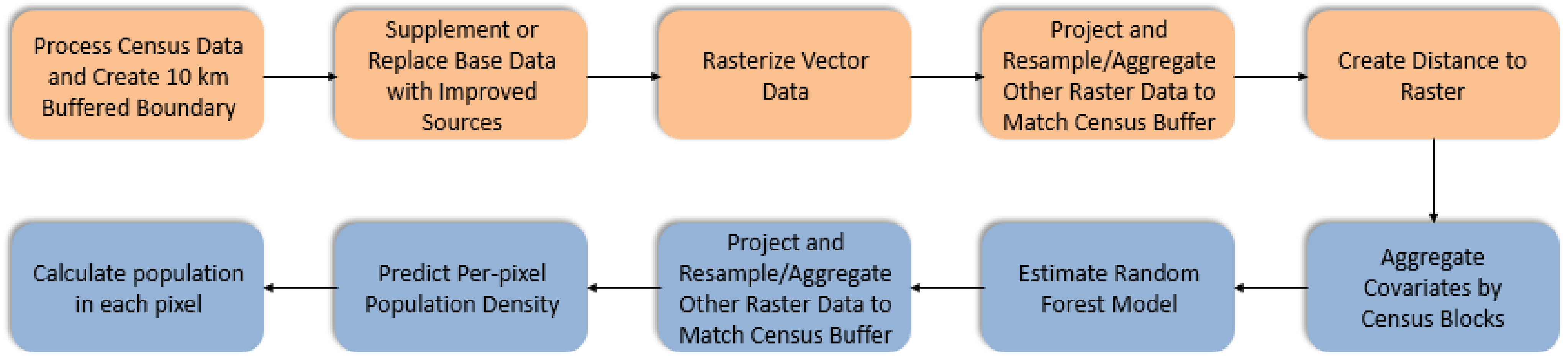
Remotely-Sensed and Ancillary Data

Data summary

Type	Description	Date Collected	Resolution	Source
Census	2014, Admin-level 2	2014		HDX
Landcover	2015, Landcover	2015	500 m	ICPAC
	Built	2015	250 m	GHSL
Raster-Format	Lights at Night	2014	500 m	NOAA
	Mean Temperature, 1950-2000	1950-2000	1 km	WorldClim
	Mean Precipitation, 1950-2000	1950-2000	1 km	WorldClim
	Elevation	2018	30 m	RCMRD
	Slope	2018	30 m	RCMRD-Derived

Type	Description	Date Collected	Resolution	Source
Vector-Format	Distance to Roads	24-SEP-2015		HDX
	Distance to Rivers	14-AUG-2017		ICPAC
	Water facilities	7-NOV-2017		ICPAC
	Health Facilities	28-JAN-2005		HDX
	Schools	2004		HDX
Survey	2018, Internally displaced people	2018		CCCM

General Structure of the Project


Process Census Data and Create 10 km Buffered Boundary
Supplement or Replace Base Data with Improved Sources
Rasterize Vector Data
Project and Resample/Aggregate Other Raster Data to Match Census Buffer
Create Distance to Raster
Aggregate Covariates by Census Blocks
Estimate Random Forest Model
Project and Resample/Aggregate Other Raster Data to Match Census Buffer
Predict Per-pixel Population Density
Calculate population in each pixel

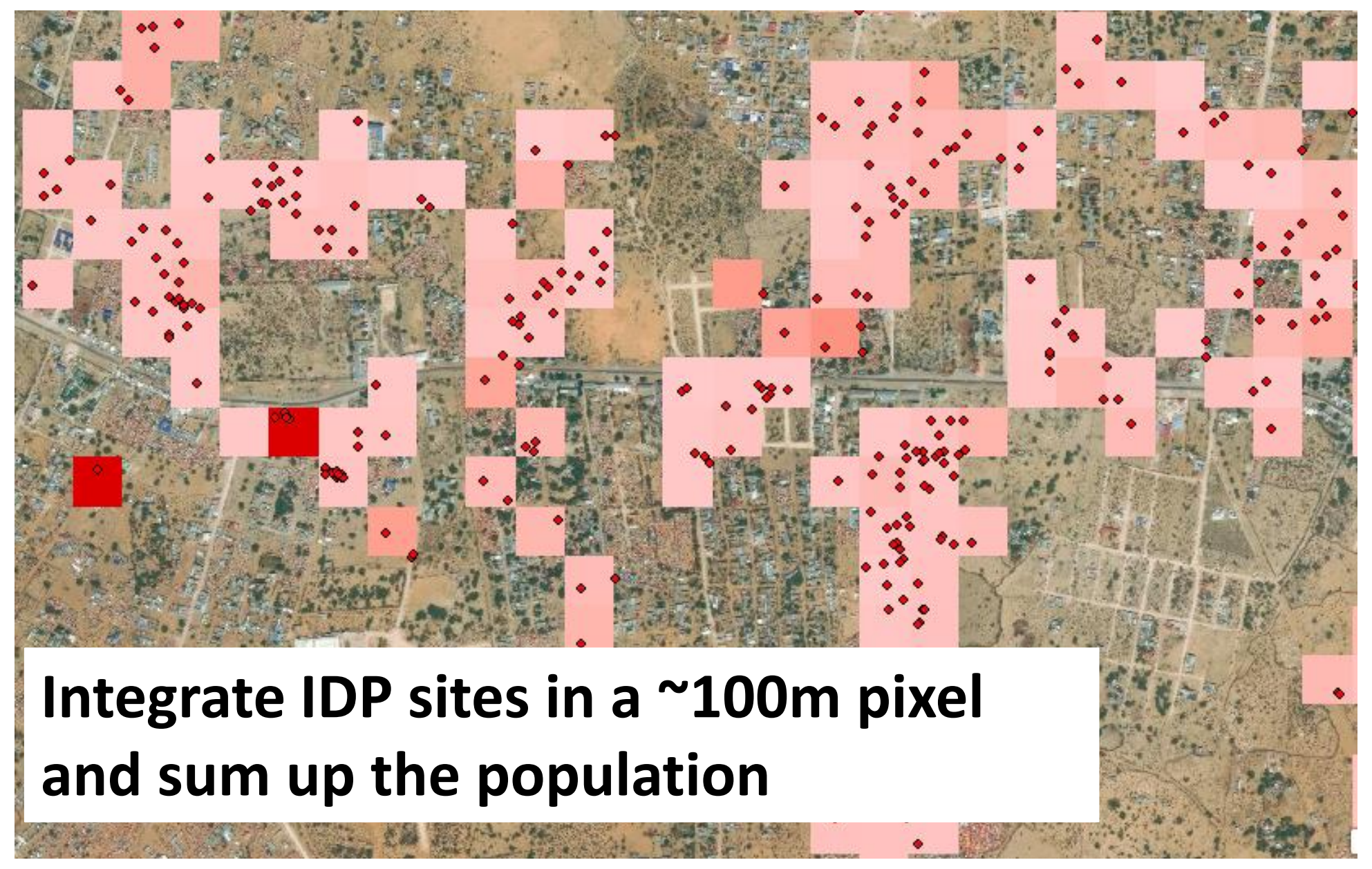
This figure represents the general structure of the data processing and map production procedure. The orange boxes represent data pre-preparation stages. Items in blue represent Random Forest model estimation, per-pixel prediction and redistribution of census counts. In the initial method, the log population density of each level 2 administrative unit is used as a dependent variable. In the improved one, the training unit will be changed.

Data Preparation

- In order to minimize edge effects associated with near-boarder populated areas, the census data are buffered 10 km.
- Also, we need to project all data into a conformal projection most appropriate for each country (e.g. UTM).
- The vector data is first projected and subset to match the buffered national borders, then is rasterized.
- Distance to the point data is generated as our covariates in the model.
- All raster covariates are resampled to match the rasterized census data and its buffer.
- We use a simple nearest neighbor filling approach to extend the edge of data sets and fill any gaps prior to model estimation.

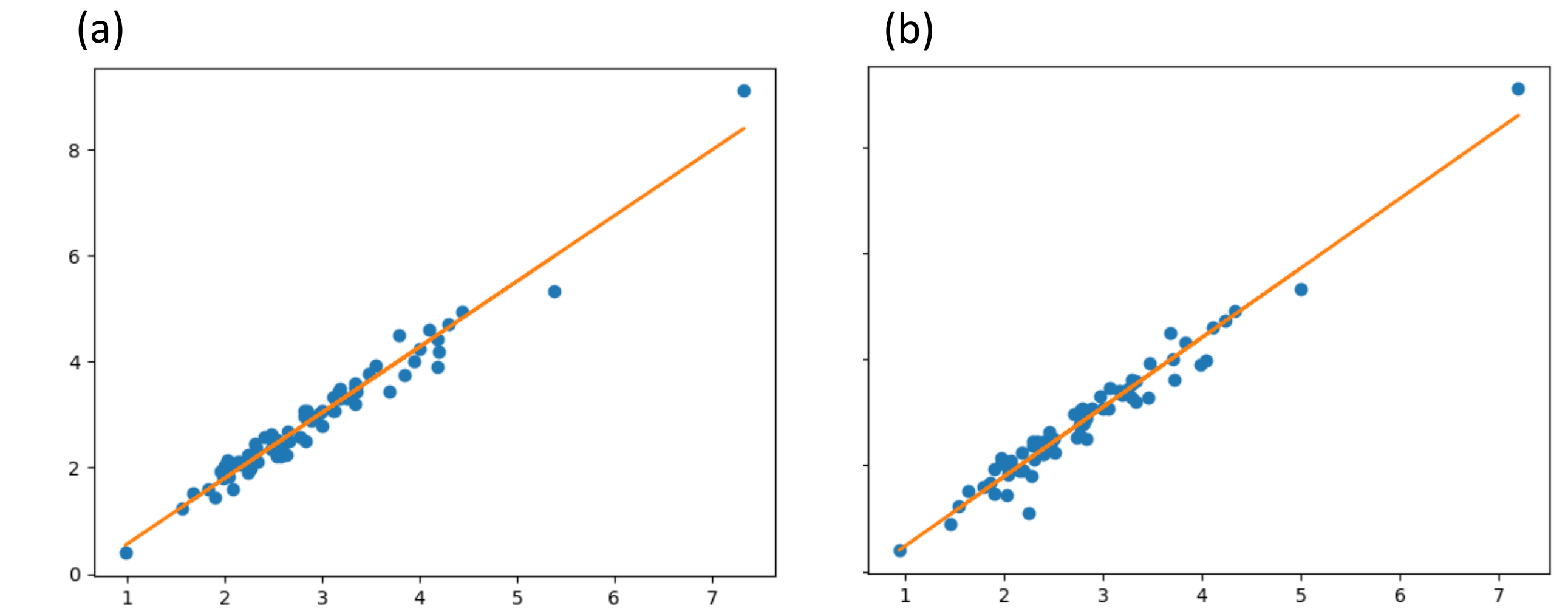
Estimation by Random Forest

- In the initial method, the training units are level 2 administrative units.
- In the improved method, we Integrate IDP sites in pixel and sum up the population. And, we use this pixels as pseudo administrative boundaries, the training unit will subtract the area and population of each pixel in the unit.
- Aggregate covariates by census blocks: we calculate zonal means for each continuous dataset.
- We log-transform the population density in order to create a more normal and even distribution of population density values with respect to other covariates.
- Due to the limited number of data, we use the out-of-bag score to assess the accuracy.
- The result of Random Forest is used to predict a country wide, pixel-level map of log population densities.

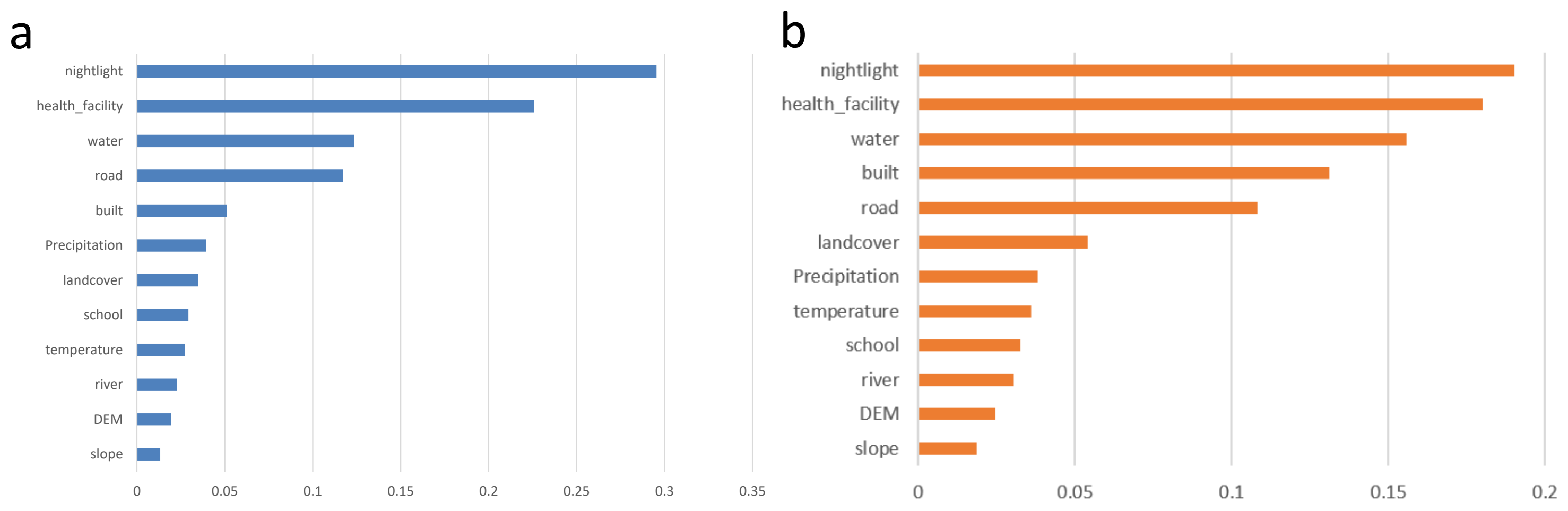


Integrate IDP sites in a ~100m pixel and sum up the population

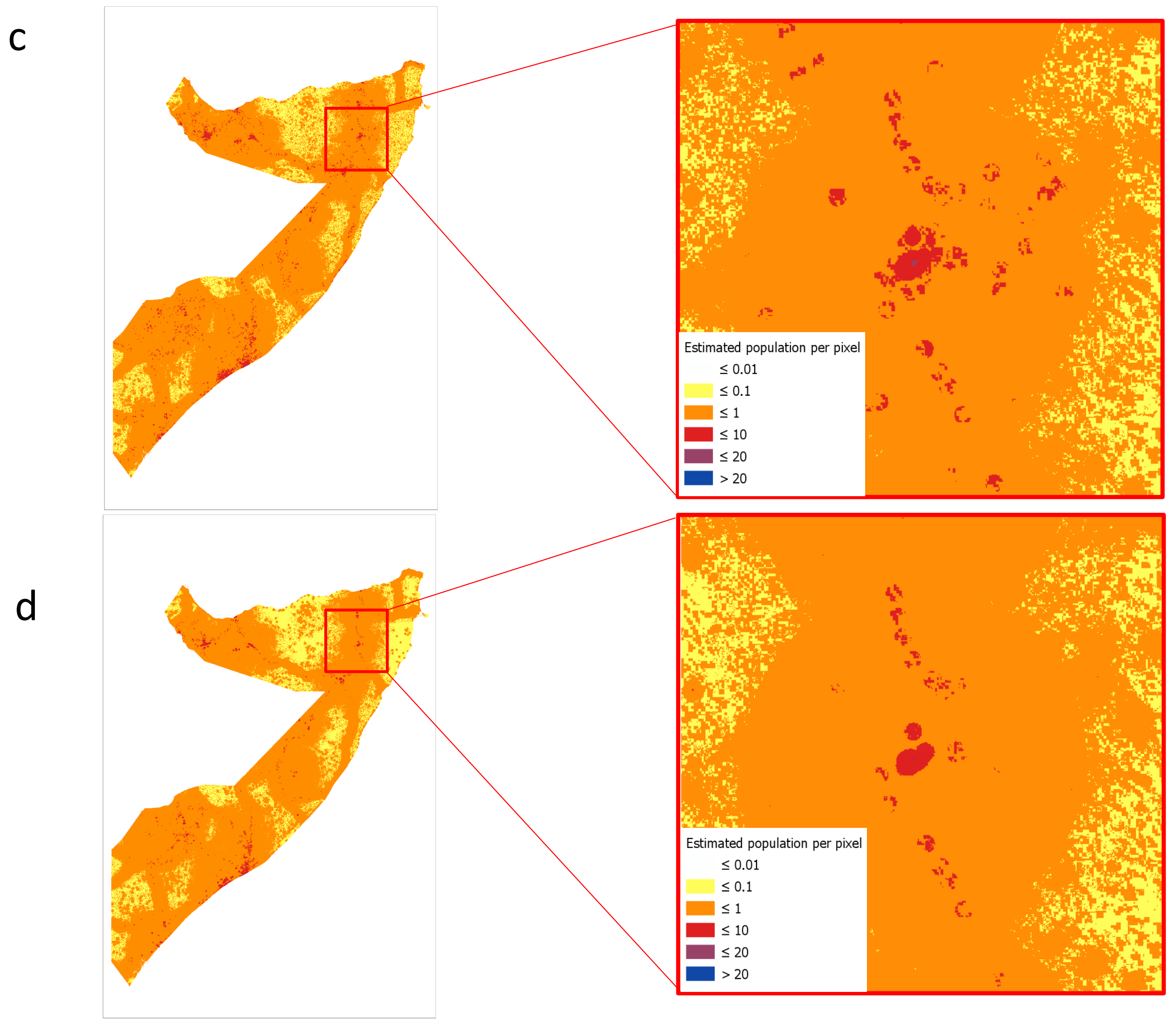
(a) The prediction and true population density in each unit using the initial method (OOB_score = 0.451); (b) The prediction and true population density in each unit using the improved method (OOB_score = 0.221);



Results and Analysis



(a) Variable Importance for Random Forest regression using the initial method; (b) Variable Importance for Random Forest regression using the new method;



(c) The redistributed map for the Somalia in 2018 using the initial method; (d) The redistributed map for the Somalia 2018 estimated census data by using the improved method;

- Nightlight is the most important variable to explain the population density in the all features in two methods.
- Because a population of high density is in one or two pixels, other pixels in the same level 2 administrative unit don't have such high population density as the mean population density in the whole unit. So, Comparing two maps using initial method and the using improved method (with IDP data) in this project, the high density population is shown more obviously in 2018. This can avoid False alarm situations, which means that high population density in map, but not in reality.
- However, when we exclude the IDP data in the unit, the explanatory power of the variable Nightlight decreases, which means the IDP camps is highly related with the night light in Somalia, so that the OOB–score decreases too.
- Estimated census data in 2018 will affect the accuracy of the prediction .

Summary and Recommendation

Summary

- Random Forest performs well in disaggregating census data for population mapping.
- After subtracting the population of each IDP site from the training unit, the accuracy of the population density model will be enhanced.
- The improved method will decrease the explanatory powers of variables.

Recommendation

- If IDP data is accessible with no outline data, integrating IDP sites in the unit pixel is a good solution to improve the accuracy of prediction model.