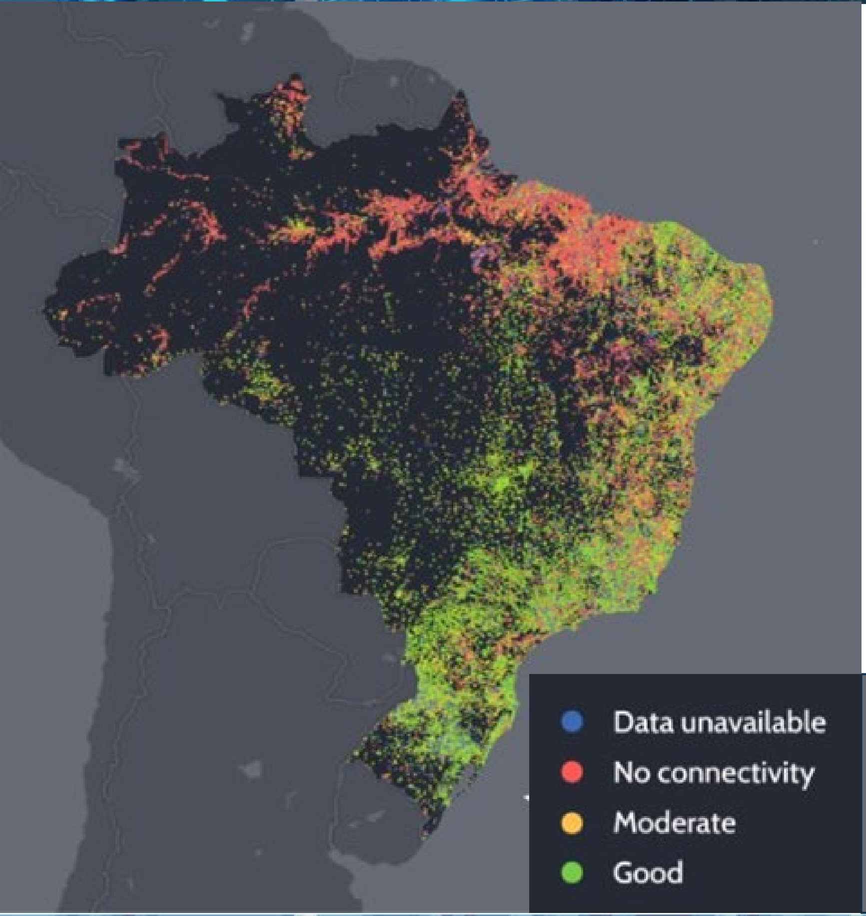
Modelling Priority Areas for Improving Global School Access

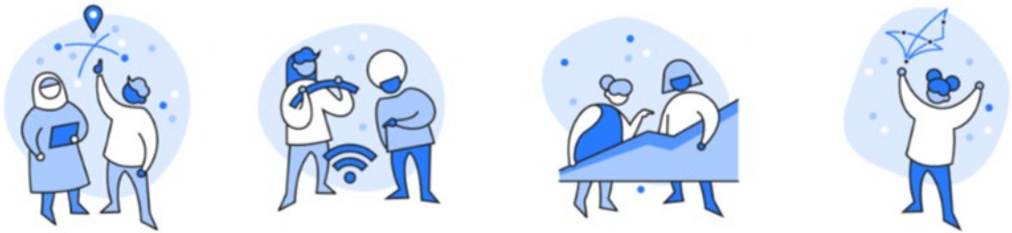
A Geostatistical Machine Learning Approach

Abi Riley
UNICEF-ESA Research Intern
PhD Student at Imperial College London
a.riley21@imperial.ac.uk

Casper Fibæk, Kelsey Doerksen, Do-Hyung Kim, Alessandro Sebastiane, Rochelle Schneider



The Giga Initiative



Map *Connect* *Finance* *Empower*

"Connecting every school, and every community to the Internet"

giga | unicef for every child | ITU

The UNICEF-ESA Internship Project

Data Processing

Combining datasets from governments and open sources

Geocoding and validation of school locations

Processing other variable datasets, e.g. GHSL, nightlights, images

Access to Schools

Modelling the spread of schools

Linking to population counts, land cover and urbanicity

Spatial modelling

Statistical analysis of results to identify key areas

Finding New Schools

Computer vision methods

High-resolution satellite imagery

Modelling Resources

Connectivity of Schools

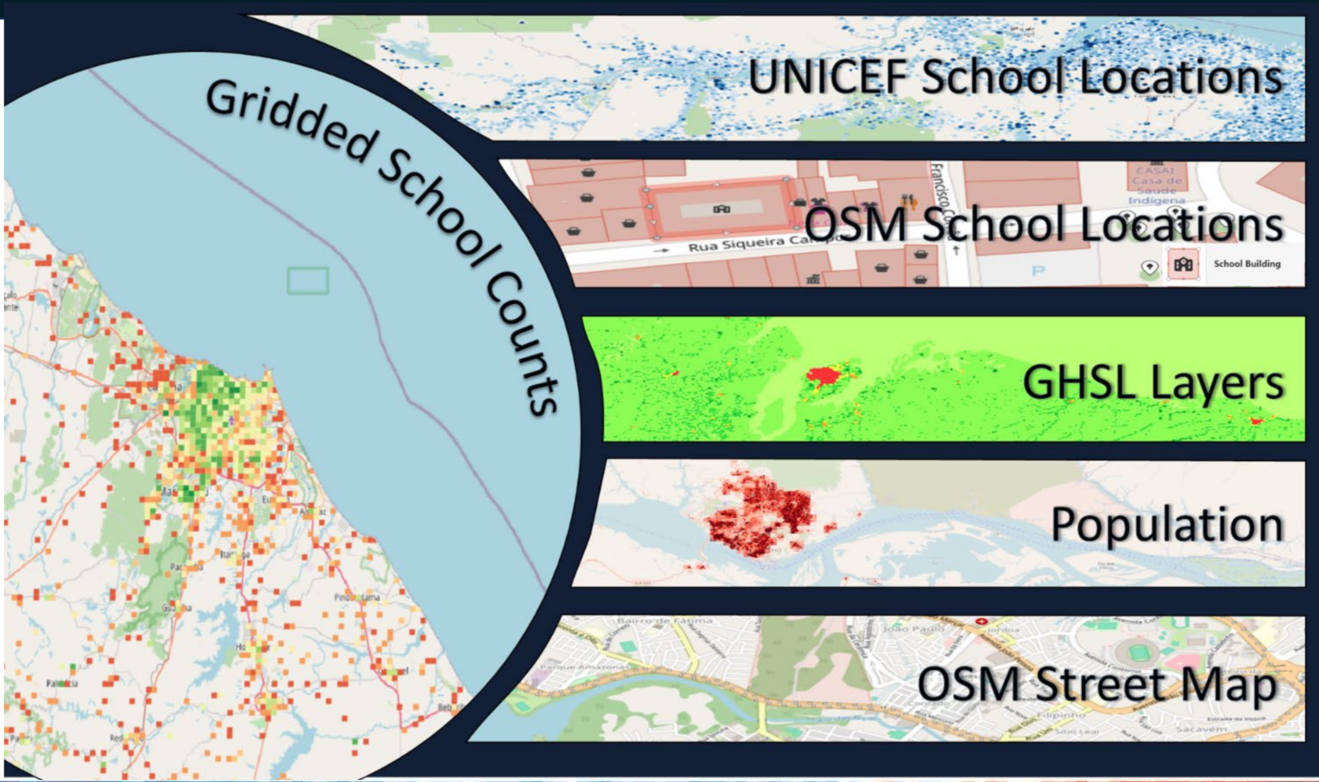
Geographically-Aware Models

Data

Schools Data:
- UNICEF, Gov, OSM sources
- Merging and validating

Gridded School Counts:
- 1km and 10km grids
- Counts and indicator

Covariates:
- GHSL settlement info
- Population and others…



Methods

Random Forest Classification for Gridded School Indicator Data

- **Outputs**: grid of predicted indicator variable, grid of predicted RF probability

Non-Spatial Random Forest Regression for Gridded School Count Data

- **Outputs:** grid of predicted school counts

Spatial Random Forest Regression for Gridded School Count Data

- **Outputs:** grid of predicted school counts

Additional Methods:

- Variable selection
- Variable importance
- Interaction terms
- Spatial autocorrelation

Case Study: North-East Brazil

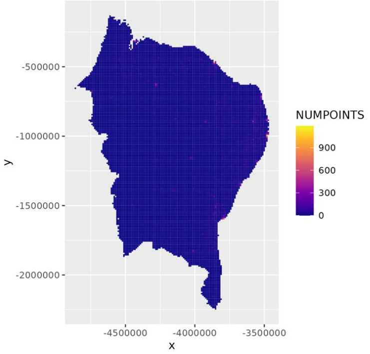
School Indicator Model

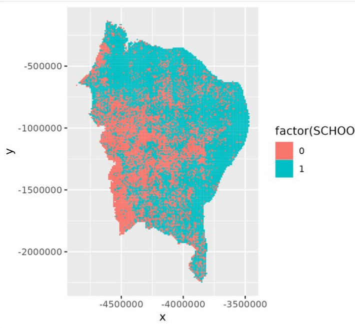
1. Fit Random Forest classification model
2. Parameter tuning: grid search
3. Prediction
4. Identify false positives and get RF prediction probabilities

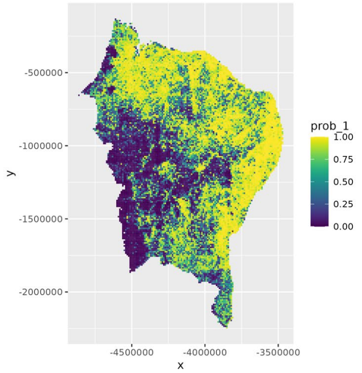
Accuracy: 0.9441

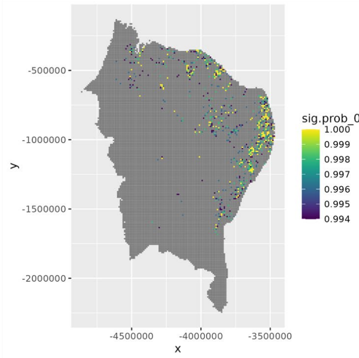
False Positive Rate: 3.11%

	0	1
0	26200	920
1	1151	8791





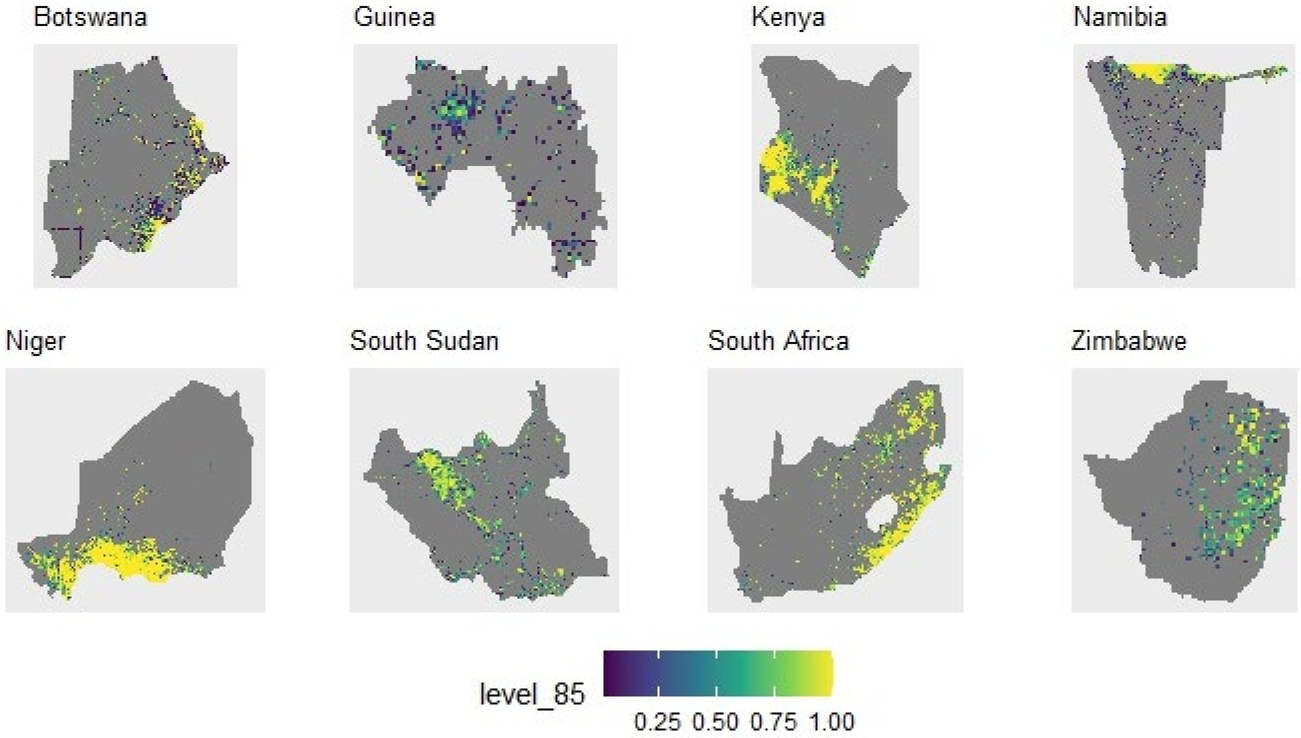




Case Study: Africa

School Indicator Model Probabilities

Example countries with top 15% of false positives

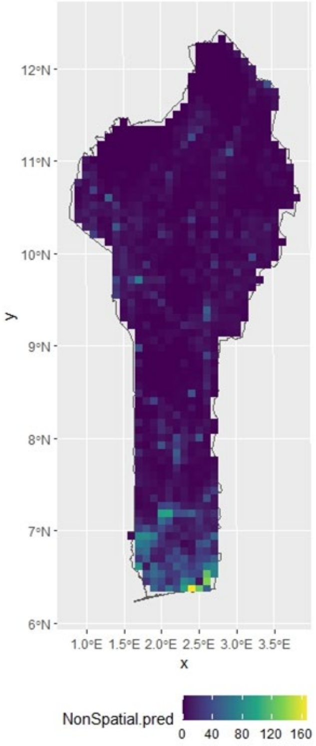


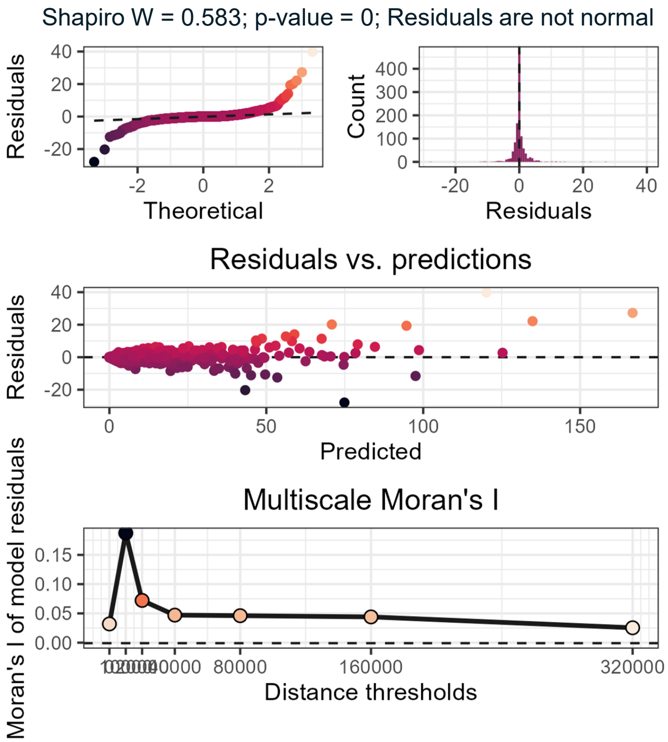
Case Study: Benin

Non-Spatial School Counts

1. Assess potential multicollinearity
2. Variable selection
3. Fit non-spatial Random Forest regression model
4. Parameter tuning
5. Cross-validation on spatial folds
6. Diagnostics of residuals

	Fit	Pred
R squared	0.973	0.837



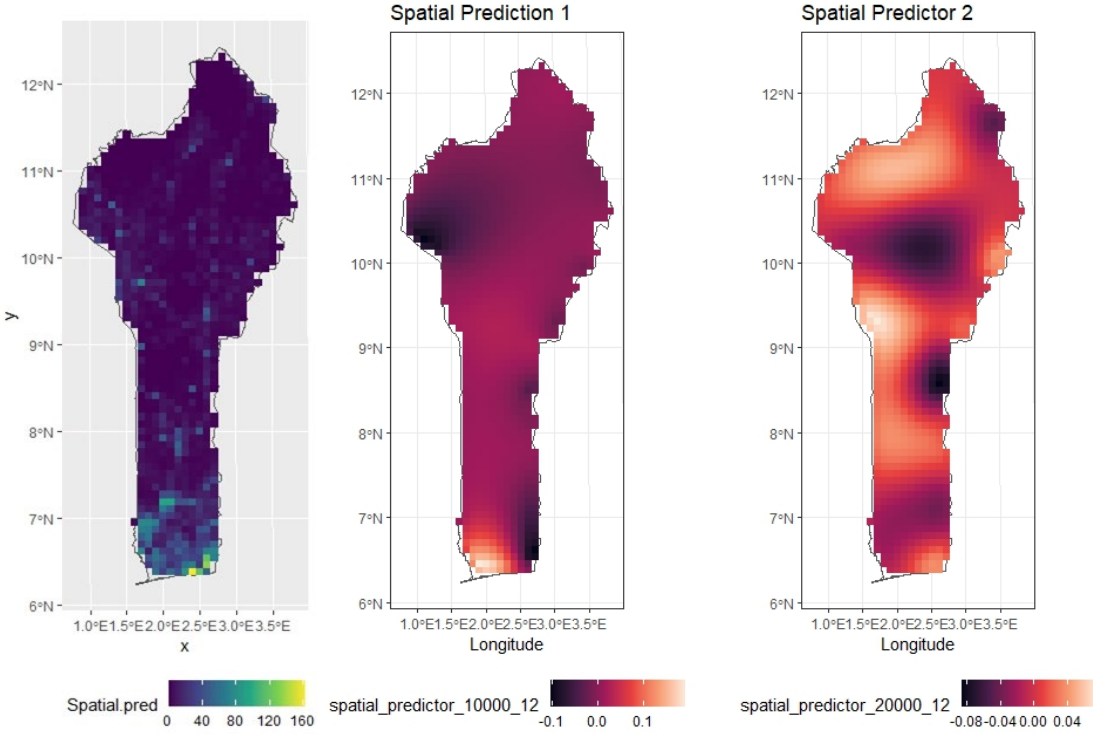


Case Study: Benin

Spatial School Counts

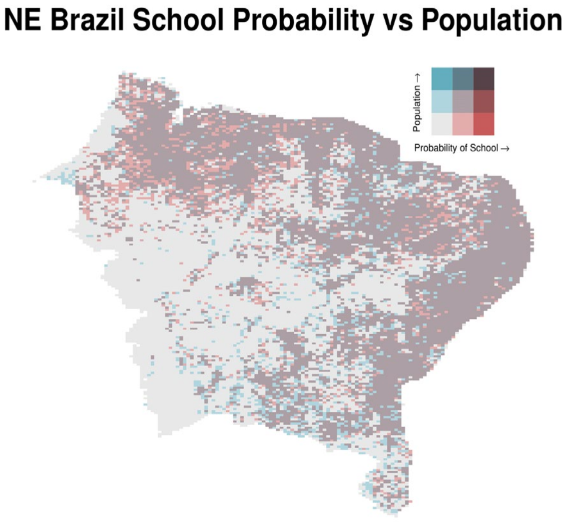
Motivated by and using variable selection and importance from non-spatial model

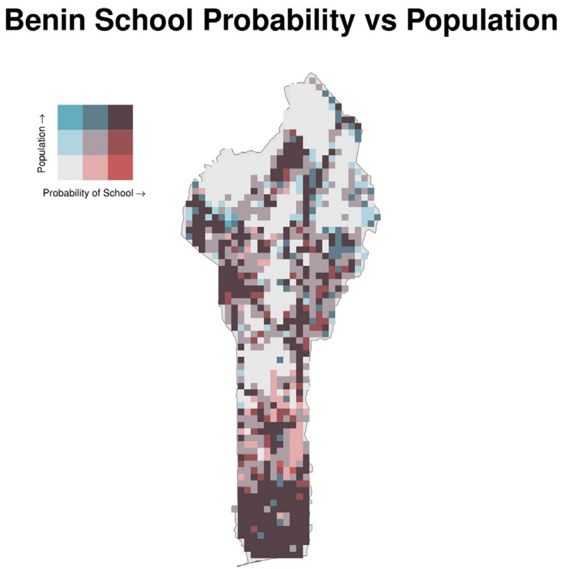
	Spatial		Non-Spatial	
	Fit	Pred	Pred	Fit
R squared	0.973	0.627	0.837	0.978

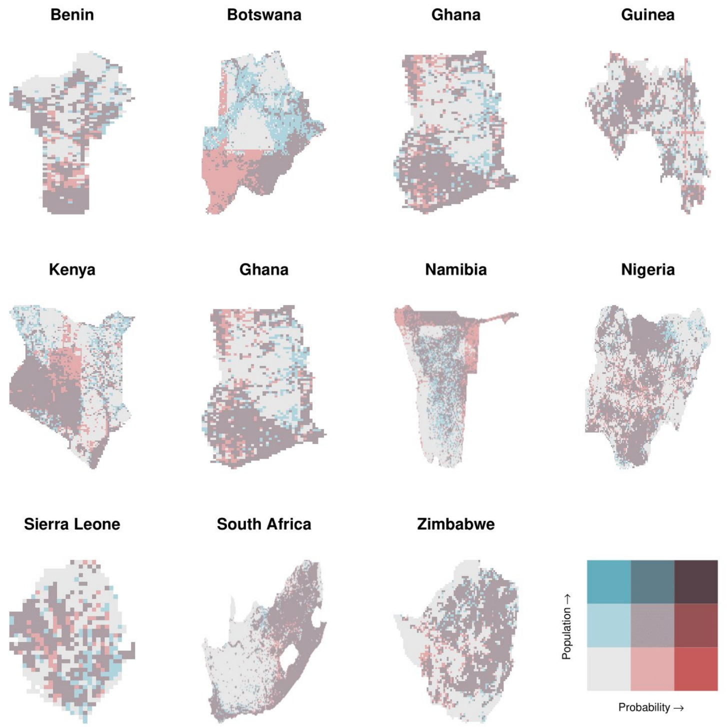


Identifying Priority Regions

Example: Balance between RF probability and population







Discussion

Conclusions

The use of low-resolution global satellite-derived data

Motivating more complex model approaches

Best performance and efficiency using Random Forest methods

Linking to Further Work

Priority areas to find new schools using high-resolution satellite imagery (Casper)

Linking to modelling connectivity (Kelsey)

My PhD Work

Spatiotemporal statistics for the effects of air pollution on mental health

Including using satellite data to model air pollution

Also using greenspaces and built-up areas, from GHSL

Global AI-Powered School Connectivity Prediction with Earth Observation

Kelsey Doerksen[1,2,3*]
PhD Student at University of Oxford
kelsey.doerksen@cs.ox.ac.uk

Casper Fibæk[3], Abi Riley[3] Rochelle Schneider[3], Isabelle Tingzon[1], Do-Hyung Kim[1]
[1]UNICEF, [2]University of Oxford Department of Computer Science, [3]European Space Agency Φ-lab

Giga: An initiative to connect every school to the Internet and every young person to information, opportunity and choice

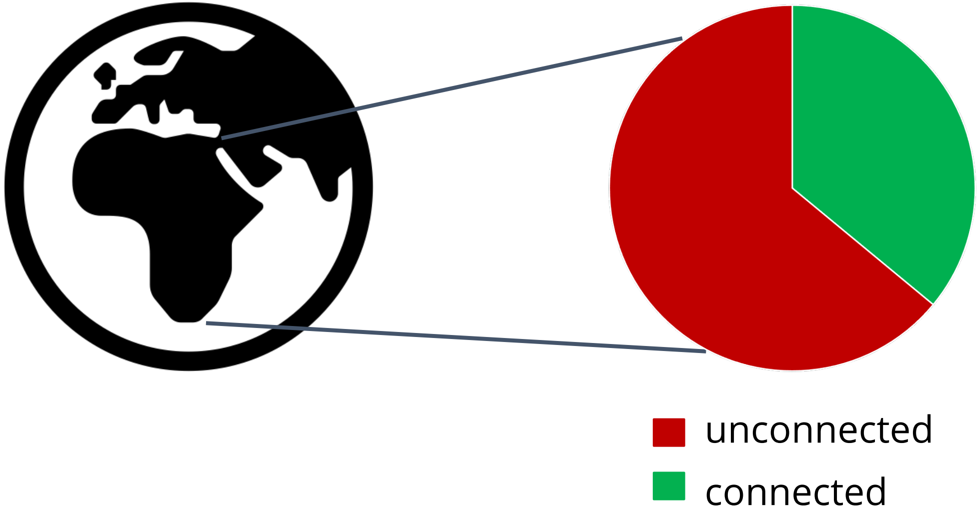
Ensure inclusive and equitable quality education and promote lifelong learning opportunities for all

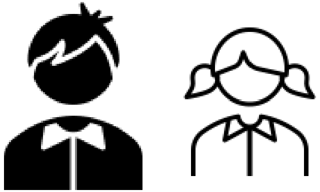
The Digital Div ide

Over **500,000,000** students worldwide don't have access to the internet

In 2022, only 36% of Africa's population had broadband internet

Africa has one of the world's widest **digital gender gaps** 35% vs 24% in 2020





What do Connected Schools look like?

Methodology - Leveraging Geospatial Data to Predict Internet Connectivity

Accepted: ICLR Machine Learning for Remote Sensing Workshop: *AI-powered School Mapping and Connectivity Status Prediction using Earth Observation*

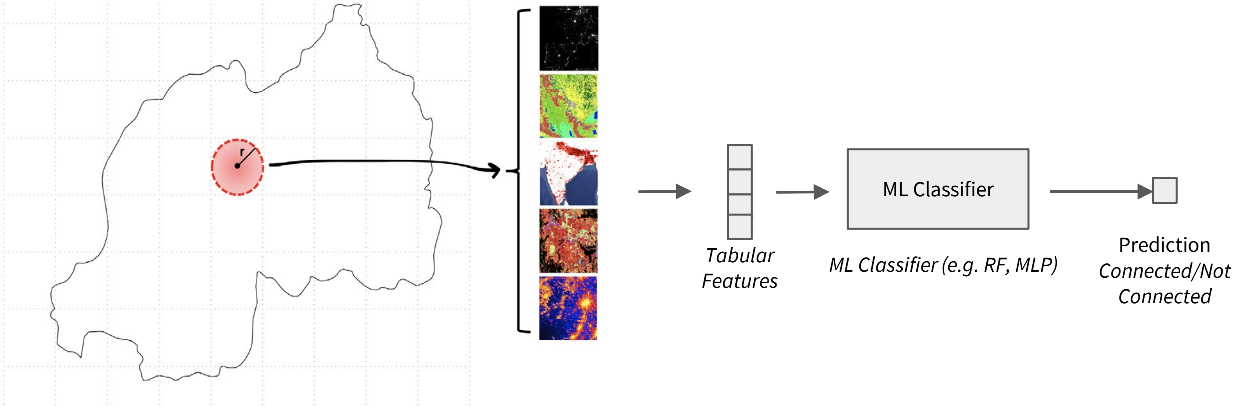
github.com/kelsdoerksen/airPy

★61

Problem Setup

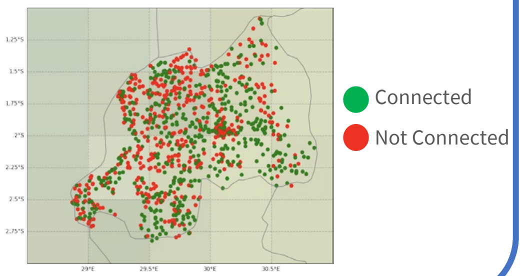
Binary classification task targeting Connected (1) or Unconnected (0) schools.
70/30 train/test split and 5-fold cross-validation with hyperparameter tuning.

Engineered Features



Feature engineering based on satellite images, electric grid information, and speedtest data

School Connectivity Data



Geospatial Data

MODIS Land Cover, VIIRS Nightlight Gridded Population of the World, Global Human Settlement Layer, Global Human Modification, Transmission Line Network, Ookla Speedtest, Regional Encoders

Results - Leveraging Geospatial Data to Predict Internet Connectivity

Table 1: Per-country macro-averaged **F1-scores** of ML classifiers for Bosnia and Herzegovina (**BIH**), Belize (**BLZ**), Botswana (**BWA**), Guinea (**GIN**), and Rwanda (**RWA**)

ML Classifier	BIH	BLZ	BWA	GIN	RWA
RF	0.82	**0.92**	**0.73**	**0.74**	**0.72**
SVM	**0.83**	0.89	0.72	0.69	0.69
LR	**0.83**	0.88	0.71	0.66	0.70
GB	0.82	0.90	**0.73**	0.70	0.69
MLP	**0.83**	0.86	0.68	0.68	0.71

Table 2: Class distribution across training and testing sets for BIH, BLZ, BWA, GIN & RWA

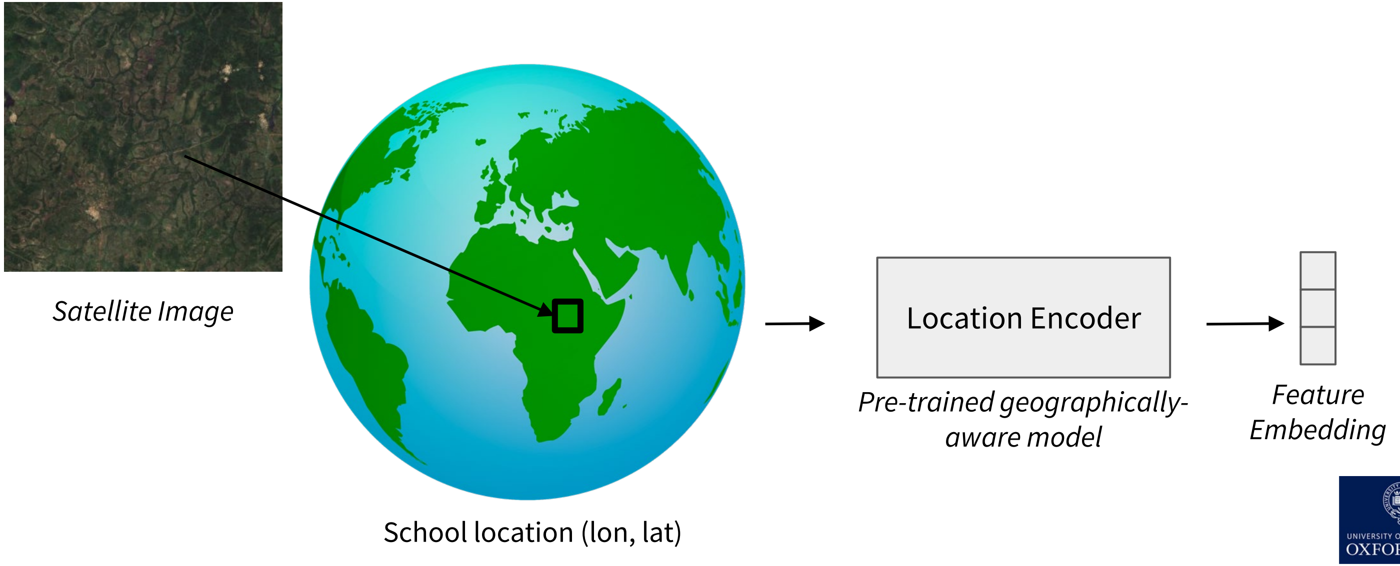
	Training Set (70%)			Test Set (30%)			Total
	Connected	Not Connected	Total	Connected	Not Connected	Total	
BIH	651	284	**935**	278	123	**401**	**1336**
BLZ	168	52	**220**	75	20	**95**	**315**
BWA	327	307	**634**	149	124	**273**	**907**
GIN	286	373	**659**	113	170	**283**	**942**
RWA	1337	1011	**2348**	551	456	**1007**	**3355**

University of Oxford giga esa Φ-lab

Methodology - Leveraging Geospatial Data + *Geographically-Aware models* to Predict Internet Connectivity

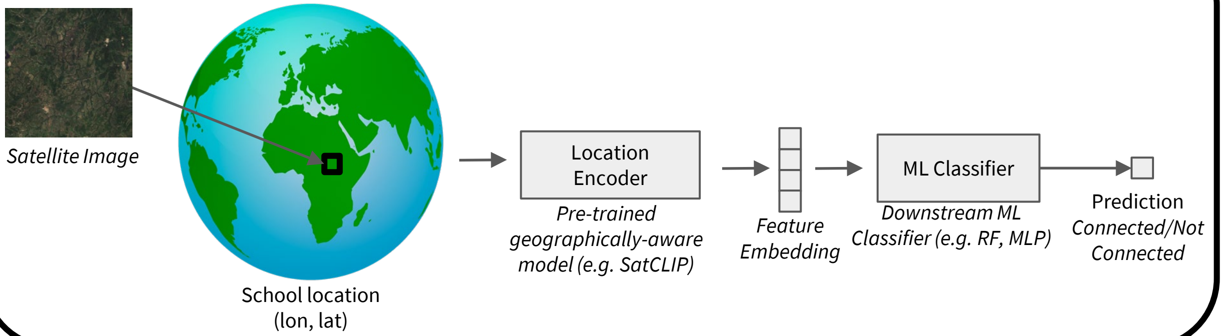
Under Review IJCAI AI and Social Good Track: *Investigating Machine Learning-Powered School Connectivity Prediction with Earth Observation and Geographically-Aware models*

→ CLIP (Contrastive Language-Image Pre-training) models are trained on a variety of (image, text) pairs, extending this to a geographic context whereby instead of training text to image encoders, **location encoders are trained to learn implicit representations of locations from satellite imagery**

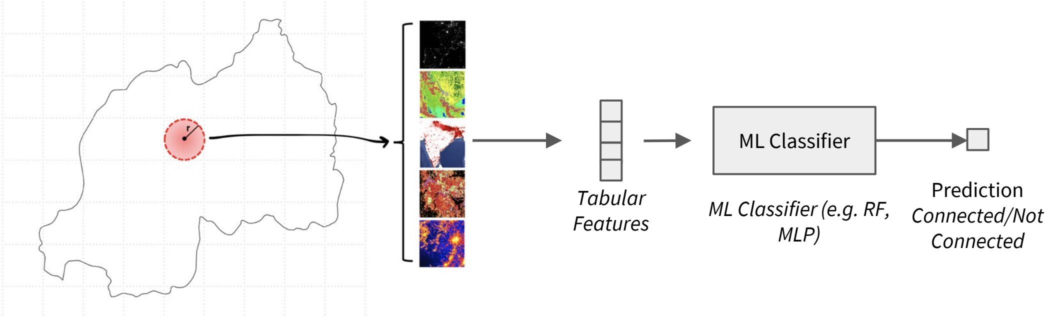


Methodology - Leveraging Geospatial Data + *Geographically-Aware models* to Predict Internet Connectivity

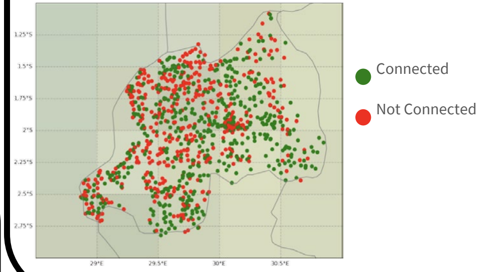
Location Encoder Embeddings



Engineered Features



School Connectivity Data



Results - Leveraging Geospatial Data + *Geographically-Aware models* to Predict Internet Connectivity

Table 3: Location Encoder Characteristics

Model	Dataset	Embedding Size
SatCLIP	Sentinel-2	256
GeoCLIP	MediaEval Placing Tasks 2016	512
CSP	Functional Map of the World	256
PhilEO VHR	ESA Very-High Resolution (VHR) Collection	1024

Table 4: Comparison of model performance scores given by per-country binary F1 and accuracy of the ML classifiers RF, MLP, GB using Engineered Featured, SatCLIP (SC), GeoCLIP (GC), CSP, PhilEO, and PhilEO + Engineered

		SC-R18-l10		SC-R18-l40		SC-R50-l10		SC-R50-l40		SC-ViT16-l10		SC-ViT16-l40		GeoClip		CSP		Engineered		PhilEO VHR		PhilEO + Eng	
		Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
BWA	RF	0.56	0.60	0.56	0.60	0.56	0.60	0.55	0.58	0.54	0.58	0.49	0.54	0.55	0.63	0.56	0.61	**0.70**	**0.76**	0.57	0.65	0.67	0.74
	MLP	0.55	0.63	0.58	0.59	0.55	0.63	0.55	0.52	0.56	0.51	0.53	0.61	0.54	0.58	0.55	0.66	**0.65**	**0.71**	0.54	0.55	0.53	0.54
	GB	0.53	0.58	0.52	0.57	0.52	0.55	0.54	0.59	0.54	0.60	0.54	0.59	0.50	0.58	0.55	0.59	0.68	0.72	0.57	0.62	**0.76**	**0.78**
RWA	RF	0.65	0.69	**0.66**	0.69	**0.66**	0.69	0.64	0.68	0.64	0.68	0.65	0.69	0.65	0.69	0.64	0.69	0.65	**0.71**	0.55	0.64	0.54	0.64
	MLP	0.56	0.67	0.57	0.65	0.57	0.65	0.57	0.59	0.55	0.68	0.58	0.59	0.63	0.67	0.56	0.68	**0.65**	**0.71**	0.53	0.56	0.51	0.49
	GB	0.53	0.58	0.52	0.57	0.52	0.55	0.54	0.59	0.64	0.60	0.54	0.59	0.63	0.67	0.60	0.64	**0.66**	**0.70**	0.52	0.60	0.60	0.66

Results - Leveraging Geospatial Data + *Geographically-Aware models* to Predict Internet Connectivity

Figure: Random Forest Classifier Top 10 Feature Importance
Botswana

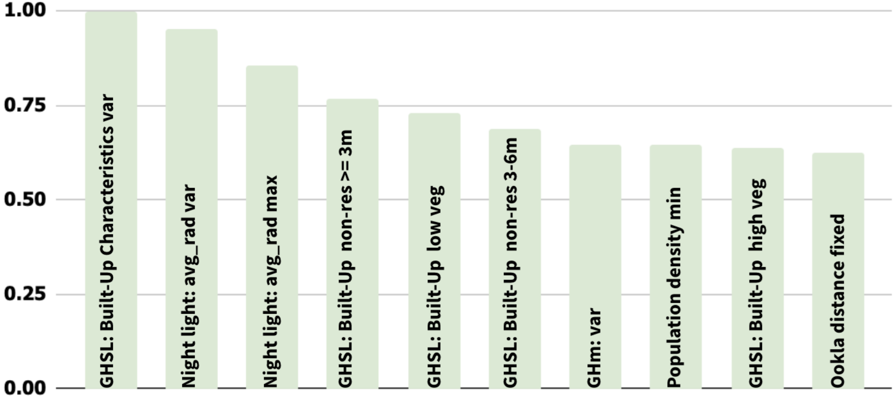
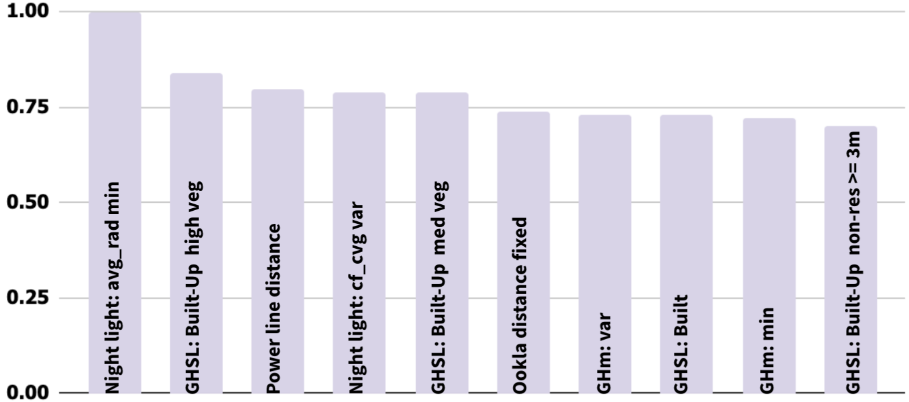
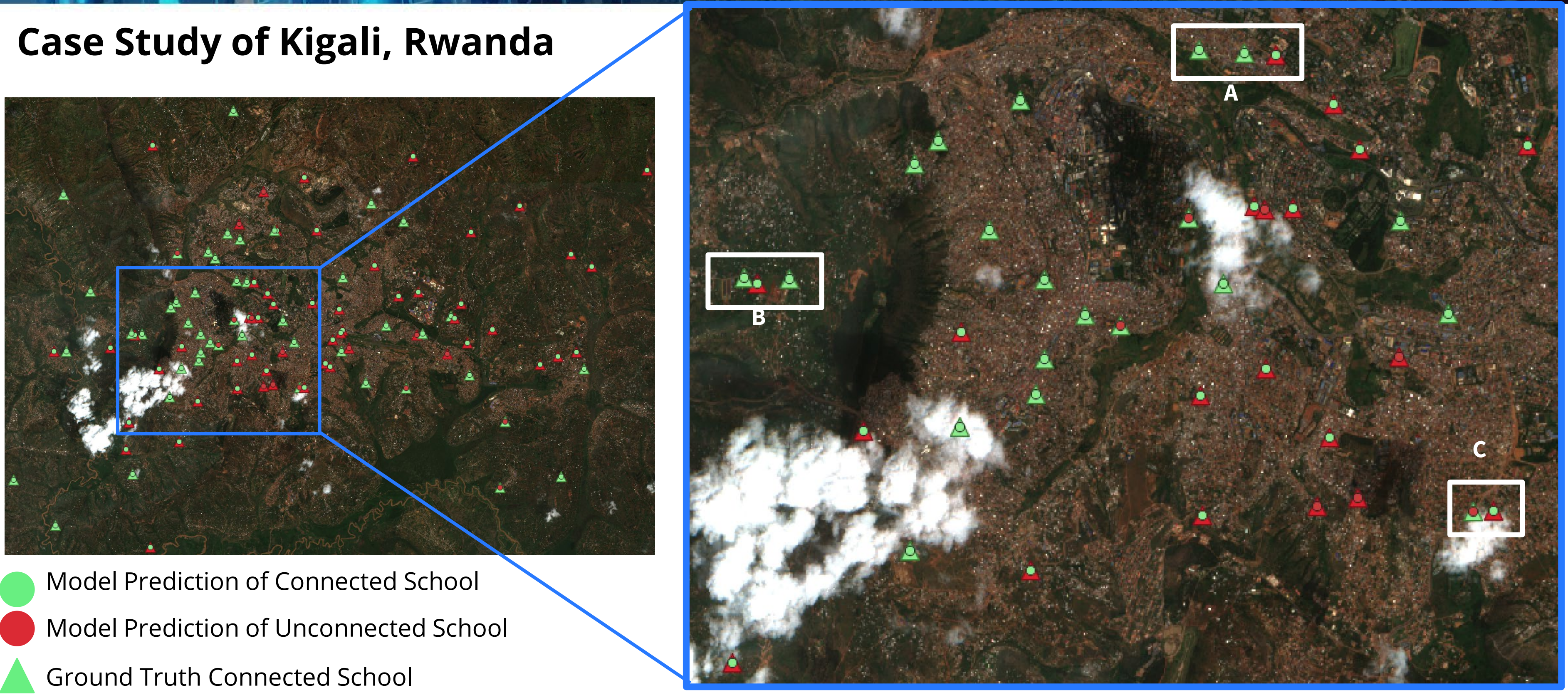


Figure: Random Forest Classifier Top 10 Feature Importance
Rwanda



giga
esa
Φ-lab

Case Study of Kigali, Rwanda



Model Prediction of Connected School

Model Prediction of Unconnected School

Ground Truth Connected School

Ground Truth Unconnected School

Limitations & Future Directions

Label Quality. Unidentified latency between connection and labeling, inconsistency of label quality

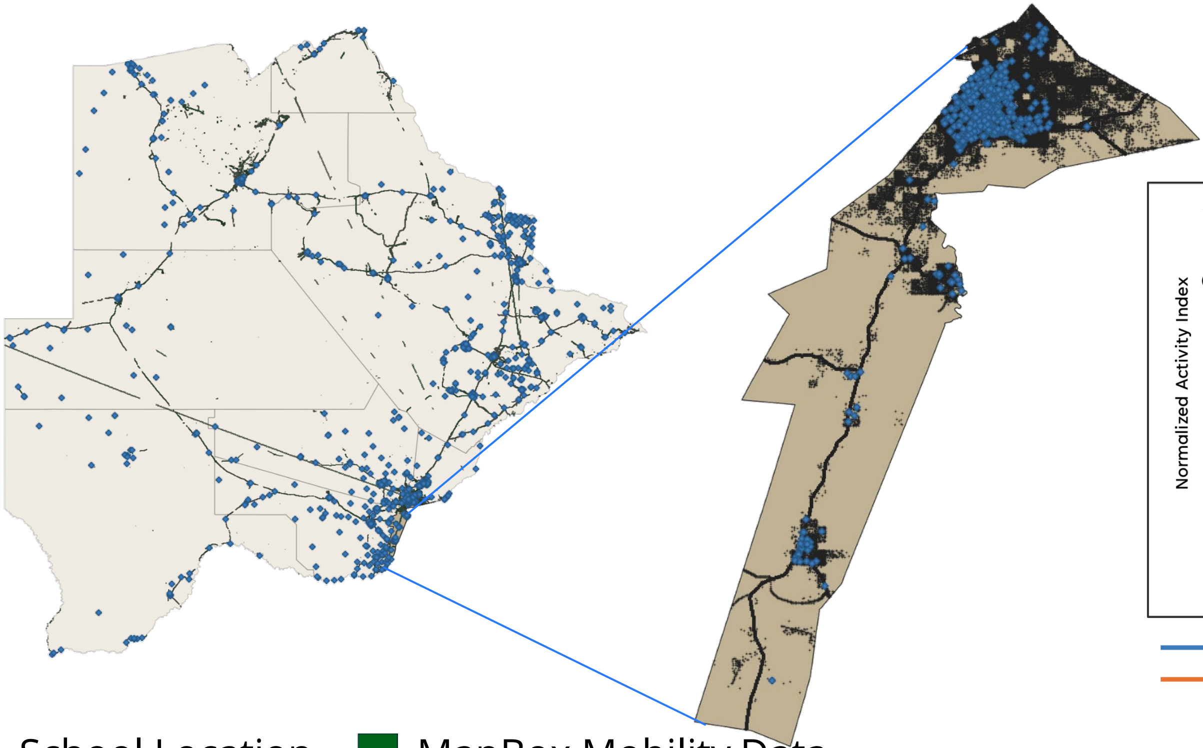
Auxiliary Information Needed. Ground-based survey information may be necessary to improve performance.

Connectivity Quality and Infrastructure. Identifying infrastructure to support digital capacity building.

School Mapping with Human Mobility Data. MapBox mobility data for distinguishing building type (school/non-school).

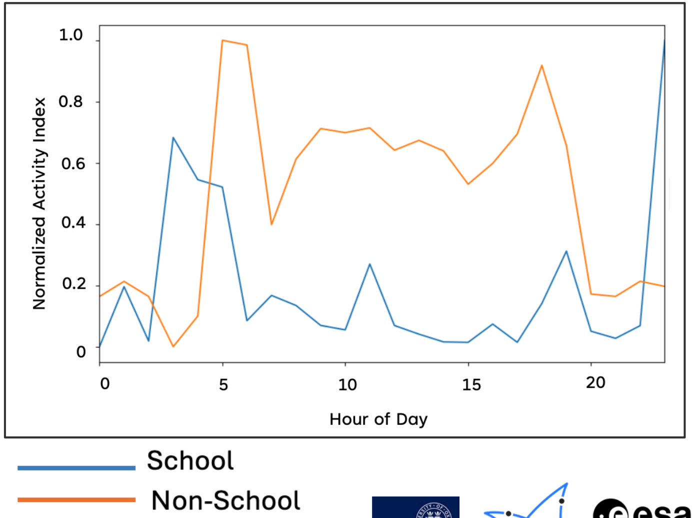
School Mapping with Mapbox Mobility Data

Case Study of South-East Botswana



School Location
MapBox Mobility Data

Figure: Mapbox Mobility Weekday Time Series for Schools and Non-Schools



UNIVERSITY OF OXFORD
giga
esa
Φ-lab

Thank you!

kelsey.doerksen@cs.ox.ac.uk
www.linkedin.com/kelsey-doerksen

a.riley21@imperial.ac.uk
www.linkedin.com/in/abiriley/