

---

# Creating a well-being data layer using open source data

# Contents

The problem

The Data

Solution

Conclusion

---

# The problem

# Demographic Health Survey is hard to obtain

## Problem Statement

Conducting economic surveys require huge amount of resources.

## Goals

Using Open-Source available data, is it possible to replace current processes?



---

# The Data

# Demographic Health Survey (DHS)

## Description

DHS national household surveys conducted in 2015.

## Usage

Different targets to predict, with the lowest aggregation being the DHS cluster.

	Wealth	Toilet Facility	Roof Material	Population_Density
DHSClust				
10001	richer	Flush	Metal	Average
10002	middle	Flush	Metal	Very Low
10003	middle	Flush	Metal	Very Low
10004	middle	Flush	Metal	Very Low
10005	middle	Flush	Metal	Very Low

# Open Street Maps

## Description

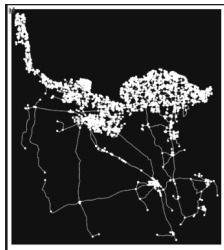
Open Source geospatial data —map tiles (geotiff) and vector data

## Usage

Location about the hospitals, cafe's and colleges.

## Status

Python modules implemented to use the OSM API to collect vector and geotiff data given a district boundary.



# Night Time Lights

## Description

NASA's VIIRS/NPP Lunar  
BRDF-Adjusted Nighttime Lights Daily  
L3 Global 500m Linear Lat Lon Grid.

## Usage

To be used as a proxy measure for  
economic activity

## Status

Python script to download hdf5 files  
based on region and time range and  
convert to geotiff format.





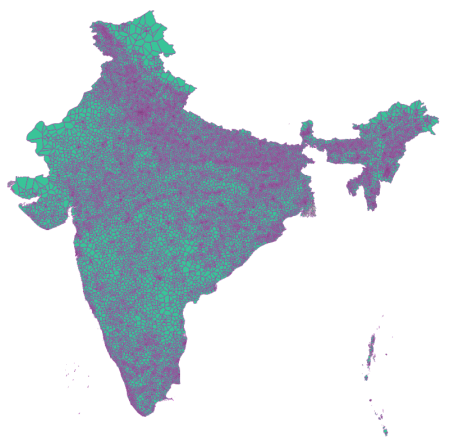
# Joining DHS, OSM and NTL data

## Voronoi Polygons

Given a finite set of DHS cluster points the map is divided into cells where each cell covers the region closest to a particular cluster point.

## Combine DHS and OSM data

Every OSM data point gets the same id as the Voronoi Cell which contains it.



---

# Solution

# Data Integration

## Description

Integrating all the data in a pipeline:

- **DHS**(Target): Wealth Indicators
- **OSM**(Features): Describing amenities in the Cluster

## Usage

Different targets to predict, with the lowest aggregation being the DHS cluster.

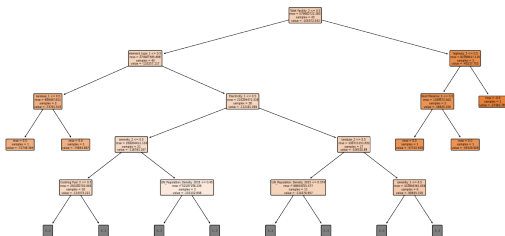
amenity	highway	landuse	building	Toilet Facility	Roof Material
hospital	crossing	forest	yes	No Facility	Metal
clinic	crossing	residential	yes	Flush	Cement
parking	crossing	residential	yes	No Facility	Metal
college	motorway_junction	residential	yes	No Facility	Metal

# Modeling

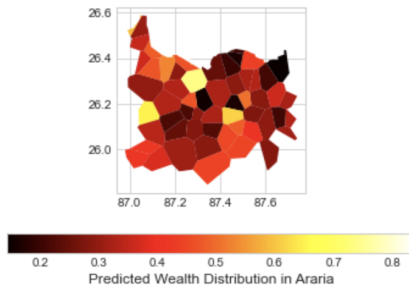
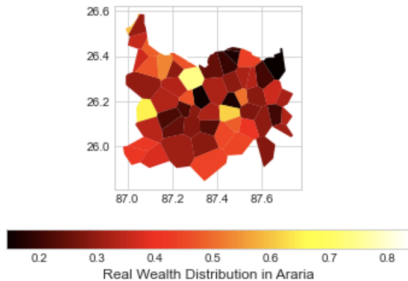
## The label

Mean Absolute Error of 0.64

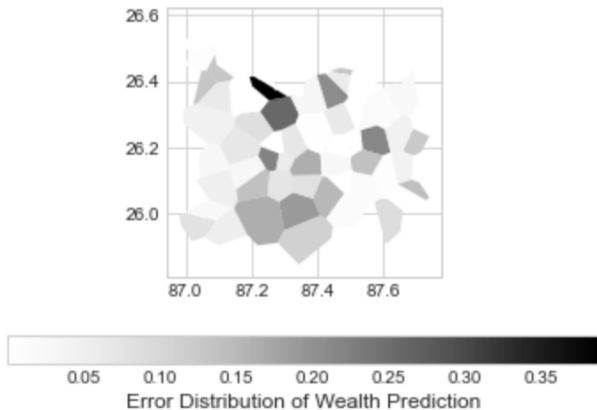
- 1: Poorest
- 2 : Poor
- 3 : middle
- 4 : richer
- 5 : richest



# Modeling



# Modeling



---

# Conclusion

# Conclusion

## Results

Predictive Model that is computationally cheap

- Open Source Data
- No cloud computing resources

Modeling remains explainable and accountable while preserving accuracy

- Explainable and Interpretable machine learning
- Accountable
- High Generalization



# Future Work

## Steps

- Scale Up
- Integrate with NTL
- Temporal Evaluation

## Scope

- Towards an application?
- Towards deployment?
- Towards research?